Robust Methods for Causal Inference Using Penalized Splines

by

Tingting Zhou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Michael Elliott, Co-Chair
Professor Roderick Little, Co-Chair
Professor Sarah Burgard
Associate Professor Lu Wang

Tingting Zhou

tkzhou@umich.edu

ORCID iD: 0000-0002-4872-9138

To my parents, brothers and family

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisors, Prof. Roderick Little and Prof. Michael Elliott, for sharing their insights and expertises, for encouraging me and steering me in the right direction when I was stuck, and for reminding me to critically evaluate my own work. I am very grateful for their support, guidance and patience over the last four years. They have made me a better researcher and statistician.

I would also like to thank my other committee members, Prof. Lu Wang and Prof. Sarah Burgard for taking the time to serve on my committee and providing valuable feedbacks to make my dissertation better. I also want to thank Prof. Sarah Burgard for baking us really delicious treats for both my proposal and defense. I am also very grateful to Prof. Goncalo Abecasis for supporting me during the first three years of graduate school.

I would like to thank the friends I've met in Ann Arbor, especially Wenting Cheng, Cui Guo, Lili Wang, Sayantan Das, and Alan Kwong. They helped me make so many fond memories during the past six years. I would like to thank Sai for being so supportive and always believing in me. I also want to thank Alan for always being such a great friend and helping me so much. Finally, I would like to thank my grandpa, my parents, my brothers, and my family for their constant support and endless love, especially my brother Luwei for helping me so much with school. Thank you all so much!

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Observational studies are important for evaluating treatment effects, especially when randomization of treatments is unethical or expensive. Without randomization, valid inferences about treatment effects can only be drawn by controlling for confounders. Propensity scores (PS) – the probability of treatment assignment as a function of covariates – are often used to control for confounders. PS-based methods are vulnerable to bias and inefficiency when outcome or propensity score models are misspecified or there is limited overlap in the propensity score distributions between treatment groups. In this dissertation, we develop new robust methods for estimating causal effects from observational studies and address two closely related topics on causal inference – the problem of limited overlap and variable selection for propensity score model.

In Chapter 2, we propose a robust multiple imputation based approach to causal inference called Penalized Spline of Propensity Methods for Treatment Comparison (PENCOMP). PENCOMP estimates causal effects by imputing missing potential outcomes with flexible spline models, and draws inference based on imputed and observed outcomes. Under the standard causal inference assumptions, PENCOMP is doubly robust, that is, yields consistent estimates of causal effects if either the propensity or the outcome model is correctly specified. Simulations suggest that it tends to outperform doubly-robust marginal structural modeling, especially when the weights are highly variable. We apply our method to the Multicenter AIDS Cohort study (MACS) to estimate the short term effect of antiretroviral treatment on CD4 counts in HIV+ patients.

In Chapter 3, we address the issue of limited overlap in the propensity score distributions across treatment groups. We investigate appropriate restrictions of the causal estimand, and compare alternative estimation methods, including various simple and augmented inverse propensity weighting approaches, matching and PENCOMP. We demonstrate the flexibility of PENCOMP for estimating different estimands. We apply these methods to the MACS dataset to estimate the effects of antiretroviral treatment on CD4 counts in HIV+ patients.

In Chapter 4, we consider variable selection techniques that seek to restrict predictors in the propensity model to true confounders, thus improving overlap in the propensity distributions and increasing efficiency. We also propose a new version of PENCOMP via bagging, which can be advantageous when the data are noisy. We examine by simulation studies the impact of various variable selection techniques, including an extension of the adaptive lasso, on inferences from PENCOMP and weighting methods. We demonstrate our methods and variable selection techniques using the MACS dataset.

CHAPTER I

Introduction

Randomized experiments allow researchers to measure the impact of an intervention on the outcome of interest since it can balance the covariate distributions across treatment groups. Unfortunately, randomization is not always feasible or ethical. In such cases, observational studies can provide some valuable information about the effectiveness of an intervention. However, without randomization, valid inference about causal effects can only be drawn by controlling for confounders. In 1983, Rosenbaum and Rubin introduced the idea of using propensity scores to estimate causal effects from observational studies. Since then, propensity scores (PS) – the probability of treatment assignment as a function of covariates – are often used. The propensity score has the balancing property: conditional on the propensity score, the observed covariates and treatment assignment are conditionally independent. The balancing property of propensity score implies that adjusting for the propensity score can remove bias due to the observed covariates (Rosenbaum and Rubin, 1983).

Inference about causal effects involves speculation about what would have happended if a subject receives some other treatment that's different from the assigned. Suppose there are two treatments denoted as 1 or 0, a subject has both an outcome under treatment 1 and an outcome under treatment 0. This describes the widely adopted Rubin's (1974) potential outcome framework in causal inference literature, which was first introduced by Neyman's (1923). Potential outcomes are defined as potentially observable outcomes under different treatments or exposure groups. Individual causal effects are defined as comparisons of the potential outcomes for that subject. However, only the potential outcome corresponding to the treatment actually assigned is observed for any subject. This is the fundamental problem of causal inference (Holland, 1986). Thus, to make causal inference, three assumptions are required: 1) SUTVA (Angrist, Imbens and Rubin, 1996) states that a) the observed outcome under the assigned treatment is the same as the potential outcome associated with that treatment, and b) the potential outcomes for a given subject are not influenced by the treatment assignments of other subjects (Rubin, 1980; Angrist, Imbens, Rubin, 1996); 2) positivity states that each subject has a positive probability of being assigned to each of the compared treatments; 3) ignorability states that treatment assignment is as if randomized conditional on all the past histories.

PS-based methods are based on estimating the propensity of treatment assignment, given potential confounding variables, and then using the estimated propensity to match, stratify or weight subjects. In matching, the treated and control subjects are selected to form matched pairs and simple matched pair analyses can be performed to obtain causal effects. In stratification, subjects are divided into strata based on their propensity scores and comparisons are performed within each stratum and causal effects are estimated by averaging across strata. Weighting each subject by the inverse of the propensity of receiving the observed treatment can also adjust for confounding variables because the weights in effect create a pseudo-population that is free of treatment confounders.

For PS-based methods to work correctly, the propensity score model should be correctly specified. Thus, more robust methods such as the ones that incorporate the outcome models can protect against misspecification of the propensity score model. One difficult but less addressed problem in causal inference is controlling for time-dependent confounders. For example, in a longitudinal study, subjects are observed over time and intermediate outcomes are measured. If these intermediate outcomes are also used to determine concomitant treatment assignments, they are both intermediate outcomes of past treatments and confounders of future treatment assignments- the phenomenon known as confounding by indication. Including these variables in standard regression models to control them as confounders does not work since they are also mediators of earlier treatment effects. For example, the Multicenter AIDS cohort study (MACS) (Kaslow et al, 1987) saw the introduction of the first antiretrovial therapy (zidovudine or AZT) at a time when no effective treatment for human immunodeficiency virus existed. Hence, early administration was based on availability and biomarkers of disease severity such as CD4 count, with sicker patients more likely to be treated. As HIV infection progresses, the number of CD4 cells decreases, and when the CD4 count was too low, patients started antiretroviral treatment to control the virus and increase the CD4 count. The CD4 count is a time-dependent confounder because it is both an intermediate outcome of past treatments and a confounder of future treatments.

The existing methods for controlling time-dependent confounders include the inverse probability treatment weighted (IPTW), the augmented IPTW (AIPTW), and g computation. The IPTW estimators are consistent if the propensity score models are correct. The AIPTW estimators are doubly robust, that is, consistent if the propensity models or all the outcome and intermediate outcome models are correct (Scharfstein, Rotnitzky, and Robins 1999; Yu and van der Laan, 2006). Finally, the g-computation provides a consistent estimator of potential outcomes and thus causal effects if all the conditional distributions relating outcomes to covariates are correctly specified (Robins, 1987). The IPTW and AIPTW estimators can result in highly variable estimates when there are extreme weights, which are common in observational studies. This is a particularly serious issue with longitudinal data with many possible treatment combinations. In addition, the weighted estimators are arguably hard to understand for applied researchers. The AIPTW estimators are doubly robust, but very hard to implement for applied researchers. The g-computation is more intuitive but not doubly robust.

Whether in a single or multiple time-point treatments, for PS methods to work reliably, there should be a sufficient overlap in the propensity score distributions between the compared treatment groups. This avoids extrapolating outside the overlap region and hence is less vulnerable to model misspecification. Restricting estimation of causal effects to a subpopulation where there is more balance in the propensity distributions between the treatment groups could reduce the sensitivity of causal effect estimates to model misspecification (Rosenbaum and Rubin, 1984). Most literature focuses on using propensity scores for assessing overlap. Cochran and Rubin (1973) suggest caliper matching when some units are left unmatched due to poor match quality based on some criteria. Gutman and Rubin (2013, 2015) propose dropping units outside of the overlap region of estimated propensity scores between the treatment groups. Dehejia and Wahba (1999) drop all control units whose estimated propensity scores are less than the smallest estimated propensity scores among the treated. Ho, Imai, King and Stuart (2005) propose a two-stage approach. In the first stage, all the treated units are paired with their closest control units, and only the matched units are included in the second stage. Crump et al (2009) propose trimming off extreme propensity values below α and above $1 - \alpha$. Li, Morgan and Zaslavsky (2017) define an estimand that weights cases to balance the weighted distributions of the covariates between treatment groups that minimizes the asymptotic variance of the estimated treatment effect.

In addition to sufficient overlap, one assumption needed for PS methods to make valid inference about causual effects is that all the confounders are observed and included in the propensity model. Since excluding important confounders in the model can lead to biased estimates, many covariates are often included, for fear of excluding some important confounders. Rubin (2007) notes that only pretreatment covariates should be included in the propensity model and argues that the model should be selected without accounting for the relationship between covariates and outcome. This approach helps maintain objectivity when making inference from nonrandomized studies. Furthermore, the variables included in the model can directly affect the degree of overlap. For example, including strong predictors of the treatment that are not predictive of the outcome in the propensity model could potentially shrink the overlap region. Recent work has also shown that including such covariates can inflate the variance of the causal estimate and may also induce bias (Brookhard et al, 2006). On the contrary, including covariates that are associated only with the outcome can improve efficiency, since it reduces random covariate imbalance in finite samples (Brookhard et al, 2006). Glymour et al (2008) argues for controlling only common causes of the treatment and outcome. VanderWeele and Shpitser (2011) propose controlling for covariates that are causes of the treatment and/or outcome. Thus, a propensity model based only on the treatment can be inefficient, as it prioritizes variables associated with treatment but not necessarily with outcome. Balancing such covariates using propensity score is unnecessary since these covariates are not confounders.

In this dissertation, we develop new statistical methods for estimating causal effects from nonrandomized studies and address two closely related topics on causal inference – the problem of limited overlap in the propensity score distributions between treatment groups and variable selection for propensity score models. In Chapter 2, we propose a simple and straightforward approach to causal inference that does not rely on weighting, is less sensitive to extreme weights, and has a double robustness property for causal effects, called Penalized Spline of Propensity Methods for Treatment Comparison (PENCOMP). PENCOMP estimates causal effects by imputing missing potential outcomes with flexible spline models, and draws inference based on imputed and observed outcomes. We compare PENCOMP with the existing weighting methods and g computation in simulation studies. We apply our method to the Multicenter AIDS Cohort study (MACS) to estimate the effect of antiretroviral treatment on CD4 counts in HIV infected patients.

In Chapter 3, we address the issue of limited overlap in the propensity score distributions across treatment groups. We investigate appropriate restrictions of the causal estimand, and compare alternative estimation methods, including various simple and augmented inverse propensity weighting approaches, matching and PENCOMP. We demonstrate the flexibility of PENCOMP for estimating different estimands when necessary. We apply these methods to the MACS dataset to estimate the effects of antiretroviral treatment on CD4 counts in HIV+ patients.

In Chapter 4, we turn our focus to model selection for the propensity score model. We consider variable selection techniques that seek to restrict predictors in the propensity model to true confounders, thus improving overlap in the propensity distributions and increasing efficiency. We also propose a new version of PENCOMP via bagging that also incorporates the variability of model selection, which can be advantageous when the data are noisy. We examine by simulation studies and the MACS dataset the impact of various variable selection techniques, including an extension of the adaptive lasso, on inferences from both versions of PENCOMP, AIPTW and IPTW. Finally in Chapter 5, we summarize our findings and suggest future directions to explore for PENCOMP.

CHAPTER II

Penalized Spline of Propensity Methods for Treatment Comparison

2.1 Introduction

Observational studies are important for evaluating treatment effects, particularly when randomization of treatments is unethical or expensive. In the absence of randomization, valid inferences about treatment effects can only be drawn by controlling for confounders. However, controlling for time-dependent confounders using standard regression methods can fail. For example, in a longitudinal study, subjects are observed over time and intermediate outcomes are measured. If these intermediate outcomes are also used to determine concomitant treatment assignments, they are both intermediate outcomes of past treatments and confounders of future treatment assignments-the phenomenon known as confounding by indication. Including these variables in standard regression models to control them as confounders does not work since they are also mediators of earlier treatment effects. Similar issues arise in studies with sequential randomization.

We adopt Rubin's (1974) potential outcome framework for estimating causal effects. Potential outcomes are defined as potentially observable outcomes under different treatments or exposure groups. Individual causal effects are defined as comparisons of the potential outcomes for that subject. Only the potential outcome corresponding to the treatment actually assigned is observed for any subject. Therefore we estimate causal effects by imputing the potential outcomes that are not observed.

We propose a robust multiple imputation based approach to causal inference in this setting, called Penalized Spline of Propensity Methods for Treatment Comparison (PENCOMP), which builds on the Penalized Spline of Propensity Prediction method (PSPP) for missing data problems (Little and An, 2004; Zhang and Little, 2009). We first illustrate our approach for the simple case of assessing the causal effect of two treatments, $Z_1 = 0$ or 1 and a function of subject level covariates X_1. Our approach estimates the propensity to be assigned Z_1 given the observed covariates X_1, using a method such as logistic regression appropriate for a binary outcome Z_1. It then estimates regression models for the potential outcome $Y^{Z_1=z_1}$ under each treatment Z_1 on (a) a spline of the logit of the propensity to be assigned that treatment, and (b) other covariates predictive of Y. These regression models are then used to predict the individual outcomes of treatments not assigned. We then draw inferences based on comparisons of the imputed and observed outcomes between treatment groups. Our approach shares some similarities with the MITSS method (Gutman and Rubin, 2015). At the first stage, they partition the subjects into subclasses based on estimated propensity scores and ensures that at least three units from each treatment group are in each subclass. At the second stage, they fit a regression spline with knots fixed at the borders of the subclasses and impute the missing potential outcomes for all the subjects and estimate the causal effects by combining the imputed datasets with Rubin's combining rule. We extend PENCOMP to longitudinal treatments, which is not considered in Gutman and Rubin (2015).

As discussed in Section 2.2 and in Appendix A.1, under the stable unit treatment value (SUTVA), positivity and ignorability assumptions, PENCOMP has a double robustness property, resulting from the balancing property of the propensity score

(Rosenbaum and Rubin, 1983). Specifically, if the relationship between Y and the logit of the propensity score is modeled correctly, the relationship between Y and other covariates can be misspecified without biasing estimates of marginal parameters of interest, namely the marginal means of Y under each treatment. This idea can be generalized to multiple time points, including the situation where variables are both mediators of initial treatments and confounders of later treatments.

Our motivating dataset is from the Multicenter AIDS Cohort study (MACS) (Kaslow et al, 1987). The MACS was started in 1984, and a total of 4,954 gay and bisexual men were enrolled in the study and followed up semi-annually. At each visit, data from physical examination, questionnaires about medical and behavioral history, and blood test results were collected. The primary outcome of interest was the CD4 count, a continuous measure of how well the immune system functions. As HIV infection progresses, the number of CD4 cells decreases, and when the CD4 count was too low, patients started antiretroviral treatment to control the virus and increase the CD4 count. The CD4 count is a time-dependent confounder because it is both an intermediate outcome of past treatments and a confounder of future treatments. The MACS public data set was released by the Center for Analysis and Management of Multicenter AIDS Cohort Study. We used this dataset to analyze the short term (1 year) effects of using antiretroviral treatment on the disease progression between visit 7 and 21, the period after the first antiretroviral drug, zidovudine, was available, and before the advent of highly active antiretroviral therapy (HAART).

Throughout this paper, we consider longitudinal data at $T+1$ discrete time points. For subject i at time $t = 1, \dots, T+1$, let $X_t(i)$ denote the vector of covariates observed, and $Z_t(i)$ the binary treatment indicator. $\bar{X}_t(i)$ and $\bar{Z}_t(i)$ are the covariate and treatment history, up to and including time t. The final outcome of interest $Y(i)$ is observed at time point $T+1$, after the last treatment $Z_T(i)$. For example, in the application, we are interested in estimating the final CD4 count $Y(i)$ after 1 year, i.e, in a three-visit window. $X_t(i)$ contains, for example, the blood count measures, such as CD4 count, at time t, for $t = 1, 2$, and 3. $Y(i) = X_3(i)$ is the final outcome of interest for subject i measured at time $t = 3$, a year from baseline. We compare results from PENCOMP with results from three versions of marginal structural models (MSMs): inverse-probability-treatment-weighted estimators, and augmented IPTW (AIPTW) estimators (Yu and van der Laan, 2006), and g-computation (Robins, 1987). The extended nature of the MACS trials allows comparison of methods on a set of causal estimands, allowing some capability of observing patterns of performance.

The IPTW method controls for confounding by weighting subjects by the inverse of the probability of receiving the observed treatment sequence. The weights in effect create a pseudo-population that is free of treatment confounders, providing the capability for the MSMs to adjust for both time-dependent and time-independent confounders. As for PENCOMP, this method assumes SUTVA, positivity, and ignorability. The IPTW estimators are consistent if the treatment propensity model is correct. On the other hand, g-computation directly simulates counterfactuals of interest of each treatment sequence based on conditional distribution of covariates and outcomes estimated from the data, so provides a consistent estimator of potential outcomes and thus causal effects if all the conditional distributions relating outcomes to covariates are correctly specified (Robins, 1987). Finally, the AIPTW estimator consistently estimates causal effects if the treatment propensity models are correct, or all the conditional distributions relating outcomes to covariates are correctly specified (Scharfstein, Rotnitzky, and Robins 1999; Yu and van der Laan, 2006).

As in g-computation, PENCOMP draws the counterfactuals of interest for each treatment sequence. However, PENCOMP utilizes the observed outcomes and only imputes the missing potential outcome to draw inference on causal effects. Also, PENCOMP has the double robustness property that g-computation lacks, since PENCOMP, like AIPTW, incorporates both the propensity and prediction models.

The compared methods are valid alternative approaches, but we argue that PENCOMP has the following attractive properties. First, it avoids weighting, which may require careful monitoring to avoid a small number of cases receiving very high weights, resulting in highly variable estimates. This is a particularly serious issue with longitudinal data sets with many possible treatment combinations. Second, PENCOMP is conceptually simple since it relies purely on regression models for prediction, with the prediction of potential outcomes addressing the issue of confounding by indication. Third, Bayesian versions of PENCOMP allow for inferences that are not asymptotic, and properly reflect uncertainty in parameter estimates. Saarela et al. (2015) propose an approach to confounding by indication that has Bayesian aspects, but since it involves weighting we regard it as a hybrid approach – see the discussion in Elliott and Little (2015).

The rest of the paper is structured as follows. In Section 2.2, we first briefly introduce PSPP, the method on which PENCOMP was built. We then describe PENCOMP for the simple case of treatment assigned at a single point in time, and for the situation where treatments are assigned at two time points, and intermediate outcomes after the first time point are used to assign treatments at the second time point. In Section 2.3, we briefly describe IPTW, AIPTW and g-computation. In Section 2.4, we compare PENCOMP with the MSM approaches in simulation studies, assessing empirical bias, root mean squared error, 95% confidence interval coverage, and width of confidence intervals. In Section 2.5, we apply our method to the MACS dataset to evaluate the short term effect of antiretroviral treatment on CD4 counts in HIV+ infected patients. In Section 2.6, we presents conclusions and topics for future research. In particular, for simplicity we restrict attention here to the situations with up to two treatment assignments, one at baseline and one at an intermediate time point. In Section 6, we also outline how PENCOMP might be applied in cases with more than two assignments, as when assessing longer term treatment impacts in the

MACS study.

2.2 Penalized Spline of Propensity Methods for Treatment Comparisons

2.2.1 Penalized Spline of Propensity Prediction (PSPP) for Missing Data

Zhang and Little (2009), refining earlier work by Little and An (2004), proposed the following Penalized Spline of Propensity Prediction (PSPP) method for missing-data problems. The objective is to estimate the mean, say μ, of a variable Y with missing values. Let R denote the response indicator for Y, taking the value 1 if Y is observed and 0 if Y is missing. Let $X = (X_1, ..., X_p)$ denote a set of p fully-observed variables. PSPP first estimates the propensity to respond given X, using a method appropriate for a binary outcome such as logistic regression. The method then predicts the missing values of Y using a linear model that includes as predictors a penalized spline of the estimated propensity to respond and a linear function of other covariates X that are predictive of Y.

Assuming the missing data are missing at random (Rubin, 1976; Little and Rubin, 2002), Zhang and Little (2009) show that this method has the following double robustness property for normal linear models: the estimate of μ is consistent if either (a) the regression model for Y is correctly specified, or (b) the model for the propensity to respond and the relationship between Y and the propensity are correctly specified. The latter assumption can be met under relatively weak conditions by regressing Y on the spline of the logit of the propensity, since the spline does not impose strong assumptions on the functional form of the relationship between Y and the propensity. Zhang and Little (2009) and Yang and Little (2015) describe simulation studies suggesting that PSPP compares favorably with alternative doubly-robust methods.

The PSPP method has three principle variants: (a) maximum likelihood (ML)

(PSPP-ML), where parameters are estimated by ML and standard errors computed using the information matrix or the bootstrap; (b) Bayes (PSPP-B), where parameters are drawn from the posterior distribution and inference about μ is based on draws from its posterior distribution; and (c) multiple imputation (MI) (PSPP-MI), where draws of the missing values are multiply imputed, and inferences based on Rubin's (1987) MI combining rules. In the next section we describe adaptations of PSPP for causal inference problems.

2.2.2 PENCOMP for Treatments at a Single Time Point

We first consider PENCOMP in the simple setting of a trial where treatments are assigned at a single time point. Suppressing indexing by subject, $Z_1 \in \{0, 1\}$ denotes assignment to control (0) or treatment (1), Y^{Z_1} denotes the potential outcome associated with a given level of Z_1, measured after treatment Z_1, and X_1 denotes the vector of pretreatment covariates. Our inferential goal is to obtain the marginal average effect of treatment on the outcome, denoted $\Delta = E(Y^1 - Y^0)$, where expectation is taken with respect to a specified population of interest. Figure 1 frames inference about Δ as a missing data problem (Rubin, 1974; Elliott and Little, 2015): note that X_1 and Z_1 are fully observed, but Y^0 is observed only for the n_0 subjects assigned to control, while Y^1 is observed only for the n_1 subjects assigned to treatment. Table 2.1 thus emphasizes the fundamental problem of causal inference (Holland, 1986): since Y^1 and Y^0 are never observed simultaneously, inference about Δ based on directly observing $Y^1 - Y^0$ is impossible.

To make progress in the face of this missing data problem, we make the following three assumptions. First, the stable unit treatment value assumption (SUTVA), assumes $Y = Z_1 Y^{Z_1} + (1 - Z_1) Y^{1-Z_1}$, so that a) the observed outcome Y under a specific treatment is equal to the potential outcome associated with that treatment, and b) the potential outcomes for a given subject are not influenced by the treatment assignment of other subjects (Rubin, 1980; Angrist, Imbens, Rubin, 1996). Next, we make the positivity assumption: $0 < P(Z_1 = 1|X_1) < 1$ for all subjects, so that all subjects have a non-zero probability of being assigned to treatment or control. In practice, this assumption is satisfied by restricting the analysis to treatments with enough cases to make the relevant regressions estimable and excluding subjects with extreme propensity, for example. Finally, we make the ignorable treatment assumption $(Y^1, Y^0) \perp\!\!\!\perp Z_1|X_1$, so that, given covariates, treatment assignment is independent of the potential outcomes of interest, i.e. no unmeasured confounders. The plausibility of SUTVA assumption can usually be assessed in a given context, while the ignorable treatment assumption may or may not be reasonable given the study design and the set of available covariates. Taken together, these assumptions allow the unobserved potential outcomes for subjects receiving treatment $Z_1 = z_1$ in Figure 1 to be imputed using the observed outcomes from subjects receiving treatment $Z_1 = 1 - z_1$. Specifically, we can use an imputation approach with bootstrapping to propagate uncertainty in parameter estimates (Heitjan and Little, 1991).

A potential shortcoming of the prediction approach is that it assumes correct specification of the model for the distribution of the outcome conditional on the covariates. Our proposed PENCOMP method weakens this assumption by exploiting the double robustness property of penalized spline propensity prediction, PSPP (Little and An, 2004; Zhang and Little, 2009). PENCOMP applies the idea of PSPP to the causal inference setting, with the propensity of response replaced by the propensity of treatment assignment and the missing data being the outcomes under unassigned treatments. We estimate the propensity to be assigned to each treatment by a regression method suitable for a categorical outcome, for example by logistic regression if there are two treatments, or polytomous regression if there are more than two treatments. We then predict the potential outcomes for the treatments not assigned to subjects using regression models that include splines on the logit of the propensity to be assigned that treatment and other covariates that are predictive of the outcome; separate models are fitted for each treatment group. Under the assumptions stated above, PENCOMP has a double robustness property for causal effects, as shown in Appendix A.1.

As with PSPP, there are ML, Bayesian and MI versions of PENCOMP: PENCOMP-ML estimates parameters by ML and calculated standard errors using an information matrix or the bootstrap, and PENCOMP-B simulated draws of the parameters and missing observations from their posterior distributions. PENCOMP-MI is analogous to the PSPP-MI algorithm for missing data, and is given as follows:

(a) For $d = 1, \cdots, D$, generate a bootstrap sample $S^{(d)}$ from the original data S by sampling units with replacement, stratified on treatment group. Then carry out steps (b)-(d) for each sample $S^{(d)}$:

(b) Estimate a logistic regression model for the distribution of Z_1 given X_1, with regression parameters γ_{z_1}. Estimate the propensity to be assigned treatment $Z_1 = z_1$ as $\hat{P}_{z_1}(X_1) = \Pr(Z_1 = z_1 | X_1, \hat{\gamma}_{z_1}^{(d)})$, where $\hat{\gamma}_{z_1}^{(d)}$ is the ML estimate of γ_{z_1}. Define $\hat{P^*}_{z_1}$ =log$[\hat{P}_{z_1}(X_1)/(1 - \hat{P}_{z_1}(X_1))]$.

(c) For each $z_1 = 0, 1$, using the cases assigned to treatment group z_1, estimate a normal linear regression of Y^{z_1} on X_1, with mean

$$E(Y^{z_1} | X_1, Z_1 = z_1, \theta_{z_1}, \beta_{z_1}) = s(\hat{P^*}_{z_1} | \theta_{z_1}) + g_{z_1}(X_1; \beta_{z_1}), \tag{2.1}$$

where $s(\hat{P^*}_{z_1} | \theta_{z_1})$ denotes a penalized spline with fixed knots (Eilers and Marx, 1996; Ngo and Wand, 2004; Wand, 2003), with parameters θ_{z_1}, and $g_{z_1}()$ represents a parametric function of other covariates predictive of the outcome, indexed by parameters β_{z_1}. One of the covariates might need to be omitted to avoid collinearity in the covariates in Eq. (2.1). A simple form is to assume linear additive function of the covariates X_1, but models with interactions between the covariates and $\hat{P^*}_{z_1}$ are also allowed. Other forms of splines are possible in Eq. (2.1), as are generalized linear mixed models for non-normal outcomes Y^{z_1}. Note that a different spline function in Eq. (2.1) is fitted for each treatment group, since there is no a priori reason to assume that the relationship between the potential outcomes under different treatment arms and the propensity of treatment assignment is the same.

In particular, for a penalized spline with truncated linear basis, $s(\hat{P}^*{}_{z_1}|\theta_{z_1}) = \theta_0 + \theta_1 \hat{P}^*{}_{z_1} + \sum_{k=1}^{K} \theta_{1k}(\hat{P}^*{}_{z_1} - K_k)_+$, where $K_1, \cdots, K_K$ are fixed knots, and $(\hat{P}^*{}_{z_1} - K_k)_+ = (\hat{P}^*{}_{z_1} - K_k)$ if $\hat{P}^*{}_{z_1} > K_k$; and $= 0$ if $\hat{P}^*{}_{z_1} \leq K_k$.

In the linear additive form for g, define the design matrices $C_1 = [1, \hat{P}^*{}_{z_1}, x_1]$, $C_2 = [(\hat{P}^*{}_{z_1} - K_1)_+, \cdots, (\hat{P}^*{}_{z_1} - K_K)_+]$, and $C = [C_1, C_2]$. Then spline model can be expressed as a linear mixed model (Wand, 2003),

$$Y^{z_1} = C_1\beta + C_2\theta + \epsilon, \begin{bmatrix} \theta \\ \epsilon \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\theta^2 I & 0 \\ 0 & \sigma_\epsilon^2 I \end{bmatrix} \right), \tag{2.2}$$

where $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ denote fixed effects, and $\theta = (\theta_{11}, \cdots, \theta_{1K})$ are random basis coefficients. REML estimates of the parameters of this model can be easily fitted in statistical software, such as PROC MIXED in SAS or lme in R. The fitted values of Y^{z_1} are $\hat{y}^{z_1} = C(C^TC + \hat{\lambda}D)^{-1}C^Ty$, where $\hat{\lambda} = \hat{\sigma}_\epsilon^2/\hat{\sigma}_\theta^2$ is the REML estimator of λ and

$$D = \begin{pmatrix} 0_{(p+1)\times(p+1)} & 0 \\ 0 & I_{K\times K} \end{pmatrix}$$

(d) For $z_1 = 0, 1$, impute the values of Y^{z_1} for subjects in treatment group $1 - z_1$ in the original data set with draws from the predictive distribution of Y^{z_1} given X_1 from the regression in (c), with ML estimates $\hat{\theta}_{z_1}^{(d)}, \hat{\beta}_{z_1}^{(d)}$ substituted for the parameters $\theta_{z_1}, \beta_{z_1}$, respectively. Let $\hat{\Delta}^{(d)}$ and $W^{(d)}$ denote the difference in treatment means and associated pooled variance estimate, based on the observed and imputed values of Y in each treatment group.

(e) The MI estimate of Δ is then $\bar{\Delta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\Delta}_d$, and the MI estimate of the variance of $\bar{\Delta}_{\mathrm{D}}$ is $T_{\mathrm{D}} = \bar{W}_D + (1+1/D)B_D$, where $\bar{W}_D = \sum_{d=1}^{D} W^{(d)}/D, B_D = \sum_{d=1}^{D}\left(\hat{\Delta}^{(d)} - \bar{\Delta}_D\right)^2/(D-1)$. The estimate Δ is t distributed with degree of freedom v, $(\Delta - \bar{\Delta}_D)T_D^{\frac{-1}{2}} \sim t_v$, where $v = (D-1)(1+\bar{W}_D/((D+1)\times B_D))^2$.

We apply this PENCOMP-MI method in the application and simulations in this article.

Table 2.1: Observed and missing outcomes for treatment at a single time point

Subjects	X_1	Z_1	Y^0	Y^1
1	■	0	■	?
2	■	0	■	?
$\cdots$	■	0	■	?
n_0	■	0	■	?
n_0+1	■	1	?	■
$\cdots$	■	1	?	■
$n = n_0 + n_1$	■	1	?	■

2.2.3 PENCOMP with Longitudinal Treatment Assignments

We now consider a longitudinal study with treatments assigned at multiple time points $t = 1, \ldots, T$. Suppressing indexing by subject, let $\bar{X}_t$ and $\bar{Z}_t$ denote the covariate and treatment history, respectively, up to and including time point t. Let $X_{t+1}^{\bar{Z}_t}$ denote the potential intermediate outcome under treatment regime $\bar{Z}_t = (Z_1, \cdots, Z_t)$. Let $Y^{\bar{Z}_T}$ denote the final potential outcome under the entire treatment regime $\bar{Z}_T = (Z_1, \cdots, Z_T)$, measured at time point $T+1$ after the assignment of last treatment Z_T. Assume at each time $t \geq 2$, the intermediate outcome X_t is both an outcome of treatment Z_{t-1} and confounder for treatment Z_{t+1}. Supposed we want to estimate the overall treatment effects as a function of treatment regime $\bar{Z}_T$, relative to $\bar{Z}_T'$. To estimate causal effect $\Delta_{\bar{Z}_T} = E(Y^{\bar{Z}_T}) - E(Y^{\bar{Z}_T'})$, we make the following assumptions.

1) SUTVA (Angrist, Imbens and Rubin, 1996) states that a) the observed outcomes under a specific treatment regime is equal to the potential outcomes associated

with that treatment regime, and b) the potential outcomes for a given subject are not influenced by the treatment assignments of other subjects (Rubin, 1980; Angrist, Imbens, Rubin, 1996)

2) Positivity states that each subject has a positive probability of being assigned to each treatment z_t at each time point t: $0 < \Pr(Z_t = z_t | \bar{X}_{t-1}, \bar{Z}_{t-1}) < 1$.

3) Sequential ignorable treatment assumption states that

$$(Y^{\bar{Z}_T}, X_{t+1}^{\bar{Z}_t}) \perp\!\!\!\perp Z_t | (\bar{Z}_{t-1}, \bar{X}_t)$$

for every $\bar{z}_T \in A$: at every time t, where A denote the set of all possible treatment combinations, that is, at each time t, treatment assignment Z_t is as if randomized conditional on all the past treatment and covariate history.

For simplicity, we illustrate a longitudinal study with two time points and binary treatments. In such setting, there are four possible treatment regimes. Let $X_2^{Z_1}$ denote the potential intermediate outcome if subject received treatment Z_1, and $Y^{\bar{Z}_2}$ the potential outcome of interest if subject received treatment regime $\bar{Z}_2$. Our inferential goal is to estimate the overall treatment effects as a function of Z_1 and Z_2, relative to no treatment at both time points, namely $\Delta_{\bar{z}_2} = E(Y^{\bar{Z}_2} - Y^{00})$, where expectation is taken with respect to a specified population of interest. In this case, we are interested in inference about Δ_{11},Δ_{10}, and Δ_{01}. Table 2.2 frames inference about the causal effects as a missing-data problem (Rubin, 1974; Elliott and Little, 2015). In this setting, values of the intermediate and final outcomes are only observed for the treatment combination actually assigned. Thus, for example, values of X_2^1 are missing for cases assigned to $Z_1 = 0$, and values of Y^{10}, Y^{01} and Y^{11} are missing for cases assigned to $(z_1, z_2) = (0, 0)$; and similarly for the other treatment combinations.

The missing values of the intermediate outcomes X_2^0 and X_2^1 are imputed using the method described in Section 2.2.2. Conditional on the values of X_1, Z_1 and the observed or imputed values of X_2, the propensity that $Z_2 = 1$ given $\bar{X}_2, Z_1$ is estimated based on a logistic regression of Z_2 on $\bar{X}_2, Z_1$. The missing values of Y^{jk} are draws from the regression model of Y^{jk} on $\bar{X}_2, \bar{Z}_2$, and a spline on the logit of the propensity score. A distinct regression model is fitted for each outcome Y^{jk}. More specifically, the steps for PENCOM-MI are as follows:

(a) For $d = 1, \cdots, D$, generate a bootstrap sample $S^{(d)}$ from the original data S by sampling units with replacement, stratified on treatment group. Then carry out steps (b)-(g) for each sample d:

(b) Estimate a logistic regression model for the distribution of Z_1 given baseline covariates X_1, with regression parameters γ_{z_1}. Estimate the propensity to be assigned treatment $Z_1 = z_1$ as $\hat{P}_{z_1}(X_1) = \Pr(Z_1 = z_1 | X_1, \hat{\gamma}_{z_1}^{(d)})$, where $\hat{\gamma}_{z_1}^{(d)}$ is the ML estimate of γ_{z_1}. Define $\hat{P}^*_{z_1}$ =log $[\hat{P}_{z_1}(X_1)/(1 - \hat{P}_{z_1}(X_1))]$.

(c) Using the cases assigned to treatment group $Z_1 = z_1$, estimate a normal linear regression of $X_2^{z_1}$ on X_1, with mean

$$E(X_2^{z_1} | X_1, Z_1 = z_1, \theta_{z_1}, \beta_{z_1}) = s(\hat{P^*}_{z_1} | \theta_{z_1}) + g_{z_1}(X_1; \beta_{z_1}), \qquad (2.3)$$

where $s(\hat{P^*}_{z_1} | \theta_{z_1})$ denotes a penalized spline with fixed knots with parameters θ_{z_1}, and $g_{z_1}()$ represents a parametric function of other predictors of the outcome, indexed by parameters β_{z_1}. As for PSPP, one of the covariates might be omitted to avoid collinearity in the covariates in Eq. (2.3). Note that a different spline model of the form (2.3) is fitted for each treatment regimen.

(d) For $z_1 = 0, 1$, impute the values of $X_2^{z_1}$ for subjects in treatment group $1 - z_1$ in the original data set with draws from the predictive distribution of $X_2^{z_1}$ given X_1 from the regression in (c), with ML estimates $\hat{\theta}_{z_1}^{(d)}, \hat{\beta}_{z_1}^{(d)}$ substituted for the parameters $\theta_{z_1}, \beta_{z_1}$.

(e) Estimate a logistic regression model for the distribution of Z_2 given $X_1, Z_1, X_2 =$

(X_2^0, X_2^1), with regression parameters γ_{z_2} and missing values of X_2 imputed from step (d). Estimate the propensity to be assigned treatment $Z_2 = z_2$ given $Z_1, \bar{X}_2$ as $\hat{P}_{z_2}(\bar{X}_2, Z_1) = \Pr(Z_2 = z_2 | \bar{X}_2, Z_1 = z_1, \hat{\gamma}_{z_2}^{(d)})$, where $\hat{\gamma}_{z_2}^{(d)}$ is the ML estimate of γ_{z_2}. The probability of treatment regimen $(Z_1 = z_1, Z_2 = z_2)$ is denoted as $\hat{P}_{\bar{z}_2} = \hat{P}_{z_1}(X_1)\hat{P}_{z_2}(\bar{X}_2, Z_1)$, and define $\hat{P}_{\bar{z}_2}^{*}$ =log$[\hat{P}_{\bar{z}_2}/(1 - \hat{P}_{\bar{z}_2})]$.

(f) Using the cases assigned to treatment group (z_1, z_2), estimate a normal linear regression of $Y^{\bar{z}_2}$ on $\bar{X}_2, \bar{Z}_2$, with mean

$$\begin{aligned} &E(Y^{\bar{z}_2} | \bar{X}_2, Z_1 = z_1, Z_2 = z_2, \theta_{\bar{z}_2}, \beta_{\bar{z}_2}), \\ &= s(\hat{P}_{\bar{z}_2}^{*} | \theta_{\bar{z}_2}) + g_{\bar{z}_2}(\bar{X}_2, \bar{Z}_2; \beta_{\bar{z}_2}) \end{aligned} \tag{2.4}$$

where $s(\hat{P}_{\bar{z}_2}^{*} | \theta_{\bar{z}_2})$ denotes a penalized spline with fixed knots with parameters $\theta_{\bar{z}_2}$, and $g_{\bar{z}_2}()$ represents a parametric function of other predictors indexed by parameters $\beta_{\bar{z}_2}$. One of the covariates might need to be omitted from $g_{\bar{z}_2}()$ to avoid collinearity in the covariates. Note that a distinct model of form (2.4) is fitted for each treatment regimen.

(g) For each combination of $\bar{z}_2 = (z_1, z_2)$, impute the values of $Y^{\bar{z}_2}$ for subjects not assigned this treatment combination in the original data set with draws from the predictive distribution of $Y^{\bar{z}_2}$ in (f), with ML estimates $\hat{\theta}_{\bar{z}_2}^{(d)}, \hat{\beta}_{\bar{z}_2}^{(d)}$ substituted for the parameters $\theta_{\bar{z}_2}, \beta_{\bar{z}_2}$. Let $\hat{\Delta}_{jk}^{(d)}$, $(j, k) = (1, 1), (1, 0)$ and (0,0) denote the average treatment effects, with associated pooled variance estimates $W_{jk}^{(d)}$, based on the observed and imputed values of Y for each treatment regimen.

(h) The MI estimate of Δ_{jk} is then $\bar{\Delta}_{jkD} = \sum_{d=1}^{D} \hat{\Delta}_{jk}^{(d)}$, and the MI estimate of the variance of $\bar{\Delta}_{jkD}$ is $T_{\mathrm{D}} = \bar{W}_{jkD} + (1+1/D)B_{jkD}$, where $\bar{W}_{jkD} = \sum_{d-1}^{D} W_{jk}^{(d)}/D, B_{jkD} = \sum_{d=1}^{D} \left(\hat{\Delta}_{jk}^{(d)} - \bar{\Delta}_{jkD}\right)^2/(D-1)$. As described in (e) of single treatment setting, draw inference about Δ_{jk} by assuming a t-distribution.

Table 2.2: Observed and missing intermediate and final outcomes for treatment at two time points

Subjects	X_1	Z_1	X_2^0	X_2^1	Z_2	Y^{00}	Y^{01}	Y^{10}	Y^{11}
1	■	0	■	?	0	■	?	?	?
$\cdots$	■	0	■	?	0	■	?	?	?
n_{00}	■	0	■	?	0	■	?	?	?
$n_{00}+1$	■	0	■	?	1	?	■	?	?
$\cdots$	■	0	■	?	1	?	■	?	?
$n_0 = n_{00} + n_{01}$	■	0	■	?	1	?	■	?	?
$n_0 + 1$	■	1	?	■	0	?	?	■	?
$\cdots$	■	1	?	■	0	?	?	■	?
$n_0 + n_{10}$	■	1	?	■	0	?	?	■	?
$n_0 + n_{10} + 1$	■	1	?	■	1	?	?	?	■
$\cdots$	■	1	?	■	1	?	?	?	■
$n = n_0 + n_{10} + n_{11}$	■	1	?	■	1	?	?	?	■

In a longitudinal study with more than two time points, the procedures are similar to those described in the two time points setting. PENCOMP imputes the first missing intermediate outcomes X_2 first and continues forward to the final outcome Y. Specifically, to impute the missing intermediate outcomes $X_{t+1}^{\bar{z}_t}$ for the subjects whose treatment sequence did not match $\bar{z}_t$, we draw values from a mean model of $E(X_{t+1}^{\bar{z}_t}|\bar{X}_t, \bar{Z}_t = \bar{z}_t, \theta_{\bar{z}_t}, \beta_{\bar{z}_t}) = s_{x_{t+1}}(\hat{P}^*_{\bar{z}_t}; \theta_{\bar{z}_t}) + g_{\bar{z}_t}\left[X_1, \cdots, X_t; \beta_{\bar{z}_t}\right]$, where X_t can be observed or imputed in the previous steps, and $\hat{P}^*_{\bar{z}_t}$ =log$[\prod_{k=1}^{t} P(Z_k = z_k|\bar{Z}_{k-1} = \bar{z}_{k-1}, \bar{X}_k)/(1 - \prod_{k=1}^{t} P(Z_k = z_k|\bar{Z}_{k-1} = \bar{z}_{k-1}, \bar{X}_k))]$, where $\prod_{k=1}^{t} P(Z_k = z_k|\bar{Z}_{k-1} = \bar{z}_{k-1}, \bar{X}_k)$ represents the propensity of being assigned the treatment sequence $\bar{z}_t$ conditional on the past treatment and covariate history. As before, the propensity of being assigned z_k at time $t = k$, $P(Z_k = z_k|\bar{Z}_{k-1} = \bar{z}_{k-1}, \bar{X}_{k-1}, \gamma_{z_k})$ can be estimated based on a logistic regression model. Under the assumptions stated above in section 2.2.3, PENCOMP has a double robustness property for causal effects in a longitudinal study setting. The proof is outlined in Appendix A.1. The marginal mean from the imputation model is consistent if

1) All the prediction models for the intermediate and final outcomes at each time point $t = 1, \cdots, T+1$, conditional on the covariate and treatment history, denoted as $g_{\bar{z}_t}$, are correctly specified. OR

2) The propensity models are correctly specified, and the relationship between X_{t+1} and $\hat{P}^*_{\bar{z}_t}$ are correctly specified at each time point $t = 1, \cdots, T+1$. Note $Y = X_{T+1}$. Again, this assumption can be weakened by assuming only a smooth functional form, such as a penalized spline as in PENCOMP.

2.2.4 Restricting cases in a treatment comparison to reduce disparity in the distribution of estimated assignment propensities

The positivity assumption requires that cases have a propensity to be assigned to any of the compared treatments that lie between zero and one. However, when there are extreme propensity scores, the propensity score distributions tend to have limited overlap. Some techniques have been proposed to address this issue. Cochran and Rubin (1973) suggest caliper matching when some units are dropped due to poor match quality. Rubin (1977) suggests dropping units with covariate values that have either no treated or no control and estimate causal effects for the range of covariate values that have both treated and control units. Dehejia and Wahba (1999) drop control units whose estimated propensity scores are less than the smallest estimated propensity scores among the treated when estimating the average treatment effects for the treated. Crump et al (2009) propose a minimum variance approach to select an optimal subpopulation for which the estimated causal effects have the least variance, where the optimal subpopulation is obtained by excluding cases with propensity scores outside of a range $[\alpha, 1-\alpha]$. Gutman and Rubin (2015) propose restricting included cases to the overlap region of estimated propensity scores between the treatment groups.

Comparison of the performance of those methods for dealing with limited overlap, especially in the longitudinal treatments where lack of overlap can be very severe, is a topic for future research. However, here we restrict the overlap region to avoid extrapolation of the prediction model outside the range of estimated propensities and extend the overlap rule to longitudinal treatments. To illustrate in the general case of $\Delta_{\overline{Z}_T}$, relative to the null treatment regime 0_T, at a given time $1 \leq t \leq T$, we first obtain the set of observations $A_t = A_{\overline{Z}_t}$ such as

$$A_{\overline{Z}_t} = \left\{ i : \{\overline{z}_{ti} = \overline{Z}_t, \overline{z}_{ti} \neq \overline{Z}_t\}, \min_{j:\overline{z}_{ti}=\overline{Z}_t} (\hat{P^*}_{j,\overline{Z}_t}) \leq \hat{P^*}_{i,\overline{Z}_t} \leq \max_{j:\overline{z}_{ti}=\overline{Z}_t} (\hat{P^*}_{j,\overline{Z}_t}) \right\}$$

A_t corresponds to the set of observations that have an estimated propensity score for treatment regime $\overline{Z}_t$ that lies within the range of the observed propensities of subjects who actually received $\overline{Z}_t$. We then obtain $B_t = B_{0_t}$ as

$$B_{0_t} = \left\{ i : \{\overline{z}_{ti} = 0_t, \overline{z}_{ti} \neq 0_t\}, \min_{j:\overline{z}_{ti}=0_t} (\hat{P^*}_{j,0_t}) \leq \hat{P^*}_{i,0_t} \leq \max_{j:\overline{z}_{ti}=0_t} (\hat{P^*}_{j,0_t}) \right\}$$

B_t corresponds to the set of observations that have an estimated propensity score for null treatment regime 0_t that lies within the range of the observed propensities of subjects who actually received the treatment regime 0_t. Finally, we restrict our analysis to the set of observations given by $A_1 \cap B_1 \cap \cdots \cap A_T \cap B_T$. In this way we assure that all observations used in the analysis have a common set of overlapping estimated propensities that are actually observed in the data.

2.3 G-computation, IPTW and AIPTW

2.3.1 G-computation

In a longitudinal treatment scenario with $T+1$ time points, let $O = (\bar{X}_T, \bar{Z}_T, Y)$ denote the observed data, as above. The likelihood of the observed data can be factored into two components $P(O) = Q_0 g_0$, where $Q_0 = P(Y|\bar{X}_T, \bar{Z}_T = \bar{z}_T) \times \prod_{t=1}^{T} P(X_t|\bar{X}_{t-1}, \bar{Z}_{t-1} = \bar{z}_{t-1})$ and $g_0 = \prod_{t=1}^{T} P(Z_t = z_t|\bar{Z}_{t-1} = \bar{z}_{t-1}, \bar{X}_{t-1})$. Under SUTVA, positivity and ignorability assumptions, for a fixed treatment regime $\bar{z}_T = (z_1, \cdots, z_T)$, $E(Y^{\bar{z}_T}) = \sum_{X_1,\cdots,X_T} E(Y|\bar{X}_T, \bar{Z}_T = \bar{z}_T) \times P(X_1) \times P(X_2|X_1, Z_1 =$

$z_1) \cdots \times P(X_T|\bar{X}_{T-1}, \bar{Z}_{T-1} = \bar{z}_{T-1})$. For continuous Xs, the expectation can be solved by using a Monte-Carlo algorithm (Robins 1987). For example, in a two-time point setting with binary treatment at each time point, there are four possible treatment combinations $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$. First, draw baseline covariate x_1^* from the empirical distribution of X_1. Set $Z_1 = z_1$ and generate a draw x_2^* from $\hat{P}(X_2|X_1 = x_1^*, Z_1 = z_1)$. Then setting $Z_1 = z_1$ and $Z_2 = z_2$, generate draws y^* from $\hat{P}(Y|X_1 = x_1^*, Z_1 = z_1, X_2 = x_2^*, Z_2 = z_2)$. Repeat the procedure many times to get the marginal distribution of the outcome of interest under each counterfactual treatment history. The marginal treatment effects between $(Z_1 = z_1, Z_2 = z_2)$ and $(Z_1 = z_1^{'}, Z_2 = z_2^{'})$ can be estimated by the sample mean of the draws y^* under $(Z_1 = z_1, Z_2 = z_2)$ and the sample mean of the draws under $(Z_1 = z_1^{'}, Z_2 = z_2^{'})$. If all the models are correctly specified, the g-computation estimator is consistent.

2.3.2 Inverse Probability Treatment Weighted Estimator

The IPTW estimator provides a consistent estimator of the parameter of the marginal mean of $E(Y^{\bar{z}_T}) = f(\bar{z}_T, \beta)$ by solving the estimating equations:

$$D_{IPTW}(O|\beta, g_0) = \frac{df(\bar{z}_T, \beta)}{d\beta} \left\{ \prod_{t=1}^{T} P(Z_t = z_t|\bar{Z}_{t-1} = \bar{z}_{t-1})/g_0 \right\} (Y^{\bar{z}_T} - f(\bar{z}_T, \beta)) = 0,$$

where g_0 is defined in Section 2.3.1.

Under the assumptions stated in Section 2.2, the IPTW estimator is consistent if the propensity score models that make up g_0 are correctly specified. For example, in a two time points treatment, the marginal structural model of interest is $E(Y^{\bar{Z}_2}) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2$. Let $h(\bar{Z}_2) = \frac{dE(Y^{\bar{Z}_2})}{d\beta} P(Z_1 = z_1) P(Z_2 = z_2|Z_1 = z_1)$, where $P(Z_2 = z_2|Z_1 = z_1)$ can be modeled as a logistic regression conditional on past treatment history. We solve the following estimating equation:

$$D_{IPTW}(O|\beta, g_0) \; = \; \{h(\bar{Z}_2)/g_0\}\Big(Y^{\bar{Z}_2} \; - \; (\beta_0 \; + \; \beta_1 Z_1 \; + \; \beta_2 Z_2 \; + \; \beta_3 Z_1 Z_2)\Big) \; = \; 0,$$

where $g_0 = P(Z_1 = z_1|X_1)P(Z_2 = z_2|Z_1 = z_1, \bar{X}_2)$

2.3.3 Augmented Inverse Probability Treatment Weighted Estimator

With treatments assigned at two time points, the AIPTW estimator is obtained by solving the following estimating equation.

$$\begin{aligned} D_{AIPTW}(O|\beta, g_0, Q_o) = D_{IPTW}(O|\beta, g_0) - \\ \sum_{t=1}^{t=2} E_{Q_0,g_0}[D_{IPTW}(O|\beta, g_0)|\bar{Z}_t, \bar{X}_t] - E_{Q_0,g_0}[D_{IPTW}(O|\beta, g_0)|\bar{X}_t] = 0 \end{aligned}$$

Under the assumptions stated in Section 2.2, the AIPTW estimator is consistent if 1) the propensity score models are correctly specified or 2) all the conditional distributions of the covariates and the outcomes are correctly-specified (Scharfstein, Rotnitzky, and Robins 1999).

See appendix A for more detailed descriptions of our implementations of IPTW and AIPTW.

2.4 Simulation Studies

2.4.1 Introduction

We conducted simulations to assess the finite sample performance of PENCOMP-MI, compared with g-computation, IPTW and a Monte-Carlo AIPTW method (Yu and van der Laan, 2006) in estimating treatment effects.

Our simulation study design considered five factors: a single point in time and a two-point in time treatment with the second treatment confounded by indication; three levels of confounding (low, moderate and high); linear vs. non-linear regression models for the outcomes; three sample sizes (200, 500 and 1000); and two forms of model misspecification. We considered three sets of models for the AIPTW and PENCOMP estimators: (A) correctly-specified propensity and prediction models, (B) a correctly-specified propensity model only, and (C) a correctly-specified prediction model only. The case with both models misspecified was not considered since none of the compared methods yields consistent estimates in that case, and conclusions from particular simulation conditions have limited generalizability. For the IPTW estimator, there is no prediction model so we considered only a correctly-specified or misspecified propensity model. One thousand simulated data sets were created for sample size of 500, but to reduce computation burden, only 500 simulated data sets were used for sample sizes of 200 and 1000 in the two-time point situation. For PENCOMP, 200 complete datasets were created to estimate treatment effects and the associated standard errors and confidence intervals. For IPTW and g-computation, 500 bootstrap samples were used to estimate standard errors and 95% confidence intervals. For AIPTW, 500 bootstraps were used to calculate standard errors and confidence intervals for sample size of 500, but to reduce computational burden, only 200 bootstraps were used for sample size 200 and 1000 in the two-time point case. For the single time point treatment, 35 equally spaced knots were used, and for the two-time point treatment, 15 equally spaced knots were used. A truncated linear basis was used in both.

We compared performance in terms of bias, RMSE, average 95% confidence interval width, and 95% confidence interval (non) coverage. To provide a more interpretable scale for bias and RMSE, we present the ratio of the bias and RMSE to the RMSE of IPTW for the correct propensity model. We also scaled the 95% confidence interval width to the width of IPTW with the correct propensity model. In the main paper, we presented the results for RMSE and 95% non-coverage. The complete results are included in Appendix A.3.

2.4.2 Simulations for a Treatment Assigned at a Single Time Point

Our simulation scenarios are the same as those in Glynn and Quinn (2010). Each simulated data set contains five variables: X_{1a}, X_{1b} and X_{1c} are baseline covariates, independently and normally distributed as $N(0,1)$. The treatment is denoted as Z_1 and is Bernoulli distributed with treatment assignment probability that depends on X_{1a} and X_{1b}. The outcome of interest is denoted as Y and is normally distributed with a mean that depends only on X_{1b} and X_{1c} and a variance of 1, so that X_{1b} confounds treatment and outcome. We considered two outcome models: linear and nonlinear. The correctly-specified and misspecified treatment assignment mechanism and the outcome models are described in Table 2.3. The data were generated based on the true models shown in Table 2.3. The treatment effects under linear and nonlinear outcome models were 5 and 9, respectively.

Table 2.3: Single Time Point Treatment Simulation Scenarios: $\gamma = c(1.5, 1.5, 0.75)$, $c(1, 1, 0.5)$, $c(0.1, 0.1, 0.05)$ corresponds to high, moderate, and low confounding, respectively. The true coefficients associated with each model are listed next to each model.

	Linear Outcome	
True	$logit(P(Z_1 = 1\|\bar{X}, \gamma) = \gamma_1 X_{1a} + \gamma_2 X_{1b} + \gamma_3 X_{1a} X_{1b}$ $E(Y_1\|\bar{X}, \beta_1) = \beta_{10} + \beta_{11} X_{1b} + \beta_{12} X_{1c}$ $E(Y_0\|\bar{X}, \beta_0) = \beta_{00} X_{1b} + \beta_{01} X_{1c}$	$\gamma = c(\gamma_1, \gamma_2, \gamma_3) = c(1.5, 1.5, 0.75)$, $c(1, 1, 0.5)$, or $c(0.1, 0.1, 0.05)$ $\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12}) = (5, 3, 1)$ $\beta_0 = (\beta_{00}, \beta_{01}) = (1, 1)$
Misspecified	$logit(P(Z_1 = 1\|\bar{X}, \lambda) = \lambda_0 + \lambda_1 X_{1a}$ $E(Y_1\|\bar{X}, \alpha_1) = \alpha_{10} + \alpha_{11} X_{1c}$ $E(Y_0\|\bar{X}, \alpha_0) = \alpha_{00} + \alpha_{01} X_{1c}$	
	NonLinear Outcome	
True	$logit(P(Z_1 = 1\|\bar{X}, \gamma)) = \gamma_1 X_{1a} + \gamma_2 X_{1b} + \gamma_3 X_{1a} X_{1b}$ $E(Y_1\|\bar{X}, \beta_1) = \beta_{10} + \beta_{11} X_{1b} + \beta_{12} X_{1c} + \beta_{13} X_{1b}^2 + \beta_{14} X_{1c}^2$ $E(Y_0\|\bar{X}, \beta_0) = \beta_{00} X_{1b} + \beta_{01} X_{1c}$	$\gamma = c(\gamma_1, \gamma_2, \gamma_3) = c(1.5, 1.5, 0.75)$, $c(1, 1, 0.5)$, or $c(0.1, 0.1, 0.05)$ $\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}) = (5, 3, 1, 2, 2)$ $\beta_0 = (\beta_{00}, \beta_{01}) = (1, 1)$
Misspecified	$logit(P(Z_1 = 1\|\bar{X}, \lambda)) = \lambda_1 X_{1a}$ $E(Y_1\|\bar{X}, \alpha_1) = \alpha_{10} + \alpha_{11} X_{1c}$ $E(Y_0\|\bar{X}, \alpha_0) = \alpha_{00} + \alpha_{01} X_{1c}$	

Results for sample size 500 are shown in Figures 2.1-2.2 and Tables A.5-A.8 in Appendix A.3. The RMSEs of the methods are shown in Figure 2.1, expressed as a proportion of the RMSE of IPTW with a correct propensity model. Both AIPTW and PENCOMP generally had substantially lower RMSEs than IPTW, especially for the linear outcome, with the ratio of RMSE to RMSE of IPTW with a correct propensity model varying from 0.3 to 1 and with most of the ratios below 0.8. AIPTW and PENCOMP had similar RMSE under low confounding or correctly specified prediction models in the linear model, but PENCOMP had substantially lower RMSE than AIPTW when the prediction model was misspecified and as the degree of confounding increased and the weights became more variable. In the nonlinear outcome model, PENCOMP and AIPTW had similar RMSE under all scenarios. Lastly, PENCOMP had similar RMSEs to g-computation when the prediction model was correctly specified.

The 95% confidence interval non-coverage rates are shown in Figure 2.2. PENCOMP generally had close to nominal coverage of 95% when the prediction model was correctly specified, and conservative (over-) coverage when the prediction model was misspecified, especially for linear outcome model, with coverage rates close to 99%. One exception is that in the nonlinear model under high confounding, PENCOMP slightly undercovered, with a coverage rate of 90%. On the other hand, AIPTW and IPTW displayed more evidence of undercoverage, especially in the linear outcome model under high confounding, with coverage rates less than 90%.

Table A.5 in the Appendix displays the empirical bias of the three methods as a fraction of RMSE of IPTW with a correct propensity score model. The IPTW estimator had close to zero empirical bias when the propensity model was correctly specified, but was substantially biased, with relative bias greater than 20% under high confounding, when the propensity model was misspecified. G-computation had negligible bias, when the prediction model was correct, but had substantial bias, with relative bias over 20% in some scenarios, when the prediction model was misspecified. Both AIPTW and PENCOMP had small empirical bias, especially when the prediction model was correctly specified or when confounding was low. The empirical biases tended to be larger when the prediction model was misspecified, with AIPTW having slightly less empirical bias than PENCOMP in some scenarios. In general, empirical bias for both PENCOMP and AIPTW represented a small fraction of the RMSE of IPTW with a correct propensity score model.

The 95% confidence width are shown in Table A.8 in the Appendix. When the prediction model was correctly specified, both AIPTW and PENCOMP had similar confidence interval widths, which were smaller than those for IPTW. However, when the prediction model was misspecified, PENCOMP tended to have a wider confidence interval under low confounding, compared to AIPTW and IPTW with the correct propensity model, a finding consistent with the over-coverage of PENCOMP in Figure 2.2. As confounding increased, both PENCOMP and AIPTW had similar confidence interval widths as IPTW with the correct propensity model in the linear outcome. In the nonlinear outcome, with the prediction model misspecified, both PENCOMP and AIPTW had similar interval widths than IPTW with correct propensity model. In addition, for all the estimators, the confidence intervals were wider when the prediction model was misspecified.

The simulation results for sample sizes 200 and 1000 are given in Table A.1-A.4 and A.9-A.12 in Appendix A. As one would expect, the empirical biases of correctly-specified IPTW, AIPTW and PENCOMP estimators decreased with increasing sample size, whereas the bias of the misspecified IPTW estimator was less dependent on sample size. PENCOMP's relative gains in RMSE over the other methods tended to increase with increasing sample size, especially under moderate or high confounding. Interval widths for PENCOMP decreased more dramatically when the prediction model was misspecified as sample size increased. Confidence coverage of the methods tended to be closer to nominal as sample size increased.

In summary, IPTW performed worse than AIPTW and PENCOMP, particularly when confounding was high, since the doubly-robust estimators rely on both the prediction model and the propensity model. PENCOMP had comparable performance to AIPTW when confounding was low and the prediction model was correct, and tended to perform better than AIPTW when the prediction model was misspecified and weights were highly variable.

Logit-transforming the propensity scores before fitting the PENCOMP model works well in general, since the weight distribution is typically highly skewed, and the logit transformation yields a more uniform distribution of propensity scores for the fitting of the spline models. However, in cases where the weight distribution is more uniformly distributed on the original scale, the logit transformation can actually skew the weight distribution, leaving data points thinly distributed in some regions so that it becomes harder to fit the model and make predictions. This is the cause of the undercoverage of PENCOMP in the nonlinear model under high confounding. In practice, examining the distribution of the propensity score with and without the logit transformation is recommended. This issue becomes moot as sample size increases, allowing for sufficient data to be available to fit the splines, as indicated by the fact that coverage is approximately correct for the nonlinear model under high confounding with sample sizes of 1000 (see Table A.11 in the Appendix). Lastly, including a covariate that is a strong predictor of the treatment but not of the outcome can lead to bias and inefficiency.

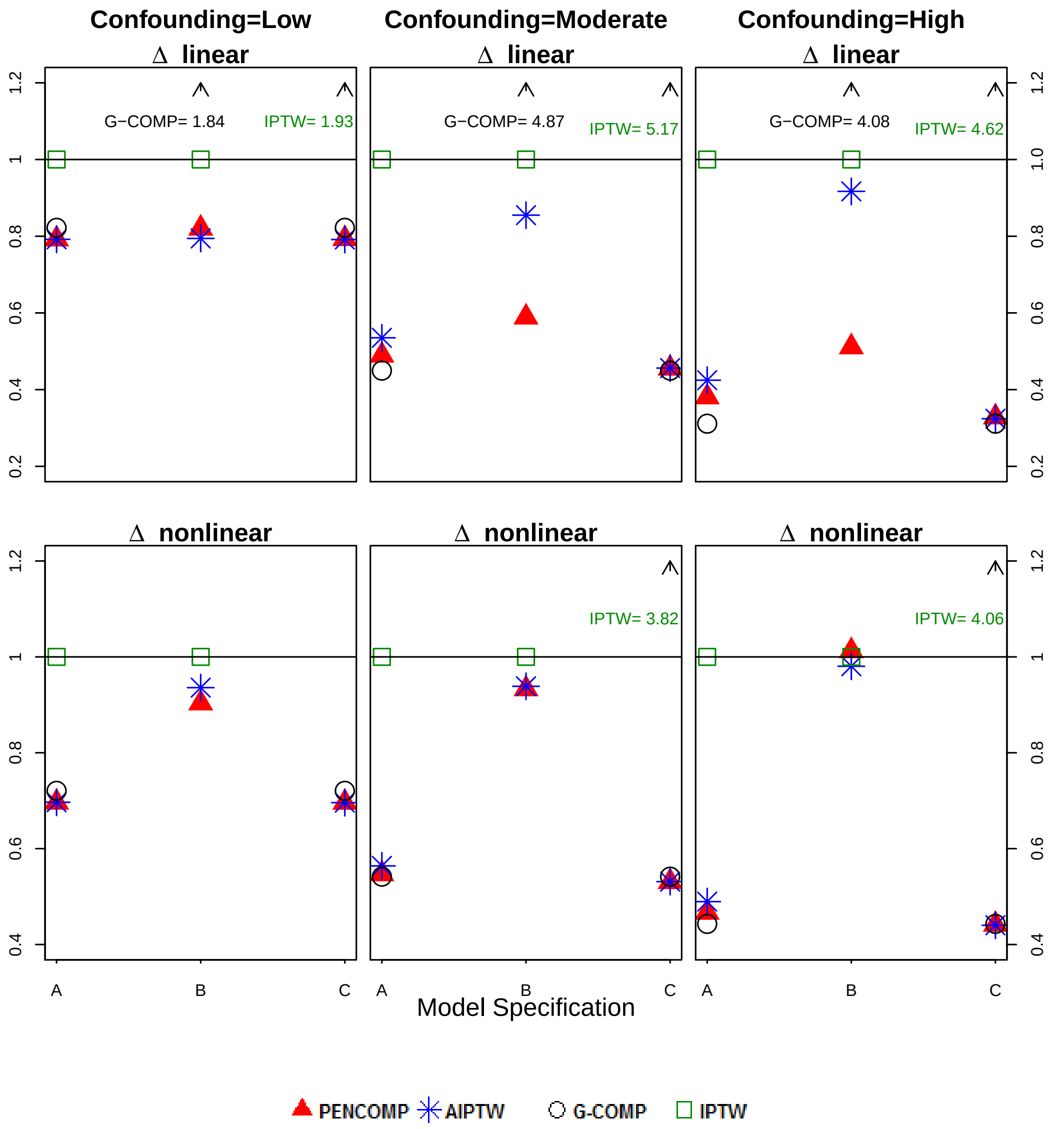

Figure 2.1: Ratio of RMSE over RMSE of IPTW(A) with correct propenisty score model across four methods-PENCOMP, AIPTW, IPTW and g-computation for treatment effect Δ in a linear and nonlinear outcome model. (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

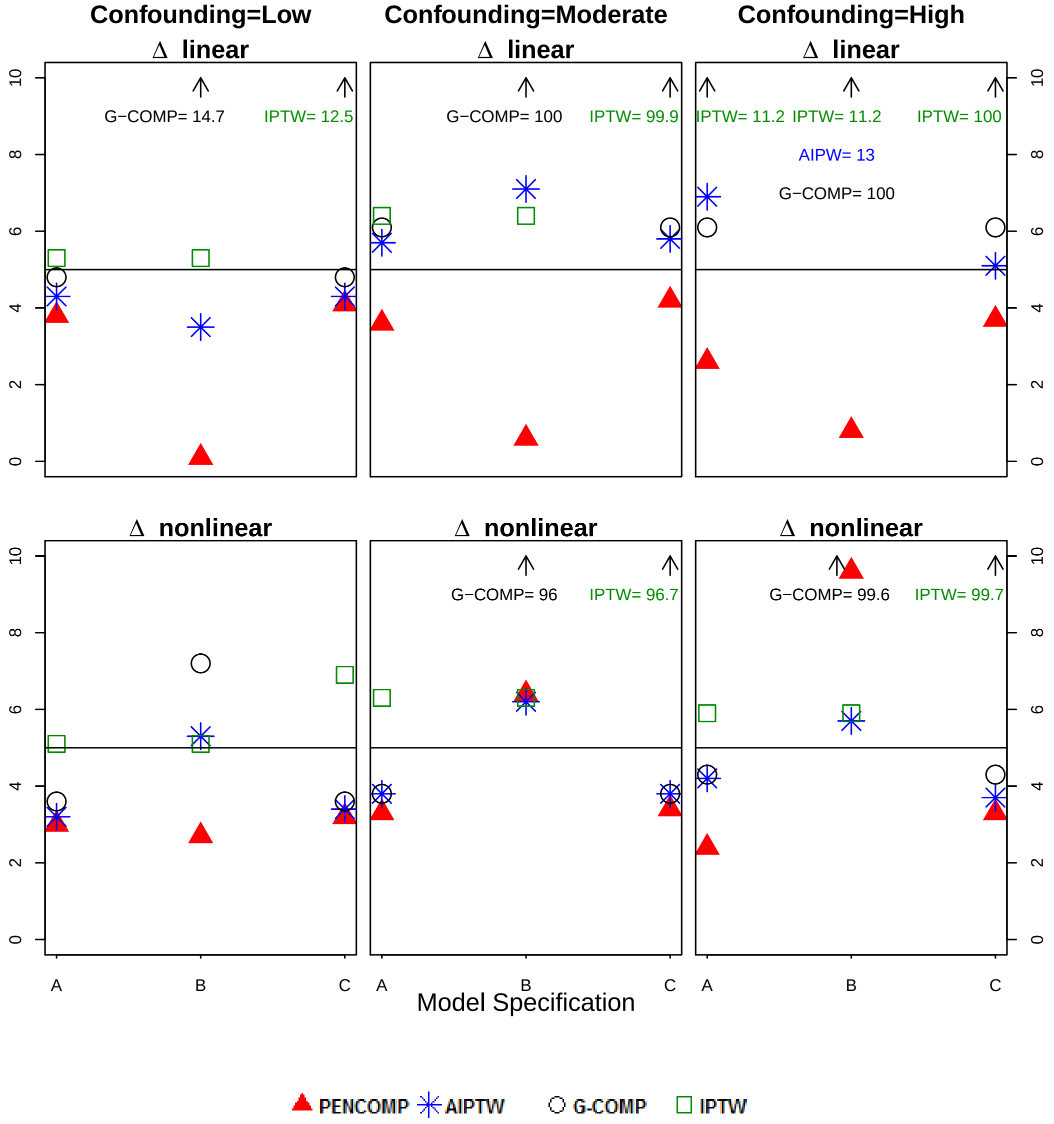

Figure 2.2: 95% noncoverage rate across four methods-PENCOMP, AIPTW, IPTW and g-computation for treatment effect Δ in a linear and nonlinear outcome model. (A) Correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

2.4.3 Simulations for Treatments Assigned at Two Time Points

In the two time-point treatment scenario, each simulated data set contains X_{1a}, X_{1b}, Z_1, X_{2a}, X_{2b}, Z_2, and Y. X_{1a} and X_{1b} are two baseline covariates and normally distributed with mean 0.2, variance 1. The first treatment Z_1 is Bernoulli distributed with success probability that depends on the two baseline variables. The intermediate outcome X_{2a} is normally distributed with a mean that depends on X_{1a}, X_{1b} and Z_1 and with a residual variance of 1. The other intermediate outcome X_{2b} is normally distributed with a mean that depends on X_{1b}, X_{2a} and Z_1, and with a residual variance of 1. The second treatment Z_2 is Bernoulli distributed with success probability that depends on all the covariate and treatment histories. Thus, X_{2a} and X_{2b} both mediate and confound the relationship between Z_1, Z_2, and Y. The coefficients in the second treatment assignment are varied to create three levels of variability of the IPTW weights: low, moderate and high. The true first and second treatment probability models are described in Table 2.4. Each outcome model is normally distributed with a mean that depends on the covariate and treatment histories, and a residual variance of 1, as shown in Table 2.4. The data were generated based on the true models in Table 2.4. Under the linear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(22.35, 11.17, 10.45)$, respectively. Under the nonlinear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(25.31, 12.69, 10.57)$, respectively.

Table 2.4: Two Time Point Treatment Simulation Scenarios: setting $(\gamma_{11},\gamma_{21},\gamma_{22},\gamma_{24})$ equal to $(-0.5,-0.1,0.2,0.2)$, $(-0.8,-0.1,0.6,0.6)$, and $(-0.8,-0.5,1.1,1.1)$ which corresponds to high, moderate, and low confounding, respectively.

.

	Linear Outcome	
True	$X_{1a} \sim N(0.2,1)$	
	$X_{1b} \sim N(0.2,1)$	
	$\text{logit}(P(Z_1=1 \mid X_1,\gamma_1)) = \gamma_{10}+\gamma_{11}X_{1a}+\gamma_{12}X_{1b}$	$\gamma_1 = (\gamma_{10},\gamma_{11},\gamma_{12}) = (-0.01,\gamma_{11},-0.3)$
	$\text{logit}(P(Z_2=1 \mid \bar{X}_2,Z_1,\gamma_2)) = \gamma_{20}+\gamma_{21}(X_{2a}-X_{1a})+\gamma_{22}Z_1(X_{2a}-X_{1a})+\gamma_{23}(X_{2b}-X_{1b})+\gamma_{24}Z_1(X_{2b}-X_{1b})$	$\gamma_2 = (\gamma_{20},\gamma_{21},\gamma_{22},\gamma_{23},\gamma_{24}) = (-0.01,\gamma_{21},\gamma_{22},-0.1,\gamma_{24})$
	$(X_{2a} \mid Z_1=0,X_{1a},X_{1b},\omega_0) \sim N(\omega_{00}X_{1a}+\omega_{01}X_{1b},1)$	$\omega_0 = (\omega_{00},\omega_{01}) = (1,0.5)$
	$(X_{2a} \mid Z_1=1,X_{1a},X_{1b},\omega_1) \sim N(\omega_{10}X_{1a}+\omega_{11}Z_1+\omega_{12}X_{1a}*Z_1+\omega_{13}X_{1b},1)$	$\omega_1 = (\omega_{10},\omega_{11},\omega_{12},\omega_{13}) = (1,0.5,0.5,0.5)$
	$(X_{2b} \mid Z_1=0,X_{1b},\alpha_0) \sim N(\alpha_{00}X_{2a}+\alpha_{01}X_{1b},1)$	$\alpha_0 = (\alpha_{00},\alpha_{01}) = (0.3,1)$
	$(X_{2b} \mid Z_1=1,X_{1b},\alpha_1) \sim N(\alpha_{10}X_{2a}+\alpha_{11}X_{1b},1)$	$\alpha_1 = (\alpha_{10},\alpha_{11}) = (0.4,1)$
	$E(Y_{11} \mid \bar{X}_2,\beta_{11}) = \beta_{110}+\beta_{111}X_{1a}+\beta_{112}X_{2a}+\beta_{113}X_{1b}+\beta_{114}X_{2b}$	$\beta_{11} = (\beta_{110},\beta_{111},\beta_{112},\beta_{113},\beta_{114}) = (25,2,2,1.5,1.5)$
	$E(Y_{10} \mid \bar{X}_2,\beta_{10}) = \beta_{100}+\beta_{101}X_{1a}+\beta_{102}X_{2a}+\beta_{103}X_{1b}+\beta_{104}X_{2b}$	$\beta_{10} = (\beta_{100},\beta_{101},\beta_{102},\beta_{103},\beta_{104}) = (15,2,1,1.5,1)$
	$E(Y_{01} \mid \bar{X}_2,\beta_{01}) = \beta_{010}+\beta_{011}X_{1a}+\beta_{012}X_{2a}+\beta_{013}X_{1b}+\beta_{014}X_{2b}$	$\beta_{01} = (\beta_{010},\beta_{011},\beta_{012},\beta_{013},\beta_{014}) = (15,1,2,1,1.5)$
	$E(Y_{00} \mid \bar{X}_2,\beta_{00}) = \beta_{000}+\beta_{001}X_{1a}+\beta_{002}X_{2a}+\beta_{003}X_{1b}+\beta_{004}X_{2b}$	$\beta_{00} = (\beta_{000},\beta_{001},\beta_{002},\beta_{003},\beta_{004}) = (15,1,1,1,1)$
Misspecified	$logit(P(Z_2=1 \mid \bar{X}_2,Z_1,\lambda)) = \lambda_0+\lambda_1X_{1a}+\lambda_2X_{2a}+\lambda_3X_{1b}$	
	$E(Y_{11} \mid \bar{X}_2,\alpha_{11}) = \alpha_{110}+\alpha_{111}X_{1a}+\alpha_{112}X_{1b}$	
	$E(Y_{10} \mid \bar{X}_2,\alpha_{10}) = \alpha_{100}+\alpha_{101}X_{1a}+\alpha_{102}X_{1b}$	
	$E(Y_{01} \mid \bar{X}_2,\alpha_{01}) = \alpha_{010}+\alpha_{011}X_{1a}+\alpha_{012}X_{1b}$	
	$E(Y_{00} \mid \bar{X}_2,\alpha_{00}) = \alpha_{000}+\alpha_{001}X_{1a}+\alpha_{002}X_{1b}$	
	NonLinear Outcome	
True	$X_{1a} \sim N(0.2,1)$	
	$X_{1b} \sim N(0.2,1)$	
	$\text{logit}(P(Z_1=1 \mid X_1,\gamma_1)) = \gamma_{10}+\gamma_{11}X_{1a}+\gamma_{12}X_{1b}$	$\gamma_1 = (\gamma_{10},\gamma_{11},\gamma_{12}) = (-0.01,\gamma_{11},-0.3)$
	$\text{logit}(P(Z_2=1 \mid \bar{X}_2,Z_1,\gamma_2)) = \gamma_{20}+\gamma_{21}(X_{2a}-X_{1a})+\gamma_{22}Z_1(X_{2a}-X_{1a})+\gamma_{23}(X_{2b}-X_{1b})+\gamma_{24}Z_1(X_{2b}-X_{1b})$	$\gamma_2 = (\gamma_{20},\gamma_{21},\gamma_{22},\gamma_{23},\gamma_{24}) = (-0.01,\gamma_{21},\gamma_{22},-0.1,\gamma_{24})$
	$(X_{2a} \mid Z_1=0,X_{1a},X_{1b},\omega_0) \sim N(\omega_{00}X_{1a}+\omega_{01}X_{1b},1)$	$\omega_0 = (\omega_{00},\omega_{01}) = (1,0.5)$
	$(X_{2a} \mid Z_1=1,X_{1a},X_{1b},\omega_1) \sim N(\omega_{10}X_{1a}+\omega_{11}Z_1+\omega_{12}X_{1a}*Z_1+\omega_{13}X_{1b},1)$	$\omega_1 = (\omega_{10},\omega_{11},\omega_{12},\omega_{13}) = (1,0.5,0.5,0.5)$
	$(X_{2b} \mid Z_1=0,X_{1b},\alpha_0) \sim N(\alpha_{00}X_{2a}+\alpha_{01}X_{1b},1)$	$\alpha_0 = (\alpha_{00},\alpha_{01}) = (0.3,1)$
	$(X_{2b} \mid Z_1=1,X_{1b},\alpha_1) \sim N(\alpha_{10}X_{2a}+\alpha_{11}X_{1b},1)$	$\alpha_1 = (\alpha_{10},\alpha_{11}) = (0.4,1)$
	$E(Y_{11} \mid \bar{X}_2,\beta_{11}) = \beta_{110}+\beta_{111}X_{1a}+\beta_{112}X_{2a}+\beta_{113}X_{1b}+\beta_{114}X_{2b}+\beta_{115}X_{2a}*X_{2b}$	$\beta_{11} = (\beta_{110},\beta_{111},\beta_{112},\beta_{113},\beta_{114},\beta_{115}) = (25,2,2,1.5,1.5,1.6)$
	$E(Y_{10} \mid \bar{X}_2,\beta_{10}) = \beta_{100}+\beta_{101}X_{1a}+\beta_{102}X_{2a}+\beta_{103}X_{1b}+\beta_{104}X_{2b}+\beta_{105}X_{2a}*X_{2b}$	$\beta_{10} = (\beta_{100},\beta_{101},\beta_{102},\beta_{103},\beta_{104},\beta_{105}) = (15,2,1,1.5,1,1)$
	$E(Y_{01} \mid \bar{X}_2,\beta_{01}) = \beta_{010}+\beta_{011}X_{1a}+\beta_{012}X_{2a}+\beta_{013}X_{1b}+\beta_{014}X_{2b}+\beta_{015}X_{2a}*X_{2b}$	$\beta_{01} = (\beta_{010},\beta_{011},\beta_{012},\beta_{013},\beta_{014},\beta_{015}) = (15,1,2,1,1.5,0.8)$
	$E(Y_{00} \mid \bar{X}_2,\beta_{00}) = \beta_{000}+\beta_{001}X_{1a}+\beta_{002}X_{2a}+\beta_{003}X_{1b}+\beta_{004}X_{2b}+\beta_{005}X_{2a}*X_{2b}$	$\beta_{00} = (\beta_{000},\beta_{001},\beta_{002},\beta_{003},\beta_{004},\beta_{005}) = (15,1,1,1,1,0.7)$
Misspecified	$logit(P(Z_2=1 \mid \bar{X}_2,Z_1,\lambda)) = \lambda_0+\lambda_1X_{1a}+\lambda_2X_{2a}+\lambda_3X_{1b}$	
	$E(Y_{11} \mid \bar{X}_2,\alpha_{11}) = \alpha_{110}+\alpha_{111}X_{1a}+\alpha_{112}X_{1b}$	
	$E(Y_{10} \mid \bar{X}_2,\alpha_{10}) = \alpha_{100}+\alpha_{101}X_{1a}+\alpha_{102}X_{1b}$	
	$E(Y_{01} \mid \bar{X}_2,\alpha_{01}) = \alpha_{010}+\alpha_{011}X_{1a}+\alpha_{012}X_{1b}$	
	$E(Y_{00} \mid \bar{X}_2,\alpha_{00}) = \alpha_{000}+\alpha_{001}X_{1a}+\alpha_{002}X_{1b}$	

Results for RMSE and 95% confidence interval noncoverage for sample size 500 are shown in Figure 2.3-2.6; other results are given in Tables A.17-A.20 in Appendix A. The RMSEs of the methods are presented in Figure 2.3 for the linear outcome model and in Figure 2.4 for the nonlinear outcome model, expressed as a proportion of the RMSE of IPTW with a correct propensity model. The AIPTW and PENCOMP methods had substantially lower RMSEs than IPTW, with the ratio of RMSEs less than 0.7 in most scenarios. The RMSEs for PENCOMP were similar to or lower than the corresponding RMSEs for AIPTW, with some substantial gains over AIPTW when the prediction models were misspecified. Lastly, g-computation had similar RMSE to PENCOMP when the prediction model was correctly specified, but markedly, higher RMSE than PENCOMP when the prediction model was misspecified.

Non-coverage rates of the 95% intervals are shown in Figure 2.5-2.6. Coverage for IPTW was markedly below nominal when the prediction models were misspecified. PENCOMP tended to have close to nominal or conservative coverages. AIPTW had close to nominal or anti-conservative coverages, and tended to undercover in situations with high confounding, particularly when the prediction model was severely misspecified and the weights were highly variable. For example, for estimation of Δ_{10} in the nonlinear regressions, as confounding increased, AIPTW and IPTW's coverage rates dropped dramatically to about 60%, while PENCOMP maintained a coverage rate of 97%.

Table A.17 displays empirical biases as a fraction of RMSE of IPTW with correctly specified propensity score model for the linear and nonlinear outcome models, respectively. As in the one time point case, IPTW had moderate empirical bias when the propensity model was correctly specified under high confounding, and was highly biased when the propensity model was misspecified, especially with moderate and high degrees of confounding. On the other hand, g-computation had negligible biases, with relative bias of less than 1% when the prediction model was correctly specified, but was highly biased when the prediction model was misspecified. AIPTW and PENCOMP had lower empirical bias under low confounding scenarios or when the prediction model was correctly specified. As confounding increased, the estimated biases became larger. However, both AIPTW and PENCOMP had relative bias of less than 5% in most cases. In terms of the RMSE of IPTW with a correct propensity model, the bias of AIPTW and PENCOMP represented a very small fraction of the RMSE, with the fractions varying from approximately 0 to 0.25.

The 95% confidence intervals widths are shown in Table A.20. In both linear and nonlinear outcome models, both AIPTW and PENCOMP had similar confidence interval widths, which were substantially smaller than IPTW. As confounding increased, PENCOMP tended to have smaller confidence interval widths than IPTW with correctly-specified propensity model and still covered better. On the other hand, AIPTW tended to undercover under high confouding. Lastly, PENCOMP tended to have similar RMSEs and mean confidence interval widths as g-computation with correctly specified prediction models.

Results for sample size 200 and 1000 are in Table A.13-A.16, A.21-A.24 in Appendix A. In general, changes in sample sizes had similar effects on the two-time point simulations as for the single time point simulations, with the finite sample bias for the robust estimators decreasing as the sample size increased. Changes in sample size had very little impact on RMSE comparisons. Coverage rates for the robust estimators were slightly improved under larger sample sizes. Confidence interval widths for PENCOMP tended to shrink as sample sizes increased, while other interval widths remained the same.

Overall, PENCOMP outperforms the other methods in terms of RMSE and coverage probability and efficiency in these simulations, although it has slightly larger bias than AIPTW in some cases-though very small as a fraction of RMSE of IPTW with a correct propensity model.

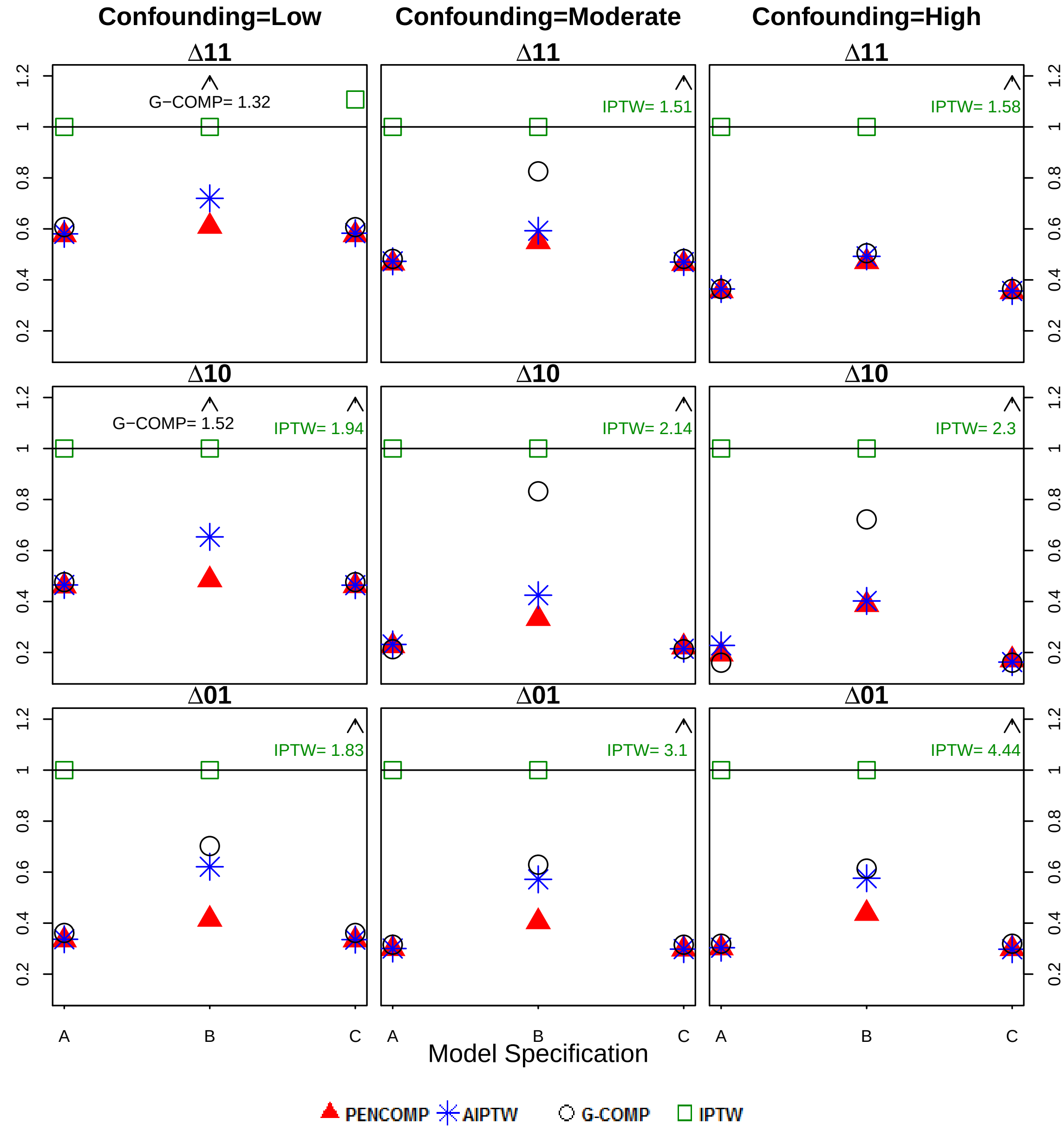

Figure 2.3: Ratio of RMSE over RMSE of IPTW(A) with correct propenisty score model across four methods-PENCOMP, AIPTW, IPTW and g-computation for three treatment effects Δ_{11}, Δ_{10}, and Δ_{01} in a linear outcome model. (A) Correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

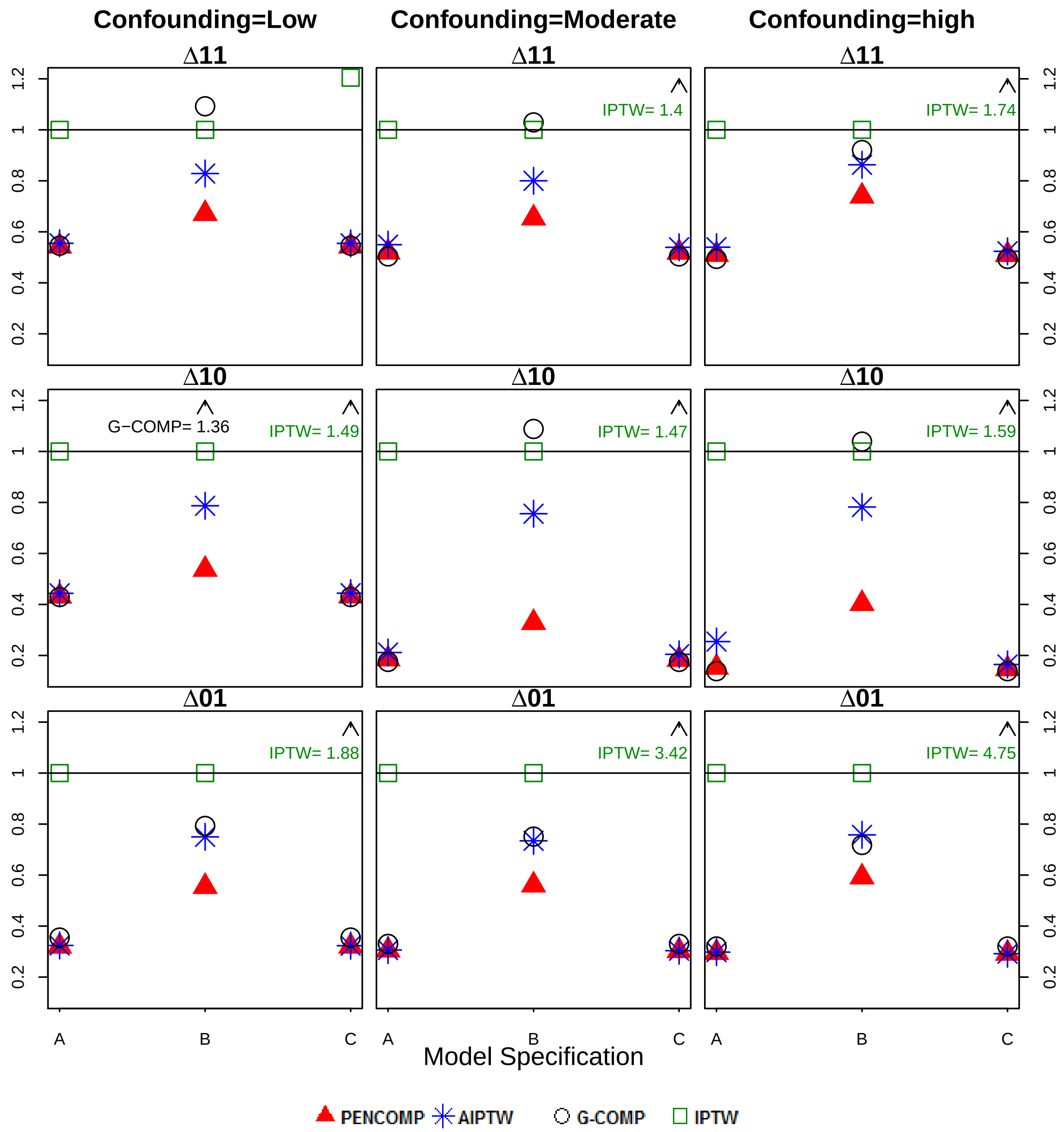

Figure 2.4: Ratio of RMSE over RMSE of IPTW(A) with correct propenisty score model across four methods-PENCOMP, AIPTW, IPTW and g-computation for three treatment effects Δ_{11}, Δ_{10}, and Δ_{01} in a nonlinear outcome model. (A) Correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

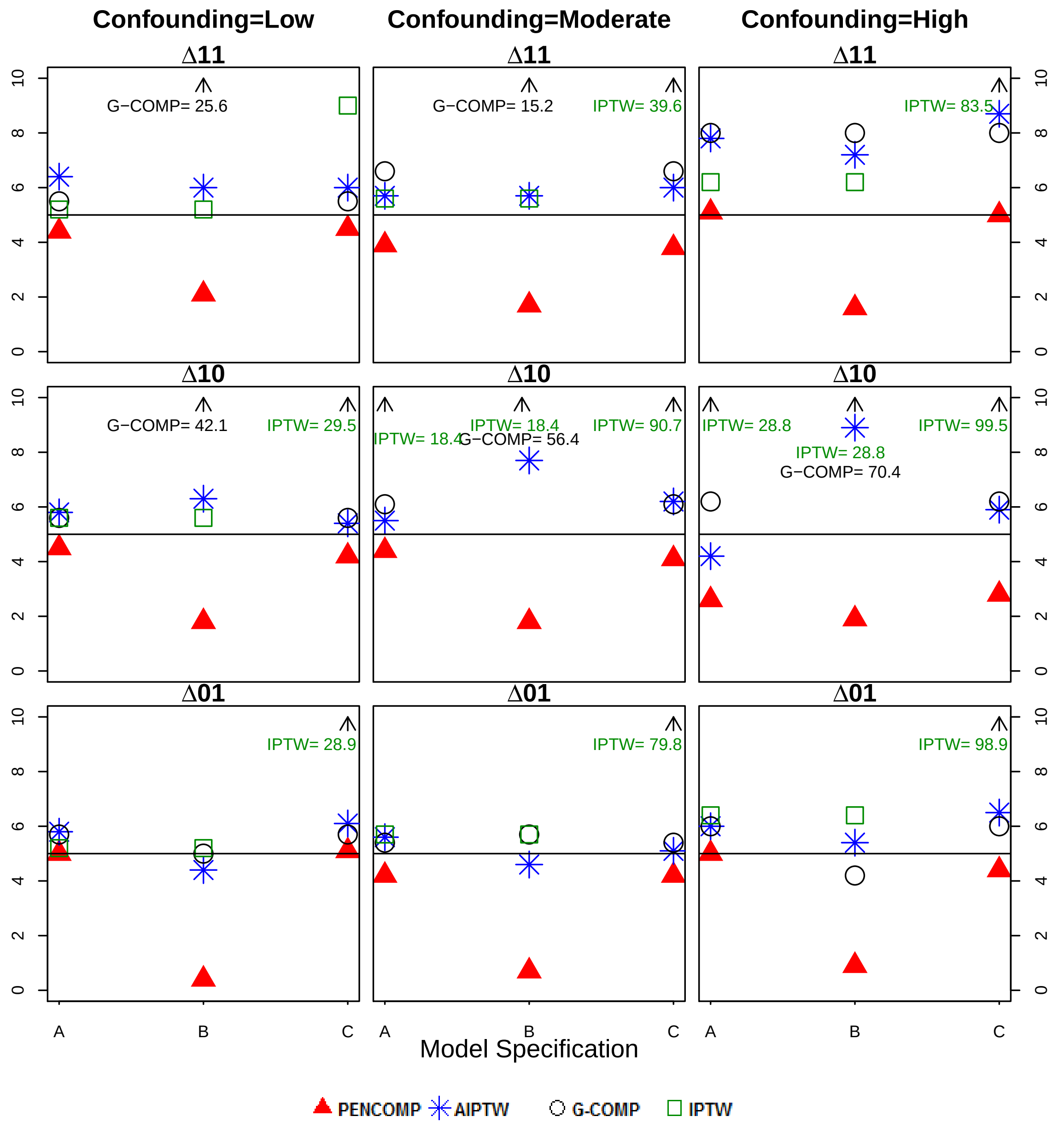

Figure 2.5: 95% noncoverage rate across four methods-PENCOMP, AIPTW, IPTW and g-computation for three treatment effects Δ_{11}, Δ_{10}, and Δ_{01} in a linear outcome model. (A) Correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

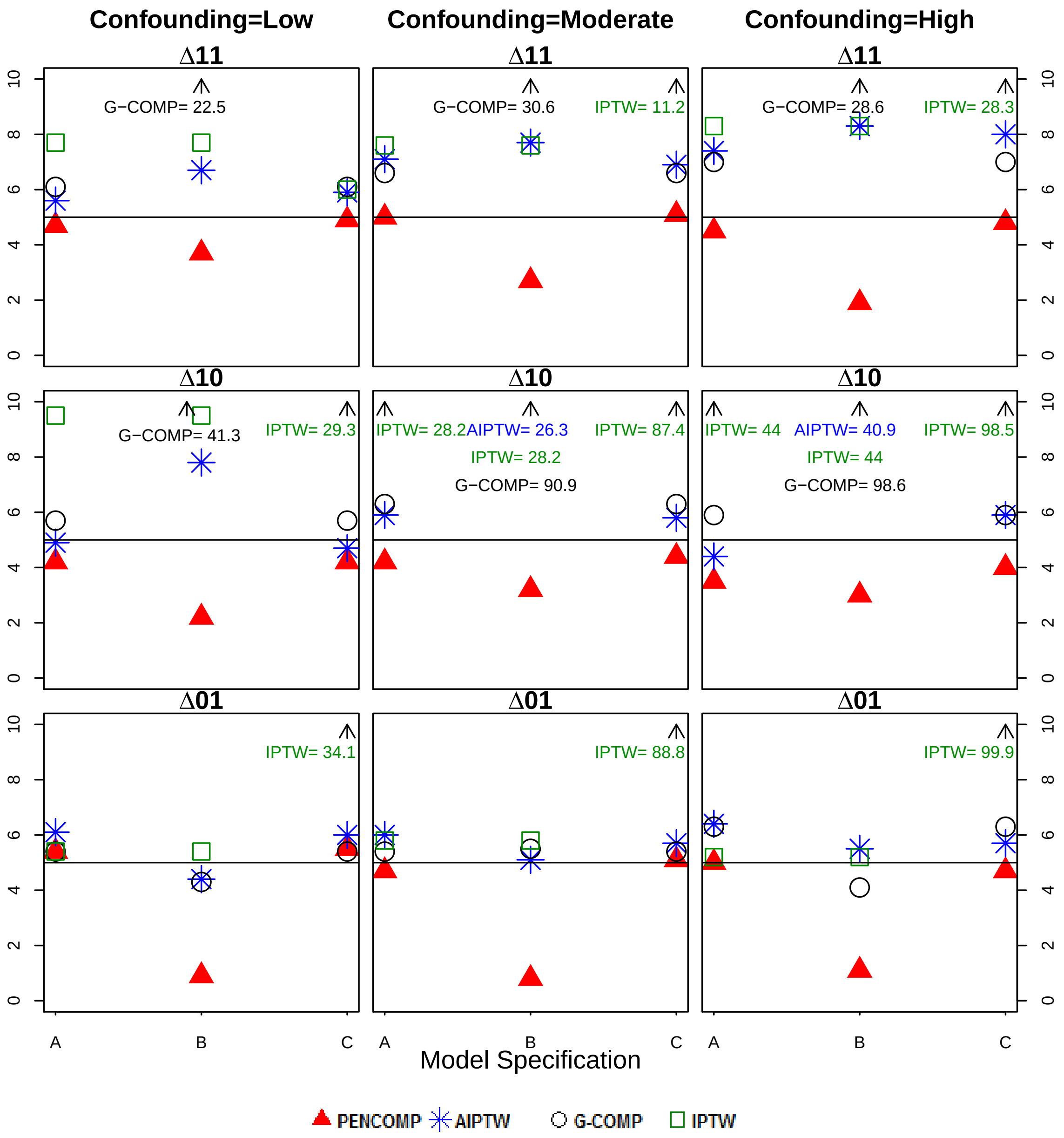

Figure 2.6: 95% noncoverage rate across four methods-PENCOMP, AIPTW, IPTW and g-computation for three treatment effects Δ_{11}, Δ_{10}, and Δ_{01} in a nonlinear outcome model. (A) Correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

2.5 Application

We applied our method to the Multicenter AIDS Cohort study (MACS) to analyze the effect of antiretroviral treatment on CD4 counts. We restrict our analyses to the period between visit 7 and 21, after the first antiretroviral treatment zidovudine (AZT) was approved for use and before the advent of highly active antiretroviral therapy (HAART). During the period between visit 14 and 17 didanosine (ddI) and zalcitabine (ddC) also became available. Then around visit 21, new treatments-stavudine (d4t) and lamivudine (3tc) were approved. We estimate the short-term (1 year) effects of using any antiretroviral treatment for HIV+ subjects. Treatment was coded to 1 if the patient reported taking any of the four mentioned antiretroviral treatment (ART) or enrolling in clinical trials of such drugs. That is, starting with visit 7, for every three-visit window we estimated the effects of using ART drugs on CD4 counts. We excluded subjects with missing values on any of the covariates included in the models. We also used the square root of the blood count variables in this analysis.

For each three-visit window, we denoted time $t = 1, 2$, and 3. Let $X_t(i)$ denote square root of subject i's blood count measures at time t, and $Z_t(i)$ be one if subject i received antiretroviral treatment during the period between time t and $t + 1$, and zero if otherwise, for $t = 1, 2$. Let $Y(i) = X_3(i)$ be the square root of CD4 count for subject i measured a year after baseline at time $t = 3$. We defined dosage as the number of times a subject went on treatment previously, i.e. from the start of enrollment to the baseline at time $t = 1$ of each three-visit window. For the outcome and propensity models, we considered baseline blood count measures, dosage, and intermediate CD4 count as potential covariates. The baseline blood count measures included CD4 count, CD8 count, white blood cell count (WBC), red blood cell count (RBC), and platelets. Specifically, the intermediate outcome models included all the baseline blood measures. The final outcome model included the baseline blood count measures and intermediate CD4 count. For sample size less than 50, especially for treatment regimes $(1, 0)$ and $(0, 1)$, the final outcome models included only prior CD4 count. The first treatment assignment Z_1 was modeled as a logistic regression with baseline blood count measures and dosage. Race, age, and education level were not included because including them seemed to increase the variance of the estimates while the estimates stayed about the same. The second treatment Z_2 was modeled as a logistic regression with the same baseline covariates as those in the first treatment model, intermediate CD4 count and Z_1. The models used to estimate the numerator of the stabilized weights excluded all covariates, except treatment indicator Z_1 and intercept. When calculating the total dosage for subjects, we assumed that subjects with missed visits did not change treatment at the missing time points. For each three-visit window starting with visit 7, we estimated the treatment effects Δ_{11}, Δ_{10}, and Δ_{01}, provided sufficient data were available to model the relevant outcomes. The number of subjects with observed treatment sequence $(Z_1, Z_2) = (1, 0)$ was very small for some of the three-visit windows, as shown in Table A.25 in Appendix A.4. The data suggested that patients tended to stay on treatment once they started. As the three-visit window moved across time, more patients got on treatments, and fewer patients switched off treatment, since there were more treatment options available if resistance or severe side effects developed with one treatment. Consequently, the number of subjects with treatment sequence $(1, 0)$ was much smaller than that with $(1, 1), (0, 1)$, or $(0, 0)$.

In both the Monte Carlo steps of AIPTW and the imputation steps in PENCOMP, we replaced the simulated/imputed transformed CD4 values that were < 0 with 0 (i.e. below detection level). The stabilized weights were still highly variable, as shown in Table A.26, so we truncated the weights at the 1st and 99th percentiles when calculating the estimates of AIPTW and IPTW estimators. Although the variances of the estimates reduced, the estimates became more biased toward the naive estimates, as seen in Figure A.1 in Appendix A.4. The results without truncation are in Figure 2.8 (See Zubizarreta (2015) for an alternative that minimizes weight variance while retaining covariate balance). For PENCOMP, we chose a mininum of 35 and 1/4 of unique data points as the number of knots. Equally spaced knots and truncated linear basis were used. In addition, for estimating outcomes for a particular treatment regimen $\overline{Z}_t$, we excluded cases where the propensity of $\overline{Z}_t$ lay outside the observed ranges of the propensity of $\overline{Z}_t$ as described in Section 2.2.4, to avoid extrapolating the regression model predictions outside the shared range of propensities.

For example, to calculate the treatment effect of $\Delta_{z_1 z_2}$, we estimated the probability of getting treatment $Z_1 = z_1$ conditional of the baseline covariate history, denoted as $\hat{P}(Z_1 = z_1|\bar{X}_1)$, and the probability of receiving treatment $Z_2 = z_2$, conditional the past covariate history and $Z_1 = z_1$, denoted as $\hat{P}(Z_2 = z_2|Z_1 = z_1, \bar{X}_1)$. Denote the probability of treatment regime (z_1, z_2) as $\hat{P}_{\bar{z}_2} = \hat{P}(Z_1 = z_1|\bar{X}_1) * \hat{P}(Z_2 = z_2|Z_1 = z_1, \bar{X}_2)$. At $t = 1$, subjects were divided into two groups using indicators $I(Z_1^{obs} = z_1)$ and $I(Z_1^{obs} \neq z_1)$. We removed subjects whose estimated propensity scores $\hat{P}(Z_1 = z_1|\bar{X}_1)$ lay outside the overlapping regions of the propensity scores. Similarly, at $t = 2$, subjects were divided into two groups using indicators $I\{(Z_1^{obs}, Z_2^{obs}) = (z_1, z_2)\}$ and $I\{(Z_1^{obs}, Z_2^{obs}) \neq (z_1, z_2)\}$. Again, we removed subjects whose estimated propensity scores $\hat{P}_{\bar{z}_2}$ lay outside of the overlapping regions of the propensity scores. We then repeated this process for $z_1 = z_2 = 0$, and took for analysis the set of observations that had not been dropped as a result of all of these comparisons. Figure 2.7 illustrates the overlapping regions of the propensity scores for one window. We repeat the same procedures for each set of time points and each treatment. The fraction of subjects that was included in each analysis varied from 25% to 89% of the total sample, shown in Table A.25. One possible reason for fewer subjects being included in later windows was that later windows included more newly infected subjects, as well as infected subjects who had survived for years; these two groups of subjects were probably very different.

One important step in building the propensity score models is to check for balance in the covariates. At $t = 1$, for the two groups of subjects $I(Z_1^{obs} = z_1)$ and $I(Z_1^{obs} \neq z_1)$, we checked whether the distributions of the baseline covariates were similar between the two groups. Similarly, at $t = 2$, we checked whether the distributions of the baseline and the intermediate covariates were similar between the two groups $I\{(Z_1^{obs}, Z_2^{obs}) = (z_1, z_2)\}$ and $I\{(Z_1^{obs}, Z_2^{obs}) \neq (z_1, z_2)\}$. As a measure of imbalance, we used the standardized difference between the two groups, which is the difference in means between the two groups divided by an estimate of the pooled standard deviation:

$$d = \left| \left(\bar{x}_{(z_1 z_2)} - \bar{x}_{\neq (z_1 z_2)} \right) \Big/ \sqrt{\frac{s^2_{(z_1 z_2)} + s^2_{\neq (z_1 z_2)}}{2}} \right|$$

If the propensity score models are adequately specified, the covariate distributions between the $(z_1 z_2)$ and $\neq (z_1 z_2)$ groups should be similar, conditional of the estimated propensity scores. Specifically, to check the balance of covariate x, we regressed x on the spline of the propensity scores and compared the residuals between treatment groups using t-test. Table 2.5 shows an example for covariate balance before and after adjusting for propensity scores. The standardized differences between treatment groups for most blood count measures and the t statistics were reduced dramatically. In addition, we assessed the degree of overlap in the propensity score distributions between treatment groups (Imbens and Rubin, 2015). For example, we measured the proportion of subjects in the $\neq (z_1, z_2)$ group whose propensity scores of (z_1, z_2) are between the $1 - \alpha$ and α quantiles of the propensity score distribution of (z_1, z_2) group, denoted as $\pi^{1-\alpha}_{(z_1,z_2)} = F_{\neq (z_1,z_2)}(F^{-1}_{(z_1,z_2)}(1 - \alpha)) - F_{\neq (z_1,z_2)}(F^{-1}_{(z_1,z_2)}(\alpha))$, where F is the cumulative distribution. Inside this region it is easier to impute missing potential outcomes $Y^{z_1 z_2}$ because there are more observations. The low degree of overlap for this dataset suggested some difficulty in imputing the missing potential outcomes, as shown in Figure 2.7 and Table A.27 in Appendix A.4.

We estimated the short term effect of antiretroviral treatment on CD4 count using four methods: naive crude estimate, g-computation, IPTW, AIPTW, and PENCOMP. The results are summarized in Figure 2.8. The standard errors were obtained using 500 bootstrap samples. For PENCOMP, 200 complete datasets were created. For all the three-visit windows, the naive estimators were negative, suggesting a harmful effect of antiretroviral treatment on CD4 count. This is likely due to uncontrolled confounding by indication, in that sicker subjects with lower CD4 counts were more likely to be assigned to treatment. The treatment effects estimated by IPTW, AIPTW and PENCOMP all suggest less harmful effects, with PENCOMP in particular having slightly negative to slightly positive effects, and IPTW having positive effects in most windows. When the weights were not variable in window 1-3, and 15-16 and the means of the stablized weights were close to one, the treatment effects obtained from all four methods were similar. The similarity of PENCOMP to the other estimates indicate that our proposed method is addressing the bias from confounding by indication. Further, when the weights became variable, the PENCOMP estimates were more stable across time, and generally had smaller standard errors than either AIPTW or IPTW, a finding that is consistent with the findings in the simulation study. Lack of stronger positive effects of treatment may be due to the inability of the observed covariates to remove all confounding.

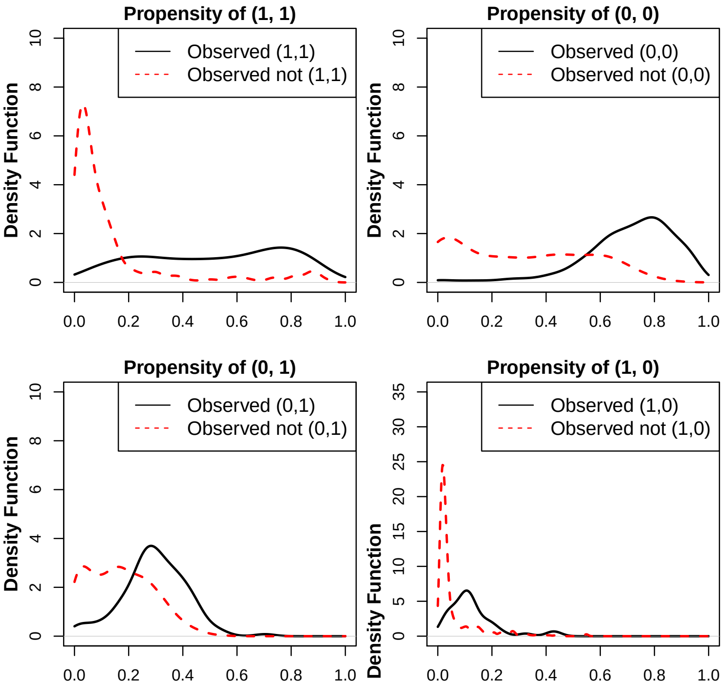

Figure 2.7: Distributions of the propensity scores in subjects whose observed treatment sequence is (z_1, z_2) and subjects whose observed treatment sequence is not (z_1, z_2) for window 4.

Table 2.5: Balance of covariates between subjects with observed treatment sequence $(1, 1)$ and everybody else before and after adjusting for propensity scores for window 8, without removing subjects outside of the overlapping regions. We regressed each covariate on the spline of the logit of the propensity score, $\hat{P}^*_{11}$. Truncated linear basis with 10 equally spaced knots was used. $**$ significant at 0.005 level, and $*$ significant at 0.05 level.

	Before Adjusting		After Adjusting	
Covariate	d	T Stats	d	T stats
RBC	1.83	25.23**	0.016	0.22
CD4	1.11	15.28**	0.0048	0.067
WBC	0.59	8.11**	0.028	0.39
CD8	0.0012	0.017	0.032	0.44
PLATE	0.10	1.37	0.044	0.61
CD4 at $t = 2$	1.12	15.28**	0.017	0.23

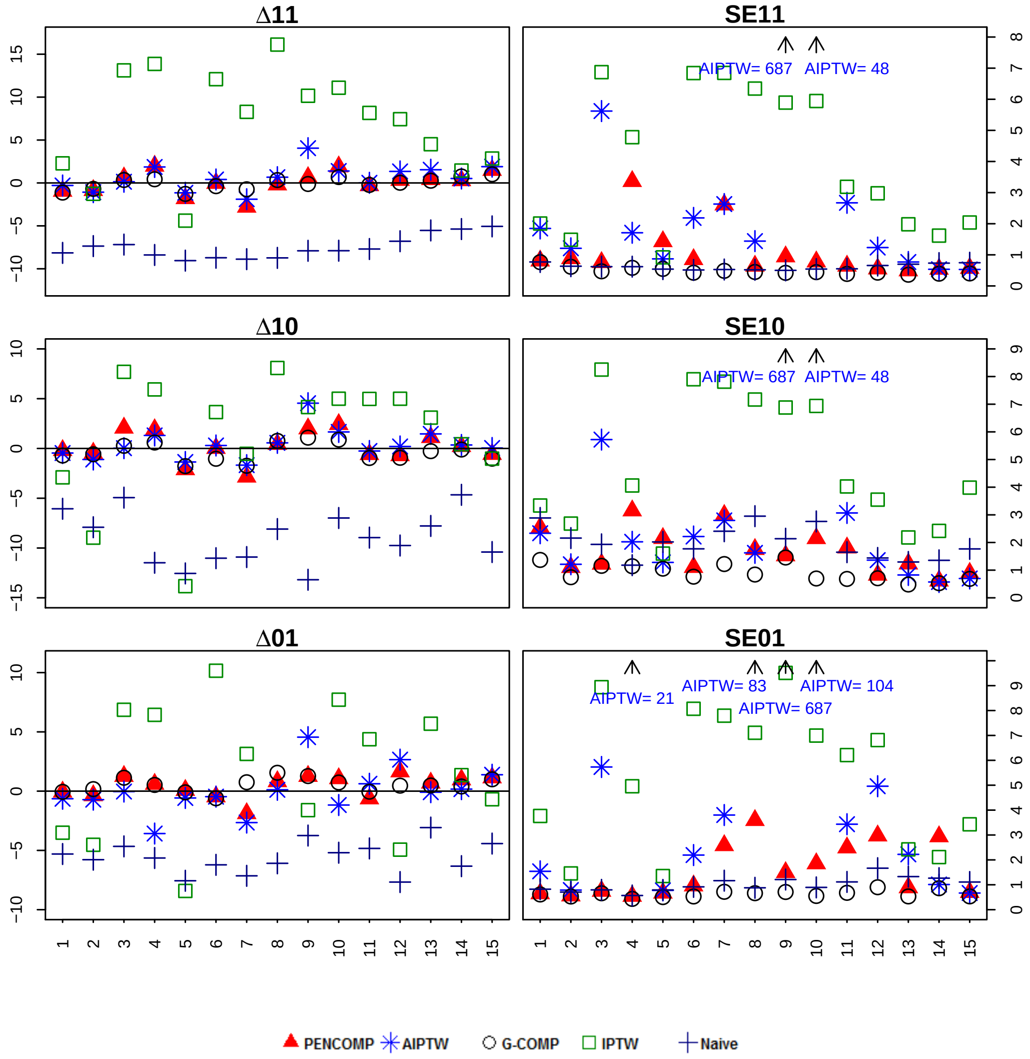

Figure 2.8: For each of the three-visit windows $1, \cdots, 15$, the estimates and standard errors (SE) of the treatment effects Δ_{11}, Δ_{10}, and Δ_{01} of the four methods: PENCOMP, AIPTW, IPTW, and Naive. For some windows, AIPTW had very large bootstrap standard errors because of a few extreme bootstrap estimates.

2.6 Discussion

We have proposed PENCOMP as a new, straightforward method to estimate treatment effects in point treatment situations and in two-time point treatment situations with time dependent confounders. The method uses the doubly-robust imputation methodology of Zhang and Little (2009) to impute the unobserved potential outcomes and compute the causal treatment effects of interest. As with other doubly-robust methods, PENCOMP offers the analyst two chances to make correct inferences about treatment effects, either by correctly specifying the propensity score model or by correctly specifying the prediction models. The robustness of PENCOMP to model misspeciffication is borne out by our simulation studies.

Three main versions of PENCOMP are PENCOMP-ML, which is based on ML with information-based or bootstrap standard errors, PENCOMP-B, which based inference on posterior distributions of the causal parameters, and PENCOMP-MI, which multiply imputes the outcomes for treatments not assigned, and uses MI combining rules for inference. For PENCOMP we considered distinct outcome models for each treatment combination in this paper. Specifically, suppose we are interested in treatment sequence $\bar{z}_T$, at each time point t, the outcome model was fitted using only the subjects with $\bar{z}_t$ that matched with $\bar{z}_T$ up to time point t. However, when the observed data are sparse, outcome models with interactions between treatment and covariates, as well as interactions between treatment and splines (Coull, 2001), can be fitted to borrow strength across different treatment sequences. However, adding interactions between treatment and splines could increase complexity when there are many treatment sequences. We fitted the spline on the propensity score on the probability scale and on the logit scale but found that the logit scale worked much better in most cases, especially when the propensity scores were too extreme on the probability scale. Lastly, we considered PENCOMP-MI in our empirical work, but it would be interesting to compare it with the alternative versions, particularly PENCOMP-B, which as a Bayesian method might have attractive small-sample properties.

A natural competitor to PENCOMP is the AIPTW estimator, which like PENCOMP has a double robustness property. In our simulation studies, the performance of PENCOMP is similar to that of AIPTW estimator when the confounding is low. However, when the confounding is moderate or high and the weights in AIPTW are highly variable, PENCOMP tends to outperform the version of AIPTW considered in this study with respect to mean square error, interval coverage, and interval width. Kang and Schafer (2007) also show drawbacks of AIPTW in small samples, especially when the weights are highly variable. The version of AIPTW we considered is based on Monte Carlo simulations and is computationally intensive. Consequently, PENCOMP is not only statistically more efficient, but is also computationally more efficient than this AIPTW estimator. Other versions of AIPTW have been suggested, and we have not compared our method with these versions; however, we expect that instability from highly-variable weights is likely to be an issue with other forms of AIPTW as well. The PENCOMP method avoids this problem by using the propensity as a predictor, rather than as a weight.

We have focused here on situations with treatment assignments at just two time points. An important question is how PENCOMP can be applied to longitudinal data sets with more than two assignment points. In the MACS data we analyzed, data are available at 16 time points, so there are over 30,000 (2^{15}) possible treatment combinations, nearly all of which are not seen in the data; providing simple and interpretable causal conclusions in such a setting requires careful thought and modeling. An initial step is to analyze the set of treatment combinations that arise in the data set, and restrict inference to the subset of ”relevant combinations” judged to have sufficient data to provide meaningful estimates. Propensity models can then be fitted sequentially over time on historical data, including prior treatment assignments and outcomes as potential covariates. The outcomes of relevant combinations can then be imputed as a function of a spline of the propensity and other predictive covariates in the history, with the propensity for each relevant combination obtained by multiplying the sequence of propensities at the set of earlier time points. Some modeling of the resulting treatment effects is likely to be needed to provide parsimonious inferences; for example a plot of treatment effects against the number of prior

”dosages” may suggest a model with a parametric form for the treatment effect as a function of dosage. To maintain stable estimates and enhance interpretability, some form of dimension reduction and variable selection, for example, a summary measure of treatments and other time varying covariates, will typically required. Implementing such strategies is outside the scope of this article, and a topic for future research. We note that proliferation of treatment regimens is a characteristic of the problem, not the statistical method; MSM models are faced with similar challenges.

In our simulation study, we considered the standard g-computation based on the full covariate history. However, when the dimension of the covariate is high, such as in longitudinal treatments, it becomes hard to check and fix the models, if they are misspecified. Achy-Brou et al (2010) proposed using a g-computation approach based on the longitudinal propensity scores as regressors, instead of the full covariate history, exploiting the fact that the sequential ignorability assumption remains true given the longitudinal propensity score history. They stratified patients based on quintiles of the propensity scores at each time point, and fitted a proportional odds logistic regression models based on the propensity quintiles for the transition probabilities between strata. PENCOMP is similar to Achy Brou’s method in the sense that both methods model the outcome based on propensity scores. However, while Achy Brou’s approach uses the propensity scores in quintiles, PENCOMP uses a penalized spline to model the relationship between the outcome and the propensity score. This relaxes the parametric assumptions between the outcome and the propensity score and gives PENCOMP the double robustness property. PENCOMP also includes other variables in the prediction models to improve efficiency.

Here we considered a smooth relationship between the outcome and the propensity score. If there are thought to be discontinuities, approaches that allow for this possibility might improve on PENCOMP. Koo (1997) considers models that allow discontinuities at the knots. An adaptive regression spline approach to PENCOMP could potentially address the issues of jump discontinuity and sharp jumps (Di Matteo, Genovesem and Kass, 2001). In addition, we have focused on estimating causal effects for a continuous and normally distributed outcome. Extensions to non-normal outcomes are straightforward in principle, by replacing the normal linear mixed models discussed here with generalized linear mixed models. For example, a logistic mixed effects regression with random effects for the spline on the propensity could be fitted when Y is a binary outcome. Gutman and Rubin (2012) examine the performace of a similar spline method for binary outcome in one time point treatment. However, the performance of such extensions to non-normal outcomes for time-dependent confounding is a topic for future research.

In summary, our simulation studies suggest that PENCOMP is a viable alternative to IPTW and AIPTW estimators. Although we focus on observational studies in this study, PENCOMP can also be used in randomized trials, where randomization at later time points are based on intermediate outcomes from earlier randomized treatments, sequential multiple assignment: randomized trials or SMART (Murphy, 2005; Nahum-Shani et al, 2012). Correct methods typically use the semi-parametric likelihood approach similar to that employed in AIPTW; use of a robust fully model-based approach similar to that of PENCOMP might provide advantages similar to those described here.

CHAPTER III

Addressing Disparities in the Assignment Propensity Distributions for Treatment Comparisons from Observational Studies

3.1 Introduction

Observational studies for inference about causal effects are valuable when randomization is not feasible or unethical. Valid causal inferences in this setting requires adjustment for differences in the distribution of confounders between the treatment groups. For example, the Multicenter AIDS cohort study (MACS) (Kaslow et al, 1987) saw the introduction of the first antiretrovial therapy (zidovudine or AZT) at a time when no effective treatment for human immunodeficiency virus existed. Hence, early administration was based on availability and biomarkers of disease severity such as CD4 count, with sicker patients more likely to be treated. To deal with confounding, propensity score - the probability of treatment assignment as a function of covariates - is often used. The balancing property of propensity score implies that adjusting for the propensity can remove the bias due to differences in all observed confounders between the treatment groups (Rosenbaum and Rubin, 1983). Propensity score-based methods to estimate causal effects from observational studies include inverse propensity weighting, matching, stratification and regression adjustment on the propensity score. However, for these methods to work reliably, there should be a sufficient overlap in the propensity score distributions for the compared treatment groups. Estimating the causal effects for units outside the overlap region depends entirely on extrapolation, and hence is vulnerable to model misspecification. Furthermore, restricting estimation of causal effects to a subpopulation where there is more balance in the propensity distributions between the treatment groups could reduce the sensitivity of causal effect estimates to model misspecification (Rosenbaum and Rubin, 1984).

Techniques have been proposed to address disparities in covariate distributions. Rubin (1977) considers a single covariate setting and suggests dropping all units with covariate values that have either no treated or no control units and restricting causal effects to covariate values that have both treated and control units. Gutman and Rubin (2013, 2015) propose dropping units outside of the overlap region of estimated propensity scores between the treatment groups. Cochran and Rubin (1973) and Dehejia and Wahba (1999) propose discarding unmatched subjects. Ho, Imai, King and Stuart (2005) propose a two-stage approach. In the first stage, all the treated units are paired with their closest control units, and only the matched units are included in the second stage for further adjustment. Similarly, Rosenbaum (2012) proposes an algorithm for choosing an optimal set of treated subjects, where some treated subjects are dropped due to poor matching quality. Crump et al (2009) propose restricting the analysis to an optimal subpopulation defined by trimming off extreme propensity values below α and above $1 - \alpha$. Li, Morgan and Zaslavsky (2017) define an estimand that weights cases to balance the weighted distributions of the covariates between treatment groups in a fashion that minimizes the asymptotic variance of the estimated treatment effect.

The propensity scores are often used to determine the common support region and subjects are simply discarded or down-weighted. Discarding units reduces the effective sample size and thus increases the variance of the estimated treatment effect. However, a subject with low probability of selection in a given arm is usually not well estimated on that arm if unobserved, since it is likely that there are only a few observed subjects on that arm with similar covariate distributions as the subject. Trimming off subjects with extreme propensities changes the estimand, since the causal effect for the subpopulation is usually not the same as that for the entire population. As the sample size increases, there are more observed subjects in the treatment and control groups with similar covariate distributions and thus the casual effects can be estimated more accurately. Therefore, it is intuitive that the range of propensities where causal effects can be estimated should depend on sample size.

The remainder of the paper is structured as follows. In Section 3.2, we discuss alternative definitions of the estimands to address limited overlap. In Section 3.3, we describe six methods for estimating causal effects. In Section 3.4, we describe propensity score estimation procedures and diagnostic checks for balance. In Section 3.5, we study in simulation studies the performance of alternative propensity score-based estimation methods, for a variety of estimands chosen to reduce covariate imbalance. In Section 6, we illustrate our methods to the MACS data and provide guidance for practice.

3.2 Alternative Causal Estimands

In a study with treatments administered as a single time point, let X_i and Z_i denote the vector of baseline covariates and a binary treatment for subject $i = 1, \cdots, N$, respectively. Let $Z_i \in (0, 1)$ denote a binary treatment with $Z_i = 1$ for treatment and $Z_i = 0$ for control. Let $Y_i^{Z_i}$ denote the potential outcome under Z_i for subject i. Suppose we are interested in making causal inference about a population from which the sample is drawn. Under Rubin's causal model, the treatment effect for a subject is defined as the difference between the potential outcomes under the two treatments. The average treatment effect defined on the entire population is the ATE estimand, $E(Y^1 - Y^0)$. The ATE estimand is widely used and the target population by the estimand is easy to interpret.

Since only one potential outcome is observed for each subject, to estimate the causal effects, we make the following three assumptions.

1) Stable Unit Treatment Value Assumption, SUTVA (Angrist, Imbens and Rubin, 1996): a) the observed outcome under the assigned treatment is the same as the potential outcome under that treatment, and b) the potential outcomes for a given subject are not influenced by the treatment assignments of other subjects (Rubin, 1980; Angrist, Imbens, Rubin, 1996)

2) Positivity: each subject has a positive probability of being assigned to either treatment of interest: $0 < \Pr(Z_i = z_i | X_i) < 1$.

3) Ignorable treatment assignment: $(Y^1, Y^0) \perp\!\!\!\perp Z | X$; that is, treatment assignment is independent of the potential outcomes, given the covariates.

In this paper, we focus attention on the positivity assumption (2) by defining restricted definitions of the target populations and analysis methods to ensure that this condition holds and robust causal inferences are possible. The positivity assumption is violated when there exists neighborhoods of covariate space where there are subjects belonging to just one of the treatment groups being compared. Causal estimation for the subjects in this neighborhood depends on extrapolation, and can be imprecise and highly sensitive to model specifications. In order to obtain more credible causal estimates, we restrict analysis to a subpopulation where there is overlap in the propensity distributions of the treatment groups. We define such subpopulations as follows.

One alternative estimand is based on truncation of propensity score. For unit

i in the population with covariate values X_i, let $P_z(X_i) = \Pr(Z_i = z|X_i)$ denote the propensity of receiving treatment z, for $z = \{0, 1\}$. The positivity condition holds for the set of units $S(0)$ where $S(0) = \{i : P_z(X_i) > 0\}$. To eliminate units where the propensity to receive one treatment is small, we may further restrict the subpopulation to units within $S(0)$ where propensities for all treatments are above the αth quantile of the propensity distribution. Specifically, within S(0), let $F_z()$ denote the cumulative distribution of the propensity of receiving treatment z, that is, $F_z(a) = \Pr(P_z(X_i) \leq a)$. Then we restrict inferences to the subpopulation $S(\alpha)$ of $S(0)$ where $S(\alpha) = \{i : P_z(X_i) > F_z^{-1}(\alpha), \text{ for } z = \{0, 1\}\}$. In addition, we can also restrict inferences to the subpopulation $S^*(\alpha)$ of $S(0)$, where the probability of all treatment assignments is greater than a pre-defined level of α directly, that is $S^*(\alpha) = \{i : P_z(X_i) > \alpha, \text{ for } z = \{0, 1\}\}$. We can assess the sensitivity of causal effect estimates to changes in the α level. When the sample size increases, the α level can be reduced since there would be more subjects in the tails of the distributions and there would be more subjects with similar covariates available in the other treatment groups even at the tails.

Samuels (2017) formally defines an estimand called ATM as the average treatment effect on a evenly matchable set, M. An unit is called evenly matchable if, within a small propensity score stratum centered around the unit, there are at least as many units from the other group as from its own group. Suppose we divide the range of the propensity score into many small strata. Within each stratum, if there are equal numbers of units from both groups, all the units are evenly matchable; otherwise, only the units from the least prevalent group are evenly matchable. The evenly matched set is the union of all the matchable units from all the strata. The estimand ATM is defined as the average treatment effect on the evenly matchable set M, $E(Y^1 - Y^0|M)$. ATM can also be defined as the weighted average treatment effect $E[W_i\delta_i]/E[W_i]$, where the weight $W_i = \min\{P_1(X_i), P_0(X_i)\}$, and δ_i is the individual conditional treatment effect for subject i (Li and Greene, 2013).

Li et al (2017) defines another estimand called ATO, the average treatment effect on the overlap population. The overlap population is created by down-weighting the units with extreme propensity scores and up-weighting the units with propensity score close to 0.5. The target population is "the units whose combination of characteristics could appear with substantial probability in either treatment group." The ATO estimand is defined as the weighted average treatment effect $E[W_i\delta_i]/E[W_i]$, where the weight $W_i = Z_iP_0(X_i)+(1-Z_i)P_1(X_i)$, and δ_i is the individual conditional treatment effect for subject i. Although the population targeted by ATO is theoretically more balanced in the covariates between the treated and control groups, it is arguably less interpretable than the original population.

The ATM and ATO estimands are fixed regardless of sample size. As the sample size increases, more units with extreme propensity scores appear in the sample. This suggests reducing the α level for the truncated estimand as the sample size increases. Thus, the estimand defined by trimming off the tails would eventually approaches to the ATE. The estimands can be very different when there are heterogeneous treatment effects.

3.3 Methods

We consider three methods for utilizing propensity scores in combination with matching and truncation methods: 1) the inverse-probability-treatment-weighted estimator (IPTW), 2) the augmented IPTW (AIPTW) estimator (Scharfstein, Rotnitzky, and Robins, 1999), 3) penalized spline of propensity method for treatment comparison (PENCOMP) (Zhou, Elliott and Little, 2018). Under the assumptions stated in Section 3.2, the IPTW estimators are consistent if the propensity models are correct. Under the same assumptions, the latter two methods are "doubly robust".

The AIPTW estimators are consistent if either the propensity models or the outcome models are correctly specified. PENCOMP consistently estimates the causal effect of the treatment if either 1) the model for the propensity score and the relationship between the outcome and the propensity score are correctly specified through penalized spline or 2) the outcome model is correct. We then implement each of these methods in combination with either pair matching or propensity score truncation. We also consider the standard and doubly robust matching weight estimators (Li and Greene, 2013), and the overlap weight estimator (Li, Morgan, and Zaslavsky, 2017).

Next we describe the estimation procedures for the methods and the estimands targeted by each method.

3.3.1 PENCOMP and Rubin's Combining Rules

PENCOMP is a robust multiple imputation based approach to causal inference. Since each subject receives one treatment, we observe the potential outcome under the observed treatment but not the potential outcome under alternative treatment, as described in Chapter 2. We estimate causal effects by imputing the potential outcomes that are not observed using regression models that include splines on the logit of the propensity to be assigned that treatment as well other covariates that are predictive of the outcome. We then draw inferences based on comparisons of the imputed and observed outcomes between treatment groups. Here we describe the implementation of PENCOMP.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample $S^{(d)}$ from the original data S by sampling units with replacement. Then carry out steps (b)-(d) for each sample $S^{(d)}$:

(b) Estimate the propensity score model for the distribution of Z given X, with regression parameters γ_z. The propensity to be assigned treatment $Z = z$ is denoted as $\hat{P}_z(X) = \Pr(Z = z|X, \hat{\gamma}_z^{(d)})$, where $\hat{\gamma}_z^{(d)}$ is the ML estimate of γ_z. Define

$\hat{P}*_z$ =log$[\hat{P}_z(X)/(1-\hat{P}_z(X))]$.

(c) Check for balance and assess whether the propensity score model is adequate as described below in Section 3.4. The best propensity score model can be selected based on how well it balances the observed covariates between treatment groups. In addition, include the covariates and/or higher order terms in the prediction models to account for residual confounding.

(d) For each $z = 0, 1$, using the cases assigned to treatment group z, estimate a normal linear regression of Y^z on X, with mean

$$E(Y^z|X, Z = z, \theta_z, \beta_z) = s(\hat{P}*_z|\theta_z) + g_z(X; \beta_z),$$

where $s(\hat{P}*_z|\theta_z)$ denotes a penalized spline with fixed knots (Eilers and Marx, 1996; Ngo and Wand, 2004; Wand, 2003), with parameters θ_z, and $g_z()$ represents a parametric function of covariates predictive of the outcome, including covariates that are adequately balanced by the estimated propensity score models, indexed by parameters β_z. A different spline function is fitted for each treatment group, since there is no a priori reason to assume that the relationship between the potential outcomes under different treatment arms and the propensity of treatment assignment is the same. We consider a penalized B spline, which can be easily fitted with gam function in the R package mgcv.

(e) For $z = 0, 1$, impute the values of Y^z for subjects in treatment group $1 - z$ in the original data set with draws from the predictive distribution of Y^z given X_1 from the regression in (c), with ML estimates $\hat{\theta}_z^{(d)}, \hat{\beta}_z^{(d)}$ substituted for the parameters θ_z, β_z, respectively.

(f) Let $\hat{\Delta}^d$ and V^d denote the difference in treatment means and associated pooled variance estimate, based on the observed and imputed values of Y in each treatment group. The MI estimate of Δ is then $\bar{\Delta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\Delta}_d$, and the MI estimate of the variance of $\bar{\Delta}_D$ is $T_D = \bar{V}_D + (1 + 1/D)B_D$, where $\bar{V}_D = \sum_{d=1}^{D} V^d/D, B_D = \sum_{d=1}^{D}(\hat{\Delta}^d - \bar{\Delta}_D)^2/(D-1)$. The estimate Δ is t distributed with degree of freedom v, $(\Delta - \bar{\Delta}_D)T_D^{\frac{-1}{2}} \sim t_v$, where $v = (D-1)(1+\bar{V}_D/((D+1) \times B_D))^2$.

3.3.2 Estimands with PENCOMP

By using matching and truncation, we can obtain the ATE, ATM, ATO and both truncated estimands. Matching and truncation can be viewed as preprocessing techniques to reduce model dependence and avoid extrapolation outside the region that is supported by the data. This could be an important step when there is limited overlap in the distributions across treatment groups. If the entire sample is used in the analysis, PENCOMP estimate the ATE estimand. Otherwise, a restricted estimand is computed based on either truncation, pair matching or ATO weight.

3.3.2.1 Truncation

The truncation method restricts the sample to the set of cases defined by either $S(\alpha)$ (based on the quantile of the propensity distributions) or $S^*(\alpha)$ (based on the propensity score itself). The PENCOMP is then computed on this restricted subsample to obtain the truncated estimand.

3.3.2.2 Matching

The ATM estimand is obtained by selecting the treated and control subjects to form matched pairs. Each treated subject is paired with the closest control that is within the prespecified caliper and has not been matched yet. The caliper size governs the bias-variance tradeoff. If the caliper size is too large, the matched pairs would not be comparable so would increase the bias of the causal estimate. On the other hand, when the caliper size is too small, many subjects are dropped and the variance of the estimate increases. Here we set the caliper size at 0.25 times the logit of propensity scores. Although matching can be based on covariates, here we focus on propensity score-based matching. PENCOMP estimates are calculated on the matched set. By combining with pair matching, PENCOMP method can improve upon pair matching by adjusting for residual imbalance in the matched set.

3.3.2.3 ATO

Since the ATO estimand targets the a population that is a combination of both the treated and control populations, it can only be obtained by weighting the individual treatment effects by the ATO weights. Each treated subject is weighted by $W_{i1}(X_i) = 1 - P_{z_{i1}=1}(X_i)$ and the control by $W_{i0} = P_{z_{i1}=1}(X_i)$. The treatment effect is the weighted mean of the individual treatment effects after imputation. Specifically, let $\delta_i = Y^1 - Y^0$ denote the treatment effect for subject i, where Y^1 or Y^0 can be imputed or observed. In Section 3.3.1, in step f, $\hat{\Delta}^d = \sum_{i=1}^n W_i \times \delta_i / \sum_{i=1}^n W_i$ and the associated variance of the weighted mean is $V^d = \sum_{i=1}^n W_i^2 \times \sigma_{\delta_i}^2 / (\sum_{i=1}^n W_i)^2$.

3.3.3 Weighting Estimators: IPTW, Matching Weight and ATO

Each subject i is weighted by the balancing weight $W_i = \omega_i / \Bigg\{ Z_i P_{z_i=1}(X_i) + (1 - Z_i)(1 - P_{z_i=1}(X_i)) \Bigg\}$. The treatment effect Δ for a population of interest is defined as follows (Mao, Li and Greene, 2018):

$$\hat{\Delta}_{weighted} = \frac{\sum_{i=1}^n W_i Z_i Y_i}{\sum_{i=1}^n W_i Z_i} - \frac{\sum_{i=1}^n W_i (1 - Z_i) Y_i}{\sum_{i=1}^n W_i (1 - Z_i)}$$

Different specifications of ω_i yield average treatment effects for different subpopulations. For the estimand ATE, ω_i is 1 which defines the IPTW estimator. For the estimand ATO, ω_i is $P_{z_i=1}(X_i) \times P_{z_i=0}(X_i)$. For the truncated estimands, ω_i is set as $I\{i \in S(\alpha) \text{ or } i \in S^*(\alpha)\}$, where I is the indicator. For the estimand ATM, ω_i is set as $\min\Bigg(P_{z_i=1}(X_i), P_{z_i=0}(X_i) \Bigg)$. In addition to using the balancing weight ω_i, another way to obtain the ATM estimand is by combining IPTW with pair matching. When calculating the IPTW estimates on the matched set, the propensity scores are reestimated after matching.

The $\hat{\Delta}$ for each estimand is computed on the original data S. The standard errors are estimated using bootstraps. The procedures are as follows.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample S^d from the original data S by sampling units with replacement. Then carry out steps (b)-(d) for each sample S^d:

(b) Select and estimate the propensity score model as described below.

(c) Check for balance and assess whether the propensity score model is adequate as described in Section 3.4. The best propensity score model can be selected based on how well it balances the weighted covariates.

(d) Estimate the weighted estimator on each bootstrap sample, Δ^d.

(e) The standard errors $\hat{sd}_D$ for $\hat{\Delta}$ based on D bootstrap samples are computed as follows.

$$\hat{sd}_D^2 = \sum_{d=1}^{D} (\hat{\Delta}_d - \hat{\Delta}_{\cdot}^{*})^2/(D-1)$$

where $\hat{\Delta}_{\cdot}^{*} = \sum_{d=1}^{D} \hat{\Delta}_d/D$. The 95% confidence intervals are computed as $\hat{\Delta} \pm 1.96\hat{sd}_D$.

3.3.4 Augmented Weighted Estimators

For each weighting estimator as described in Section 3.3.3, an augmented weighting estimator can be defined as follows (Mao, Li and Greene, 2018) :

$$\begin{aligned}\hat{\Delta}_{aug} = \frac{\sum_{i=1}^{n} \omega_i \{m_1(X_i, \alpha_1) - m_0(X_i, \alpha_0)\}}{\sum_{i=1}^{n} \omega_i} &+ \frac{\sum_{i=1}^{n} W_i Z_i \{Y_i - m_1(X_i, \alpha_1)\}}{\sum_{i=1}^{n} W_i Z_i} \\ &- \frac{\sum_{i=1}^{n} W_i (1 - Z_i) \{Y_i - m_0(X_i, \alpha_0)\}}{\sum_{i=1}^{n} W_i (1 - Z_i)}\end{aligned}$$

where $m_1(X_i, \alpha_1) = E(Y_i|X_i, Z_i = 1)$ and $m_0(X_i, \alpha_1) = E(Y_i|X_i, Z_i = 0)$. Throughout the paper, we refer the augmented estimator with $\omega_i = 1$ as AIPTW. Similar procedures based on bootstrap samples are used to estimate the standard error for $\hat{\Delta}_{aug}$.

3.4 Balance Checking

To assess whether the propensity score model is adequately specified, we assess whether the covariate distributions between the treated and control are balanced, after conditioning on the propensity scores, such as weighting, matching or regressing. One measure of balance is the absolute standardized mean difference in each covariate between the treated and control groups. The methods we consider here have different ways of assessing balance. For matching, the absolute standardized mean difference in covariate x is calculated on the matched set:

$$d_{match} = \left| \bar{x}_1 - \bar{x}_0 \right| \Big/ \sqrt{\frac{s_1^2 + s_0^2}{2}}$$

where s_z^2 is the variance of the original covariate in the entire treated or control groups before adjusting for propensity scores. For the weighting estimators, covariate balance is assessed by the absolute standardized weighted mean difference as follows:

$$d_{weight} = \left| \frac{\sum_{i=1}^{n} w_{i1} z_i x_i}{\sum_{i=1}^{n} w_{i1} z_i} - \frac{\sum_{i=1}^{n} w_{i0}(1-z_i) x_i}{\sum_{i=1}^{n} w_{i0}(1-z_i)} \right| \Big/ \sqrt{\frac{s_1^2 + s_0^2}{2}}$$

where the weights w_{i1} and w_{i0} are different across the weighting methods and defined in Section 3. As an analog to the above measures, we assess balance as follows with PENCOMP:

$$d_{pencomp} = \left| \bar{x}_{res1} - \bar{x}_{res0} \right| \Big/ \sqrt{\frac{s_1^2 + s_0^2}{2}}$$

where x_{res} is the residual after regressing the original covariates on the spline of propensity score. Here for comparison across the different measures, we use the same s_1^2 and s_0^2, which are calculated on the original dataset.

3.5 Simulation

In this section we explore the performance of our proposed approach combining with truncation or matching to alternative weighting approaches and matching, as discussed in Section 3.3. We propose 1) combining PENCOMP(IPTW, AIPTW) with truncation at an α quantile level or at propensity α level, referred to as PENCOMP(IPTW, AIPTW)α and PENCOMP(IPTW, AIPTW)$\alpha *$; 2) combining PENCOMP (IPTW, AIPTW) with caliper matching, referred to as PENCOMP(IPTW, AIPTW)+match. The three additional weighting approaches we compare with are 3) matching weights, both standard and the doubly robust version, referred as match weight and match weight DR respectively; and 4) the overlap weights (ATO). We compare the methods using empirical bias, root mean squared error (RMSE), 95% coverage, ratio of empirical bias as a fraction of empirical RMSE, and mean 95% confidence interval width. We compare the methods when the prediction and/or propensity models are correct. Specifically, we compare the following cases: A) correctly specified prediction and propensity models; B) incorrectly specified prediction model but correctly specified propensity model; and C) correctly specified prediction model but incorrectly specified propensity model.

In our simulation, we assess the influence of these three factors on the relative performance of the methods. The first factor is the degree of overlap in the propensity score distributions between the treatment groups. The second factor is the relative importance of each covariate in predicting the treatment assignment and the outcome. There are three types of covariates: covariates are predictive of only the treatment or the outcome, and true confounders-covariates that are predictive of both the treatment and outcome. We consider two scenarios: 1) aligned-the same set of covariates, and 2)misaligned-different set of covariates predicting the outcome and treatment. The third factor considers whether treatment effects are heterogeneous or not. In the case of homogeneous treatment effects, all the methods estimate the same quantity. Otherwise, the estimands are different and each method is evaluated based on its own truth.

For the heterogeneous treatment effects case, we simulate each dataset as described below. Each simulated dataset contains three baseline covariates, $X = [X_1, X_2, X_3]$, which are independently and normally distributed as $N(0, 1)$. The treatment Z is Bernoulli distributed with probability of being assigned $Z = 1$ depending on X_1 and X_2. The outcomes Y^1 and Y^0 are normally distributed with variance of 1 and means that depend on X_1 and X_2 in the aligned case and X_2 and X_3 in the misaligned case. Table 3.1 details the simulation scenarios.

	Intercept	X_1	X_2	X_1X_2	X_1^2	X_2^2	X_3	X_3^2
Treatment Assignment								
Low	0	1.5	1.5	0.75				
High	0	0.1	0.1	0.05				
Aligned and Parallel								
Y_0	0	1	3		2	2		
Y_1	5	1	3		2	2		
Aligned and Not Parallel								
Y_0	0	1	3					
Y_1	5	1	3		2	2		
misaligned and Parallel								
Y_0	0		3		2		1	2
Y_1	5		3		2		1	2
misaligned and Not Parallel								
Y_0	0		3				1	
Y_1	5		3		2		1	2

Table 3.1: Simulation Scenarios: logistic regression model for treatment assignment, and linear outcome model parameters.

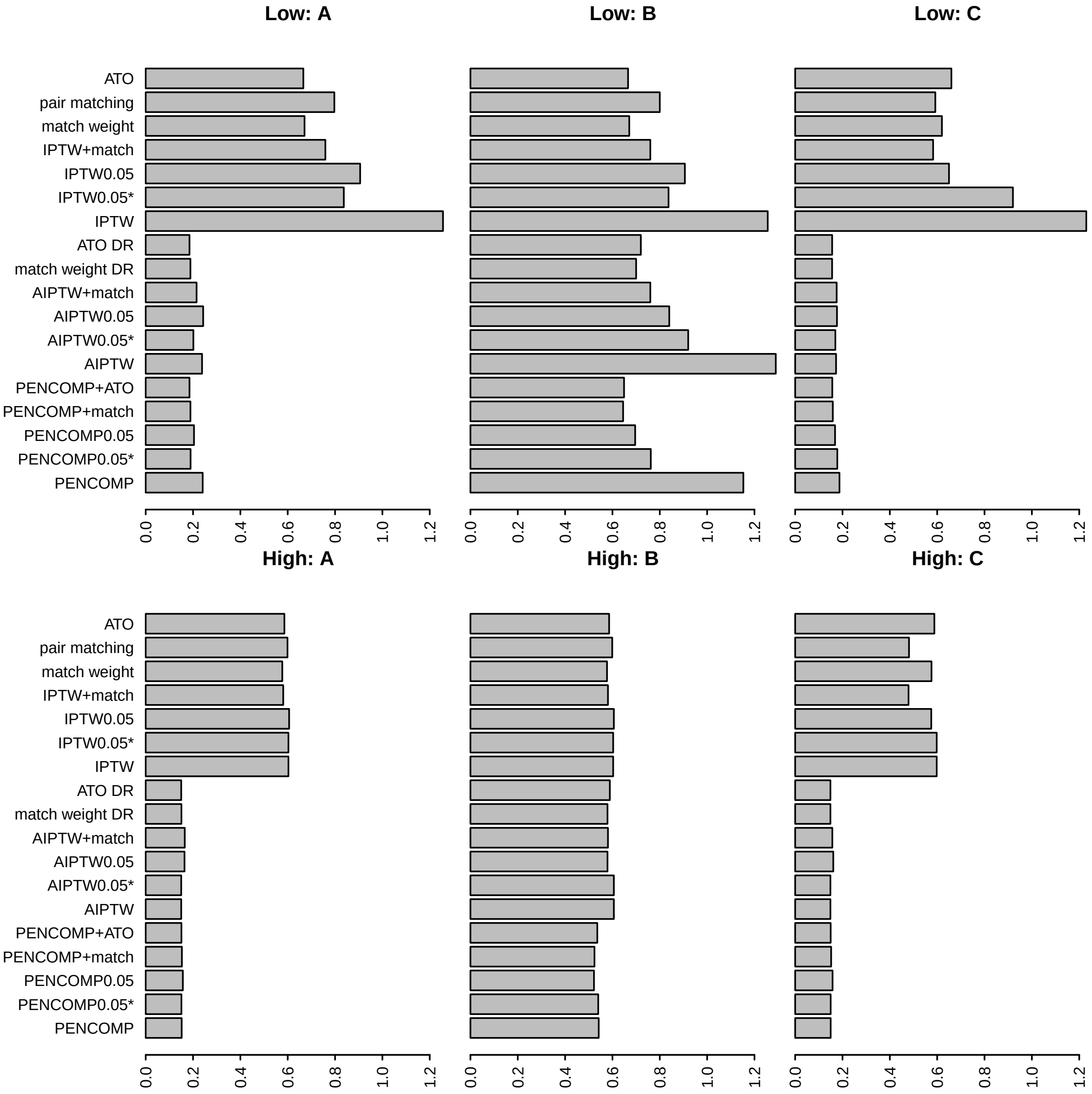

Figure 3.1: Parallel surface and Misaligned: Empirical RMSE, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

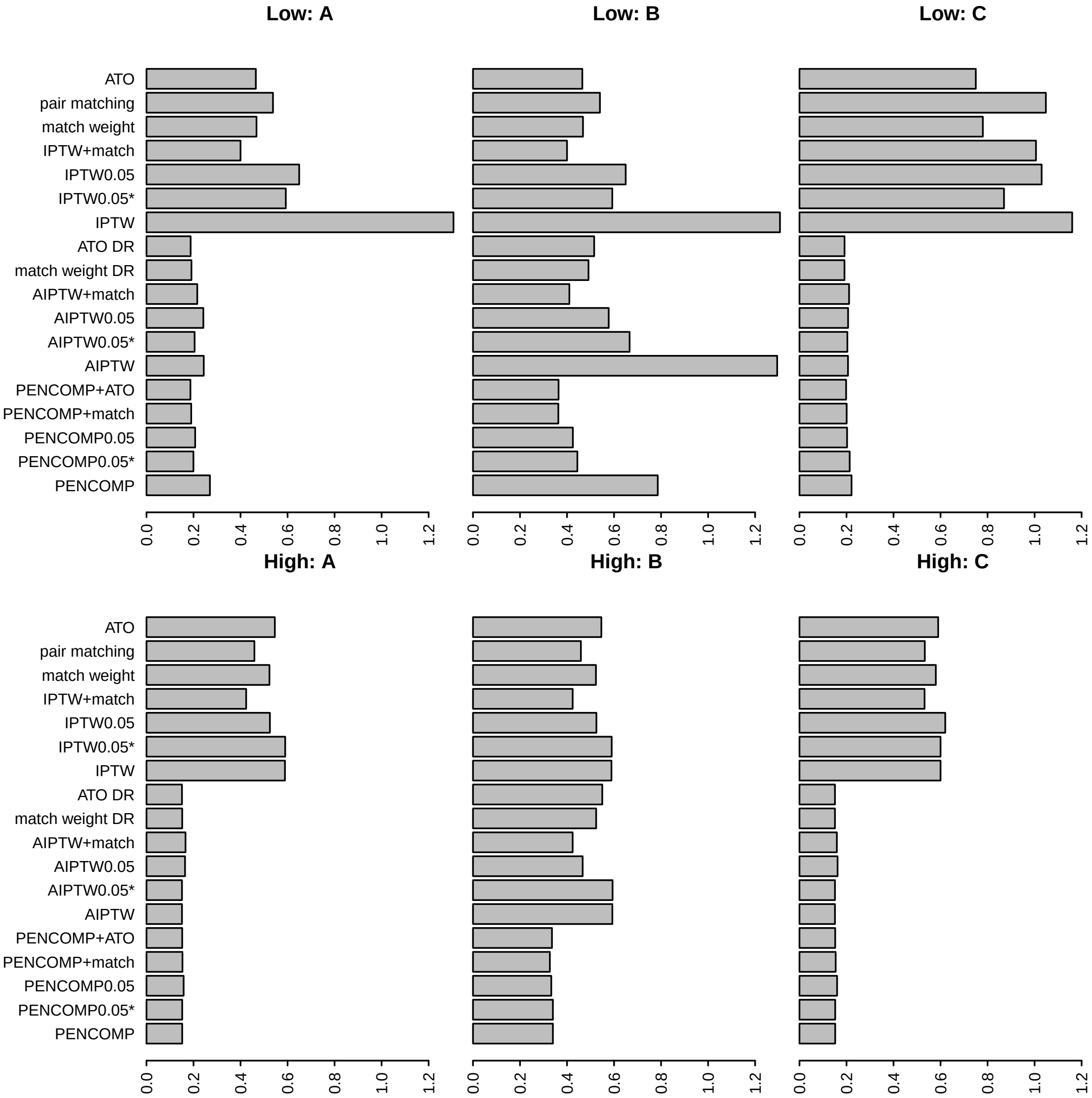

Figure 3.2: Parallel surface and Aligned: Empirical RMSE, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

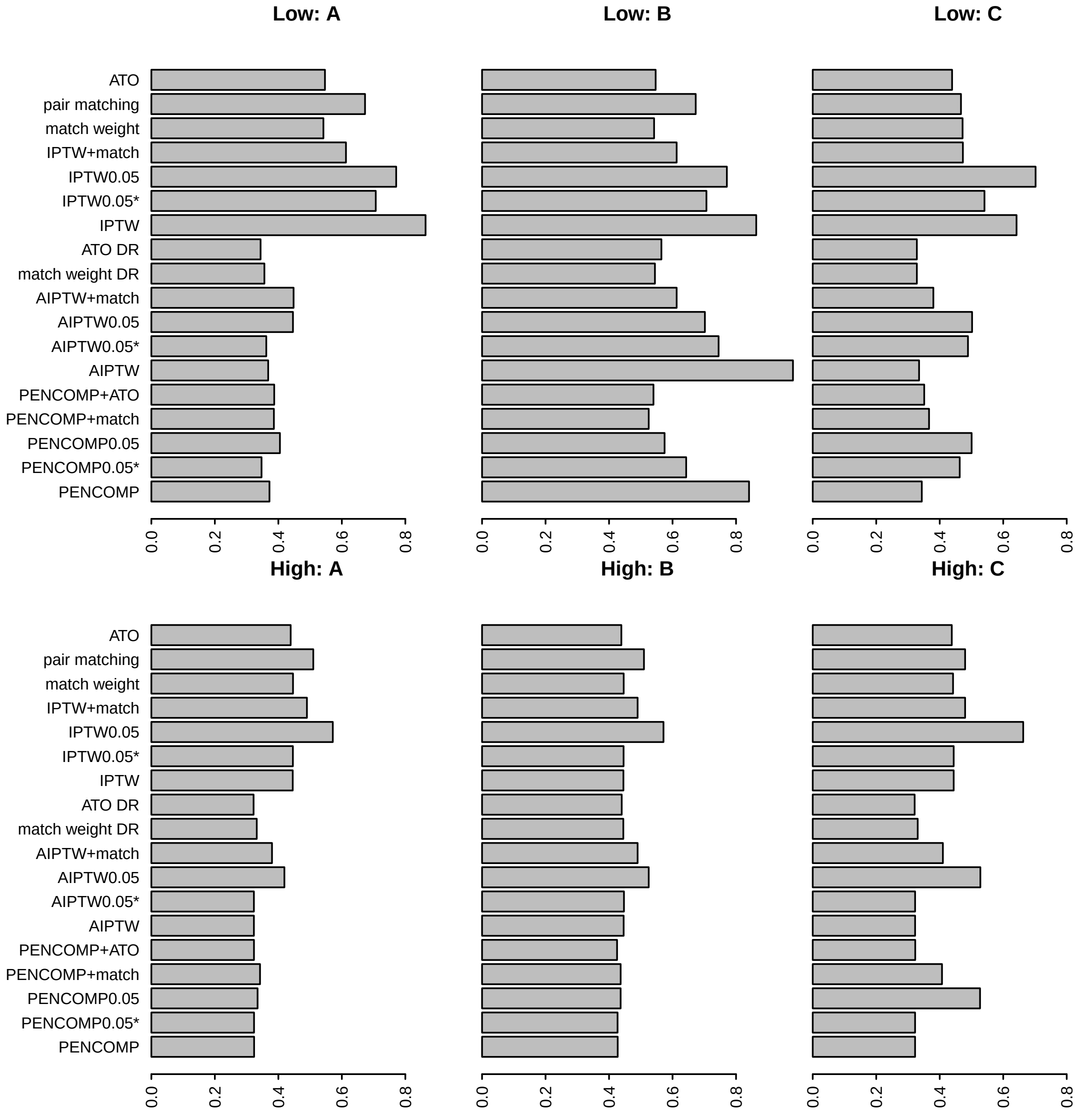

Figure 3.3: Nonparallel surface and Misaligned: Empirical RMSE, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

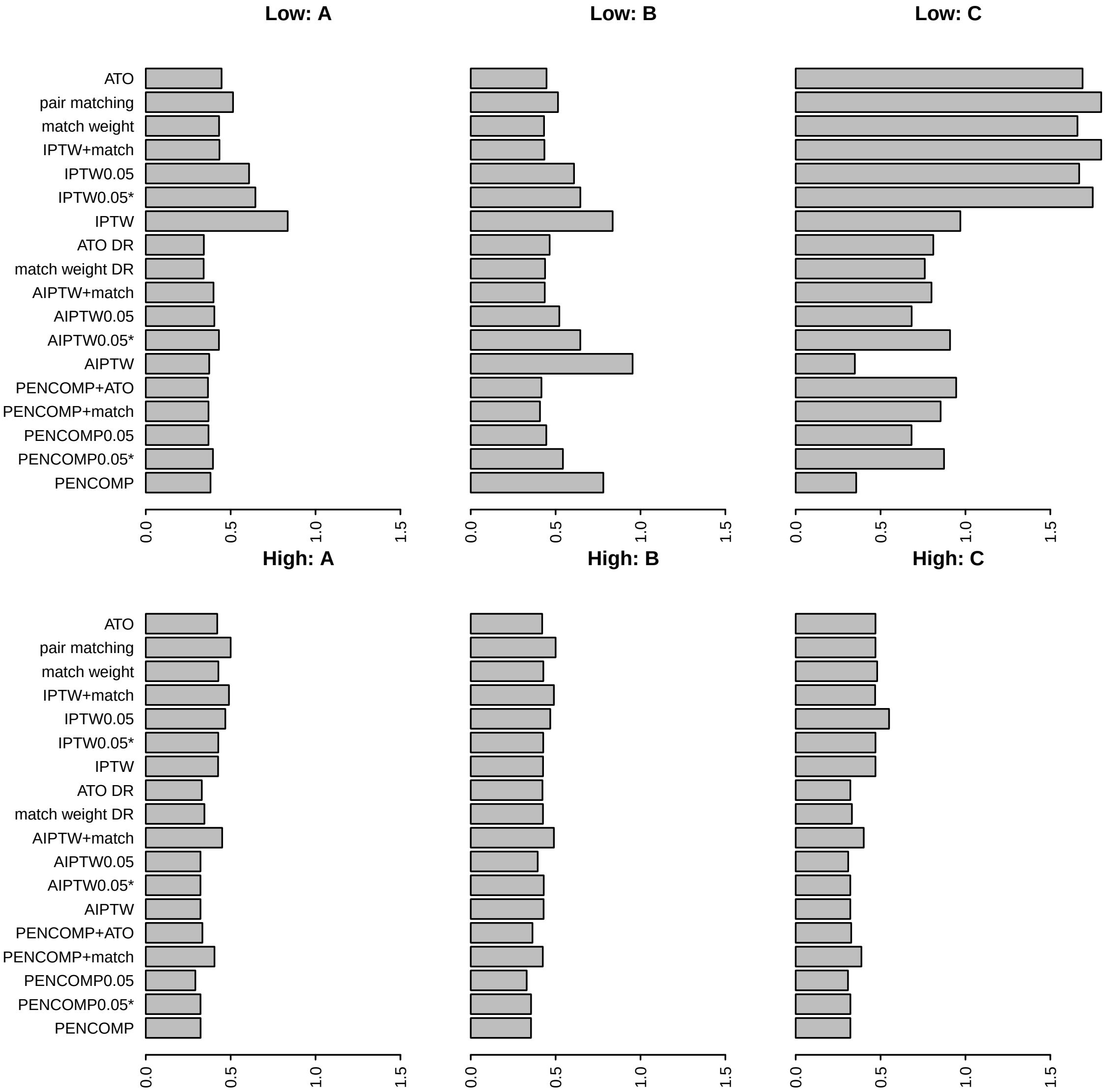

Figure 3.4: Nonparallel surface and Aligned: Empirical RMSE, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

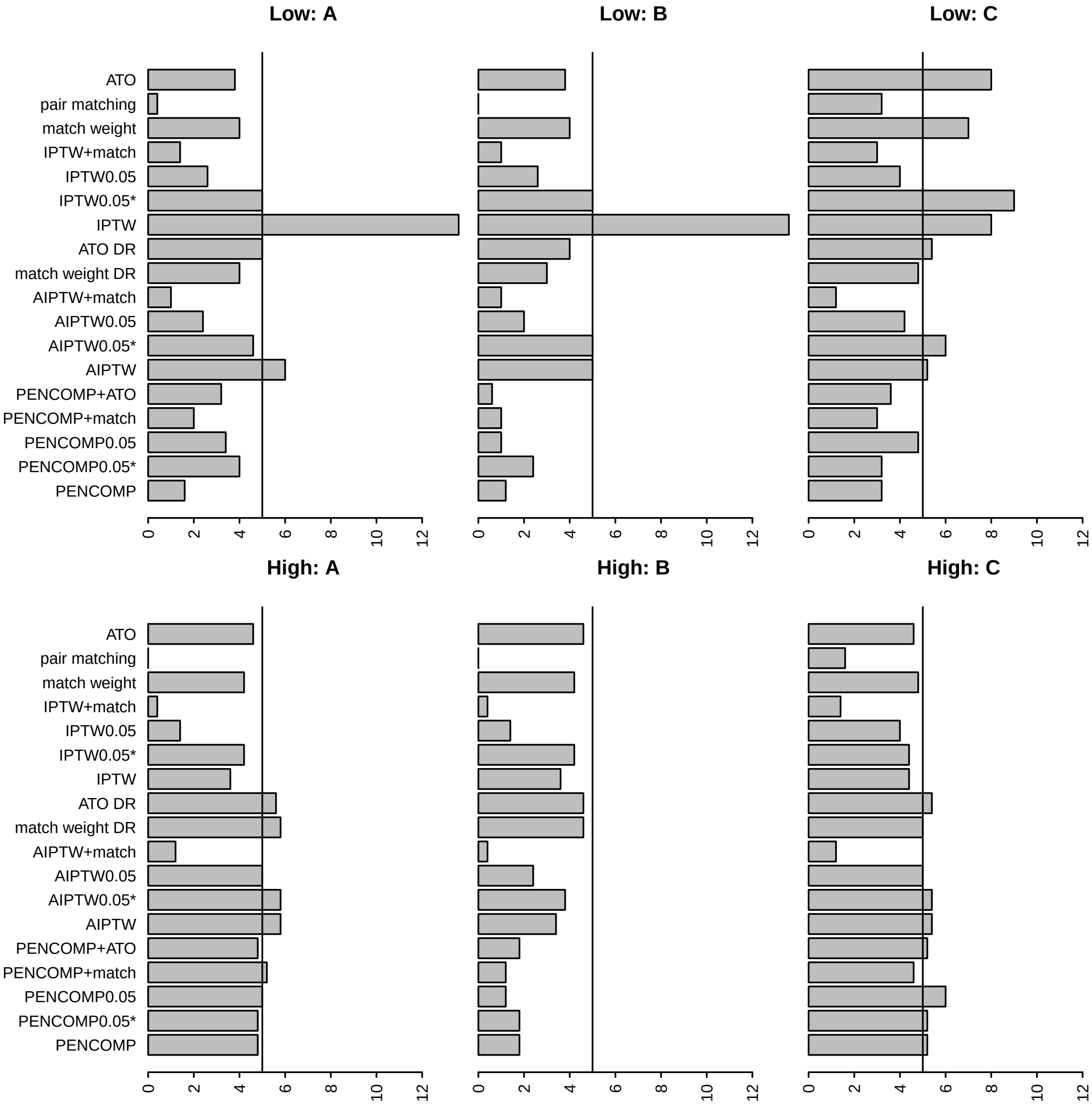

Figure 3.5: Parallel surface and Misaligned: 100 * 95% non coverage rate, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

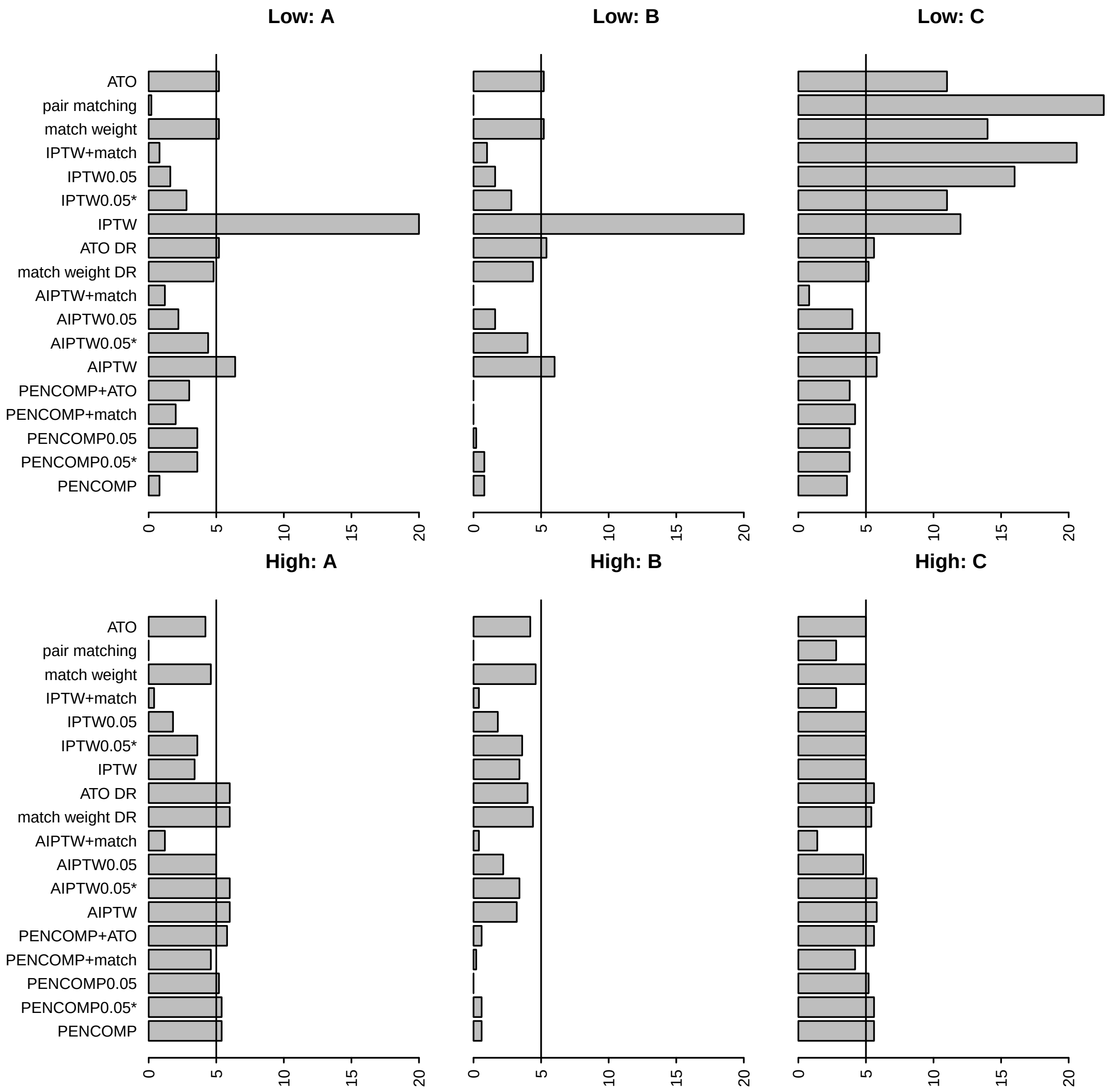

Figure 3.6: Parallel surface and Aligned: 100 * 95% non coverage rate, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

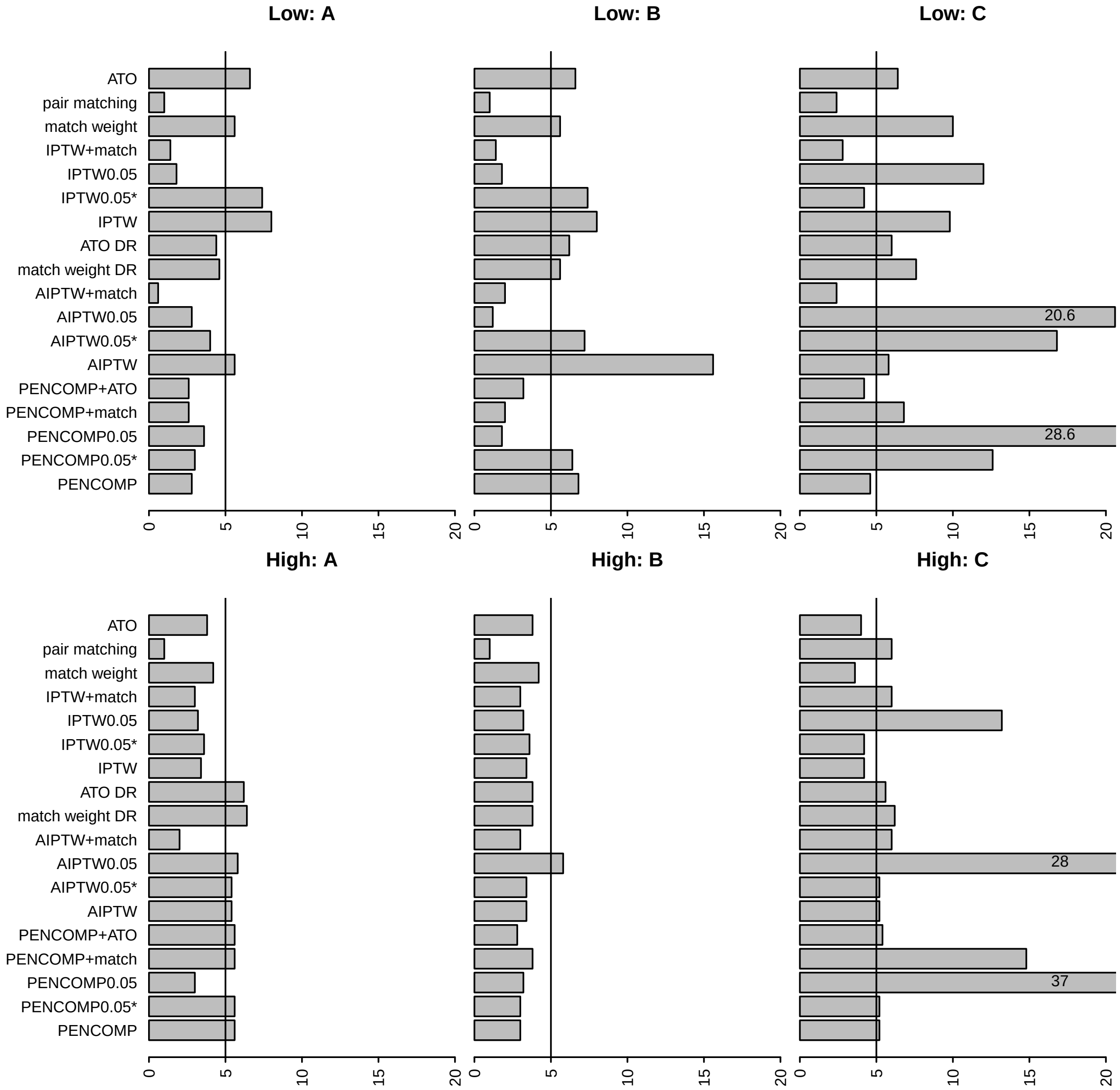

Figure 3.7: Nonparallel surface and Misaligned: 100 * 95% non coverage rate, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

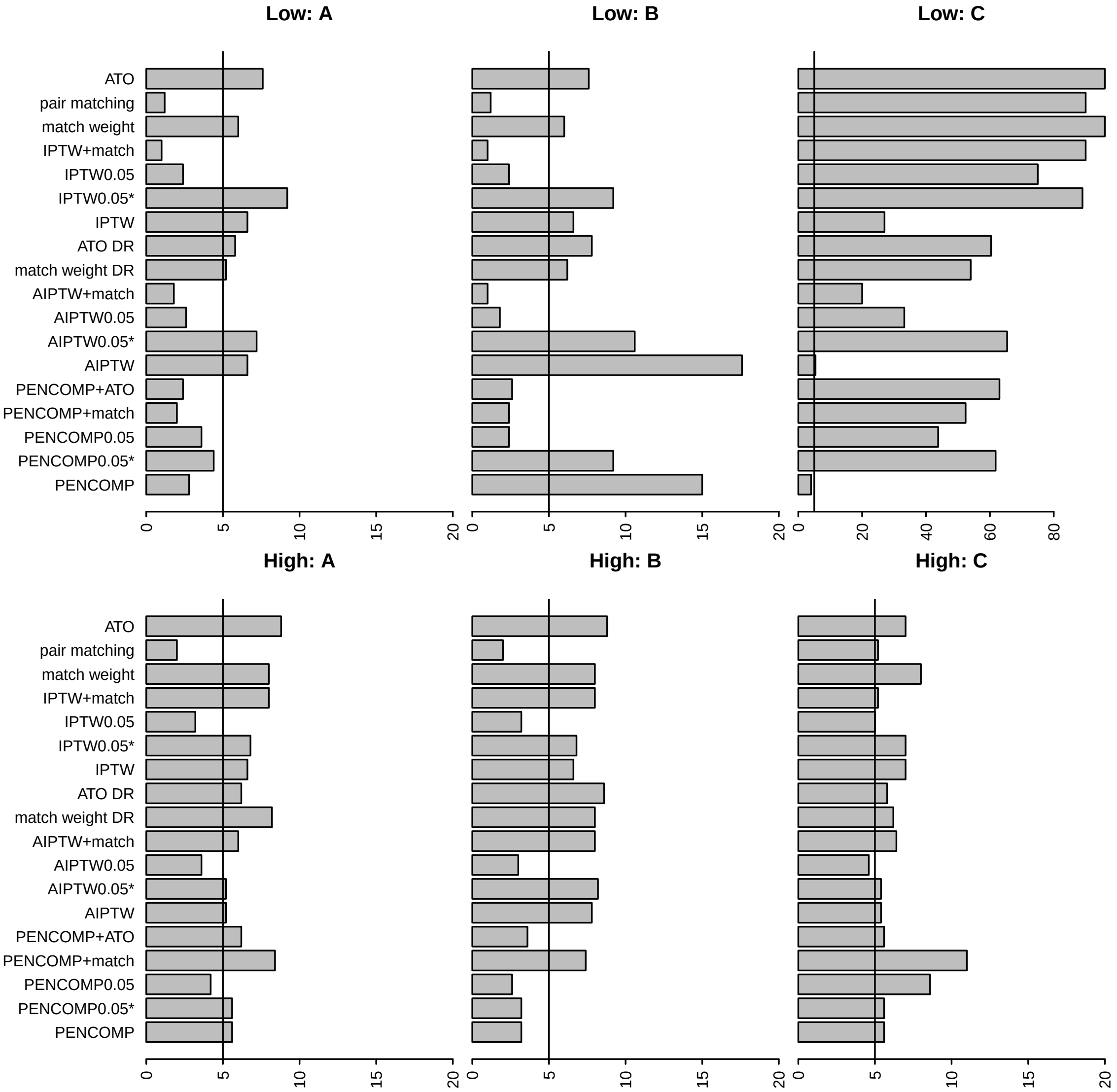

Figure 3.8: Nonparallel surface and Aligned: 100 * 95% non coverage rate, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

Figures 3.1-3.8 show the results for sample size of 200. The results on empirical RMSEs are shown in Figures 3.1-3.4. For Figures 3.1-3.2, the outcome surfaces were parallel so the ATE and restricted estimands were the same. When there was a high degree of overlap in the propensity distribution, restricted estimands such as ATM, ATO and truncated, didn't improve the RMSE much, and as expected, the

ATE estimand could be estimated reliably. The methods that didn't incorporate the outcome model–pair matching, IPTW, truncated IPTW, ATO, and match weight–performed similarly when the compared groups overlapped sufficiently. The robust methods that incorporate the outcome models had smaller RMSEs than the methods that did, but less so when the outcome models were misspecified. When the overlap was low as shown in the top panels in Figures 3.1-3.2, the performance of the methods varied more greatly, especially when the outcome models were misspecified. The RMSEs of the restricted estimands had smaller RMSE than that of the ATE estimand, especially when the overlap was low and the outcome models were misspecified. For example, in Figure 3.2, the RMSE of IPTW for the ATE estimand reduced from over 1.2 to less than 0.8 for restricted estimands. Similarly, the RMSE of AIPTW for the ATE estimand went from over 1.2 to less than 0.8, and PENCOMP went from 0.8 to less than 0.5. When the outcome models were correct, as seen in (A) and (C) in Figures 3.1-3.2, the ATE estimands had similar RMSE as the retricted estimands. Overall, PENCOMP had comparable or smaller RMSEs than the augmented weighted estimators for both the ATE and restricted estimands.

Figures 3.3-3.4 show the results for nonparallel surfaces. Similar patterns were observed: restricting inference to subpopulations improved RMSE when the overlap was low, especially when the outcome models were misspecifed as well. Unlike the parallel surfaces, the ATE and restricted estimands were different. Furthermore, the restricted estimands also changed with the specifications of the propensity score models. Hence, in Figure 3.3-3.4 (C), restricting estimands could increase the RMSEs since misspecifying the propensity score models altered the estimands.

Figures 3.5-3.8 show the noncoverage rates of all the methods. As expected, the coverage rates for all the methods were close to the nominal coverage when the overlap in the propensity distributions was high, compared to when the overlap was low. In the presence of low overlap, the IPTW for the ATE estimand had very low coverage rates and restricting estimands improved the coverage rates significantly. Low overlap could also affect the coverage rates of the robust weighting methods but less so, since the outcome models attenuated some of the effects of low overlap. As seen in our previous studies, PENCOMP tended to have more conservative coverage rates than the weighting estimators. Furthermore, when the propensity models were misspecified, dropping subjects yielded poor coverage rates since the empirical biases were larger and the subpopulations were not correctly defined, as seen in Figure 3.7-3.8 (C).

The results on empirical bias are shown in the Appendix B.1-B.4. Overall, the empirical biases associated with restricted estimands tended to be smaller than that of the ATE estimand. As expected, when the outcome models were correct, the biases were negligible. When the overlap was high, all the methods had very small empirical biases, regardless whether the outcome models were incorporated or not. For Figures B.3-B.4, the restricted estimands under the misspecified propensity models were different from those under the correctly specified propensity scores. Hence, the empirical biases for the restricted estimands increased significantly, as seen in Figures B.3-B.4 (C).

Lastly, in the Appendix Figure B.5-B.16 present the results on the RMSEs, empirical bias, and coverage rates for sample size of 1000. Similar patterns as before were observed. Overall, the coverage rates and RMSEs were better when the sample sizes increased. PENCOMP tended to have comparable or smaller RMSE than AIPTWs. When the overlap between the compared treatment groups was low, restricting inference to subpopulations that were more supported by the data tended to perform better. PENCOMP provides a viable alternative for estimating both the ATE and restricted estimands considered here.

3.6 Application

The Multicenter AIDS Cohort study (MACS) was started in 1984 (Kaslow et al, 1987). A total of 4,954 gay and bisexual men were enrolled in the study and followed up semi-annually. At each visit, data from physical examination, questionnaires about medical and behavioral history, and blood test results were collected. The primary outcome of interest was the CD4 count, a continuous measure of how well the immune system functions. We used this dataset to analyze the short term (1 year) effects of using antiretroviral treatment on disease progression. Here we restrict our analyses to the period between visit 7 and 12, after the first antiretroviral treatment zidovudine (AZT) was approved for use and before the advent of highly active antiretroviral therapy (HAART). Treatment was coded to 1 if the patient reported taking any of antiretroviral treatment (ART) or enrolling in clinical trials of such drugs. We estimate the short-term (6-month) effects of using any antiretroviral treatment for HIV+ subjects. We excluded subjects with missing values on any of the covariates included in the models. We log-transformed the blood counts in this analysis.

Here we treat each visit as a single time point treatment. Let $t = 1$ denote the time when the treatment was administered, and $t = 2$ the time 6-month later when the outcome was measured. In addition, let $t = -1, -2, -3$ denote 1, 2, and 3 visits away from the current visit $t = 1$. Let $X(t = 1, -1, -2, -3)$ denote the blood count histories prior to treatment assignment. Let Z be the binary treatment indicator. Let $Y(t = 2)$ be the CD4 count 6 months after the treatment. For the propensity score model, we considered blood counts-CD4, CD8, white blood cell (WBC), red blood cell (RBC), and platelets and treatment histories from the most recent 4 visits, as well as demographic variables-college education, age, and race. The treatment assignment Z was modeled as a logistic regression. For the outcome model, we considered the last two CD4 counts and their squared terms. We estimated the mean CD4 count difference between the treated and the control at each visit, denoted as Δ, from visit 7 to visit 12. For PENCOMP, we replaced the simulated/imputed transformed CD4 values that were < 0 with 0 (i.e. below detection level). A total of 15 equally spaced knots and B spline were used.

As shown in Figure 3.9, we see that over time the treated and control subjects became more disimilar. The propensity score distributions became more and more skewed, as the treated had propensity of treatment close 1 and the control close to 0. We measured the proportion of subjects in the control group whose propensity scores were between the $1 - \alpha$ and α quantiles of the propensity score distribution of the treated group, denoted as $\pi_{z=0}^{1-\alpha} = F_{z=0}(F_{z=1}^{-1}(1-\alpha)) - F_{z=0}(F_{z=1}^{-1}(\alpha))$, where F is the cumulative distribution. Inside this region it is easier to impute missing potential outcomes Y^0 because there are more observations. Similarly, for $\pi_{z=1}^{1-\alpha}$. The small proportions suggested difficulty in imputing missing potential outcomes. Since the propensity score distributions were extreme and the overlap was low, the sample sizes after matching were much smaller than before, as seen in Table 3.2.

Table 3.2: Sample sizes before and after trimming and matching. The measure of overlap of the original data at each visit: $\pi_{z=1}^{1-\alpha}$ and $\pi_{z=0}^{1-\alpha}$ for $\alpha = 5\%$.

				Trimming						
	All		quantile $\alpha = 0.02$		$\alpha = 0.02$		Matching		Overlap	
Visit	treat	control	treat	control	treat	control	treat	control	$\pi_{z=1}^{0.95}$	$\pi_{z=0}^{0.95}$
visit 7	98	575	86	404	56	467	69	69	0.42	0.51
visit 8	127	468	102	375	44	256	58	58	0.28	0.51
visit 9	160	418	136	362	55	236	65	65	0.25	0.42
visit 10	194	412	159	378	81	269	69	69	0.24	0.38
visit 11	227	287	227	283	177	198	116	116	0.67	0.54
visit 12	302	209	125	193	69	87	65	65	0.17	0.25

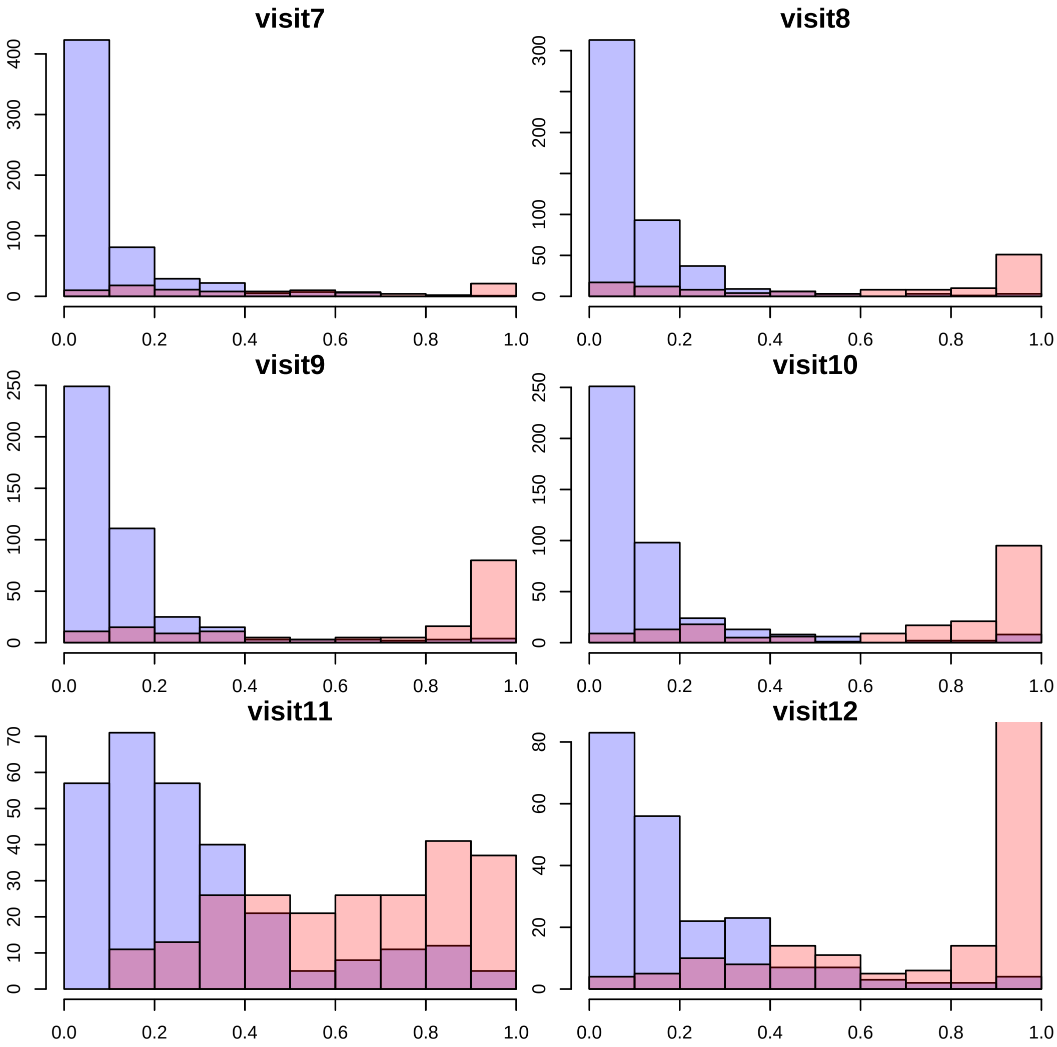

Figure 3.9: Propensity distribution for visit 7-12.

One important step in building the propensity score models is to check for balance in the covariates. At each visit, we checked for covariate balance between the treated and control groups, as described in Section 3.4. As shown in Figure 3.10, trimming or matching first reduced the standardized mean differences in the baseline covariates between the treatments groups. As expected, the ATO weights achieved the best balance.

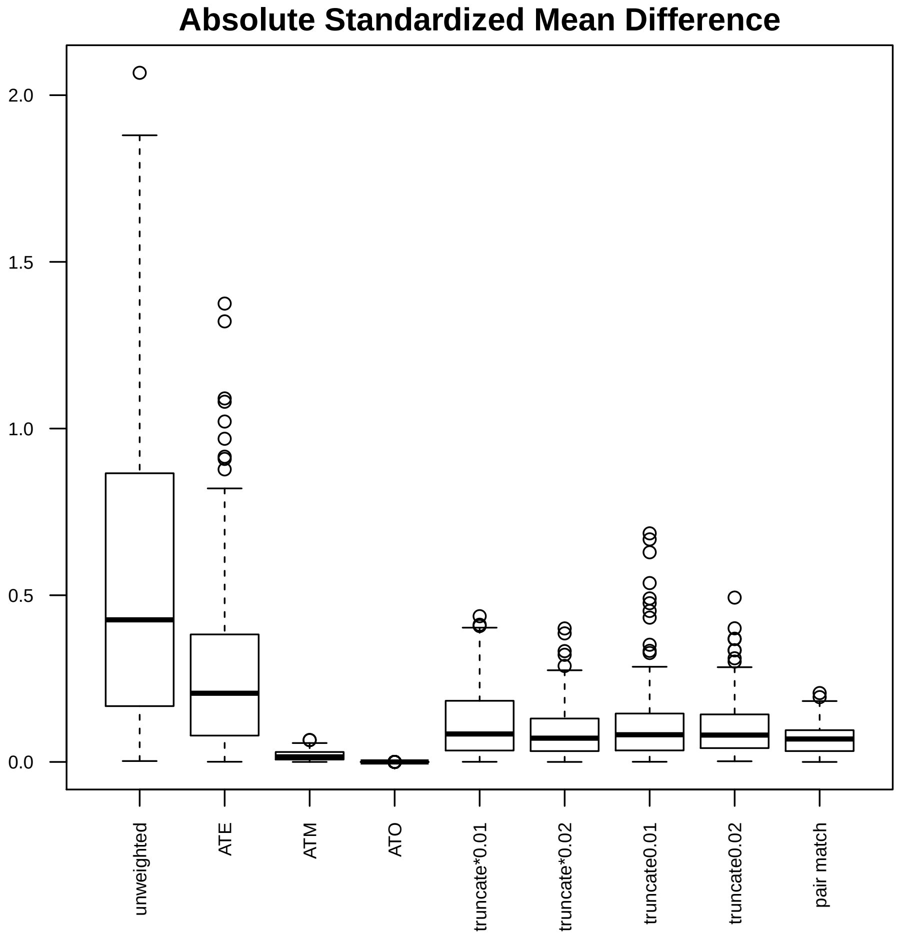

Figure 3.10: Absolute standardized mean differences across all the visits 7-12.

We estimated the short term effect of antiretroviral treatment on CD4 count using pair matching, PENCOMP, both the weighted and augmented weighted estimators. The results for visit 7-12 are summarized in Figure 3.11. The standard errors were obtained using 1000 bootstrap samples. For PENCOMP, 1000 complete datasets were created. The naive estimators were negative, suggesting a harmful effect of antiretroviral treatment on CD4 count. This is likely due to uncontrolled confounding by indication, in that sicker subjects with lower CD4 counts were more likely to be assigned to treatment. All the treatment effect estimates, seen in the first column of Figure 3.11, suggested less harmful effects. The weighted estimators and pair matching performed worse than the robust methods–the augmented estimators and PENCOMP, especially for the ATE estimand. For the ATE estimand, PENCOMP had smaller standard errors than the augmented weighted estimator when the weights were variable, as found in Chapter 2. With pair matching, many subjects were dropped due to extreme propensity scores and the sample sizes became very small, as shown in Table 3.2. The re-estimated propensity scores could potentially became more extreme so the IPTW(AIPTW)+match estimators performed worse in terms of standard errors. Thus, here we used the match weights as described in Section 3.3.3. For PENCOMP, we also used the match weights to weight the individual causal effects. The match weight estimators performed better than pair matching because there was low overlap in the propensity scores and many subjects were dropped. For the alternative causal estimands–ATM, ATO and truncated estimands, PENCOMP had a comparable performance to the augmented weighted estimators, with PENCOMP having slightly smaller standard errors than the augmented weighted estimators for the truncated estimands.

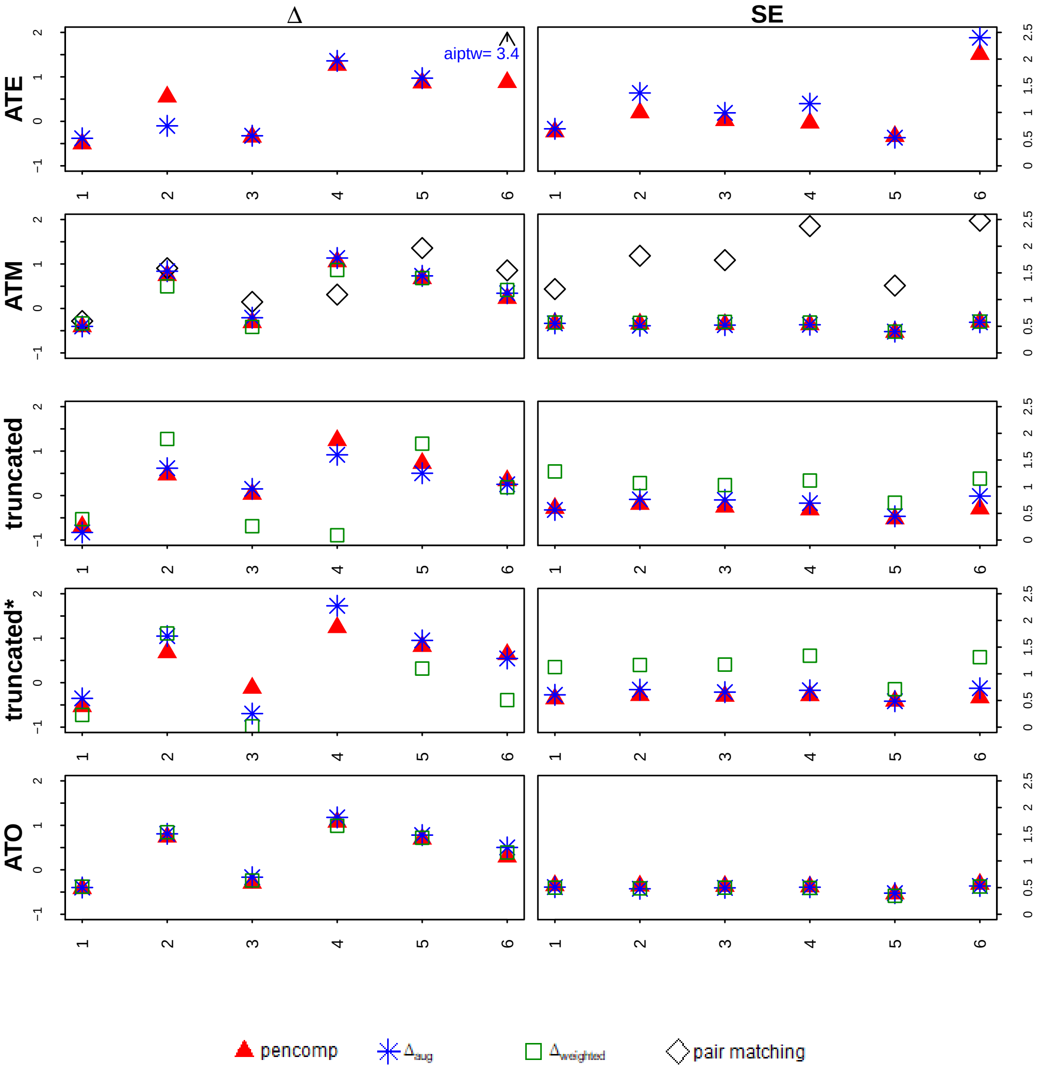

Figure 3.11: Treatment effect estimates and standard error (SE) for the ATE, ATM, and truncated estimands for visit 7-12. Truncated*: truncating at quantile level $\alpha = 0.02$ of the propensity score distributions. Truncated: truncating at α level of propensity score. Naive estimates(SE) for visit 7-12 were -7.7(0.7), -7.0(0.6), -6.8(0.6), -6.3(0.6), -5.5(0.6), and -8.2(0.6), respectively. The IPTW estimates(SE) for ATE estimand were 1.9(1.7), 2.0(2.8), 2.2(3.0), 2.7(3.0), 0.2(0.9), and 7.3(5.1), respectively.

3.7 Discussion

Here we show that PENCOMP has the flexibility of estimating different estimands when needed and its performances can improve for restricted estimands when the overlap is low. In general, it tends to outperform the weighted estimator for the ATE estimand and has comparable performance for restricted estimands, in terms of RMSE and coverage rate.

All the previous approaches described above rely on estimated propensity scores to determine the subpopulation. Trimming off extreme propensities to increase precision of an estimator could be harmful if deleted subjects are of interest to investigators (Lechner 2008). One alternative to trimming is to provide nonparametric bounds (Lechner 2008). Because lack of overlap is a small sample problem, subjects with extreme propensities might still be relevant and if more samples were taken, there would be subjects in the other treatment group who have similar propensity scores. In addition, defining the subpopulation in term of estimated propensity scores might not be meaningful to investigators who are more interested in identifying that subpopulation in term of observed covariates.

The propensity score plays an important role in identifying the common support region. However, its performance depends on what variables are included in the model. A small set of covariates $W \in X$ might exist such that $0 < \Pr(Z = z|W) < 1$ and the ignorability assumption $(Y^1, Y^0) \perp\!\!\!\perp Z|W$ still hold. Hill and Su (2013) define $0 < \Pr(Z = z|W) < 1$ as common causal support, where W is the set of covariates such as $(Y^1, Y^0) \perp\!\!\!\perp Z|W$ holds. Thus, it might not be necessary to require common support on all the covariates X. For example, a propensity score model based only on treatment assignment can be inefficient, as it prioritizes variables predictive of the treatment but not necessarily predictive of the outcome. Common support on such predictors are not relevant since these predictors are not confounders. Furthermore, dropping subjects due to lack of overlap on such predictors could increase variance of the estimate, since including such predictors in the propensity model could potentially shrink the overlap region of the propensity scores, especially in a high dimensional setting. It is much harder to have overlap in many covariates, so it is more important to consider a more sparse propensity score model that satisfies the assumption of unconfoundedness. Chapter 4 addresses the issue of model selection in causal inference.

CHAPTER IV

Variable Selection in Causal Inference

4.1 Introduction

The propensity score, which is defined as the probability of treatment assignment given covariates, plays an important role in bias reduction for estimation of causal effects from nonrandomized studies. The propensity score has the balancing property: conditional on the propensity score, the observed covariates and treatment assignment are conditionally independent (Rosenbaum and Rubin, 1983). The balancing property of propensity score implies that adjusting for the propensity score can remove the bias due to differences in all observed confounders between the treatment groups (Rosenbaum and Rubin, 1983). One important assumption needed for propensity score-based methods to make valid inference about causual effects is that all the confounders are observed and included in the propensity model. Since excluding important confounders in the model can lead to biased estimates, many covariates are often included, for fear of excluding some important confounders. Rubin (2007) notes that only pretreatment covariates should be included in the propensity model and argues that the model should be selected without accounting for the relationship between covariates and outcome. This approach helps maintain objectivity when making inference from nonrandomized studies.

However, for propensity score-based methods to work reliably, there should be sufficient overlap in the propensity score distributions for the compared treatment groups. Estimating the causal effects for units outside the overlap region depends entirely on extrapolation, and hence is vulnerable to model misspecification. The variables included in the propensity score model influence the degree of overlap. For example, including strong predictors of the treatment that are not predictive of the outcome in the propensity score model could potentially shrink the overlap region. Removing such predictors from the model could increase the overlap region. Recent work has also shown that including such covariates can inflate the variance of the causal estimate and may also induce bias (Brookhard et al, 2006). On the contrary, including covariates that are associated with the outcome only can improve efficiency, since it reduces random covariate imbalance in finite samples (Brookhard et al, 2006). Glymour et al (2008) argues for controlling only common causes of the treatment and outcome. VanderWeele and Shpitser (2011) propose controlling for covariates that are causes of the treatment and/or outcome. Thus, a propensity score model based only on the treatment can be inefficient, as it prioritizes variables associated with treatment but not necessarily with outcome. Balancing such covariates using propensity score is unnecessary since these covariates are not confounders. Recently researchers have started looking at how to select variables for the propensity model by taking into account the relationship between the covariates and outcome (Shortreed and Ertefaie 2017, de Luna, Waernbaum and Richardson 2011). In this paper, we extend the same idea to a recently proposed propensity score-based multiple imputation based approach, called penalized spline of propensity method for treatment comparison (PENCOMP), and propose a new variant of PENCOMP via bagging, and compare the performances of PENCOMP with that of inverse probability treatment weighted approach (IPTW) in the presence of variable selection.

A useful class of propensity score-based methods is based on estimating the propensity of treatment assignment, given potential confounding variables, and then using the estimated propensity as a weight, or as a predictor in regression models for the outcome under alternative treatment assignments. The IPTW method controls for confounding by weighting subjects by the inverse of the probability of receiving the observed treatment sequence. The weights in effect create a pseudo-population that is free of treatment confounders. PENCOMP controls for confounding by including a penalized spline of the logit of the propensity to be assigned that treatment in regression models. It has both the propensity and prediction models and is robust to misspecification in the propensity model or the prediction model.

Here we focus on the issue of model selection for our proposed method PENCOMP and IPTW. We compare the performance of two confounder selection methods: with and without considering the outcome-covariate relationship. Furthmore, often a propensity score model is selected based on how well it balances the observed covariates and inferences are made based on a single model. This simple approach ignores the model uncertainty regarding what variables should be included. Failure to account for uncertainty could affect estimation accuracy. Hence, we also address the issue of model uncertainty when making inference and propose a new version of PENCOMP based on bagging. For PENCOMP, we consider two methods for estimating standard errors and confidence intervals: (a) bootstrap method that takes into account model selection, proposed by Efron (2014), and (b) multiple imputation based on Rubin's combining rules.

The outline of this paper is as follows. In Section 4.2, we describe the estimands and assumptions of the approach we consider. In Section 4.3, we describe two versions of (PENCOMP) for estimating causal effects: one based on multiple imputation and the other based on bootstrap smoothing, as well as a review of inverse probability treatment weighted and adaptive lasso. In Section 4.4, we describe variable selection techniques for both the propensity and prediction models. In Section 4.5, we examine using simulation studies how variable inclusion affects the performance of propensity score-based methods-PENCOMP, AIPTW and IPTW. We also evaluate the impact of accounting for model uncertainty in propensity score and prediction models. In Section 4.6, we illustrate our methods using the Multicenter AIDS Cohort study (MACS) to estimate the effect of antiretroviral treatment on CD4 counts in HIV infected patients. We conclude with a discussion of the results and present some possible future work.

4.2 Estimands and Assumptions

Let X_i denote the vector of baseline covariates and $Z_i \in (0,1)$ denote a binary treatment with $Z_i = 1$ for treatment and $Z_i = 0$ for control, for subject $i = 1, \cdots, N$, respectively. Let $Y_i^{Z_i}$ be the potential outcome under treatment Z_i. Here we focus on the estimand of interest-the average treatment effects for the entire population (ATE), denoted as $E(Y^1 - Y^0)$. Thus, we compute the subject-level causal effect as the difference between the potential outcome under treatment and the potential outcome under control for the same subject. The average treatment effect for the entire population is estimated by averaging all the subject-level causal effects across the entire population. In this chapter, we focus on the estimand ATE, but the same idea can be applied to other estimands. See Chapter 3 for other estimands.

In order to estimate the causal effects, we make the following assumptions:

1) SUTVA (Angrist, Imbens and Rubin, 1996) states that a) the observed outcomes under a specific treatment sequence is equal to the potential outcomes associated with that treatment sequence, and b) the potential outcomes for a given subject are not influenced by the treatment assignments of other subjects (Rubin, 1980; Angrist, Imbens, Rubin, 1996)

2) Positivity states that each subject has a positive probability of being assigned to either treatment of interest: $0 < \Pr(Z_i = z_i|X_i) < 1$.

3) Ignorable treatment assumption states that $(Y^1, Y^0) \perp\!\!\!\perp Z | X$; that is, treatment assignment is as if randomized conditional on the covariates. In general, it is possible that there exists a subset of covariates $W \in X$ such as $(Y^1, Y^0) \perp\!\!\!\perp Z | W$.

4.3 Methods

4.3.1 PENCOMP and Multiple Imputation

PENCOMP is a robust multiple imputation based approach to causal inference, under Rubin's potential outcome framework (1974). Since each subject receives a single treatment, we observe the potential outcome under the observed treatment but not the potential outcomes under other treatments. We assume a single binary treatment setting, although the approach could be extended to multiple treatments. We estimate causal effects by imputing the potential outcomes that are not observed using regression models that include splines on the logit of the propensity to be assigned that treatment as well as other covariates that are predictive of the outcome. We then draw inferences based on comparisons of the imputed and observed outcomes between treatment groups.

PENCOMP relies on the balancing property of propensity score, in combination with mean model for the outcome. Under the assumptions stated above, PENCOMP has a double robustness property for causal effects. Specifically, if either 1) the model for the propensity score and the relationship between the outcome and the propensity score are correctly specified through penalized spline, or 2) the outcome model is correct, the causal effect of the treatment will be consistently estimated.

Here, we describe the estimation procedures based on multiple imputation with Rubin's combining rules.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample S^d from the original data S by sampling units with replacement, stratified based on treatment group. Then carry out steps (b)-(d) for each sample S^d:

(b) Select and estimate the propensity score model as described in Section 4.4 for the distribution of Z given X, with regression parameters γ_z. The propensity to be assigned treatment $Z = z$ is denoted as $\hat{P}_z(X) = \Pr(Z = z|X, \hat{\alpha}_z^{(d)})$, where $\hat{\alpha}_z^{(d)}$ is the ML estimate of α_z. Define $\hat{P^*}_z$ =log$[\hat{P}_z(X)/(1 - \hat{P}_z(X))]$.

(c) For each $z = 0, 1$, using the cases assigned to treatment group z, estimate a normal linear regression of Y^z on X, with mean

$$E(Y^z|X, Z = z, \theta_z, \beta_z) = s(\hat{P^*}_z|\theta_z) + g_z(X; \beta_z),$$

where $s(\hat{P^*}_z|\theta_z)$ denotes a penalized spline with fixed knots (Eilers and Marx, 1996; Ngo and Wand, 2004; Wand, 2003), with parameters θ_z, and $g_z()$ represents a parametric function of covariates predictive of the outcome, including covariates that are adequately balanced by the estimated propensity score models, indexed by parameters β_z. A different spline function is fitted for each treatment group, since there is no a priori reason to assume that the relationship between the potential outcomes under different treatment arms and the propensity of treatment assignment is the same. For simplicity, a penalized spline with truncated linear basis is used, $s(\hat{P^*}_z|\theta_z) = \theta_0 + \theta_1\hat{P^*}_z + \sum_{k=1}^{K} \theta_{1k}(\hat{P^*}_z - K_k)_+$, where $K_1, \cdots, K_K$ are fixed knots, and $(\hat{P^*}_z - K_k)_+ = (\hat{P^*}_z - K_k)$ if $\hat{P^*}_z > K_k$; and $= 0$ if $\hat{P^*}_z \leq K_k$. The spline model can be formulated as a linear mixed model (Wand, 2003),

$$Y^z = C_1\beta + C_2\theta + \epsilon, \begin{bmatrix} \theta \\ \epsilon \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\theta^2 I & 0 \\ 0 & \sigma_\epsilon^2 I \end{bmatrix} \right),$$

where $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ denote fixed effects, and $\theta = (\theta_{11}, \cdots, \theta_{1K})$ are random basis coefficients. REML estimates of the parameters of this model can be easily fitted in statistical software, such as PROC MIXED in SAS or lme in R. The fitted values of Y^z are $\hat{y}^z = C(C^TC + \hat{\lambda}D)^{-1}C^Ty$, where $\hat{\lambda} = \hat{\sigma}_\epsilon^2/\hat{\sigma}_\theta^2$ is the REML estimator of λ and

$$D = \begin{pmatrix} 0_{(p+1)\times(p+1)} & 0 \\ 0 & I_{K\times K} \end{pmatrix}$$

(d) For $z = 0, 1$, impute the values of Y^z for subjects in treatment group $1 - z$ in the original data set with draws from the predictive distribution of Y^z given X from the regression in (c), with ML estimates $\hat{\theta}_z^{(d)}, \hat{\beta}_z^{(d)}$ substituted for the parameters θ_z, β_z, respectively. Repeat the above procedures to produce D complete data sets.

Let $\hat{\Delta}^{(d)}$ and $W^{(d)}$ denote the difference in treatment means and associated pooled variance estimate, based on the observed and imputed values of Y in each treatment group. The MI estimate of Δ is then $\bar{\Delta}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\Delta}_d$, and the MI estimate of the variance of $\bar{\Delta}_D$

$$T_D = \bar{W}_D + (1 + 1/D)B_D \tag{4.1}$$

where $\bar{W}_D = \sum_{d=1}^{D} W^{(d)}/D, B_D = \sum_{d=1}^{D}\left(\hat{\Delta}^{(d)} - \bar{\Delta}_D\right)^2/(D-1)$. The estimate Δ is t distributed with degree of freedom v, $(\Delta - \bar{\Delta}_D)T_D^{\frac{-1}{2}} \sim t_v$, where $v = (D-1)(1 + \bar{W}_D/((D+1) * B_D))^2$.

4.3.2 PENCOMP and Bagging

As an alternative to using multiple imputation combining rules, we can draw inference about causal effects based on bootstrap smoothing, also called bagging. The bagging estimator, a form of model averaging, accounts for model uncertainty. Efron (2014) proposes standard error and confidence interval for the bootstrap smoothed estimator.

Let $S = (S_1, S_2, \cdots, S_N)$ denote the original data for N subjects. A nonparametric bootstrap sample with replacement is denoted as $S^d = (S_1^d, S_2^d, \cdots, S_N^d)$. The causal estimate based on the original data S is $\hat{\Delta}_s$. In most cases, the nonparametric standard error $\hat{sd}_D$ for $\hat{\Delta}_s$ based on D bootstrap samples is

$$\hat{sd}_D = \sum_{d=1}^{D} (\hat{\Delta}_d - \tilde{\Delta})^2/(D-1) \tag{4.2}$$

where $\tilde{\Delta} = \sum_{d=1}^{D} \hat{\Delta}_d/D$. The standard 95% confidence intervals $\hat{\Delta}_s \pm 1.96\hat{sd}_D$ or the percentile $(\hat{\Delta}_d^{0.025}, \hat{\Delta}_d^{0.975})$ based on the 2.5th and 97.5th percentiles of the D bootstrap estimates. However, in the presence of model selection, the bootstrap estimates can be "jumpy and erratic" and the standard methods assume smooth distribution. As an alternative, the bootstrap estimate $\tilde{\Delta}$ and associated confidence interval $\tilde{\Delta} \pm 1.96\tilde{sd}_D$ are used. The standard error $\tilde{sd}_D$ is calculated as follows.

$$\tilde{sd}_D = (\sum_{j=1}^{n} \hat{cov}_j^2)^{1/2} \tag{4.3}$$

$$\hat{cov}_j^2 = \sum_{d=1}^{D} (Q_{dj}^* - Q_{.j}^*)(\hat{\Delta}_d - \tilde{\Delta})/D$$

where $Q_{.j}^* = \sum_{d=1}^{D} Q_{dj}^*/D$ and $Q_{dj}^* = \#\{S^d = S_j\}$ is the number of times that data point j of the original data S is selected in dth bootstrap sample S^d.

The procedures for PENCOMP are similar as described above, except in steps (e). In step (e), the imputations are carried out on each bootstrap sample S^d, instead of the original data S. Inference is made using the bootstrap smoothed estimator $\tilde{\Delta}$ and confidence interval $\tilde{\Delta} \pm 1.96\tilde{sd}_D$, instead of the Rubin's multiple imputation combining rules.

4.3.3 Inverse Probability Treatment Weighted Estimator IPTW

The IPTW estimator estimates the ATE and is defined as

$$\hat{\Delta}_{IPTW} = \sum_{i=1}^{N} \frac{Z_i Y_i}{\hat{P}_{z_i=1}(X_i, \hat{\alpha})} - \sum_{i=1}^{N} \frac{(1-Z_i)Y_i}{1-\hat{P}_{z_i=1}(X_i, \hat{\alpha})}$$

The causal IPTW estimate on the original data S is $\hat{\Delta}_{IPTW}$. The standard errors are estimated based on bootstraps. The procedures are as follows.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample S^d from the original data S by sampling units with replacement. Then carry out steps (b)-(d) for each sample S^d:

(b) Select and estimate the propensity score model as described in Section 4.4.

(d) Estimate Δ^d_{IPTW} for each bootstrap sample.

The standard approach for computing the standard errors $\hat{sd}_D$ is based on Eq 4.2 and the 95% confidence intervals are computed as $\hat{\Delta}_{IPTW} \pm 1.96\hat{sd}_D$. The bootstrap smoothed estimate is $\tilde{\Delta}_{IPTW} = \frac{1}{D}\sum_{d=1}^{D} \hat{\Delta}^d_{IPTW}$, and the confidence intervals $\tilde{\Delta}_{IPTW} \pm 1.96\tilde{sd}_D$, where $\tilde{sd}_D$ is computed based on Eq 4.3.

4.3.4 Augmented Inverse Probability Treatment Weighted Estimator (AIPTW)

Each subject i is weighted by the balancing weight $W_i = 1/\Big\{ Z_i P_{z_i=1}(X_i, \hat{\alpha}) + (1 - Z_i)(1 - P_{z_i=1}(X_i, \hat{\alpha}))\Big\}$. The AIPTW estimate Δ_{AIPTW} is defined as follows (Mao, Li and Greene, 2018):

$$\hat{\Delta}_{AIPTW} = \frac{\sum_{i=1}^{n} \omega_i \{m_1(X_i, \beta_1) - m_0(X_i, \beta_0)\}}{\sum_{i=1}^{n} \omega_i} + \frac{\sum_{i=1}^{n} W_i Z_i \{Y_i - m_1(X_i, \beta_1)\}}{\sum_{i=1}^{n} W_i Z_i} - \frac{\sum_{i=1}^{n} W_i (1-Z_i)\{Y_i - m_0(X_i, \beta_0)\}}{\sum_{i=1}^{n} W_i(1-Z_i)}$$

where $m_1(X_i, \beta_1) = E(Y_i|X_i, Z_i = 1)$ and $m_0(X_i, \beta_1) = E(Y_i|X_i, Z_i = 0)$. Similar procedures based on bootstrap samples are used to estimate the standard error for

$\hat{\Delta}_{AIPTW}$.

4.3.5 Adaptive Lasso

Let X denote the design matrix $X = [X_1, \cdots, X_p]$ for p predictors. We assume that the outcome of interest Y is continuous with a mean that is a linear function of the predictors: $E(Y) = \beta_1 X_1 + \cdots + \beta_p X_p$. Here we assume the data are centered so that the intercept is not included. Suppose the model is sparse, that is, the true model depends only on a small subset of the predictors. Let $A = \{j : \beta_j \neq 0\}$ and $|A| = p_0 < p$. The adaptive lasso is defined as (Zou, 2006):

$$\hat{\beta}_{AL} = argmin_{\beta} ||y - \sum_{j=1}^{p} X_j \beta_j||^2 + \lambda \sum_{j=1}^{p} \hat{w}_j(\hat{\beta}_j)|\beta_j|$$

where $w_j = 1/|\hat{\beta}_j|^{\gamma}$ and $\gamma > 0$, and $\hat{\beta}$ are from ordinary least square or ridge regression. The adaptive lasso has the oracle properties: 1) it identifies the right subset covariates with probability tending to one: $\lim_n P(A_n = A) = 1$, where $A_n = \{j : \hat{\beta}_j \neq 0\}$; 2) it estimates the nonzero coefficients as if the true model were known, i.e. $\sqrt{n}(\hat{\beta}_A - \beta_A) \rightarrow_d N(0, \Sigma)$, where Σ is the covariance matrix under the true model.

4.4 Model Selection for Propensity and Prediction

In observational studies, both the propensity and prediction models need to be estimated from the data. Including all available covariates in both models can lead to highly unstable estimates of treatment assignment and/or outcomes if sample sizes are small, and may be highly inefficient if covariates are not predictive of both treatment and outcome-that is, they are potential confounders. We consider scenarios where there are some variables that are predictors of outcome, and some that are predictors of treatment, some that are predictors of both treatment and outcome, and some that are spurious, in the sense that they affect neither the propensity or the outcome.

We assume that both the propensity and prediction models depend only on a small subset of the variables. Using the notations as in Shortreed and Ertefaie (2017), let C denote the true confounders, P predictors of outcome, I predictors of treatment, and S spurious covariates. The objective is to select out the relevant variables. We consider two strategies of building the propensity models: 1) separating the outcome from the design (Rubin 2007), and 2) taking into account the information in the outcome.

For strategy 1, one simple approach is to use the stepwise variable selection algorithm with the Bayesian Information Criterion (BIC) to select the variables that are predictive of treatment, regardless of how well they predict outcome. Separately, we use the same stepwise algorithm to select the prediction model for PENCOMP. The algorithm is abbreviated as SW. Instead of the stepwise algorithm with BIC criterion, we also carry out an adaptive lasso algorithm to select both the propensity and prediction models separately. This adaptive lasso algorithm is referred to as AL. For outcome Y and treatment Z, the adaptive lasso estimates are defined as follows:

$$\hat{\beta}_{AL} = argmin_{\beta} ||y - \sum_{j=1}^{p} X_j \beta_j||^2 + \lambda \sum_{j=1}^{p} \hat{w}_{\beta_j} |\beta_j| \tag{4.4}$$

$$\hat{\alpha}_{AL} = argmin_{\alpha} \sum_{i=1}^{n} -Z_i(X_i^T \alpha) + log(1 + e^{X_i^T \alpha}) + \lambda_n \sum_{j=1}^{p} \hat{w}_{\alpha_j} |\alpha_j| \tag{4.5}$$

where $w_{\alpha_j} = 1/|\hat{\alpha_j}|$, $w_{\beta_j} = 1/|\hat{\beta_j}|$, and $\hat{\alpha}$ and $\hat{\beta}$ are estimated from ridge regression. Both SW and AL satisfy Rubin's criterion by separating the outcome from the design.

In strategy 2, we consider taking into account the relationship between covariates and outcome when building the propensity model. Shortreed and Ertefaie (2017) propose an outcome adaptive lasso approach for variable selection. Their approach takes into account the covariate-outcome relationships when selecting propensity model. It tends to select covariates that are true confounders and predictors of the outcome and improves statistical efficiency. The outcome adaptive lasso estimates for the propensity model are defined as:

$$\hat{\alpha}_{OAL} = argmin_{\alpha} \sum_{i=1}^{n} -Z_i(X_i^T\alpha) + log(1 + e^{X_i^T\alpha}) + \lambda_n \sum_{j=1}^{p} \hat{w}_{\alpha_j}|\alpha_j| \qquad (4.6)$$

where $w_{\alpha_j} = 1/|\hat{\beta}_j|^{\gamma}$ such that $\gamma > 1$ and minimizes the mean weighted standardized difference between the treated and control. $\hat{\beta}$ are the coefficient estimates by regressing the outcome Y on the covariates and the treatment indicator. By penalizing the covariates depending on the strength of the covariate and outcome relationship, the outcome adaptive lasso selects covariates that are predictive of the outcome and does not select covariates that are associated with the treatment but not with the outcome. In our setting, the outcome adaptive lasso is designed to select the covariates denoted by P and C, i.e. $A = \{j : j \in P \cup C\}$.

De Luna, Waernbaum and Richardson (2011) show how to identify subsets of the covariates such that given the subset, the unconfoundedness assumption still holds. They propose two algorithms to identify the reduced subsets. First, remove the covariates that are not associated with outcome, given the others, and then remove the covariates that are not associated with the treatment, given a smaller subset of the covariates selected at the first step. Alternatively, reverse the order by first removing the covariates that are not associated with the treatment and then removing the covariate that are not associated with the outcome. Dimension reduction in this manner can further reduce the variance of the casual estimate and improve the overlap in the propensity score distributions between treatment groups.

Building on the two-stage approach as in de Luna, Waernbaum and Richardson (2011), we use a two-stage adaptive lasso approach. In the first stage, we select a subset of covariates that are predictive of the outcome using adaptive lasso. In the second stage, we use the subset of covariates found in the first stage in the propensity model, denoted as Step-ALY. Similarly, we can reverse the steps by performing outcome adaptive lasso for the propensity model first and then the prediction model, denoted as Step-ALT. Unlike SW and AL algorithms, OAL, Step-ALT and Step-ALY all take into account the outcome information during model selection. By using a two-stage approach, in finite samples, we can further reduce the probability of selecting any irrelevant covariates. The models from the two-stage appraoch could be more sparse that the models selected by the outcome adaptive lasso approach proposed in Shortreed and Ertefaie (2017).

4.5 Simulation

We simulate each dataset as described in Zigler and Dominici (2014) and Shortreed and Ertefaie (2017). Each simulated dataset contains n subjects and p covariates X. The treatment Z_1 is Bernoulli distributed with logit of $P(Z_1 = 1|X) = \sum_{j=1}^{p} \gamma_j X_j$. The outcome of interest Y is normally distributed with a mean of $\eta Z_1 + \sum_{j=1}^{p} \beta_j X_j$ and a variance of 1. The treatment effect η is equal to 0, without loss of generality. We set all the coefficients 0, except the first 6 covariates $X_1, \cdots, X_6$. X_1 and X_2 are the true confounders; X_3 and X_4 are predictors of the outcome but not of the treatment; and X_5 and X_6 are predictors of the treatment but not of the outcome; all the other $d - 6$ covariates are spurious. We vary the strengh of relationships between covariates, outcome and treatment. In the first scenario, β and γ are set as: $\beta = (0.6, 0.6, 0.6, 0.6, 0, 0, 0, \cdots, 0)$, and $\gamma = (1, 1, 0, 0, 1, 1, 0, \cdots, 0)$. In the second scenario, confounders X_1 and X_2 have a weaker relationship with the treatment: $\beta = (0.6, 0.6, 0.6, 0.6, 0, 0, 0, \cdots, 0)$ and $\gamma = (0.4, 0.4, 0, 0, 1, 1, 0, \cdots, 0)$. In the third scenario, confounders X_1 and X_2 have a weaker relationship with the outcome: $\beta = (0.2, 0.2, 0.6, 0.6, 0, 0, 0, \cdots, 0)$ and $\gamma = (1, 1, 0, 0, 1, 1, 0, \cdots, 0)$. We also simulate the sample sizes, n=200 and n=1000.

As in the real world setting, we consider scenarios where we treat all variables as potential confounders. We compare these variable selection techniques:
(a) SW: stepwise variable selection algorithm with the Akaike Information Criterion (BIC) separately for the propensity and prediction models.
(b) AL: adaptive lasso selection technique separately for the propensity and prediction models.
(c) OAL: outcome adaptive lasso proposed by Shortreed and Ertefaie (2017) for the propensity model, and adaptive lasso for the prediction model.
(d) Step-ALT: outcome adaptive lasso for the propensity model at the first stage and then adaptive lasso for the prediction model at the second stage using only the variables that are selected at the first stage.
(e) Step-ALY: adaptive lasso for the prediction model at the first stage and then logistic regression model with all the variables selected at the first stage for the propensity model.
(f) allLasso: all the variables that are selected for the propensity and prediction models, as described in VanderWeele and Shpitser (2011).

In addition to the variable selection techniques, we present results for four propensity (PS) models that include the same covariates across simulations: (1)True includes the true propensity models that are used to generate the data, i.e. X_1, X_2, X_5, and X_6; (2) trueConf includes only the true confounders X_1 and X_2; (3) outcomePred includes both the confounders and the predictors of outcome; (4) allPoten includes all 20 variables. For these four PS models, the prediction models for PENCOMP are also correctly specified.

For each simulation scenario and for each of the two methods PENCOMP and IPTW, we compare the performance of the variables selection techniques described above for both PENCOMP and IPTW, in terms of empirical bias (BIAS), the empirical standard error (Emp.SE), mean of estimated standard error (Est.SE), average length of 95% confidence intervals (Ave. CI), and empirical coverage rate of the 95% confidence interval (Cov) over 500 simulated data sets. For each dataset, the estimated standard errors and confidence intervals are calculated based on 1000 bootstrap samples. For PENCOMP, we compare the two methods of standard error estimations: multiple imputation (MI) as in Eq 1 and bootstrap smoothing (Boot) based as in Eq 3. For IPTW, we compare the standard approach based in Eq 2 with the bootstrap smoothing.

4.5.1 Results

Tables 4.1-4.4 show the results for sample size of 200 and Tables 4.5-4.8 for sample size of 1000. By comparing the four (PS) models that do not involve variable selections: true, trueConf, outcomePred, and allPotent, we can see that excluding variables associated only with treatment reduced the RMSE, and including variables associated only with outcome further reduced the RMSE. For IPTW estimates, outcomePred had the smallest RMSE and mean confidence interval widths. The estimated standard errors (SE) were closer to the empirical standard errors (SE) and the coverage was close to the nominal coverage of 95%. The trueConf PS model performed slightly worse than the outcomePred PS model, since including variables associated only with outcome improves efficiency. The more spurious variables were added as in allPotent model, the wider the confidence intervals got. Figure 4.3 shows that outcomePred PS model had the smallest variability across the 1000 bootstrap estimates, while allPotent had the biggest variability. This pattern was observed in PENCOMP and AIPTW but less pronounced than in IPTW, since the prediction model in PENCOMP and AIPTW attenuated the effect of including variables not associated with the outcome. In addition, PENCOMP tended to perform better than AIPTW in terms of RMSEs, when the propensity score models included many irrelevant covariates. As shown in Figure 4.2, the variables associated with the outcome were selected about 99% of the time, but in small samples, the confounders that were weakly associated with the outcome were selected less than 80% of the time, as seen in scenario 3.

Out of the five variable selection techniques, the two-stage techniques: Step-ALT and Step-ALY performed the best. For PENCOMP, AIPTW and IPTW, both Step-ALT and Step-ALY had RMSEs that were closer to the RMSEs of outcomePred PS models. Both Step-ALT and Step-ALY were more effective at excluding spurious variables and including variables associated only with outcome, compared to the outcome adaptive lasso (OAL), stepwise selection with BIC (SW), and adaptive lasso (AL) procedures, as seen in Figure 4.1. For example, for the sample size of 200 in scenario 1, all the variable selection techniques selected the confounders X_1 and X_2 about 99% of the time. Step-ALY, Step-ALT, and OAL selected the non-confounders X_3 and X_4 about 99% of the time. In contrast, AL and SW selected X_3 and X_4 about 40-60% of the time. AL and SW selected the non-confounders X_5 and X_6 about 99% of the time. In contrast, OAL selected X_5 and X_6 about 30% of the time, but Step-ALY and Step-ALT selected them around 8% of the time. Lastly, Step-ALY and Step-ALT selected the spurious variables at about 8% of the time, while SW, OAL, and AL selected them about 40%, 34% and 60%, respectively. In scenario 2, because the confounders X_1 and X_2 had a weaker relationship with treatment, they were selected about 80% of the time for sample size of 200. A larger sample size is needed to detect those confounders, as seen in Figure 4.1. In scenario 3 where the confounders had a weak relationship with the outcome, the outcome adaptive selection procedures performed worse than SW and AL. Step-ALT and Step-ALY selected the weak confounders X_1 and X_2 about 50% of the time, while OAL selected them around 80% of the time. Excluding weak confounders increases the bias as seen in Table 4.3 for Step-ALT and Step-ALY, although the reduction in variance by excluding many spurious variables was big enough that the RMSEs were still better.

In summary, excluding predictors of the treatment only and including predictors of the outcome, even the ones not associated with treatment, can improve the efficiency of the estimators without substantially increasing the bias.

As shown in Table 4.1-4.4, the bootstrap smoothing with Efron's formula tended to perform better than MI(PENCOMP) and the standard method (AIPTW, IPTW) for sample size of 200: the estimated SE were closer the empirical SE, the coverage rates were closer the nominal 95% coverage, and confidence interval widths were smaller. The gain of using Efron's formula was more pronouced for SE, OAL, and AL. This was probably due to the fact that the bootstrap estimates were more variable-many different models and causal estimates were obtained across the bootstraps. The distributions of the bootstrap estimates were thus more "jumpy and erratic". As shown in Figures 4.3-4.5, the 1000 bootstrap estimates for one simulated dataset were more variable for sample size of 200 than for sample size of 1000, especially for SW, AL and OAL selection procedures, which tended to select many more spurious variables. In the presence of high variability across the bootstrap estimates, Efron's formula provided tighter confidence intervals.

As the sample size increased to 1000, the gain of using Efron's formula disappeared, as seen in Tables 4.4-4.8. The standard procedure of calculating the confidence intervals in the case of IPTW and AIPTW, and using multiple imputation-based PENCOMP performed better than using Efron's formula. When there is no much variability in the estimates, using Efron's formula can lead to greater confidence interval widths and overcoverage. Figure 4.3-4.5 shows that for sample size of 1000, all the models had similar variability in the bootstrap estimates and the level of variability was much less, compared to that for sample size of 200. In summary, using Efron's formula is advantagous when the sample size is smaller and the data are more noisier and the model selection is more variable across bootstrap samples.

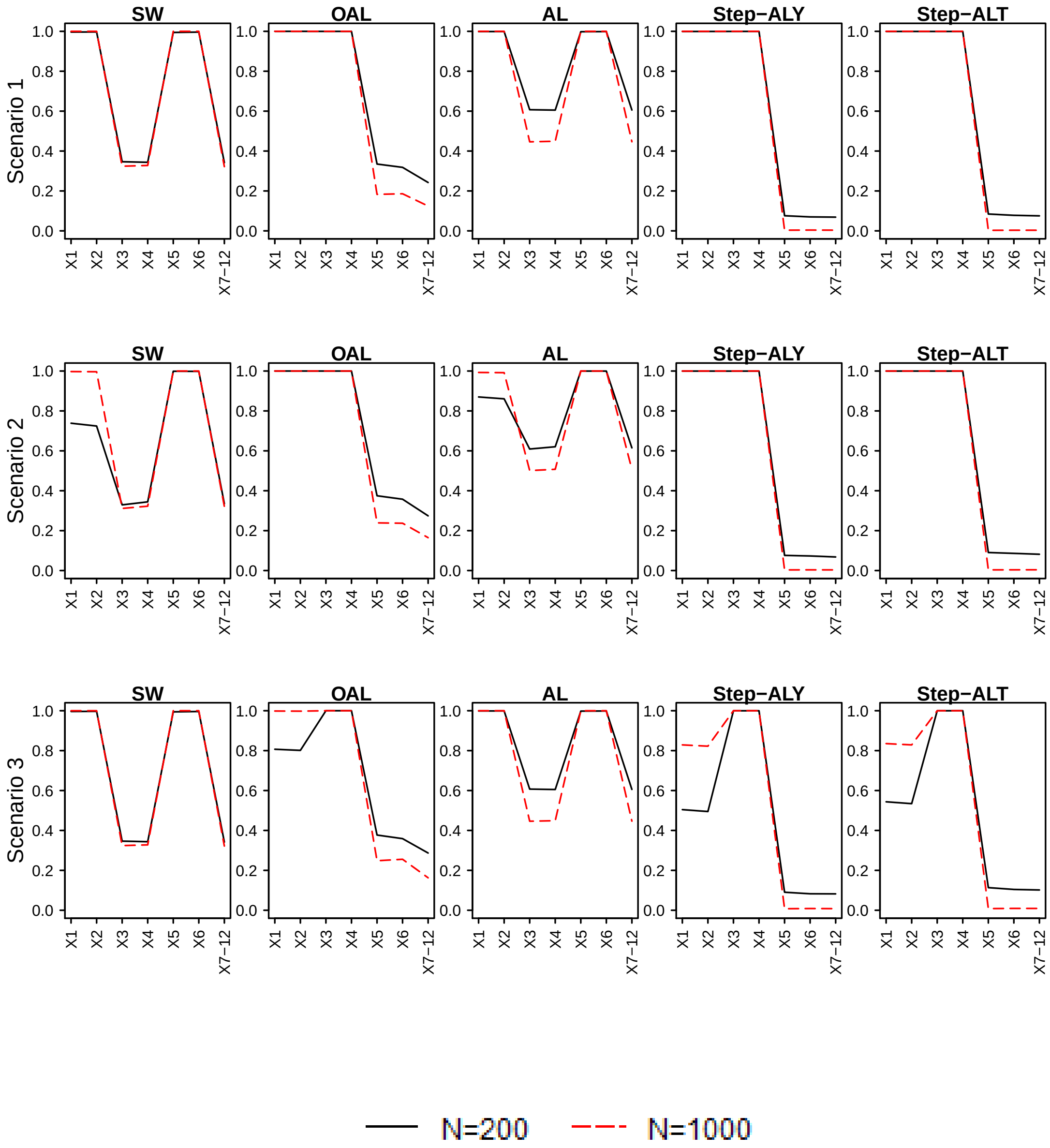

Figure 4.1: Proportions of each variable selected for propensity model across 500 simulated datasets and 1000 bootstrap samples for each simulated dataset for sample size of 200 and 1000. X_1 and X_2 are the true confounders; X_3 and X_4 are predictors of the outcome but not of the treatment; and X_5 and X_6 are predictors of the treatment but not of the outcome; all the other 14 covariates are spurious. Average across the spurious variables.

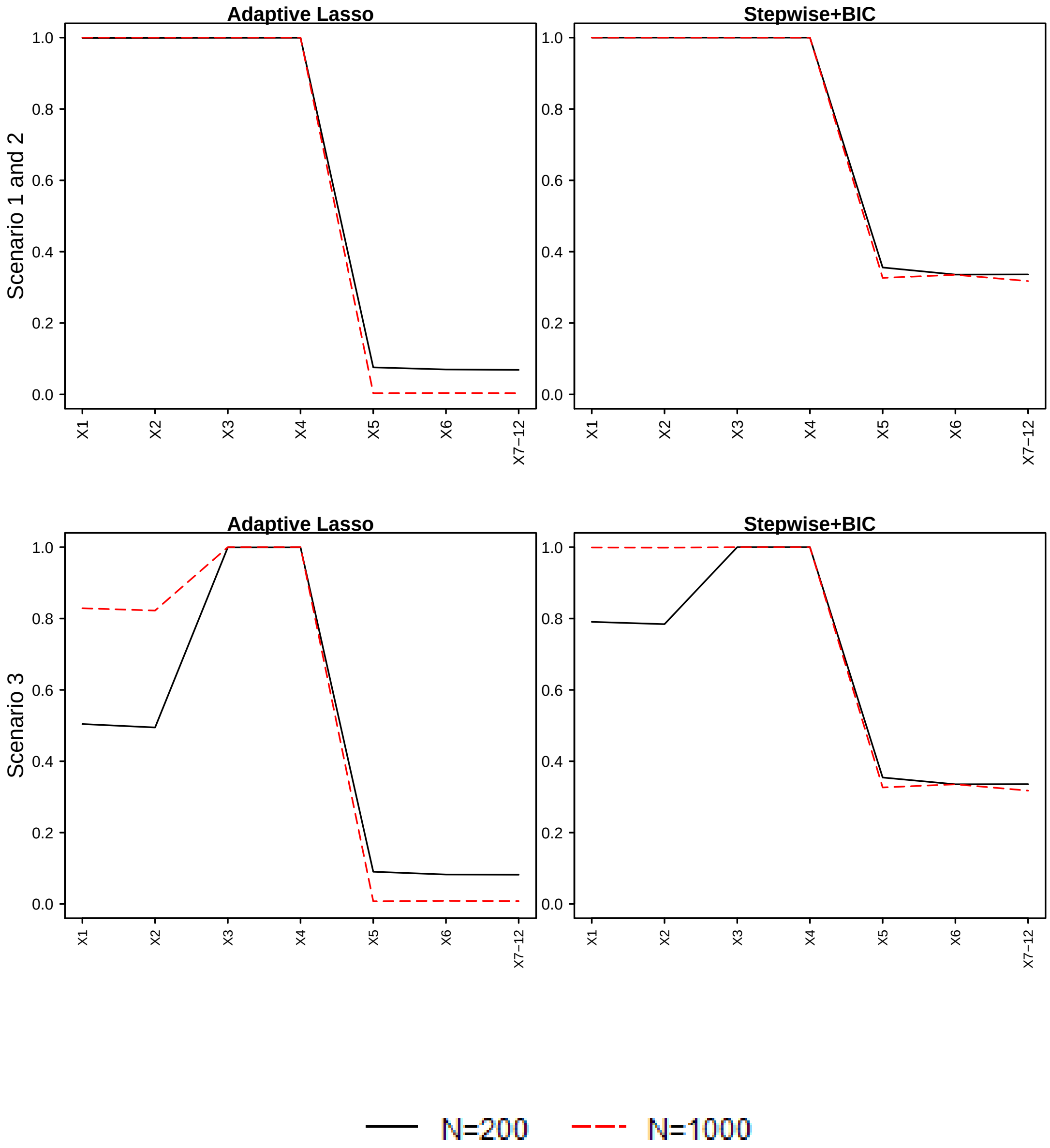

Figure 4.2: Proportions of each variable selected for prediction model across 500 simulated datasets and 1000 bootstrap samples for each simulated dataset for sample size of 200 and 1000. X_1 and X_2 are the true confounders; X_3 and X_4 are predictors of the outcome but not of the treatment; and X_5 and X_6 are predictors of the treatment but not of the outcome; all the other 14 covariates are spurious. Average across the spurious variables.

Table 4.1: 100× RMSE with sample size of 200. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

.

100× Empirical RMSE										
		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	21	19	21	28	24	28	41	31	34
Bagging	allPotent	22	20	22	25	23	25	34	29	30
Standard/Rubin	true	21	19	21	22	20	22	36	29	31
Bagging	true	21	19	21	21	19	21	33	28	29
Standard/Rubin	outcomePred	16	14	16	16	14	16	19	15	17
Bagging	outcomePred	16	14	16	16	14	16	19	15	17
Standard/Rubin	trueConf	16	14	16	16	14	16	22	19	21
Bagging	trueConf	16	14	16	16	14	16	22	19	21
Standard/Rubin	SW	21	19	21	25	23	25	38	32	33
Bagging	SW	22	20	22	23	22	23	33	28	28
Standard/Rubin	AL	21	19	21	26	24	27	39	30	32
Bagging	AL	22	19	22	24	22	24	33	28	29
Standard/Rubin	allLasso	18	17	18	18	17	19	22	18	20
Bagging	allLasso	18	17	18	18	17	19	21	18	19
Standard/Rubin	OAL	18	17	18	18	17	19	22	18	20
Bagging	OAL	18	17	18	18	17	19	21	18	19
Standard/Rubin	Step-ALT	16	15	18	17	15	23	19	15	24
Bagging	Step-ALT	16	15	18	17	15	18	19	15	19
Standard/Rubin	Step-ALY	16	15	18	16	14	24	19	15	25
Bagging	Step-ALY	16	15	18	17	15	18	19	15	19

Table 4.2: 100× noncoverage rate with sample size of 200. The nominal coverage is 95%. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

.

		100× Noncoverage Rate								
		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	1	2	1	2	2	2	1	1	1
Bagging	allPotent	3	4	3	4	5	4	6	3	4
Standard/Rubin	true	2	3	2	4	5	4	7	6	6
Bagging	true	3	4	3	3	4	3	7	6	5
Standard/Rubin	outcomePred	3	4	3	4	4	4	6	4	6
Bagging	outcomePred	2	3	2	4	3	4	5	3	5
Standard/Rubin	trueConf	3	4	3	4	5	4	5	4	3
Bagging	trueConf	2	3	2	3	4	3	4	2	3
Standard/Rubin	SW	1	2	1	1	2	1	1	1	1
Bagging	SW	3	4	3	4	5	4	6	4	5
Standard/Rubin	AL	1	2	1	1	2	1	2	1	1
Bagging	AL	4	4	4	4	5	4	6	4	4
Standard/Rubin	allLasso	1	3	2	2	3	3	2	2	2
Bagging	allLasso	3	4	3	4	4	4	4	3	4
Standard/Rubin	OAL	2	3	2	2	3	2	2	2	2
Bagging	OAL	3	4	3	4	4	4	4	3	4
Standard/Rubin	Step-ALT	2	3	3	4	4	10	4	2	11
Bagging	Step-ALT	2	2	5	4	3	6	4	3	7
Standard/Rubin	Step-ALY	2	3	3	3	3	10	4	2	11
Bagging	Step-ALY	3	2	5	4	3	6	4	3	7

Table 4.3: 1000× empirical bias with sample size of 200. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

1000× Empirical Bias										
		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	5	-2	5	2	4	2	60	11	26
Bagging	allPotent	2	-6	2	2	2	2	63	5	26
Standard/Rubin	true	6	4	6	6	3	6	61	11	27
Bagging	true	3	1	3	5	3	5	82	18	33
Standard/Rubin	outcomePred	10	7	10	8	7	8	33	9	19
Bagging	outcomePred	7	4	7	8	7	8	39	9	21
Standard/Rubin	trueConf	8	7	8	8	7	8	32	6	18
Bagging	trueConf	5	4	5	8	7	8	39	6	20
Standard/Rubin	SW	5	-2	6	4	-4	7	66	39	33
Bagging	SW	2	-5	3	1	0	5	68	37	27
Standard/Rubin	AL	6	-2	11	5	-0	19	71	15	29
Bagging	AL	6	-3	7	2	1	11	72	16	29
Standard/Rubin	allLasso	2	-3	21	4	-1	17	35	2	26
Bagging	allLasso	3	-3	20	2	-1	23	46	4	32
Standard/Rubin	OAL	6	0	25	4	-1	17	35	2	26
Bagging	OAL	5	-1	22	2	-1	23	47	5	33
Standard/Rubin	Step-ALT	2	-4	65	7	6	132	33	8	146
Bagging	Step-ALT	3	-3	65	3	-3	66	40	2	83
Standard/Rubin	Step-ALY	2	-4	70	8	7	138	33	9	160
Bagging	Step-ALY	2	-4	70	2	-3	70	36	1	90

Table 4.4: 10× mean 95% confidence interval width with sample size of 200. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

.

10× Mean 95% Confidence Width										
		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	13	10	13	13	11	13	20	16	17
Bagging	allPotent	10	8	10	10	9	10	14	12	12
Standard/Rubin	true	10	8	10	8	7	8	12	10	11
Bagging	true	9	8	9	8	8	8	12	11	11
Standard/Rubin	outcomePred	7	6	7	7	6	7	8	6	7
Bagging	outcomePred	7	6	7	7	6	7	8	6	7
Standard/Rubin	trueConf	7	6	7	7	6	7	9	8	8
Bagging	trueConf	7	6	7	7	6	7	9	8	9
Standard/Rubin	SW	13	10	13	12	10	12	18	15	15
Bagging	SW	9	8	9	9	8	9	13	11	11
Standard/Rubin	AL	13	10	13	13	11	13	19	15	16
Bagging	AL	9	8	9	9	8	9	13	11	11
Standard/Rubin	allLasso	9	8	9	8	8	9	11	9	10
Bagging	allLasso	8	7	8	8	7	8	9	8	8
Standard/Rubin	OAL	9	8	9	8	7	9	11	9	10
Bagging	OAL	8	7	8	8	7	8	9	8	8
Standard/Rubin	Step-ALT	8	7	8	7	6	8	9	7	8
Bagging	Step-ALT	7	7	8	7	6	7	8	7	7
Standard/Rubin	Step-ALY	8	7	8	7	6	8	9	7	8
Bagging	Step-ALY	7	7	8	7	6	7	8	7	7

Table 4.5: 100× RMSE with sample size of 1000. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

100× Empirical RMSE										
		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	9	8	9	13	9	13	18	11	15
Bagging	allPotent	9	8	9	12	9	12	17	11	14
Standard/Rubin	true	9	8	9	12	9	12	19	12	16
Bagging	true	9	8	9	11	9	11	18	12	15
Standard/Rubin	outcomePred	7	6	7	7	6	7	9	6	8
Bagging	outcomePred	7	6	7	7	6	7	9	6	8
Standard/Rubin	trueConf	7	6	7	7	6	7	10	8	10
Bagging	trueConf	7	6	7	7	6	7	10	8	10
Standard/Rubin	SW	9	8	9	13	9	13	19	12	16
Bagging	SW	9	8	9	12	9	12	17	11	14
Standard/Rubin	AL	9	8	9	12	9	12	18	12	15
Bagging	AL	9	8	9	11	9	11	16	11	13
Standard/Rubin	allLasso	8	7	8	8	7	8	9	7	9
Bagging	allLasso	8	7	8	8	7	8	9	7	8
Standard/Rubin	OAL	8	7	8	8	7	8	9	7	9
Bagging	OAL	8	7	8	8	7	8	9	7	8
Standard/Rubin	Step-ALT	7	6	8	7	6	9	9	6	11
Bagging	Step-ALT	7	6	8	7	6	8	9	6	9
Standard/Rubin	Step-ALY	7	6	8	7	6	9	9	6	11
Bagging	Step-ALY	7	6	8	7	6	8	9	6	9

Table 4.6: 100× noncoverage rate with sample size of 1000. The nominal coverage is 95%. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

100× Noncoverage Rate

		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	3	4	3	5	4	5	7	5	5
Bagging	allPotent	1	1	1	1	1	1	3	1	0
Standard/Rubin	true	4	3	4	6	4	6	9	5	5
Bagging	true	0	1	0	1	1	1	3	0	1
Standard/Rubin	outcomePred	5	5	5	5	4	5	6	5	6
Bagging	outcomePred	0	0	0	1	0	1	1	1	1
Standard/Rubin	trueConf	5	5	5	5	4	5	5	6	5
Bagging	trueConf	0	0	0	1	0	1	1	1	1
Standard/Rubin	SW	3	4	3	5	4	5	8	4	5
Bagging	SW	1	1	1	1	1	1	3	1	1
Standard/Rubin	AL	3	4	3	6	4	6	9	5	5
Bagging	AL	1	1	1	1	1	1	3	1	1
Standard/Rubin	allLasso	2	3	3	3	2	3	3	2	3
Bagging	allLasso	0	0	0	1	0	1	1	0	1
Standard/Rubin	OAL	2	3	3	3	2	3	3	2	3
Bagging	OAL	0	0	0	1	0	1	1	0	1
Standard/Rubin	Step-ALT	4	4	2	5	4	6	6	4	8
Bagging	Step-ALT	0	0	0	1	0	0	1	1	1
Standard/Rubin	Step-ALY	4	4	2	5	4	6	6	4	8
Bagging	Step-ALY	0	0	0	1	0	0	1	1	1

Table 4.7: 1000× empirical bias with sample size of 1000. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

.

1000× Empirical Bias										
		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	4	1	4	4	1	4	20	4	11
Bagging	allPotent	5	1	5	4	1	4	24	4	13
Standard/Rubin	true	4	2	4	5	2	5	25	5	14
Bagging	true	5	2	5	5	2	5	33	7	17
Standard/Rubin	outcomePred	0	-0	0	2	-0	2	14	1	7
Bagging	outcomePred	1	0	1	2	-0	2	16	1	7
Standard/Rubin	trueConf	0	-0	0	2	-0	2	16	0	8
Bagging	trueConf	1	0	1	2	0	2	17	1	9
Standard/Rubin	SW	4	1	4	3	1	3	17	4	9
Bagging	SW	5	1	5	4	1	4	25	5	13
Standard/Rubin	AL	4	1	5	5	2	7	27	7	15
Bagging	AL	5	2	5	4	1	5	33	8	16
Standard/Rubin	allLasso	2	1	3	4	1	5	17	2	9
Bagging	allLasso	2	0	3	3	1	5	21	3	12
Standard/Rubin	OAL	2	0	3	4	1	5	17	2	9
Bagging	OAL	3	1	4	3	1	5	21	3	12
Standard/Rubin	Step-ALT	1	-0	20	2	-0	22	14	1	39
Bagging	Step-ALT	0	-1	20	2	-0	21	16	1	36
Standard/Rubin	Step-ALY	1	-0	21	2	-0	23	14	1	40
Bagging	Step-ALY	0	-1	21	2	-0	22	16	1	36

Table 4.8: 10× mean 95% confidence interval width with sample size of 1000. The treatment effects η=2. S1, S2, and S3 denote scenario 1, 2, and 3, respectively.

.

10× Mean 95% Confidence Interval Width

		PENCOMP			AIPTW			IPTW		
	Model Select	S1	S2	S3	S1	S2	S3	S1	S2	S3
Standard/Rubin	allPotent	4	3	4	4	4	4	6	5	5
Bagging	allPotent	5	5	5	6	5	6	9	6	7
Standard/Rubin	true	4	3	4	4	3	4	6	5	6
Bagging	true	5	5	5	6	5	6	9	7	8
Standard/Rubin	outcomePred	3	3	3	3	3	3	3	3	3
Bagging	outcomePred	4	4	4	4	4	4	5	4	4
Standard/Rubin	trueConf	3	3	3	3	3	3	4	3	4
Bagging	trueConf	4	4	4	4	4	4	6	5	5
Standard/Rubin	SW	4	3	4	4	4	4	6	5	6
Bagging	SW	5	5	5	6	5	6	9	6	7
Standard/Rubin	AL	4	3	4	4	4	4	6	5	5
Bagging	AL	5	5	5	6	5	6	8	6	7
Standard/Rubin	allLasso	3	3	3	3	3	4	4	3	4
Bagging	allLasso	5	4	5	5	4	5	5	4	5
Standard/Rubin	OAL	3	3	3	3	3	4	4	3	4
Bagging	OAL	5	4	5	5	4	5	5	4	5
Standard/Rubin	Step-ALT	3	3	3	3	3	3	3	3	4
Bagging	Step-ALT	4	4	5	4	4	5	5	4	5
Standard/Rubin	Step-ALY	3	3	3	3	3	4	3	3	4
Bagging	Step-ALY	4	4	5	4	4	5	5	4	5

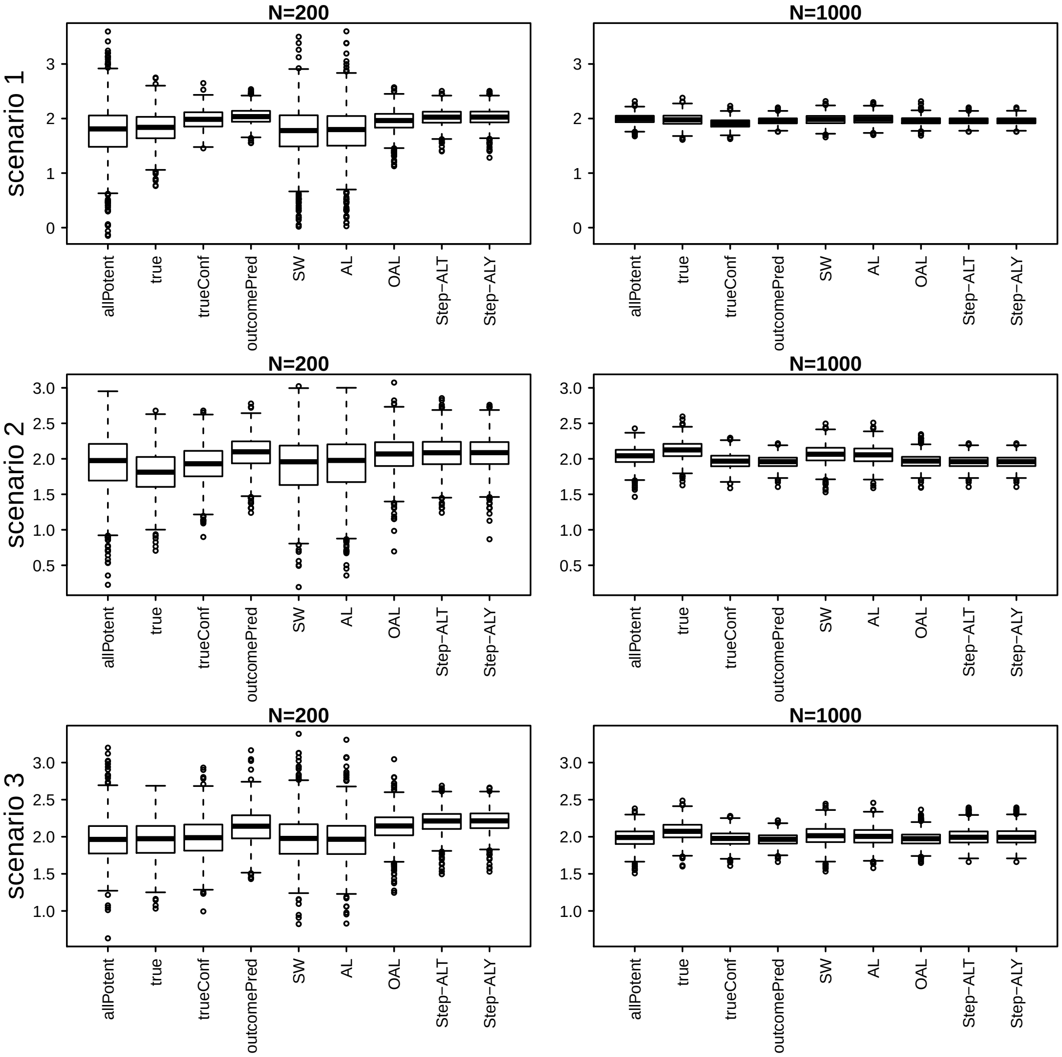

Figure 4.3: Distributions of 1000 bootstrap IPTW estimates for one simulated dataset

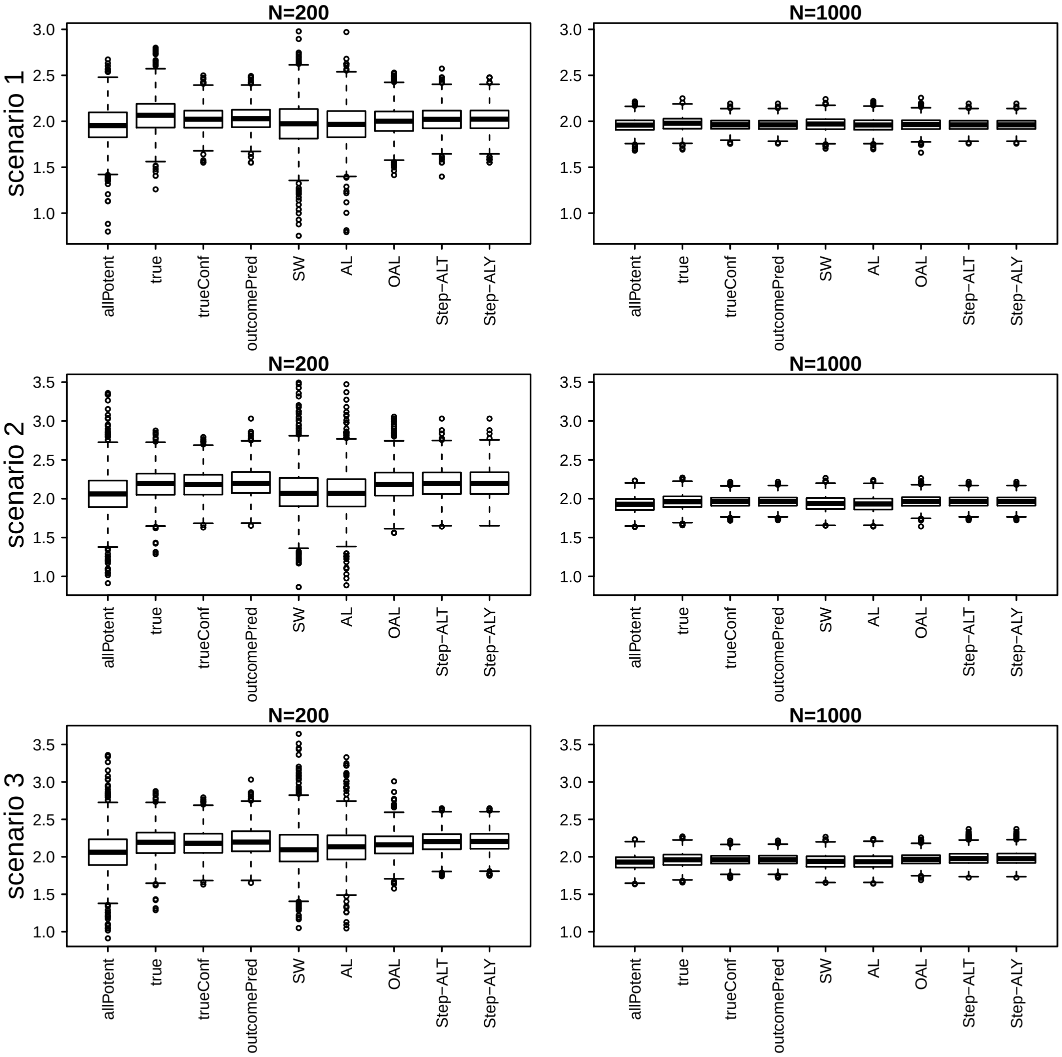

Figure 4.4: Distributions of 1000 bootstrap AIPTW estimates for one simulated dataset

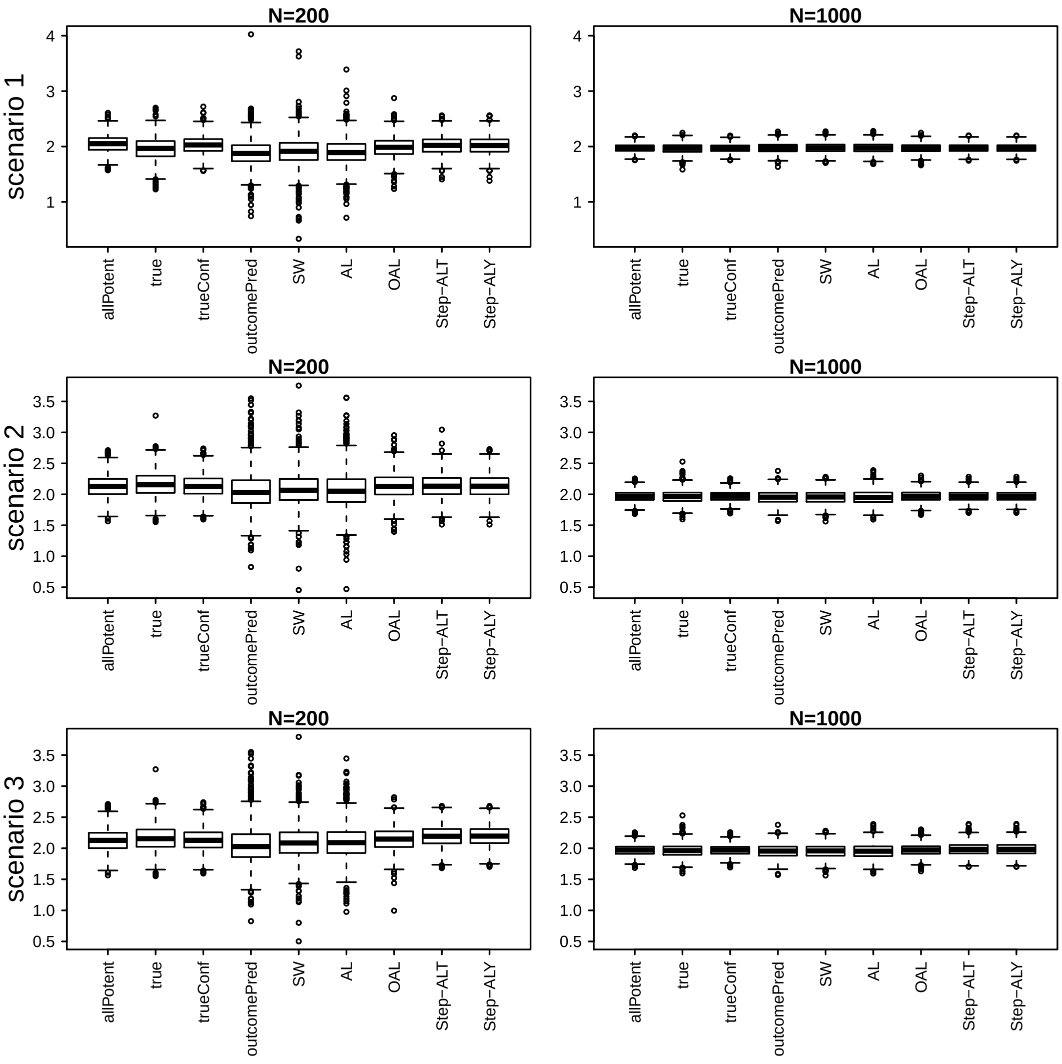

Figure 4.5: Distributions of 1000 bootstrap PENCOMP estimates for one simulated dataset

4.6 Application

The Multicenter AIDS Cohort study (MACS) was started in 1984 (Kaslow et al, 1987). A total of 4,954 gay and bisexual men were enrolled in the study and followed up semi-annually. At each visit, data from physical examination, questionnaires about medical and behavioral history, and blood test results were collected. The primary outcome of interest was the CD4 count, a continuous measure of how well the immune system functions. We used this dataset to analyze the short term effects of using antiretroviral treatment. Here we restrict our analyses to visit 12. Treatment was coded to 1 if the patient reported taking any of antiretroviral treatment (ART) or enrolling in clinical trials of such drugs. We estimate the short-term (6-month) effects of using any antiretroviral treatment for HIV+ subjects. We excluded subjects with missing values on any of the covariates included in the models. We log-transformed the blood counts in this analysis.

Here we treat each visit as a single time point treatment. Let $t = 1$ denote the time when the treatment was administered, and $t = 2$ the time 6-month later when the outcome was measured. In addition, let $t = -1, -2, -3$ denote 1, 2, and 3 visits away from the current visit $t = 1$. Let $X(t = 1, -1, -2, -3)$ denote the blood count histories prior to treatment assignment. Let Z be the binary treatment indicator. Let $Y(t = 2)$ be the CD4 count 6 months after the treatment. For the outcome model, we considered blood counts-CD4, CD8, white blood cell (WBC), red blood cell (RBC), and platelets and treatment histories from the last 4 visits. For the propensity model, we considered the same covariates as those in the outcome model, as well as demographic variables-college education, age, and race. The treatment assignment Z was modeled as a logistic regression. We estimated the mean CD4 count difference between the treated and the control at each visit, denoted as Δ. For PENCOMP, we replaced the simulated/imputed transformed CD4 values that were < 0 with 0 (i.e. below detection level). A total of 15 equally spaced knots and B spline were used.

As shown in Figure 4.6, we see that the treated and control subjects were very disimilar. The propensity score distributions were very skewed, as the treated had propensity of treatment close 1 and the control close to 0. Here we considered the variable selection methods in the simulation studies to select the relevant variables for the propensity score model. To quantify the amount of overlap, we measured the proportion of subjects in the control group whose propensity scores were between the 95^{th} and 5^{th} quantiles of the propensity score distribution of the treated group, denoted as $\pi_{z=0}^{0.95} = F_{z=0}(F_{z=1}^{-1}(0.95)) - F_{z=0}(F_{z=1}^{-1}(0.05))$, where F is the cumulative distribution. Similarly, $\pi_{z=1}^{0.95}$ denotes the proportion of the treated subjects whose propensity scores were between the 95^{th} and 5^{th} quantiles of the propensity score distribution of the control group. Including only the covariates that were selected more than 20% of times by Step_ALT among 1000 bootstrap samples improved the overlap, as shown in Figure 4.6. Table 4.9 shows the proportion that each variable was selected across 1000 bootstrap samples. Subjects who got treatment at the recent visits were more likely to receive treatments again. Thus, recent treatment histories were highly predictive of the subsequent treatment, but weakly associated with the outcome. Recent CD4 counts were much more predictive of the future CD4 counts. Thus, when we accounted for the outcome-covariate relationship during propensity model building, as in Step-ALT and Step-ALY, recent past treatment variables were selected less than 10% of the times, compared to close to 100% of the time in SW and AL, and 58% of the time in OAL. As seen in simulation studies, compared to the OAL, the two-stage selection procedures were more effective at excluding variables not or weakly associated with the outcome.

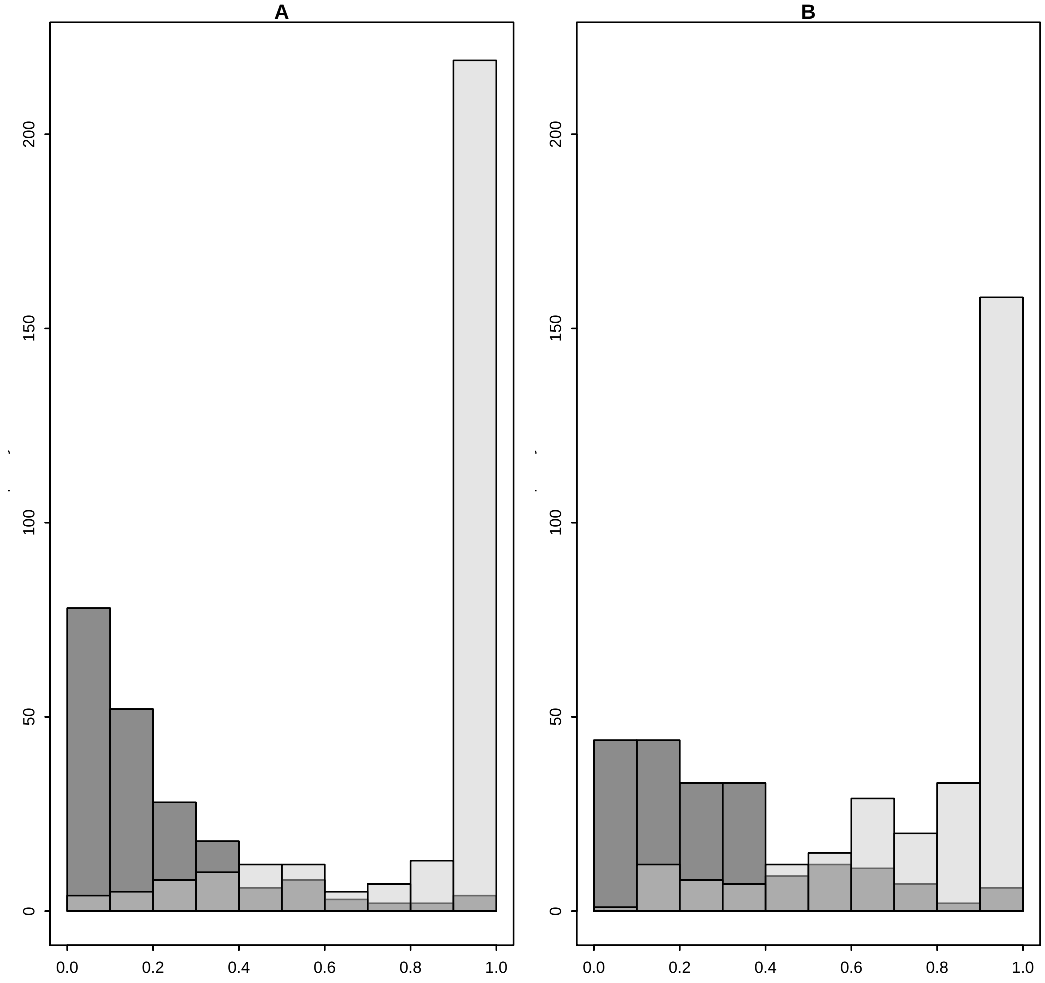

Figure 4.6: Propensity score distributions between the treated (grey) and control (black) if (A) including all covariates in the propensity score model, $\pi^{0.95}_{z=1} = 18\%$ and $\pi^{0.95}_{z=0} = 22\%$; (B) if including only the covariates that were selected more than 20% of times by Step_ALT among 1000 bootstrap samples, $\pi^{0.95}_{z=1} = 33\%$ and $\pi^{0.95}_{z=0} = 49\%$

Table 4.9: Proportion of each variable selected for prediction model across 1000 bootstrap samples.

	Outcome Model		Propensity Model				
Covariate	SW	AL	SW	AL	OAL	Step_ALT	Step_ALY
CD4 t=-1	100	100	26	47	100	100	100
CD4 t=1	100	100	100	100	100	100	100
CD8 t=-1	71	20	20	35	77	20	20
RBC t=1	65	28	35	56	76	30	28
RBC t=-2	64	7	41	64	81	8	7
WBC t=1	59	24	16	39	61	23	25
college	57	9	19	22	38	8	9
CD4 t=-2	52	36	19	48	58	32	36
platelet t=-1	49	14	37	59	65	12	14
CD8 t=1	46	13	62	59	56	14	13
treat t=-3	43	7	38	58	59	6	6
treat t=-1	42	11	100	100	58	12	11
treat t=-2	41	7	80	94	42	9	7
platelet t=-3	37	4	21	34	38	3	4
WBC t=-1	30	1	17	36	40	2	1
age	24	2	28	28	15	1	2
CD8 t=-2	23	1	11	34	35	2	1
RBC t=-1	22	3	17	44	45	5	3
white	21	1	25	21	13	1	1
platelet t=1	19	1	20	40	36	1	1
CD4 t=-3	18	3	12	40	39	3	3
CD8 t=-3	17	2	28	37	25	2	2
WBC t=-2	14	1	19	37	30	1	1
WBC t=-3	13	1	29	37	25	1	1
platelet t=-2	12	1	15	32	27	1	1
RBC t=-3	10	0	21	38	15	1	0

Table 4.10: Treatment effect estimates and 95% confidence intervals.

		IPTW	AIPTW	PENCOMP
allPotent	Rubin/standard	7.5 (-2.2, 17.1)	1.3 (-0.7, 3.3)	0.7 (-1.4, 2.7)
	Bagging	5.8 (-3.0, 14.6)	0.9 (-0.9, 2.6)	0.7 (-1.0, 2.4)
SW	Rubin/standard	11.9 (1.4, 22.4)	2.7 (-0.03, 5.4)	0.9 (-1.9, 3.7)
	Bagging	6.7 (-2.7, 16.0)	1.7 (-0.5, 3.9)	0.9 (-1.3, 3.1)
AL	Rubin/standard	11.7 (1.6, 21.9)	2.8 (-0.7, 6.3)	0.9 (-1.6, 3.4)
	Bagging	6.1 (-2.7, 15.0)	2.3 (-0.6, 5.3)	0.9 (-1.3, 3.1)
OAL	Rubin/standard	2.5 (-6.6, 11.5)	0.9 (-2.1, 3.9)	0.6 (-1.5, 2.7)
	Bagging	4.9 (-3.2, 13.0)	1.6 (-0.9, 4.1)	0.6 (-1.3, 2.5)
Step-ALT	Rubin/standard	0.5 (-6.9, 7.9)	-0.4 (-2.5, 1.7)	-0.05 (-1.8, 1.7)
	Bagging	2.0 (-5.0, 9.0)	0.4 (-1.6, 2.3)	-0.04 (-1.6, 1.5)
Step-ALY	Rubin/standard	0.5 (-7.0, 7.9)	-0.4 (-2.4, 1.6)	-0.09 (-1.8, 1.7)
	Bagging	1.9 (-5.3, 9.0)	0.3 (-1.6, 2.2)	-0.08 (-1.7, 1.6)

.

We estimated the short term effect of antiretroviral treatment on CD4 count using

PENCOMP, AIPTW and IPTW, shown in Table 4.10. The standard errors were obtained using 1000 bootstrap samples. For PENCOMP, 1000 complete datasets were created. Overall, the IPTW estimates had the biggest confidence interval widths. Incorporating the outcome models as in AIPTW and PENCOMP decreased the standard errors and interval widths significantly. PENCOMP tended to have slightly smaller interval widths than AIPTW. The IPTW bootstrap estimates were much more variable, compared to the PENCOMP or AIPTW bootstrap estimates. As seen in the simulation studies, the bagging estimators tended to have smaller standard errors and confidence interval widths than the standard approach for IPTW and AIPTW, or the MI-based approach with Rubin's combining rules for PENCOMP. Excluding irrelevant covariates from the propensity score model, as seen in Step-ALT and Step-ALY, improved the performance of IPTW significantly, in terms of the standard errors and confidence interval widths. Incorporating the outcome models in the AIPTW and PENCOMP attenuated some of the effect of including such covariates.

4.7 Discussion

We propose a new version of PENCOMP via bagging that could have better performance, in terms of SE, confidence interval width and coverage, than the original version of PENCOMP with Rubin's multiple imputation combining rules. The bagging PENCOMP estimator have smaller standard errors, confidence interval width, and better nominal coverage than the MI pencomp estimator when the data are noisy. This can occur when there is limited overlap in the propensity score distributions between the treated and control. Lastly, we modeled PENCOMP as a mixed model in our empirical work, but it would be interesting to compare it with the alternative version, particularly via Bayesian approach (PENCOMP-Bayes), which as a Bayesian method might have attractive small-sample properties. Bagging is a form of model averaging, which can improve the performance of the estimators when the data are noisy. One future topic of research would be to compare it with Bayesian model averaging combined with PENCOMP-Bayes.

Our simulation studies show that excluding strong predictors of the treatment but not of the outcome, or spurious variables, helps improve the performance of the propensity score-based methods, especially for the IPTW estimator. The doubly robust PENCOMP and AIPTW are not as heavily affected by including such variables. However, one shortcoming of using outcome adaptive approach to propensity score model building is that in small samples, it can miss many weak confounders. While the outcome adaptive approach can decrease the standard errors of the estimates, by excluding spurious variables and strong predictors of the treatment but not of the outcome, it can potentially increase bias by excluding variables that are weakly associated with the outcome, especially in small samples. This is a bias-variance trade off problem. In addition, for the IPTW and AIPTW estimators, the bagging approach incorporates model selection so performs better than the standard approach in terms of bias, since it improves the chance that weak confounders are selected in some bootstrap samples. This effect is not seen in PENCOMP, since the multiple imputation-based approach already incorporates model selection. Whether using an outcome adaptive approach can be beneficial depends on specific studies. In the presence of many weak confounders in the data, the reduction in variance from using an outcome adaptive approach might not offset the increase in bias.

On the other hand, in high dimensional setting, including all the observed variables in the propensity model can lead to highly unstable or even infeasible estimation. One criticism of focusing on confounders rather than just predictors of treatment assignment (i.e. balancing covariates between the treatment arms) is that incorporating the outcome in the estimation procedure, whether via prognostic score (Hansen, 2008) or as we have done here, violates the principle that causal inference methods using observational data should mimic as closely as possible randomized trial designs, where outcomes are not considered until the final estimation step. Following such a rule avoids both overt and inadvertent attempts to bias model building toward preferred outcomes ("the garden of forking paths" Gelman and Loken, 2013). However, with the advent of advanced "automatic" penalized regression methods such as adaptive lasso, the risk of such "model shopping" may be sufficiently reduced–though not eliminated, so that analysts that follow the approach outlined here should endeavor to pre-specify to the extent possible the covariates to be used before the analysis begins.

CHAPTER V

Summary and Future Work

In this dissertion, we have proposed PENCOMP as a new, straightforward method to estimate treatment effects in single time-point and in two-time point treatment situations with time-dependent confounders. PENCOMP has the double robustness property for causal effects, which means that PENCOMP offers the analyst two chances to make correct inferences about treatment effects, either by correctly specifying the propensity score model or by correctly specifying the prediction models. In simulation studies, we show that PENCOMP is less sensitive to extreme weights, and flexibile for estimating different estimands such as ATE, ATM and truncated, by restricting to the appropriate subpopulation. We show that excluding variables associated only with treatment reduces the RMSE, and including variables associated only with outcome further reduces the RMSE. Compared with IPTW, PENCOMP as a doubly robust method is less sensitive to the side effects of including strong predictors of the treatment only.

We propose two versions of PENCOMP: 1) PENCOMP-MI–based on the multiple imputation (MI) and MI combining rules for inference; and 2) PENCOMP-bagging–based on bagging. Through simulation studies, we have shown that PENCOMP-bagging could have better performance than PENCOMP-MI, in terms of confidence interval widths and coverage, when the data are noisy, such as in small samples and in the presence of variable selection.

Our next step would be to create a R package so that applied researchers can easily implement our method. Here we also propose some future directions to explore for PENCOMP.

5.1 Missing Data and PENCOMP

PENCOMP is built on Penalized Spline of Propensity Prediction (PSPP) for missing-data problems (Zhang and Little, 2009; Little and An, 2004). Let R denote the response indicator for Y, taking the value 1 if Y is observed and 0 if Y is missing. Let $X = (X_1, ..., X_p)$ denote a set of p fully-observed variables. PSPP first estimates the propensity to respond given X, using a method appropriate for a binary outcome such as logistic regression. The method then predicts the missing values of Y using a linear model that includes as predictors a penalized spline of the estimated propensity to respond and a linear function of other covariates X that are predictive of Y. For the applications considered in this dissertation, we analyzed the treatment effects on complete data sets to focus on the problem of causal inference. However, the estimates on the complete data sets were probably biased, since the subjects who were lost to follow up were probably sicker with lower CD4 counts. In the future, we would consider a more realistic approach that accounts for missing data. We propose using PSPP to impute the missing covariates to create D complete datasets. For each data set $d = 1, \cdots, D$, use PENCOMP-bagging or PENCOMP-MI to impute all the missing potential outcomes, and then combine the D complete datasets with all missing potential outcomes imputed for inference. PENCOMP has the advantage of easily incorporating missing data. Simluation studies would be carried out to assess the performance of such procedure.

5.2 Bayesian PENCOMP

For both versions of PENCOMP, PENCOMP-bagging and PENCOMP-MI, the spline models are fitted via REML. Instead of REML, we can estimate the spline model using a fully Bayesian approach. There we describe the Baysian penalized spline with truncated linear bases.

$$Y^{z_1} = C_1\beta + C_2\theta + \epsilon, \begin{bmatrix} \theta \\ \epsilon \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\theta^2 I & 0 \\ 0 & \sigma_\epsilon^2 I \end{bmatrix} \right),$$

where $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ denote fixed effects, and $\theta = (\theta_{11}, \cdots, \theta_{1K})$ are random basis coefficients. Specify a diffuse prior for β as $P(\beta) \sim 1$; and prior distributions for the variances σ_θ^2 and σ_ϵ^2 as $P(\sigma_\theta^2) \sim IG(A_\theta, B_\theta)$ and $P(\sigma_\epsilon^2) \sim IG(A_\epsilon, B_\epsilon)$. To have non-formative priors, the hyperparameters need to be small.

The posterior distributions for $P(\beta, \theta | Y, \sigma_\epsilon^2, \sigma_\theta^2) \sim N(\Sigma C^T Y, \sigma_\epsilon^2 \Sigma)$, where $\Sigma = (C^T C + \sigma_\epsilon^2/\sigma_\theta^2 D)^{-1}$. The posterior distribution for $P(\sigma_\epsilon^2 | Y, \beta, \theta, \sigma_\theta^2) \sim IG(A_\epsilon + n/2, B_\epsilon + 1/2||y - C_1\beta - C_2\theta||^2)$. The posterior distribution for $P(\sigma_\theta^2 | Y, \beta, \theta, \sigma_\epsilon^2) \sim IG(A_\theta + K/2, B_\theta + 1/2||\theta||^2)$. Compared to the mixed model framework, a fully Bayesian approach takes into account the variability in hyperparameters. Future studies would be done to investigate the performances of these versions of PENCOMP.

5.3 Extension of PENCOMP to Survival Outcome

Through this dissertation, we focus on continous outcome. Another important topic for future research is to extend PENCOMP to non-normal outcomes. For example, we can extend PENCOMP to address the problem of truncation by death. Suppose we are interested in estimating the effect of a treatment on Quality of Life (QOL) that is truncated by death (Rubin, 2002). Some patients die after treatment is assigned and before QOL is measured. For those patients who die, their outcomes QOL are not defined. It is not approriate to treat truncated outcome as a missing data problem, since those outcomes are neither censored or missing. In the presence of censoring by death, the framework of principal stratification can be applied to estimate causal effects (Frangakis and Rubin 2002). The idea is to stratify subjects into four principal strata: subjects who would live under both treatments (LL), subjects who live under treatment and die under control (LD), subjects who die under treatment and live under control (DL), and those who die under both treaments (DD). Since causal effects should be drawn on the same set of people, and subjects who die do not have well defined Quality of Life measure, the causal effects in this case should be defined only for the group LL.

However, since each subject can receive one treatment at a time and only one potential outcome is observed, the principal strata are not unknown. Large sample bounds for causal effects within the principal strata can be obtained with some assumptions (Zhang and Rubin, 2003). Likelihood based approach with EM algorithm was used to estimate causal effects within the principal strata (Zhang and Rubin, 2009). However, with PENCOMP, we can impute the missing survival outcomes and the missing potential outcomes of interest if survived in a single time-point and multiple time-point treatments scenarios, similar to what's done in Chapter 2.

For simplicity, we illustrate our approach to the two time-point treatments. Let $t = 1$ denote the baseline. At time $t = 2$, we first impute the missing survival status S_2 and if $S_2 = 1$ (alive), impute the missing intermediate outcome X_2. Similarly, at time $t = 3$, we first impute the survival status S_3, and if $S_3 = 1$, impute missing potential outcome Y^{jk}. More specifically, the implementations are described as follows:

(a) For $d = 1, \cdots, D$, generate a bootstrap sample $B^{(d)}$ from the original data S by sampling units with replacement, stratified on treatment group. Then carry out steps (b)-(g) for each sample d:

(b) Estimate a logistic regression model for the distribution of Z_1 given baseline covariates X_1, with regression parameters γ_1. Estimate the propensity to be assigned treatment $Z_1 = z_1$ as $\hat{P}_{z_1}(X_1) = \Pr(Z_1 = z_1|X_1, \hat{\gamma}_{z_1}^{(d)})$, where $\hat{\gamma}_{z_1}^{(d)}$ is the ML estimate of γ_{z_1}. Denote $\hat{P}^*_{z_1}$ =log $[\hat{P}_{z_1}(X_1)/(1 - \hat{P}_{z_1}(X_1))]$.

(c) Using the cases assigned to treatment group $Z_1 = z_1$, estimate a logistic regression of for the survival $S_2^{z_1}$ on X_1, with mean

$$logit(P(S_2^{z_1} = 1|X_1, Z_1 = z_1, S_1 = 1, \theta_{z_1}, \beta_{z_1})) = s(\hat{P^*}_{z_1}|\theta_{z_1}) + g_{z_1}(X_1; \beta_{z_1}), \quad (5.1)$$

and normal linear regression of $X_2^{z_1}$ on X_1, with mean

$$E(X_2^{z_1}|X_1, Z_1 = z_1, S_2 = L, \theta_{z_1}, \beta_{z_1}) = s(\hat{P^*}_{z_1}|\theta_{z_1}) + g_{z_1}(X_1; \beta_{z_1}), \quad (5.2)$$

where $s(\hat{P^*}_{z_1}|\theta_{z_1})$ denotes a penalized spline with fixed knots with parameters θ_{z_1}, and $g_{z_1}()$ represents a parametric function of other predictors of the outcome, indexed by parameters β_{z_1}. One of the covariates might be omitted to avoid collinearity. Note that a distinct model is fitted for each treatment regimen.

(d) For $z_1 = 0, 1$, impute the survival status of $S_2^{z_1}$ and for $S_2 = 1$, then impute the values of $X_2^{z_1}$ for subjects in treatment group $1 - z_1$ in the original data set with draws from the predictive distribution of $X_2^{z_1}$ given X_1 from the regression in (c), with ML estimates $\hat{\theta}_{z_1}^{(d)}, \hat{\beta}_{z_1}^{(d)}$ substituted for the parameters $\theta_{z_1}, \beta_{z_1}$.

(e) Estimate the propensity to be assigned treatment $Z_2 = z_2$ given $Z_1, \bar{X}_2$ as $\hat{P}_{z_2}(\bar{X}_2, Z_1) = \Pr(Z_2 = z_2|\bar{X}_2, Z_1 = z_1, \hat{\gamma}_{z_2}^{(d)}, S_2 = 1)$, where $\hat{\gamma}_{z_2}^{(d)}$ is the ML estimate of γ_{z_2}. The probability of treatment regimen $(Z_1 = z_1, Z_2 = z_2, S_1 = 1)$ is denoted as $\hat{P}_{\bar{z}_2} = \hat{P}_{z_1}(X_1)\hat{P}_{z_2}(\bar{X}_2, Z_1)\hat{P}_{z_1}(S_2, X_1)$, where $\hat{P}_{z_1}(S_2, X_1) = \hat{P}(S_2 = 1|X_1, Z_1 = z_1)$ Denote $\hat{P}^*_{\bar{z}_2}$ =log$[\hat{P}_{\bar{z}_2}/(1 - \hat{P}_{\bar{z}_2})]$.

(f) Using the cases assigned to treatment group $\bar{Z}_2 = \bar{z}_2$, given past covariate and treatment histories $\bar{X}_2, \bar{Z}_2$, estimate a logistic regression of $S^{\bar{z}_2}$ with mean

$$logit(P(S_3^{\bar{z}_2} = 1|\bar{X}_2, \bar{Z}_2 = \bar{z}_2, S_2 = 1, \theta_{\bar{z}_2}, \beta_{\bar{z}_2}) = s(\hat{P}^*_{\bar{z}_2}|\theta_{\bar{z}_2}) + g_{\bar{z}_2}(\bar{X}_2, \bar{Z}_2; \beta_{\bar{z}_2})$$

and a normal linear regression of $Y^{\bar{z}_2}$ with mean

$$E(Y^{\bar{z}_2}|\bar{X}_2, \bar{Z}_2 = \bar{z}_2, S_3 = 1, \theta_{\bar{z}_2}, \beta_{\bar{z}_2}) = s(\hat{P}^*_{\bar{z}_2}|\theta_{\bar{z}_2}) + g_{\bar{z}_2}(\bar{X}_2, \bar{Z}_2; \beta_{\bar{z}_2})$$

where $s(\hat{P}^*_{\bar{z}_2}|\theta_{\bar{z}_2})$ denotes a penalized spline with fixed knots with parameters $\theta_{\bar{z}_2}$, and $g_{\bar{z}_2}()$ represents a parametric function of other predictors indexed by parameters $\beta_{\bar{z}_2}$. One of the covariates might need to be omitted from $g_{\bar{z}_2}()$ to avoid collinearity in the covariates.

(g) For each combination of $\bar{z}_2 = (z_1, z_2)$, first impute the missing survival status $S_3 = 1$ and for subjects with $S_3 = 1$, impute the values of $Y^{\bar{z}_2}$ for subjects not assigned this treatment combination in the original data set with draws from the predictive distribution of $Y^{\bar{z}_2}$ from the regression in (f), with ML estimates $\hat{\theta}^{(d)}_{\bar{z}_2}, \hat{\beta}^{(d)}_{\bar{z}_2}$ substituted for the parameters $\theta_{\bar{z}_2}, \beta_{\bar{z}_2}$. From the imputed values, we can infer which principal stratum each subject belongs to.

(h) Use Rubin's Combining rule to combine all the complete datasets with potential outcomes filled in. We can compare the survival probabilities for both treatments, in addition to comparing the mean difference of Y for the revelant principal stratum.

5.4 Extension of PENCOMP to Longitudinal Treatments

In Chapter 2, we focused on two-time point treatment situation. An important question is how PENCOMP can be applied to longitudinal data sets with more than two time points. For example, in the MACS data we analyzed, there are 16 time points, so there are over 30,000 (2^{15}) possible treatment combinations, nearly all of which are not seen in the data. Providing simple and interpretable causal conclusions in such a setting requires careful thought and modeling. In such hign dimensional setting, reparametrization and some form of dimension reduction are needed. For example, restrict inference to the subset of "relevant combinations" judged to have sufficient data to provide meaningful estimates. Propensity models can then be fitted sequentially over time on historical data, including prior treatment assignments and outcomes as potential covariates. The outcomes of relevant combinations can then be imputed as a function of a spline of the propensity and other predictive covariates in the history, with the propensity for each relevant combination obtained by multiplying the sequence of propensities at the set of earlier time points. Some modeling of the resulting treatment effects is likely to be needed to provide parsimonious inferences. For example, a plot of treatment effects against the number of prior "dosages" may suggest a model with a parametric form for the treatment effect as a function of dosage. To maintain stable estimates and enhance interpretability, some form of dimension reduction and variable selection, for example, a summary measure of treatments and other time varying covariates, will typically required. Implementing such strategies is a topic for future research.

5.5 Variable Selection for Propensity Score Model

In Chapter 4, we proposed two-stage techniques: Step-ALY and Step-ALT. For Step-ALY, in the first stage, we select a subset of covariates that are predictive of the outcome using adaptive lasso. In the second stage, we use the subset of covariates found in the first stage in the propensity score model. Similarly, for Step-ALT, we reverse the steps by performing outcome adaptive lasso for the propensity model first and then the prediction model. Future studies could be conducted to see if combining the variable selection for the propensity score and prediction models into a single joint selection model would be more efficient than using these two-stage techniques.

APPENDICES

APPENDIX A

Penalized Spline of Propensity Methods for Treatment Comparison

A.1 Double Robustness of PENCOMP

A.1.1 Single Time Point Treatment Assignment

Let X_1 denote the baseline covariates that affect treatment assignment Z_1. Suppose $Z_1 \in \{0, 1\}$ denotes assignment to control (0) or treatment (1). Let Y^{Z_1} denotes the potential outcome associated with treatment Z_1.

Result 1: The ignorable treatment assignment implies that $(Y^1, Y^0) \perp\!\!\!\perp Z_1 | P_{z_1}(X_1)$ (Rosenbaum and Rubin 1983), where $P_{z_1}(X_1) = Pr(Z_1 = z_1 | X_1)$ denotes the propensity of being assigned z_1.

In the single time point treatment setting, suppose Y^0 is observed only for subjects $i = 1, \cdots, n_0$, while Y^1 is observed only for subjects $i = n_0 + 1, \cdots, n$. We are interested in estimating the causal effect $\Delta = E(Y^1 - Y^0)$. Under SUTVA, ignorability and positivity assumptions, we can estimate causal effects from the regression models on covariates X_1: $E(Y|X_1, Z_1 = 1)$ and $E(Y|X_1, Z_1 = 0)$, or from regression models on a summary measure of the covariates-propensity score $P_{z_1}(X_1)$: $E(Y|P_{z_1}(X_1), Z_1 = 1)$ and $E(Y|P_{z_1}(X_1), Z_1 = 0)$.

$$\begin{aligned}
E(Y^1 - Y^0) &= E(E(Y^1 - Y^0|X_1)) \\
&= E(E(Y^1|X_1)) - E(E(Y^0|X_1)) \\
&= E\Big(E(Y|X_1, Z_1 = 1)\Big) - E\Big(E(Y|X_1, Z_1 = 0)\Big) \text{ by ignorability} \\
&= E\Big(E(Y|P_{z_1}(X_1), Z_1 = 1)\Big) - E\Big(E(Y|P_{z_1}(X_1), Z_1 = 0)\Big)\text{by Result 1}
\end{aligned}$$

Alternatively, the mean $E(Y^1)$ can be written as $E(Y^1) = P(Z_1 = 1)E(Y^1|Z_1 = 1) + P(Z_1 = 0)E(Y^1|Z_1 = 0)$, estimated as:

$$\begin{aligned}
\hat{E}(Y^1) &= \frac{n_0}{n} * \frac{1}{n_0}\sum_{i=1}^{n_0}\hat{Y}_i^1 + \frac{n_1}{n} * \frac{1}{n_1}\sum_{i=(n_0+1)}^{n} Y_i^{obs} \\
&= \frac{1}{n} * \left(\sum_{i=1}^{n_0}\hat{Y}_i^1 + \sum_{i=n_0+1}^{n} Y_i^{obs}\right)
\end{aligned}$$

where $E(Y^1|Z_1 = 1) = Y^{obs}$ and $E(Y^1|P_{z_1}(x_1), Z_1 = 0) = \hat{Y}^1$.

PENCOMP imputes the missing potential outcomes $Y^{z_1=1}$ for subjects $i = 1, \cdots, n_0$ from the mean model $E(Y^{z_1}|X_1, Z_1 = z_1, \theta_{z_1}, \beta_{z_1}) = s(\hat{P^*}_{z_1}; \theta_{z_1}) + g_{z_1}(\hat{P^*}_{z_1}, X_1; \beta_{z_1})$, where $\hat{P^*}_{z_1} = \log\,[\hat{P}_{z_1}(X_1)/(1 - \hat{P}_{z_1}(X_1))]$. Zhang and Little (2009) showed that this imputation model is equivalent to a centered version of the form $E(Y^{z_1}|X_1, Z_1 = z_1, \theta_{z_1}, \beta_{z_1}) = s(\hat{P^*}_{z_1}; \theta_{z_1}) + g_{z_1}(\hat{P^*}_{z_1}, X_1 - s_{x_1}(\hat{P^*}_{z_1}; \omega_{z_1}); \beta_{z_1})$, where $s_{x_1}(\hat{P^*}_{z_1}; \omega_{z_1}) = E(X_1|\hat{P^*}_{z_1})$ is the spline of X_1 on the logit of the propensity score, denoted as, $\hat{P^*}_{z_1}$ as shown in Little and An (2004). Specifically, in the centered version, the residuals from the spline regressions of covariates X_1 on $\hat{P^*}_{z_1}$ enter the parametric g function. Both Zhang and Little (2009) and Little and An (2004) showed that both imputation models in the missing data context yields a consistent estimate for $E(Y^1)$. Here we show the double robustness property of PENCOMP using the centered version for simplicity.

a) When the mean model of Y^1 given $(\hat{P^*}_{z_1}, X_1)$ are correctly specified, the marginal mean of Y^1 from the imputation model is consistent, as a consequence of the properties of a well-defined regression model.

b) When the prediction model given X_1 is misspecified, and the propensity and the spline models are correctly specified, the marginal mean of Y^1 is consistent. Here we prove the case for linear g function. In the case of a nonlinear g function, we can approximate it using linear terms and the results will still hold.

$$\begin{aligned}
E\left(\hat{Y}^1|P^*_{z_1}\right) &= s_y\left(P^*_{z_1}\right) + E\left[g\left(P^*_{z_1}, X_1 - s_{x_1}(P^*_{z_1})\right)|P^*_{z_1}\right] \\
&= s_y\left(P^*_{z_1}\right) + g\left(P^*_{z_1}, E\left(X_1 - s_{x_1}(P^*_{z_1})\middle|P^*_{z_1}\right)\right) \\
&\approx s_y\left(P^*_{z_1}\right) + g\left(P^*_{z_1}, 0,\right) \\
&= s_y\left(P^*_{z_1}\right) \\
&= E\left(Y^1|P^*_{z_1}\right) \\
&= E(Y^1|P^*_{z_1}, Z_1 = 1) \\
&= E(Y^1|P^*_{z_1}, Z_1 = 0)
\end{aligned}$$

where the last two equalities again follow from Result 1.

Thus, for the subjects who actually received controls, the marginal mean of the imputed values $\hat{Y}^1$ from our imputation model is consistent even when the prediction model on covariates is misspecified: $\frac{1}{n_0}\sum_{i=1}^{n_0}\hat{Y}_i^1 \to E(Y^1|Z_1 = 0)$ as $n_0 \to \infty$. Similar approaches can be used to estimate $E(Y^0|Z_1 = 1)$ and thus estimated $E(Y^0)$.

A.1.2 Longitudinal Treatment Assignments

Suppose treatments are assigned at T discrete time points: $t = 1, \ldots, T$. Let $\bar{X}_t$ and $\bar{Z}_t$ denote the covariate and treatment history, respectively, up to and including time point t. Let $Y^{\bar{z}_T}$ denote the potential outcome under treatment regime $\bar{z}_T = (z_1, \cdots, z_T)$. The final outcome of interest $Y^{\bar{z}_T}$ is measured after time point T. Suppose, each z_t is binary treatment. For a particular treatment regime $\bar{z}_T = (z_1, z_2, \cdots, z_t, z_{t+1}, \cdots, z_T)$, under SUTVA, sequential ignorability and positivity assumptions, for all $t = 1, \cdots, T$, the following results hold.

Result 2: $Y^{\bar{z}_T} \perp\!\!\!\perp I(Z_t = z_t) | P_{z_t}(\bar{X}_t, \bar{z}_{t-1})$, where $I(.)$ is the indicator function, and $P_{z_t}(\bar{X}_t, \bar{z}_{t-1}) = P(Z_t = z_t | \bar{X}_t, \bar{Z}_{t-1})$, as a direct extension of the single time point treatment (Rosenbaum and Rubin 1983).

Result 3: $Y^{\bar{z}_T} \perp\!\!\!\perp I(\bar{Z}_t = \bar{z}_t) | P_{\bar{z}_t}$, where $I(.)$ is the indicator function, $P_{\bar{z}_t} = \prod_{k=1}^{t} P(Z_k = z_k | \bar{Z}_{k-1} = \bar{z}_{k-1}, \bar{X}_k)$, which is the propensity of being assigned treatment regime $\bar{z}_t$, conditional on the past treatment and covariate history. In other words, the treatment regime $\bar{Z}_t$ up to and including time point t is independent of potential outcomes $Y^{\bar{z}_T}$ given the propensity of receiving that treatment regime $\bar{Z}_t$, for all $t = 1, \cdots, T$. The proof is outline here.

$$\begin{aligned} P\Big(I(\bar{Z}_t = \bar{z}_t) | Y^{\bar{z}_T}, P_{\bar{z}_t}\Big) &= P\Big(I(\bar{Z}_t = \bar{z}_t) | P_{\bar{z}_t}\Big) \\ &= P_{\bar{z}_t} \end{aligned}$$

$$\begin{aligned} P\Big(I(\bar{Z}_t = \bar{z}_t) | Y^{\bar{z}_T}, P_{\bar{z}_t}\Big) &= E\Big(I(\bar{Z}_t = \bar{z}_t) | Y^{\bar{z}_T}, P_{\bar{z}_t}\Big) \\ &= E\Big[E\Big(I(\bar{Z}_t = \bar{z}_t) | Y^{\bar{z}_T}, \bar{X}_t, \bar{Z}_{t-1}, P_{\bar{z}_t}\Big) | Y^{\bar{z}_T}, P_{\bar{z}_t}\Big] \\ &= E\Big[I(\bar{Z}_{t-1} = \bar{z}_{t-1}) E\Big(I(Z_t = z_t) | \bar{X}_t, \bar{Z}_{t-1}, P_{\bar{z}_t}\Big) | Y^{\bar{z}_T}, P_{\bar{z}_t}\Big] \end{aligned}$$

by sequential ignorability assumption

$$= E\Big[I(\bar{Z}_{t-1} = \bar{z}_{t-1})P_{z_t}(\bar{X}_t, \bar{z}_{t-1})|Y^{\bar{z}_T}, P_{\bar{z}_t}\Big]$$

$$= E\Big[E\Big(I(\bar{Z}_{t-1} = \bar{z}_{t-1})P_{z_t}(\bar{X}_t, \bar{z}_{t-1})|Y^{\bar{z}_T}, \bar{X}_{t-1}, \bar{Z}_{t-2}, P_{\bar{z}_t}\Big)\Big|Y^{\bar{z}_T}, P_{\bar{z}_t}\Big]$$

$$= E\Big[I(\bar{Z}_{t-2} = \bar{z}_{t-2})P_{z_t}(\bar{X}_t, \bar{z}_{t-1})E\Big(I(Z_{t-1} = z_{t-1})|Y^{\bar{z}_T}, \bar{X}_{t-1}, \bar{Z}_{t-2}, P_{\bar{z}_t}\Big)$$

$$\Big|Y^{\bar{z}_T}, P_{\bar{z}_t}\Big]$$

$$= E\Big[I(\bar{Z}_{t-2} = \bar{z}_{t-2})P_{z_t}(\bar{X}_t, \bar{z}_{t-1})P_{z_{t-1}}(\bar{X}_{t-1}, \bar{z}_{t-2})\Big|Y^{\bar{z}_T}, P_{\bar{z}_t}\Big]$$

$$= E\Big[P_{\bar{z}_t}|Y^{\bar{z}_T}, P_{\bar{z}_t}\Big] \text{by the same argument for each } Z_t$$

$$= P_{\bar{z}_t}$$

By the same argument but without the need for the sequential ignorability assumption, $P\Big(I(\bar{Z}_t = \bar{z}_t)|P_{\bar{z}_t}\Big) = P_{\bar{z}_t}$. Thus, $P\Big(I(\bar{Z}_t = \bar{z}_t)|Y^{\bar{z}_T}, P_{\bar{z}_t}\Big) = p\Big(I(\bar{Z}_t = \bar{z}_t)|P_{\bar{z}_t}\Big)$

Suppose we want to impute the missing potential outcomes X_3^{11} for subjects $1, \cdots, n_0$ and subjects $i = n_0 + 1, \cdots n$ receive treatment combination $(1, 1)$. As shown below, we can build a model for X_3^{11} from the subjects with observed treatment sequence of $(1, 1)$ to impute missing potential outcomes X_3^{11} for other subjects. Similar to single time point treatment, we can estimate causal effects from the regression models on the covariates or on the propensity scores.

$$E\Big(X_3^{11}\Big) = E\Big[E(X_3^{11}|\bar{X}_2)\Big]$$

$$= E\Big[E(X_3|\bar{X}_2, Z_1 = 1, Z_2 = 1)\Big] \text{by sequential ignorability}$$

$$= E\Big[E(X_3|P_{\bar{z}_2=(11)}, Z_1 = 1, Z_2 = 1)\Big] \text{by result 3}$$

Alternatively, the mean $E(X_3^{11})$ can be written as $E(X_3^{11}) = P(\bar{Z}_2 = (1, 1))E(X_3^{11}|\bar{Z}_2 =$

$(1,1)) + P(\bar{Z}_2 \neq (1,1))E(X_3^{11}|\bar{Z}_2 \neq (1,1))$, estimated as:

$$\begin{aligned}
\hat{E}(X_3^{11}) &= \frac{n_{00}}{n} * \frac{1}{n_{00}} \sum_{i=1}^{n_{00}} \hat{X}_3^{\ 11} + \frac{n_{01}}{n} * \frac{1}{n_{01}} \sum_{i=n_{00}+1}^{n_{00}+n_{01}} \hat{X}_3^{\ 11} \\
&+ \frac{n_{10}}{n} * \frac{1}{n_{10}} \sum_{i=n_{00}+n_{01}+1}^{n_{00}+n_{01}+n_{10}} \hat{X}_3^{\ 11} + \sum_{i=n_{00}+n_{01}+n_{10}+1}^{n} X_3^{obs} \\
&= \frac{1}{n} * \left(\sum_{i=1}^{n_{00}+n_{01}+n_{10}} \hat{X}_3^{\ 11} + \sum_{i=n_{00}+n_{01}+n_{10}+1}^{n} X_3^{obs} \right)
\end{aligned}$$

where $\hat{X}_3^{11} = \hat{E}(X_3^{11}|P_{\bar{z}_2}, Z_1 \neq 1, Z_2 \neq 1)$.

PENCOMP imputes the first missing intermediate outcomes X_2 first, X_3, and continue forward to the final outcome Y. By induction, we can show PENCOMP has double robustness property in longitudinal study. We have shown double robustness property for the base case $t = 1$ as in the single treatment. Suppose PENCOMP has the double robustness property in imputing missing potential outcomes X_t. We want to show that the double robustness property also holds for the missing potential outcomes X_{t+1}, Suppose we are interested in estimating $X_{t+1}^{\bar{z}_t}$, where $\bar{z}_t = (z_1, \cdots, z_t)$ and subjects $i = 1, \cdots, n_0$ do not treatment sequence $\bar{Z}_t$ that match $\bar{z}_t$. Thus, to impute the missing potential outcomes $X_{t+1}^{\bar{z}_t}$ for the subjects whose treatment sequence did not match $\bar{z}_t$, we draw values from the mean model $E(X_{t+1}^{\bar{z}_t}|\bar{X}_t, \bar{Z}_t = \bar{z}_t, \theta_{\bar{z}_t}, \beta_{\bar{z}_t}, \gamma_{\bar{z}_t}) = s_{x_{t+1}}(\hat{P}_{\bar{z}_t}^*; \theta_{\bar{z}_t}) + g\left[\hat{P}_{\bar{z}_t}^*, X_1, \cdots, X_t; \beta_{\bar{z}_t}\right]$, which is equivalent to the mean model $E(X_{t+1}^{\bar{z}_t}|\bar{X}_t, \bar{Z}_t = \bar{z}_t, \theta_{\bar{z}_t}, \beta_{\bar{z}_t}, \gamma_{\bar{z}_t}) = s_{x_{t+1}}(\hat{P}_{\bar{z}_t}^*; \theta_{\bar{z}_t}) + g\left[\hat{P}_{\bar{z}_t}^*, X_1 - s_{x_1}(\hat{P}_{\bar{z}_t}^*; \omega_{\bar{z}_t^1}), \cdots, X_t - s_{x_t}(\hat{P}_{\bar{z}_t}^*; \omega_{\bar{z}_t^t}); \beta_{\bar{z}_t}\right]$, where $\hat{P}_{\bar{z}_t}^* = \log\left(\hat{P}_{\bar{z}_t}/(1 - \hat{P}_{\bar{z}_t})\right)$. Here we need to show the double robustness property of PENCOMP with the centered version.

a) When the mean model of $X_{t+1}^{\bar{z}_t}$ given the covariate history $\bar{X}_t$ are correctly specified, the marginal mean of $X_{t+1}^{\bar{z}_t}$ from the imputation model is consistent, as a consequence of well-defined regression models.

b) When the prediction model given $\bar{X}_t$ is misspecified, and all the propensity models up to and including time point t and the spline models are correctly specified, the marginal mean of $X_{t+1}^{\bar{z}_t}$ is consistent. Again we prove the case for linear g function. We can approximate a nonlinear g function with using linear terms and the results will still hold.

$$\begin{aligned}
E\Bigg(\hat{X}_{t+1}^{\bar{z}_t}|P_{\bar{z}_t}^*\Bigg) &= s_{x_{t+1}}\Bigg(P_{\bar{z}_t}^*\Bigg) + E\Bigg[g\Bigg(P_{\bar{z}_t}^*, X_1 - s_{x_1}(P_{\bar{z}_t}^*), \cdots, X_t - s_{x_t}(P_{\bar{z}_t}^*)\Bigg)|P_{\bar{z}_t}^*\Bigg] \\
&= s_{x_{t+1}}\Bigg(P_{\bar{z}_t}^*\Bigg) + g\Bigg[P_{\bar{z}_t}^*, E\Bigg(X_1 - s_{x_1}(P_{\bar{z}_t}^*)|P_{\bar{z}_t}^*\Bigg), \cdots, E\Bigg(X_t - s_{x_t}(P_{\bar{z}_t}^*)|P_{\bar{z}_t}^*\Bigg)\Bigg] \\
&\approx s_{x_{t+1}}\Bigg(P_{\bar{z}_t}^*\Bigg) + g\Bigg[P_{\bar{z}_t}^*, 0, \cdots, 0\Bigg] \\
&= s_{x_{t+1}}\Bigg(P_{\bar{z}_t}^*\Bigg) \\
&= E\Bigg(X_{t+1}^{\bar{z}_t}|P_{\bar{z}_t}^*\Bigg) \\
&= E(X_{t+1}^{\bar{z}_t}|P_{\bar{z}_t}^*, \bar{Z}_t \neq \bar{z}_t) \\
&= E(X_{t+1}^{\bar{z}_t}|P_{\bar{z}_t}^*, \bar{Z}_t = \bar{z}_t)\text{by result 4}
\end{aligned}$$

where the last two equalities follow from Result 3.

Thus, $\frac{1}{n_{0k}}\sum_{i=1}^{n_{0k}} \hat{X}_{k+1,i}^{\bar{z}_k} \rightarrow E(X_{k+1}^{\bar{z}_k}|\bar{Z}_k \neq \bar{z}_k)$ as $n_{0k} \rightarrow \infty$, where n_{0k} is the sample size of the observations for which $\bar{Z}_k \neq \bar{z}_k$, and we assume that the observations are ordered that the first n_0 corresponds to the observations for which $\bar{Z}_k \neq \bar{z}_k$. Thus, by induction, PENCOMP has double robustness property in longitudinal study.

A.2 Implementations of the IPTW and the AIPTW Estimators

A.2.1 IPTW

Let $O_i = (\bar{X}_{iT}, \bar{Z}_{iT}, Y_i)$ denote the observed data for subject i, where $i = 1, \cdots, n$. The likelihood of the observed data can be factored into two components $P(O) = Q_0 g_0$, where $Q_0 = P(Y|\bar{X}_T, \bar{Z}_T = \bar{z}_T) \prod_{t=1}^{T} P(X_t|\bar{X}_{t-1}, \bar{Z}_{t-1})$ and $g_0 = \prod_{t=1}^{T} P(Z_t|\bar{Z}_{t-1}, \bar{X}_{t-1})$. Denote the MLE of Q_0 and g_0 as Q_n and g_n respectively.

From the IPTW estimating equation $\sum_{i=1}^{n} D_{IPTW}(O_i|\beta, g_n) = 0$, we can obtain $\hat{E}(Y^{z_1}) = \sum_{i=1}^{n} \frac{I(Z_{1i}=z_{1i})}{\hat{P}(Z_{1i}=z_{1i}|X_{1i})})^{-1} \sum_{i=1}^{n} \frac{Z_{1i}Y_i}{\hat{P}(Z_{1i}|X_{1i})}$. Thus, the estimated causal effect $\hat{\Delta}$ in a single time point is

$$\hat{\Delta}^{IPTW} = (\sum_{i=1}^{n} \frac{Z_{1i}}{\hat{P}(Z_{1i}|X_{1i})})^{-1} \sum_{i=1}^{n} \frac{Z_{1i}Y_i}{\hat{P}(Z_{1i}|X_{1i})} - (\sum_{i=1}^{n} \frac{1-Z_{1i}}{1-\hat{P}(Z_{1i}|X_{1i})})^{-1} \sum_{i=1}^{n} \frac{(1-Z_{1i})Y_i}{(1-\hat{P}(Z_{1i}|X_{1i}))}$$

Similarly, in a two time points treatment, the estimated causal effects $\hat{\Delta}_{z_1 z_2}$ are

$$\begin{aligned}\hat{\Delta}^{IPTW}_{z_1 z_2} &= (\sum_{i=1}^{n} \frac{I(Z_{1i}=z_1, Z_{2i}=z_2)}{P(Z_{1i}|X_{1i})P(Z_{2i}|X_{1i}, X_{2i}, Z_{1i})})^{-1} \sum_{i=1}^{n} \frac{I(Z_{1i}=z_1, Z_{2i}=z_2)Y_i}{P(Z_{1i}|X_{1i})P(Z_{2i}|X_{1i}, X_{2i}, Z_{1i})} \\ &- (\sum_{i=1}^{n} \frac{I(Z_{1i}=0, Z_{2i}=0)}{P(Z_{1i}|X_{1i})P(Z_{2i}|X_{1i}, X_{2i}, Z_{1i})})^{-1} \sum_{i=1}^{n} \frac{I(Z_{1i}=0, Z_{2i}=0)Y_i}{P(Z_{1i}|X_{1i})P(Z_{2i}|X_{1i}, X_{2i}, Z_{1i})}\end{aligned}$$

A.2.2 AIPTW

To solve the estimating equation $\sum_{i=1}^{n} D_{AIPTW}(O_i|\beta, g_n, Q_n) = 0$ in the single treatment assignment setting, we proceeds as follows.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample $S^{(d)}$ from the original data S by sampling units with replacement, stratified on treatment group. Then carry out steps (b)-(h) for each sample d:

(b) Estimate a logistic regression model for the distribution of Z_1 given X_1, with regression parameters γ_{z_1} . Estimate the propensity to be assigned treatment $Z_1 = z_1$ as $\hat{P}(Z_1 = z_1 | X_1, \hat{\gamma_{z_1}}^{(d)})$, where $\hat{\gamma_{z_1}}^{(d)}$ is the ML estimate of γ_{z_1}.

(c) For $z_1 = 0, 1$, using the cases assigned to treatment group z_1, estimate the distribution Y given X_1 and Z_1, $\hat{P}(Y|X_1 = x_1, Z_1 = z_1)$, using a normal linear regression with mean $E(Y|X_1, Z_1 = z_1, \beta_{z_1}) = g_{z_1}(X_1; \beta_{z_1})$, where $g_{z_1}()$ represents a parametric function of X_1 and Z_1 indexed by parameters β_{z_1}.

(d) Estimate the distributions of baseline covariates $P(X_1)$ using the empirical distributions from the data, denoted as $\hat{P}(X_1)$.

e) Estimate $\hat{\beta}_n^{mc} = (\hat{\beta}_0^{mc}, \hat{\beta}_1^{mc})$ using the g-computation to generate $10,000$ number of Y_0 and Y_1 from their respective counterfactual reference distributions. Specifically, draw x_1^* from the empirical distribution of X_1, $\hat{P}(X_1)$. Set $Z_1 = z_1$ and generate draws y^* from $\hat{P}(Y|X_1 = x_1^*, Z_1 = z_1)$. Then fit the MSM model $E(Y^{Z_1}) = \beta_0 + \beta_1 Z_1$ to this collection of $(y^*, 1)$ and $(y^*, 0)$ to obtain $\hat{\beta}_n^{mc}$.

f) Using Q_n, g_n and $\hat{\beta}_n^{mc}$, estimate $E_{Q_n,g_n}[D_{IPTW}(O_i|\hat{\beta}_n^{mc}, g_n)|Z_{1i} = z_{1i}, X_{1i} = x_{1i}]$ for each subject i as follows. Given $(Z_{1i} = z_{1i}, X_{1i} = x_{1i})$, generate 2,000 draws of Y_i^{mc} from $\hat{P}(Y|X_{1i} = x_{1i}, Z_{1i} = z_{1i})$ and compute

$$D_i^{mc} = \frac{\hat{h}(Z_{1i})}{\hat{P}(Z_{1i}|X_{1i})}(Y_i^{mc} - (\hat{\beta}_0^{mc} + \hat{\beta}_1^{mc} Z_{1i}))$$

where $\hat{h}(Z_{1i}) = \frac{dE(Y^{Z_{1i}})}{d\beta}\hat{P}(Z_{1i})$. Take the mean of 2000 Monte Carlo values as the estimate.

g) Similarly estimate $E_{Q_n,g_n}[D_{IPTW}(O_i|\hat{\beta}_n^{mc}, g_n)|X_{1i} = x_{1i}]$. Given $X_{1i} = x_{1i}$, first generate draws of z_{1i}^{mc} from $\hat{P}(Z_{1i}|X_{1i} = x_{1i})$, then generate draws of Y_i^{mc} from $\hat{P}(Y|X_{1i} = x_{1i}, Z_{1i} = z_{1i}^{mc})$ and compute D_i^{mc}. Take the mean of 2000 Monte Carlo values D_i^{mc} as the estimate.

h) Let $\hat{\pi}_i = \hat{E}_{Q_n,g_n}[D_{IPTW}(O|\hat{\beta}_n^{mc}, g_n)|Z_{1i}, X_{1i}] - E_{Q_n,g_n}[D_{IPTW}(O|\hat{\beta}_n^{mc}, g_n)|X_{1i}]$. Solve (β_0, β_1) using Newton Raphson algorithm

$$\sum_{i=1}^{n} D_{AIPTW}(O_i|\beta, g_n, Q_n) = \sum_{i=1}^{n} D_{IPTW}(O_i|\beta, g_n) - \hat{\pi}_i = 0$$

The treatment effect is $\hat{\Delta}^{AIPTW} = \sum_{d=1}^{D} \hat{\beta}_1^{(d)}/D$. Estimate the variance by bootstrap and obtain the 95% confidence interval from the bootstrap samples.

Similarly to solve the AIPTW estimating equation in a two time points treatment, the steps proceeds as follows. Let $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$.

(a) For $d = 1, \cdots, D$, generate a bootstrap sample $S^{(d)}$ from the original data S by sampling units with replacement, stratified on treatment group. Then carry out steps (b)-(i) for each sample d:

(b) Estimate a logistic regression model for the distribution of Z_1 given X_1, with regression parameters γ_{z_1}. Estimate the propensity to be assigned treatment $Z_1 = z_1$ as $\hat{P}(Z_1 = z_1|X_1, \hat{\gamma}_{z_1}^{(d)})$, where $\hat{\gamma}_{z_1}^{(d)}$ is the ML estimate of γ_{z_1}.

(c) Estimate the distributions of baseline covariates $P(X_1)$ as the empirical distributions from the data, denoted as $\hat{P}(X_1)$.

(d) Using the cases assigned to treatment group $Z_1 = z_1$, estimate $\hat{P}(X_2|X_1, Z_1)$ using a normal linear regression with mean

$$E(X_2^{z_1}|X_1, Z_1 = z_1, \theta_{z_1}, \beta_{z_1}) = g_{z_1}(X_1, Z_1, \beta_{z_1}) \tag{A.1}$$

where $g_{z_1}()$ represents a parametric function of X_1, and Z_1 indexed by parameters β_{z_1}.

(e) Estimate a logistic regression model for the distribution of Z_2 given $\bar{X}_2, Z_1$, with regression parameters γ_{z_2}. Estimate the propensity to be assigned treatment $Z_2 = z_2$ given $Z_1, \bar{X}_2$ as $\hat{P}(Z_2 = z_2|\bar{X}_2, Z_1, \hat{\gamma}_{z_2}^{(d)})$, where $\hat{\gamma}_{z_2}^{(d)}$ is the ML estimate of γ_{z_2}.

(f) Using the cases assigned to treatment regime $\bar{Z}_2 = \bar{z}_2$, estimate $\hat{P}(Y|\bar{X}_2, \bar{Z}_2)$ using a normal linear regression with mean

$$E(Y^{\bar{z}_2}|\bar{X}_2, \bar{Z}_2 = \bar{z}_2, \beta_{\bar{z}_2}) = g_{z_1 z_2}(\bar{X}_2, \bar{Z}_2; \beta_{\bar{z}_2})$$

where $g_{\bar{z}_2}()$ represents a parametric function indexed by parameters $\beta_{\bar{z}_2}$.

g) Estimate $\hat{\beta}_n^{mc} = (\hat{\beta}_0^{mc}, \hat{\beta}_1^{mc}, \hat{\beta}_2^{mc}, \hat{\beta}_3^{mc})$ using the g-computation to generate $10,000$ draws of the potential outcomes Y^{00},Y^{01},Y^{11},Y^{10} from their respective counterfactual distributions. Specifically, first generate a draw x_1^* from the empirical distribution $\hat{P}(X_1)$. Set $Z_1 = z_1$ and generate a draw x_2^* from $\hat{P}(X_2|X_1 = x_1^*, Z_1 = z_1)$. Then set $Z_2 = z_2$ and generate draws y^* from $\hat{P}(Y|X_1 = x_1^*, Z_1 = z_1, X_2 = x_2^*, Z_2 = z_2)$. Then fit the model $E(Y^{\bar{Z}_2}) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1 Z_2$ to this collection of $(y^*, 0, 0)$, $(y^*, 1, 0)$, $(y^*, 0, 1)$ and $(y^*, 1, 1)$ to obtain $\hat{\beta}_n^{mc}$.

h) Using Q_n, g_n and $\hat{\beta}_n^{mc}$, estimate $E_{Q_n,g_n}[D_{IPTW}(O_i|\hat{\beta}_n^{mc}, g_n)|\bar{Z}_{2i} = \bar{z}_{2i}, \bar{X}_{2i} = \bar{x}_{2i}]$ for each subject i as follows. Given $(\bar{Z}_{2i} = \bar{z}_{2i}, \bar{X}_{2i} = \bar{x}_{2i})$, generate 2,000 draws of Y_i^{mc} from $P(Y|\bar{Z}_{2i} = \bar{z}_{2i}, \bar{X}_{2i} = \bar{x}_{2i})$ and compute D_i^{mc}. Take the mean of the 2,000 Monte Carlo values as the estimate.

$$D_i^{mc} = \frac{\hat{h}(\bar{Z}_{2i})}{\hat{P}(Z_{1i}|X_{1i})\hat{P}(Z_{2i}|Z_{1i}, \bar{X}_{2i})}(Y_i^{mc} - (\hat{\beta}_0^{mc} + \hat{\beta}_1^{mc} Z_{1i} + \hat{\beta}_2^{mc} Z_{2i} + \hat{\beta}_3^{mc} Z_{1i} Z_{2i}))$$

where $\hat{h}(\bar{Z}_{2i}) = \frac{dE(Y^{\bar{Z}_{2i}})}{d\beta}\hat{P}(Z_{1i})\hat{P}(Z_{2i}|Z_{1i})$. Follow the similar procedures to estimate the other three conditional expectations.

i) Solve the estimating equation using Newton Raphson algorithm

$$\sum_{i=1}^{n} D_{AIPTW}(O_i|\beta, g_n, Q_n) = \sum_{i=1}^{n} D_{IPTW}(O_i|\beta, g_n) - \hat{\pi}_i = 0 \qquad \text{(A.2)}$$

where,

$$\hat{\pi}^i = \sum_{j=1}^{j=2} E_{Q_n,g_n}[D_{IPTW}(O|\hat{\beta}^{mc}, g_n)|\bar{Z}_j, \bar{X}_j] - E_{Q_n,g_n}[D_{IPTW}(O|\hat{\beta}^{mc}, g_n)|\bar{X}_j]$$

The treatment effects are $\hat{\Delta}_{11}^{AIPTW} = \sum_{d=1}^{D} \hat{\Delta}_{11}^{AIPTW(d)}/D$, $\hat{\Delta}_{10}^{AIPTW} = \sum_{d=1}^{D} \hat{\Delta}_{10}^{AIPTW(d)}/D$, and $\hat{\Delta}_{01}^{AIPTW} = \sum_{d=1}^{D} \hat{\Delta}_{01}^{AIPTW(d)}/D$, where $\hat{\Delta}_{11}^{AIPTW(d)} = \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$; $\hat{\Delta}_{10}^{AIPTW(d)} = \hat{\beta}_1$; $\hat{\Delta}_{01}^{AIPTW(d)} = \hat{\beta}_2$. Estimate the variance and obtain the 95% confidence interval from D bootstrap samples.

A.3 Supplemental Tables from the Simulation Study

Table A.1: 100 * Ratio of bias over RMSE of IPTW (A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 200. The treatment effects Δs under linear and nonlinear outcome models were 5 and 9, respectively.

$\Delta = E(Y^1) - E(Y^0)$						
100 * Empirical Bias / RMSE IPTW(A)						
	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	1	14	25	4	3	4
g-computation(A)	-0	-0	-1	6	4	3
AIPTW(A)	0	0	2	2	2	2
PENCOMP(A)	0	2	2	2	3	3
IPTW(A)	1	14	25	4	3	4
g-computation(B)	79	357	303	41	225	225
AIPTW(B)	0	18	29	-1	3	5
PENCOMP(B)	11	2	5	-1	-37	-52
IPTW(C)	82	375	340	46	250	273
g-computation(A)	-0	-0	-1	6	4	3
AIPTW(C)	0	0	0	2	2	1
PENCOMP(C)	-0	0	0	2	2	1

Table A.2: 100*Ratio of empirical RMSE over RMSE of IPTW (A), denoted as RMSE/RMSE IPTW(A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly- specified prediction model only, based on 1000 simulations with sample size of 200.

$\Delta = E(Y^1) - E(Y^0)$						
100 * RMSE / RMSE IPTW(A)						
	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100
g-computation(A)	80	51	37	75	61	51
AIPTW(A)	78	59	47	73	63	55
PENCOMP(A)	78	57	46	73	62	54
IPTW(A)	100	100	100	100	100	100
g-computation(B)	168	367	307	124	246	240
AIPTW(B)	81	89	90	95	99	97
PENCOMP(B)	83	65	56	91	97	102
IPTW(C)	181	389	347	130	273	290
g-computation(A)	80	51	37	75	61	51
AIPTW(C)	78	53	39	73	60	51
PENCOMP(C)	78	54	40	73	61	51

Table A.3: Empirical 95% non-coverage rate*100 (nominal noncoverage of 5), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 200.

$\Delta = E(Y^1) - E(Y^0)$						
100 * 95% Non-coverage Rate						
	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	5	6	12	5	6	7
g-computation(A)	4	5	6	5	6	5
AIPTW(A)	4	6	6	5	6	5
PENCOMP(A)	4	3	3	4	5	3
IPTW(A)	5	6	12	5	6	7
g-computation(B)	10	99	100	6	64	81
AIPTW(B)	3	8	13	5	6	7
PENCOMP(B)	0	0	1	2	5	6
IPTW(C)	10	96	99	6	63	82
g-computation(A)	4	5	6	5	6	5
AIPTW(C)	4	5	7	5	6	6
PENCOMP(C)	4	4	5	5	5	5

Table A.4: 100 * Ratio of empirical mean 95% confidence interval width to that of IPTW (A), denoted as mean 95% interval width/mean 95% interval width IPTW(A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 200.

$\Delta = E(Y^1) - E(Y^0)$						
100 * mean 95% interval width/mean 95% interval width IPTW(A)						
	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100
g-computation(A)	80	53	42	74	64	57
AIPTW(A)	79	60	60	73	66	67
PENCOMP(A)	80	69	68	74	70	72
IPTW(A)	100	100	100	100	100	100
g-computation(B)	136	84	64	114	105	95
AIPTW(B)	84	89	93	96	97	100
PENCOMP(B)	124	102	102	113	120	130
IPTW(C)	147	103	82	118	117	113
g-computation(A)	80	53	42	74	64	57
AIPTW(C)	79	55	44	73	64	58
PENCOMP(C)	79	58	50	73	65	61

Table A.5: 100 * Ratio of bias over RMSE of IPTW (A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 500. The treatment effects Δs under linear and nonlinear outcome models were 5 and 9, respectively.

$\Delta = E(Y^1) - E(Y^0)$

100 * Empirical Bias / RMSE IPTW (A)

	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	1	4	16	-5	-8	-1
g-computation(A)	0	-0	-1	-0	-0	-0
AIPTW(A)	-0	-2	0	-2	-3	-1
PENCOMP(A)	-0	-0	0	-3	-2	-1
IPTW(A)	1	4	16	-5	-8	-1
g-computation(B)	123	482	406	58	333	327
AIPTW(B)	1	6	16	-8	-9	-4
PENCOMP(B)	15	-1	2	-2	-46	-67
IPTW(C)	125	510	458	62	367	396
g-computation(A)	0	-0	-1	-0	-0	-0
AIPTW(C)	-0	0	-0	-2	-2	-1
PENCOMP(C)	-1	-0	-0	-3	-2	-1

Table A.6: 100*Ratio of empirical RMSE over RMSE of IPTW (A), denoted as RMSE/RMSE IPTW(A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly- specified prediction model only, based on 1000 simulations with sample size of 500.

$\Delta = E(Y^1) - E(Y^0)$

100 * RMSE / RMSE IPTW(A)

	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100
g-computation(A)	82	45	31	72	54	44
AIPTW(A)	79	54	42	70	56	49
PENCOMP(A)	79	49	38	70	55	47
IPTW(A)	100	100	100	100	100	100
g-computation(B)	184	487	408	128	345	336
AIPTW(B)	79	85	92	94	94	98
PENCOMP(B)	82	59	51	90	93	101
IPTW(C)	193	517	462	134	382	406
g-computation(A)	82	45	31	72	54	44
AIPTW(C)	79	46	32	70	53	44
PENCOMP(C)	79	46	33	70	53	44

Table A.7: Empirical 95% non-coverage rate*100 (nominal noncoverage of 5), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 500.

$\Delta = E(Y^1) - E(Y^0)$						
100 * 95% Non-coverage Rate						
	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	5	6	11	5	6	6
g-computation(A)	5	6	6	4	4	4
AIPTW(A)	4	6	7	3	4	4
PENCOMP(A)	4	4	3	3	3	2
IPTW(A)	5	6	11	5	6	6
g-computation(B)	15	100	100	7	96	100
AIPTW(B)	4	7	13	5	6	6
PENCOMP(B)	0	1	1	3	6	10
IPTW(C)	12	100	100	7	97	100
g-computation(A)	5	6	6	4	4	4
AIPTW(C)	4	6	5	3	4	4
PENCOMP(C)	4	4	4	3	3	3

Table A.8: 100 * Ratio of empirical mean 95% confidence interval width to that of IPTW (A), denoted as mean 95% interval width/mean 95% interval width IPTW(A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 500.

$\Delta = E(Y^1) - E(Y^0)$

100 * mean 95% interval width/mean 95% interval width IPTW(A)

	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100
g-computation(A)	84	52	38	77	61	53
AIPTW(A)	81	58	51	74	63	60
PENCOMP(A)	81	61	54	75	64	61
IPTW(A)	100	100	100	100	100	100
g-computation(B)	141	82	57	116	101	88
AIPTW(B)	82	88	92	95	95	96
PENCOMP(B)	120	92	85	107	108	113
IPTW(C)	151	98	73	119	113	105
g-computation(A)	84	52	38	77	61	53
AIPTW(C)	81	52	39	74	60	53
PENCOMP(C)	81	55	43	74	61	55

.

Table A.9: 100 * Ratio of bias over RMSE of IPTW (A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 1000. The treatment effects Δs under linear and nonlinear outcome models were 5 and 9, respectively.

$\Delta = E(Y^1) - E(Y^0)$

100 * Empirical Bias / RMSE IPTW (A)

	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	-1	3	11	-2	-3	-1
g-computation(A)	2	2	1	-2	-1	-0
AIPTW(A)	2	1	0	-3	-2	-2
PENCOMP(A)	3	2	1	-2	-1	-1
IPTW(A)	-1	3	11	-2	-3	-1
g-computation(B)	182	674	517	92	459	420
AIPTW(B)	2	7	14	-2	-1	-0
PENCOMP(B)	21	1	3	6	-36	-61
IPTW(C)	181	706	578	95	502	505
g-computation(A)	2	2	1	-2	-1	-0
AIPTW(C)	1	1	0	-3	-2	-1
PENCOMP(C)	3	2	1	-2	-1	-1

Table A.10: 100*Ratio of empirical RMSE over RMSE of IPTW (A), denoted as RMSE/RMSE IPTW(A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly- specified prediction model only, based on 1000 simulations with sample size of 1000.

$\Delta = E(Y^1) - E(Y^0)$

100 * RMSE / RMSE IPTW(A)

	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100
g-computation(A)	85	45	29	79	58	44
AIPTW(A)	80	53	45	74	59	52
PENCOMP(A)	80	49	36	74	56	45
IPTW(A)	100	100	100	100	100	100
g-computation(B)	233	678	518	151	468	427
AIPTW(B)	81	90	94	96	94	94
PENCOMP(B)	85	59	50	92	88	96
IPTW(C)	238	711	581	153	513	512
g-computation(A)	85	45	29	79	58	44
AIPTW(C)	80	45	30	74	54	42
PENCOMP(C)	80	45	30	74	54	42

Table A.11: Empirical 95% non-coverage rate*100 (nominal noncoverage of 5), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 1000.

$\Delta = E(Y^1) - E(Y^0)$						
100 * 95% Non-coverage Rate						
	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	5	6	12	4	6	6
g-computation(A)	6	5	5	6	5	5
AIPTW(A)	5	5	6	5	6	6
PENCOMP(A)	5	5	4	5	5	5
IPTW(A)	5	6	12	4	6	6
g-computation(B)	26	100	100	12	100	100
AIPTW(B)	4	7	13	6	6	6
PENCOMP(B)	1	1	2	3	4	8
IPTW(C)	24	100	100	11	100	100
g-computation(A)	6	5	5	6	5	5
AIPTW(C)	5	5	6	5	5	5
PENCOMP(C)	5	4	4	5	5	5

Table A.12: 100 * Ratio of empirical mean 95% confidence interval width to that of IPTW (A), denoted as mean 95% interval width/mean 95% interval width IPTW(A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 1000.

$\Delta = E(Y^1) - E(Y^0)$						
100 * mean 95% interval width/mean 95% interval width IPTW(A)						
	Linear Outcome			NonLinear Outcome		
Method	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100
g-computation(A)	88	54	36	79	63	51
AIPTW(A)	81	59	50	74	63	57
PENCOMP(A)	82	60	49	74	63	56
IPTW(A)	100	100	100	100	100	100
g-computation(B)	144	83	54	118	103	84
AIPTW(B)	82	87	90	95	95	95
PENCOMP(B)	119	90	79	105	106	106
IPTW(C)	152	99	69	120	113	98
g-computation(A)	88	54	36	79	63	51
AIPTW(C)	81	52	37	74	60	50
PENCOMP(C)	81	54	39	74	61	51

Table A.13: 100 * Empirical bias over RMSE of IPTW (A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 200. Under the linear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(22.35, 11.17, 10.45)$, respectively. Under the nonlinear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(25.31, 12.69, 10.57)$, respectively.

	100 * Empirical Bias / RMSE IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	-2	7	8	-0	-26	-45	6	5	-1	-6	1	1	2	-25	-47	5	3	-2
g-computation(A)	1	1	-0	2	1	0	2	2	1	1	1	0	2	1	1	3	3	2
AIPTW(A)	-0	0	-1	1	1	-3	2	1	1	0	0	-0	1	0	-4	3	2	2
PENCOMP(A)	1	1	1	2	2	1	3	2	2	2	3	3	2	2	1	4	3	2
IPTW(A)	-2	7	8	-0	-26	-45	6	5	-1	-6	1	1	2	-25	-47	5	3	-2
g-computation(B)	-64	-34	-11	-75	-53	-55	-7	-5	-4	-52	-60	-50	-63	-83	-95	-8	-7	5
AIPTW(B)	-4	-1	2	-3	-5	-13	1	2	3	-8	-9	-9	-2	-22	-40	-1	-0	4
PENCOMP(B)	1	5	8	1	8	9	1	1	3	-4	-4	0	-2	-9	-15	2	-1	5
IPTW(C)	25	82	100	-86	-160	-193	-83	-179	-257	19	58	93	-57	-115	-148	-93	-208	-277
g-computation(A)	1	1	-0	2	1	0	2	2	1	1	1	0	2	1	1	3	3	2
AIPTW(C)	0	0	-1	1	0	-0	3	2	2	0	0	-0	1	1	0	3	3	2
PENCOMP(C)	1	1	0	2	0	-1	3	2	1	2	3	2	2	1	1	4	3	2

Table A.14: 100* Ratio of empirical RMSE to RMSE of IPTW(A), denoted as 100 * RMSE/RMSE(IPTW(A)), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 200.

	100 * RMSE / RMSE IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(A)	58	52	41	45	26	21	34	30	27	53	59	57	39	23	21	34	32	27
AIPTW(A)	56	52	41	46	31	68	33	30	28	53	62	60	40	30	105	33	31	27
PENCOMP(A)	58	52	41	46	29	27	33	30	27	53	61	59	39	24	23	32	31	27
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(B)	105	80	55	103	67	65	70	65	62	93	95	84	88	91	99	82	79	69
AIPTW(B)	71	65	54	64	45	70	66	62	59	82	84	85	74	72	78	82	78	75
PENCOMP(B)	63	63	55	50	42	51	43	44	46	69	76	85	51	40	52	61	62	62
IPTW(C)	97	116	118	140	175	200	139	213	277	108	121	139	110	128	153	145	236	292
g-computation(A)	58	52	41	45	26	21	34	30	27	53	59	57	39	23	21	34	32	27
AIPTW(C)	57	51	40	45	27	22	33	30	27	53	61	59	40	26	23	32	31	27
PENCOMP(C)	57	52	40	46	29	24	33	30	27	53	60	58	39	25	37	32	31	27

Table A.15: Empirical 95% non-coverage rate*100 (nominal noncoverage of 5), under (A) correctly- specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 200.

	100 * 95% Non-coverage Rate																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	6	7	8	4	18	29	6	4	6	11	10	10	11	30	47	6	5	4
g-computation(A)	6	7	8	5	6	6	5	4	4	5	7	8	5	6	6	5	4	4
AIPTW(A)	5	8	6	5	4	3	5	4	4	5	8	7	6	5	3	4	4	4
PENCOMP(A)	4	3	2	2	1	1	2	1	1	4	4	3	2	1	1	1	1	1
IPTW(A)	6	7	8	4	18	29	6	4	6	11	10	10	11	30	47	6	5	4
g-computation(B)	15	10	8	20	24	36	5	6	6	18	20	18	22	66	85	5	6	5
AIPTW(B)	5	7	4	5	6	5	5	6	5	11	10	9	9	26	35	5	5	6
PENCOMP(B)	2	1	0	0	0	0	0	0	0	2	2	2	1	1	1	0	0	0
IPTW(C)	6	16	39	15	69	93	11	42	80	10	7	13	20	73	94	14	52	88
g-computation(A)	6	7	8	5	6	6	5	4	4	5	7	8	5	6	6	5	4	4
AIPTW(C)	6	6	7	5	4	6	4	3	5	5	8	8	5	6	7	5	3	3
PENCOMP(C)	4	3	3	1	1	1	2	2	2	3	3	3	2	2	2	2	1	2

Table A.16: 100*Ratio of empirical 95% confidence interval width to that of IPTW (A), denoted as mean 95% interval width/that of IPTW(A), under (B) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 200.

	100 * mean 95% interval width / mean 95% interval width IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(A)	57	48	40	45	33	32	34	31	31	58	58	55	45	37	41	33	32	30
AIPTW(A)	57	49	47	46	40	78	33	32	31	61	63	61	47	51	130	32	31	30
PENCOMP(A)	63	58	56	53	62	91	39	40	40	66	73	73	53	59	86	38	40	38
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(B)	79	66	56	72	57	55	67	64	63	81	74	67	72	58	59	79	76	72
AIPTW(B)	72	63	60	66	61	87	64	61	61	84	84	86	78	78	87	79	76	75
PENCOMP(B)	83	92	108	74	108	191	67	73	78	104	119	172	93	135	228	89	99	98
IPTW(C)	90	80	66	113	88	75	107	106	98	105	107	102	101	75	63	106	103	91
g-computation(A)	57	48	40	45	33	32	34	31	31	58	58	55	45	37	41	33	32	30
AIPTW(C)	57	49	41	45	34	33	33	32	31	61	62	58	47	42	48	32	31	30
PENCOMP(C)	63	57	49	54	58	61	39	39	39	66	72	69	54	56	102	38	40	38

Table A.17: 100 * Empirical bias over RMSE of IPTW (A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 500. Under the linear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(22.35, 11.17, 10.45)$, respectively. Under the nonlinear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(25.31, 12.69, 10.57)$, respectively.

	100 * Empirical Bias / RMSE IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	-3	-0	2	-6	-17	-31	-5	-5	-5	-2	0	-0	-6	-16	-31	-3	-3	-1
g-computation(A)	2	1	1	1	0	1	-0	-0	0	2	1	1	1	0	0	1	1	1
AIPTW(A)	1	0	1	1	0	1	-1	-1	-0	1	1	1	1	0	1	0	0	1
PENCOMP(A)	3	1	1	1	-1	-0	-1	-1	-1	3	3	3	2	0	0	0	-0	0
IPTW(A)	-3	-0	2	-6	-17	-31	-5	-5	-5	-2	0	-0	-6	-16	-31	-3	-3	-1
g-computation(B)	-102	-51	-18	-132	-75	-67	-14	-14	-12	-77	-81	-72	-117	-104	-101	-10	-10	10
AIPTW(B)	0	0	1	-3	-5	-6	0	-0	1	-0	-2	-4	-5	-14	-25	3	3	7
PENCOMP(B)	3	5	6	1	6	7	0	0	1	-0	-4	-4	-3	-7	-9	1	-0	4
IPTW(C)	59	127	146	-155	-205	-226	-148	-289	-431	49	88	138	-111	-141	-157	-157	-326	-466
g-computation(A)	2	1	1	1	0	1	-0	-0	0	2	1	1	1	0	0	1	1	1
AIPTW(C)	1	1	1	1	0	1	-1	-0	0	1	1	1	1	-0	0	0	0	1
PENCOMP(C)	3	1	1	1	-0	1	-1	-1	-1	3	3	3	2	0	1	0	-0	0

Table A.18: 100* Ratio of empirical RMSE to RMSE of IPTW(A), denoted as 100 * RMSE/RMSE(IPTW(A)), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 500.

	100 * RMSE / RMSE IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(A)	61	48	36	48	21	16	36	32	32	55	50	49	43	17	14	36	33	32
AIPTW(A)	58	47	36	46	23	23	34	30	30	56	55	54	44	21	25	32	31	30
PENCOMP(A)	58	47	36	46	23	19	33	30	30	54	52	51	43	19	15	32	31	30
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(B)	132	83	50	152	83	72	70	63	61	109	103	92	136	109	104	79	75	72
AIPTW(B)	72	59	49	65	42	40	62	57	58	83	80	86	79	76	78	75	73	76
PENCOMP(B)	61	55	47	48	33	39	41	41	44	67	65	74	54	33	40	55	56	59
IPTW(C)	111	151	158	194	214	230	183	310	444	120	140	174	149	147	159	188	342	475
g-computation(A)	61	48	36	48	21	16	36	32	32	55	50	49	43	17	14	36	33	32
AIPTW(C)	58	47	36	46	21	16	34	30	30	55	54	52	44	20	16	32	30	29
PENCOMP(C)	58	46	35	46	22	17	33	30	30	54	52	51	43	18	15	32	30	29

Table A.19: Empirical 95% non-coverage rate*100 (nominal noncoverage of 5), under (A) correctly- specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 500.

	100 * 95% Non-coverage Rate																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	5	6	6	6	18	29	5	6	6	8	8	8	9	28	44	5	6	5
g-computation(A)	6	7	8	6	6	6	6	5	6	6	7	7	6	6	6	5	5	6
AIPTW(A)	6	6	8	6	6	4	6	6	6	6	7	7	5	6	4	6	6	6
PENCOMP(A)	4	4	5	5	4	3	5	4	5	5	5	5	4	4	4	5	5	5
IPTW(A)	5	6	6	6	18	29	5	6	6	8	8	8	9	28	44	5	6	5
g-computation(B)	26	15	8	42	56	70	5	6	4	22	31	29	41	91	99	4	6	4
AIPTW(B)	6	6	7	6	8	9	4	5	5	7	8	8	8	26	41	4	5	6
PENCOMP(B)	2	2	2	2	2	2	0	1	1	4	3	2	2	3	3	1	1	1
IPTW(C)	9	40	84	30	91	100	29	80	99	6	11	28	29	87	98	34	89	100
g-computation(A)	6	7	8	6	6	6	6	5	6	6	7	7	6	6	6	5	5	6
AIPTW(C)	6	6	9	5	6	6	6	5	6	6	7	8	5	6	6	6	6	6
PENCOMP(C)	5	4	5	4	4	3	5	4	4	5	5	5	4	4	4	6	5	5

Table A.20: 100*Ratio of empirical 95% confidence interval width to that of IPTW (A), denoted as mean 95% interval width/that of IPTW(A), under (B) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 1000 simulations with sample size of 500.

	100 * mean 95% interval width / mean 95% interval width IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(A)	60	49	38	48	29	26	36	32	32	58	56	54	45	28	28	36	33	31
AIPTW(A)	59	49	39	47	32	41	34	31	30	59	61	59	46	34	59	33	31	29
PENCOMP(A)	61	53	44	49	35	40	34	32	32	61	63	62	47	33	36	33	33	31
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(B)	83	67	53	75	49	44	70	63	63	81	71	65	73	47	44	83	77	74
AIPTW(B)	73	62	53	66	51	58	64	59	59	84	84	87	79	76	80	80	75	76
PENCOMP(B)	75	74	70	60	57	86	57	57	59	84	89	112	71	71	121	77	79	78
IPTW(C)	92	80	62	115	79	63	108	107	100	109	109	106	100	63	48	108	102	91
g-computation(A)	60	49	38	48	29	26	36	32	32	58	56	54	45	28	28	36	33	31
AIPTW(C)	59	49	38	47	29	26	34	31	30	59	60	58	47	32	32	33	31	29
PENCOMP(C)	61	53	43	49	34	32	35	32	32	61	63	62	47	32	33	33	32	31

Table A.21: 100 * Empirical bias over RMSE of IPTW (A), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 1000. Under the linear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(22.35, 11.17, 10.45)$, respectively. Under the nonlinear outcome model, $(\Delta_{11}, \Delta_{10}, \Delta_{01})$ were $(25.31, 12.69, 10.57)$, respectively.

	100 * Empirical Bias / RMSE IPTW (A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	-3	-0	6	5	-14	-24	-1	2	1	-3	1	2	1	-19	-28	1	3	1
g-computation(A)	6	6	3	4	2	1	-0	-1	-1	3	3	3	3	2	1	2	1	1
AIPTW(A)	5	5	3	2	2	1	-1	-2	-1	3	4	4	2	1	1	1	-1	0
PENCOMP(A)	5	5	3	2	1	1	-2	-3	-2	4	5	4	2	1	1	0	-1	-1
IPTW(A)	-3	-0	6	5	-14	-24	-1	2	1	-3	1	2	1	-19	-28	1	3	1
g-computation(B)	-146	-76	-25	-182	-110	-87	-22	-18	-14	-116	-120	-107	-156	-157	-131	-17	-13	15
AIPTW(B)	1	2	7	1	-2	-5	-3	-3	2	-1	1	0	-2	-17	-24	-2	-2	2
PENCOMP(B)	4	4	8	2	10	8	-3	-4	-1	-1	-2	-5	-3	-5	-7	-3	-5	-0
IPTW(C)	84	192	205	-203	-306	-296	-207	-405	-585	71	134	209	-142	-210	-201	-225	-467	-670
g-computation(A)	6	6	3	4	2	1	-0	-1	-1	3	3	3	3	2	1	2	1	1
AIPTW(C)	5	5	3	3	1	1	-1	-2	-2	3	4	3	2	1	1	1	-1	-0
PENCOMP(C)	5	5	3	2	1	1	-2	-3	-2	4	5	5	2	1	1	0	-1	-1

Table A.22: 100* Ratio of empirical RMSE to RMSE of IPTW(A), denoted as 100 * RMSE/RMSE(IPTW(A)), under (A) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 1000.

	100 * RMSE / RMSE IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(A)	67	56	40	49	24	16	40	35	33	61	58	58	44	20	13	41	38	36
AIPTW(A)	61	52	38	48	26	22	34	31	30	59	60	60	44	23	21	32	31	30
PENCOMP(A)	60	52	37	49	24	18	34	30	30	57	58	57	44	21	14	32	31	30
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(B)	171	106	56	197	117	91	78	65	62	143	140	127	170	161	132	92	80	79
AIPTW(B)	72	63	47	68	41	37	66	58	59	83	86	104	77	75	78	83	78	83
PENCOMP(B)	61	60	48	51	34	36	42	39	43	70	71	83	53	34	38	60	58	61
IPTW(C)	124	209	213	234	313	298	236	421	595	128	174	237	172	215	202	253	480	677
g-computation(A)	67	56	40	49	24	16	40	35	33	61	58	58	44	20	13	41	38	36
AIPTW(C)	61	52	37	49	24	16	34	32	30	59	60	59	44	22	15	33	32	31
PENCOMP(C)	60	52	37	49	24	16	34	30	29	57	57	57	44	20	14	32	31	30

Table A.23: Empirical 95% non-coverage rate*100 (nominal noncoverage of 5), under (A) correctly- specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 1000.

	100 * 95% Non-coverage Rate																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	7	5	9	4	12	26	5	5	6	7	8	8	7	21	36	3	4	5
g-computation(A)	8	9	9	5	5	8	6	6	7	8	8	9	7	6	6	7	5	6
AIPTW(A)	8	8	9	5	8	8	5	5	6	7	8	9	7	8	6	4	4	4
PENCOMP(A)	7	7	7	7	4	4	4	4	4	7	7	7	7	6	6	4	3	4
IPTW(A)	7	5	9	4	12	26	5	5	6	7	8	8	7	21	36	3	4	5
g-computation(B)	41	23	11	69	81	94	6	6	5	33	50	48	64	100	100	5	5	6
AIPTW(B)	7	6	8	6	7	9	5	6	6	6	8	10	7	20	37	5	5	6
PENCOMP(B)	4	3	5	4	3	4	1	1	1	5	5	5	4	4	6	2	2	1
IPTW(C)	17	74	98	44	98	100	46	97	100	8	23	58	39	96	100	52	98	100
g-computation(A)	8	9	9	5	5	8	6	6	7	8	8	9	7	6	6	7	5	6
AIPTW(C)	8	8	9	6	7	8	5	5	5	8	8	9	7	8	8	5	4	5
PENCOMP(C)	7	6	6	6	5	6	4	4	4	7	6	6	7	6	5	4	3	4

Table A.24: 100*Ratio of empirical 95% confidence interval width to that of IPTW (A), denoted as mean 95% interval width/that of IPTW(A), under (B) correctly-specified propensity and prediction models; (B) a correctly-specified propensity model only; (C) a correctly-specified prediction model only, based on 500 simulations with sample size of 1000.

	100 * mean 95% interval width / mean 95% interval width IPTW(A)																	
	Linear Outcome									Nonlinear Outcome								
	Δ_{11}			Δ_{10}			Δ_{01}			Δ_{11}			Δ_{10}			Δ_{01}		
Method	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High	Low	Mod	High
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(A)	63	51	39	51	27	22	38	34	33	58	57	55	46	26	23	39	37	34
AIPTW(A)	60	49	38	47	29	30	33	30	30	59	61	60	46	29	36	33	31	29
PENCOMP(A)	61	52	41	48	29	29	34	31	30	59	62	61	46	28	26	33	31	30
IPTW(A)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
g-computation(B)	86	68	52	77	45	38	71	64	63	82	71	66	72	42	36	85	79	76
AIPTW(B)	74	61	51	66	48	49	64	58	59	84	87	93	79	75	79	80	76	78
PENCOMP(B)	71	68	61	54	46	62	52	51	53	77	83	102	61	54	88	72	74	72
IPTW(C)	92	80	60	115	74	55	108	106	100	109	110	108	100	59	40	108	102	92
g-computation(A)	63	51	39	51	27	22	38	34	33	58	57	55	46	26	23	39	37	34
AIPTW(C)	60	49	37	47	27	22	33	30	30	58	60	58	46	29	26	33	31	29
PENCOMP(C)	61	52	40	48	29	25	34	31	30	59	62	60	46	27	25	33	31	30

A.4 Supplemental Table from Application

Table A.25: The number of subjects with observed treatment regimen $(1,1)$, $(1,0)$, $(0,1)$, and $(0,0)$, denoted as no11, no10, no01, and no00, respectively in each three-visit window, as well as the number of subjects kept in the estimation of Δ_{11}, Δ_{10}, and Δ_{00}, after trimming, denoted as no Δ_{11}, no Δ_{10}, and no Δ_{01}, respectively. The total is the total number of subjects with complete data on blood count measures considered in the models in each window.

	sample size observed				sample size after trimming			
Window	no11	no10	no01	no00	no Δ_{11}	no Δ_{10}	no Δ_{01}	Total
Window1	88	11	82	638	772	731	770	819
Window2	138	16	88	620	770	794	785	862
Window3	178	23	76	550	635	700	721	827
Window4	160	42	134	352	612	459	603	688
Window5	265	13	114	292	509	458	518	684
Window6	390	26	59	348	773	749	756	823
Window7	401	13	41	322	694	648	686	777
Window8	397	12	43	299	717	655	564	751
Window9	389	14	30	281	541	518	544	714
Window10	373	14	48	245	516	462	476	680
Window11	356	21	37	225	590	545	562	639
Window12	310	36	16	217	552	514	532	579
Window13	254	45	24	220	504	395	437	543
Window14	216	31	30	203	420	410	353	480
Window15	197	14	39	182	374	365	373	432

Table A.26: Summary of the stabilized weights.

	Stabilized Weights	
Window	Mean(SD)	Minimum/Maximum
Window1	1.091 (1.97)	0.1103 / 40.3
Window2	1.065 (3.20)	0.1026 / 91.5
Window3	6.160 (146.78)	0.2010 / 4220.5
Window4	4.662 (83.11)	0.1391 / 2163.2
Window5	0.966 (1.11)	0.3274 / 15.2
Window6	2.378 (37.86)	0.4039 / 1083.5
Window7	3.052 (59.23)	0.1692 / 1651.0
Window8	23.893 (618.50)	0.1102 / 16949.0
Window9	4.085 (63.72)	0.2095 / 1541.6
Window10	6.937 (106.37)	0.1468 / 2307.3
Window11	1.586 (11.11)	0.2741 / 250.8
Window12	1.731 (12.57)	0.2944 / 266.1
Window13	1.336 (7.32)	0.1705 / 164.7
Window14	1.033 (1.67)	0.1935 / 17.6
Window15	1.046 (2.05)	0.2134 / 32.0

Table A.27: Summary of overlap proportions at both time points.

	First Time Point		Second Time Point			
	$\pi_1^{0.95}$	$\pi_0^{0.95}$	$\pi_{11}^{0.95}$	$\pi_{10}^{0.95}$	$\pi_{01}^{0.95}$	$\pi_{00}^{0.95}$
Window1	81	40	73	86	89	51
Window2	60	38	63	96	83	45
Window3	50	37	41	96	86	38
Window4	36	28	34	42	69	39
Window5	45	78	45	79	92	39
Window6	40	31	40	92	84	35
Window7	36	19	34	82	87	17
Window8	21	14	22	85	61	18
Window9	22	12	23	78	70	14
Window10	8	6	12	74	31	15
Window11	27	19	33	79	62	21
Window12	38	25	40	67	51	28
Window13	22	52	34	57	73	49
Window14	42	51	38	85	52	57
Window15	32	51	34	96	82	41

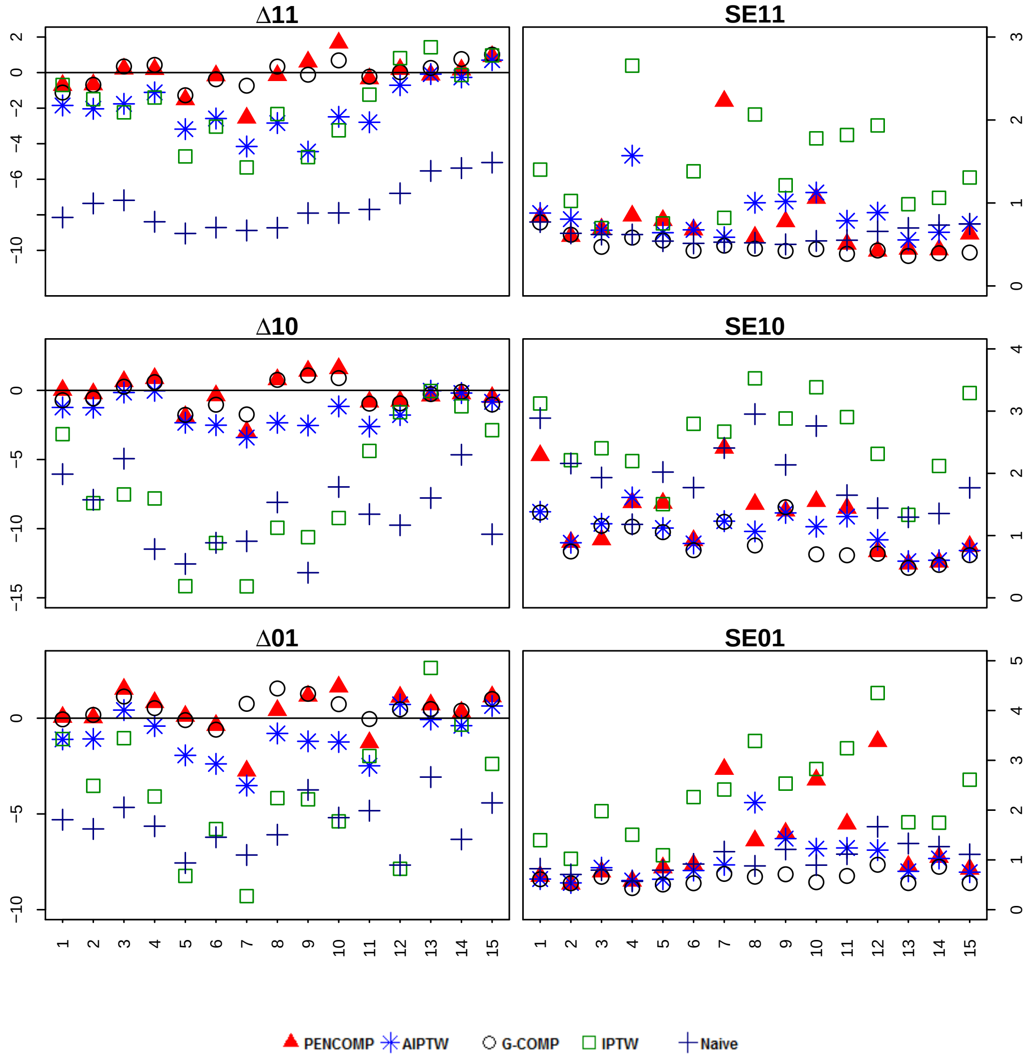

Figure A.1: For each of the three-visit windows $1, \cdots, 15$, the estimates and standard errors (SE) of the treatment effects Δ_{11}, Δ_{10}, and Δ_{01} of the four methods: PENCOMP, AIPTW, IPTW, and Naive. Here 1st% and 99th% weight truncation was done for IPTW and AIPTW. PENCOMP estimates were computed on the overlapping regions, as described in Section 2.4. Since the propensity score distributions were very skewed for some windows, restricting to the quantiles $c(\alpha, 1-\alpha)$ (for example $\alpha = 0.025$) of the propensity score distributions can significantly reduce the variances without changing the estimates much (results not shown here). Note the estimands are different.

APPENDIX B

Addressing Disparities in the Assignment Propensity Distributions for Treatment Comparisons from Observational Studies

B.1 Supplementary Tables from the Simulation Study

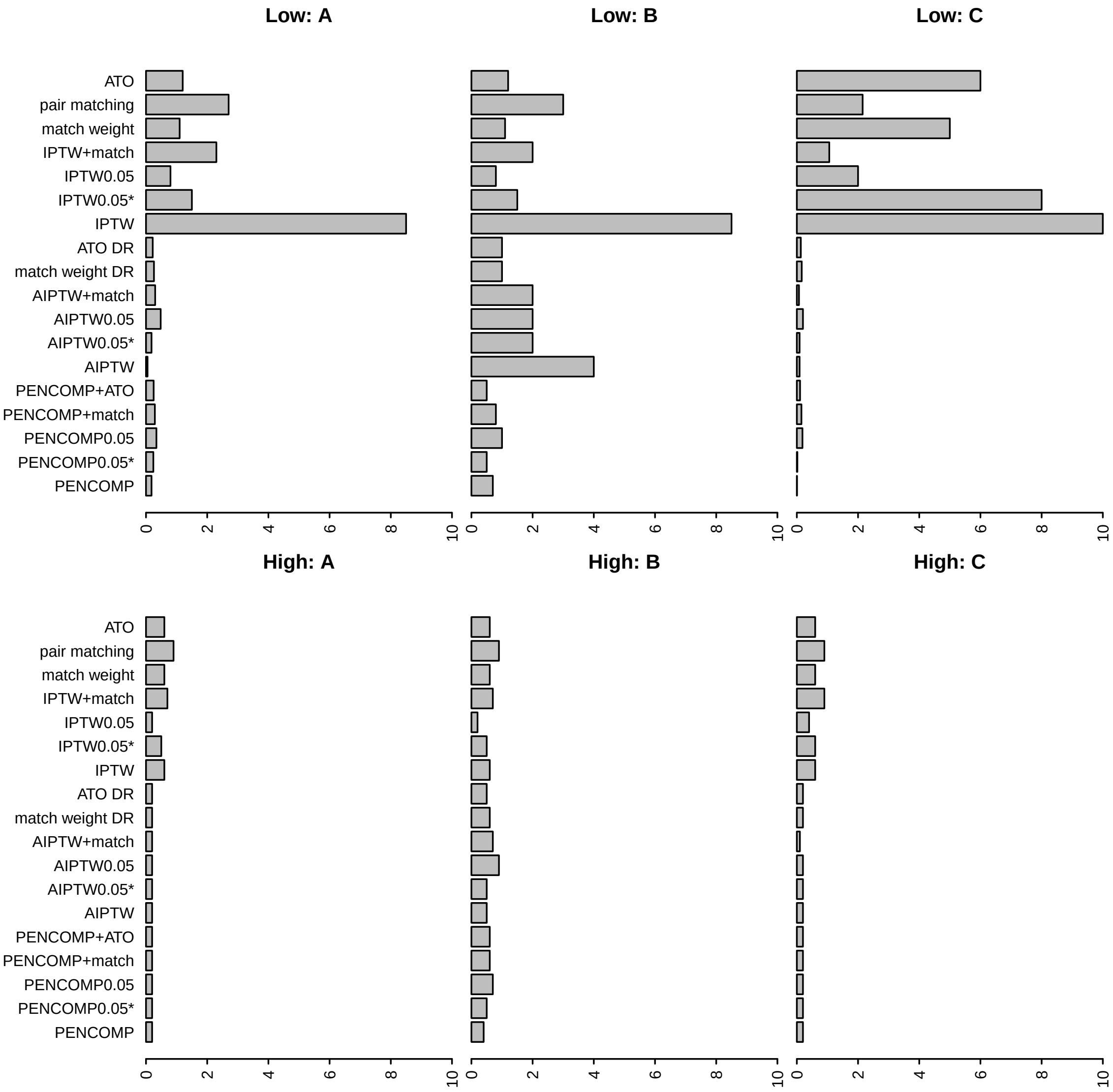

Figure B.1: Parallel surface and Misaligned: absolute bias in percentage, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

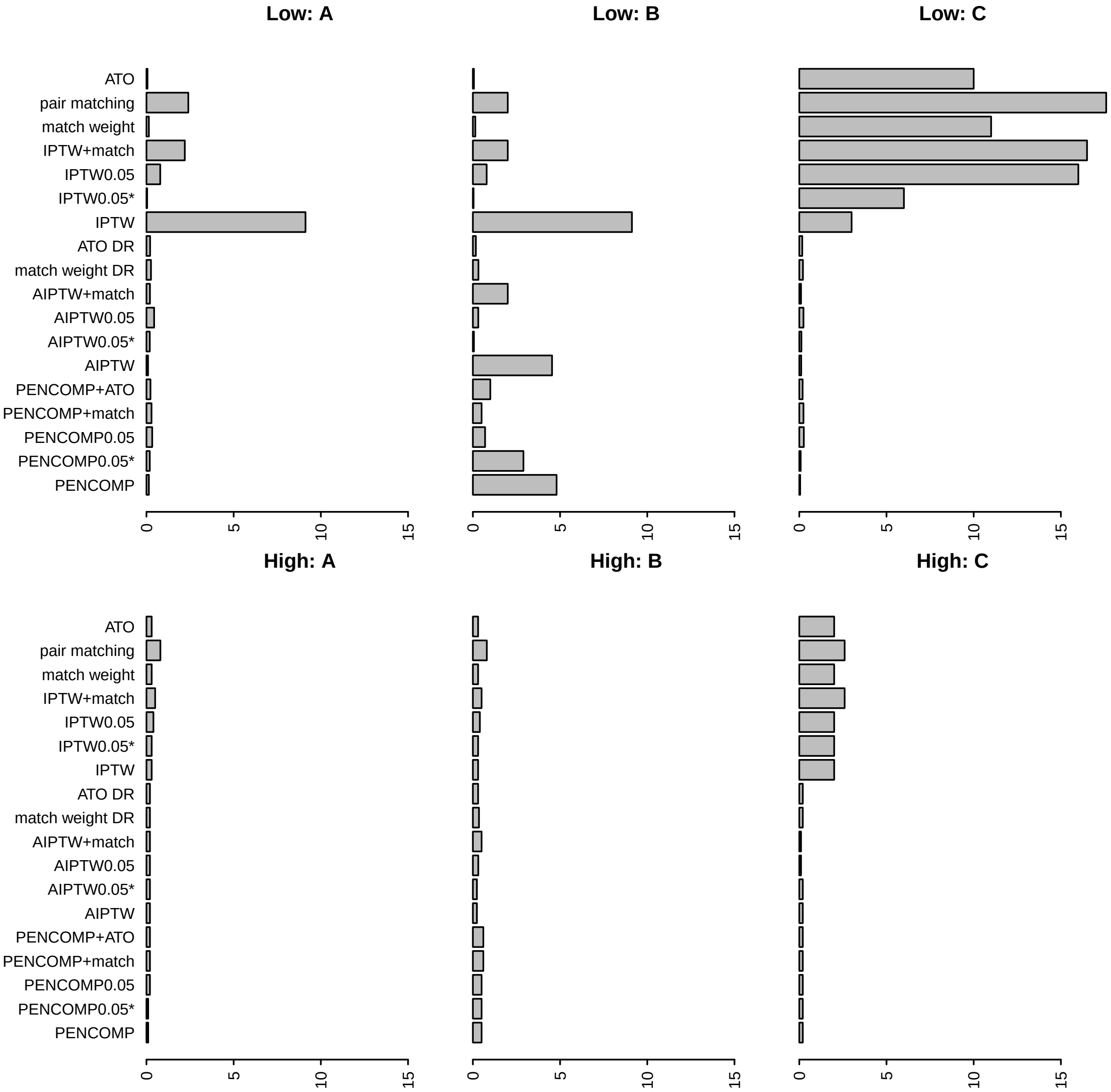

Figure B.2: Parallel surface and Aligned: absolute bias in percentage, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

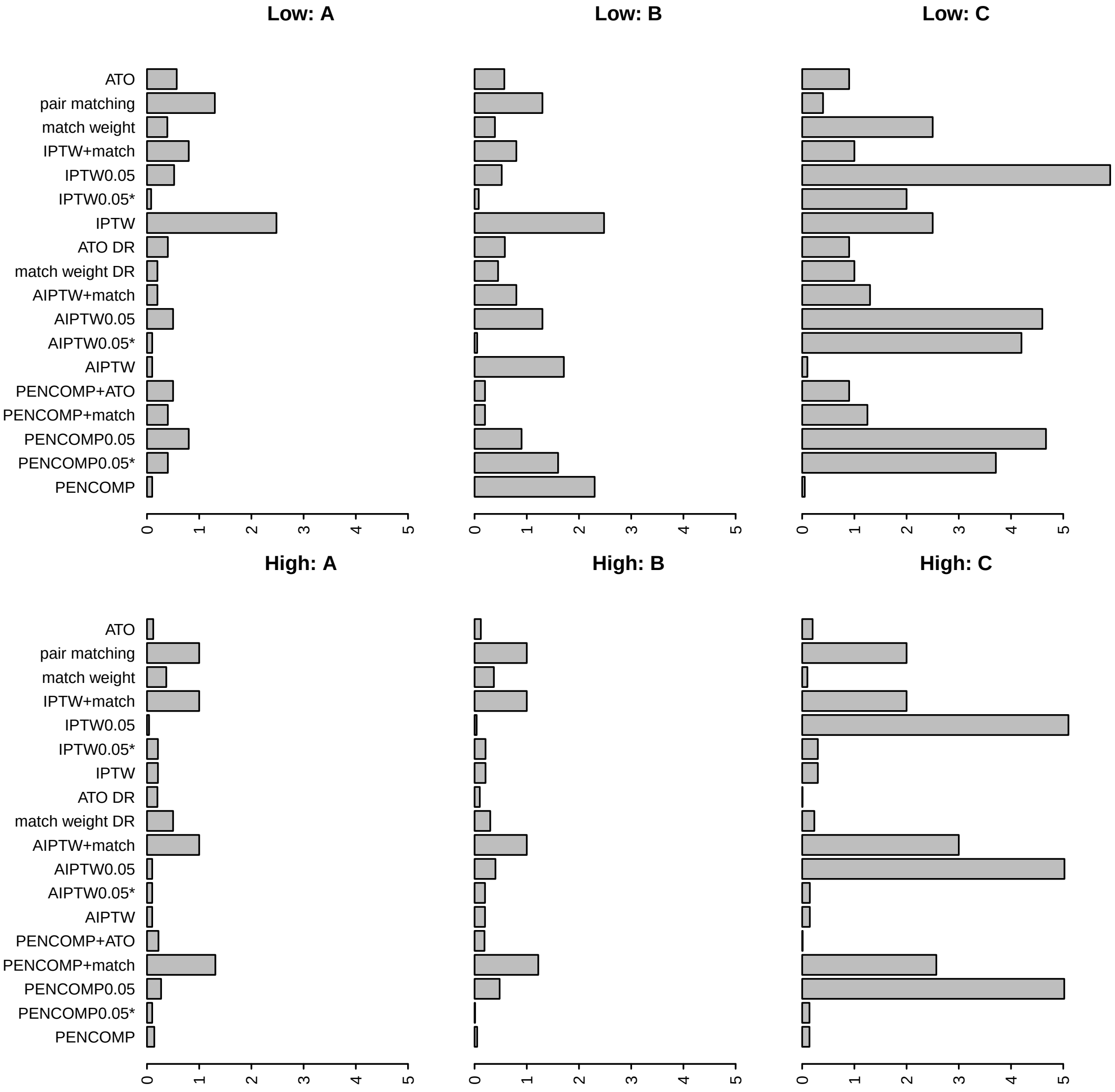

Figure B.3: Nonparallel surface and Misaligned: absolute bias in percentage, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

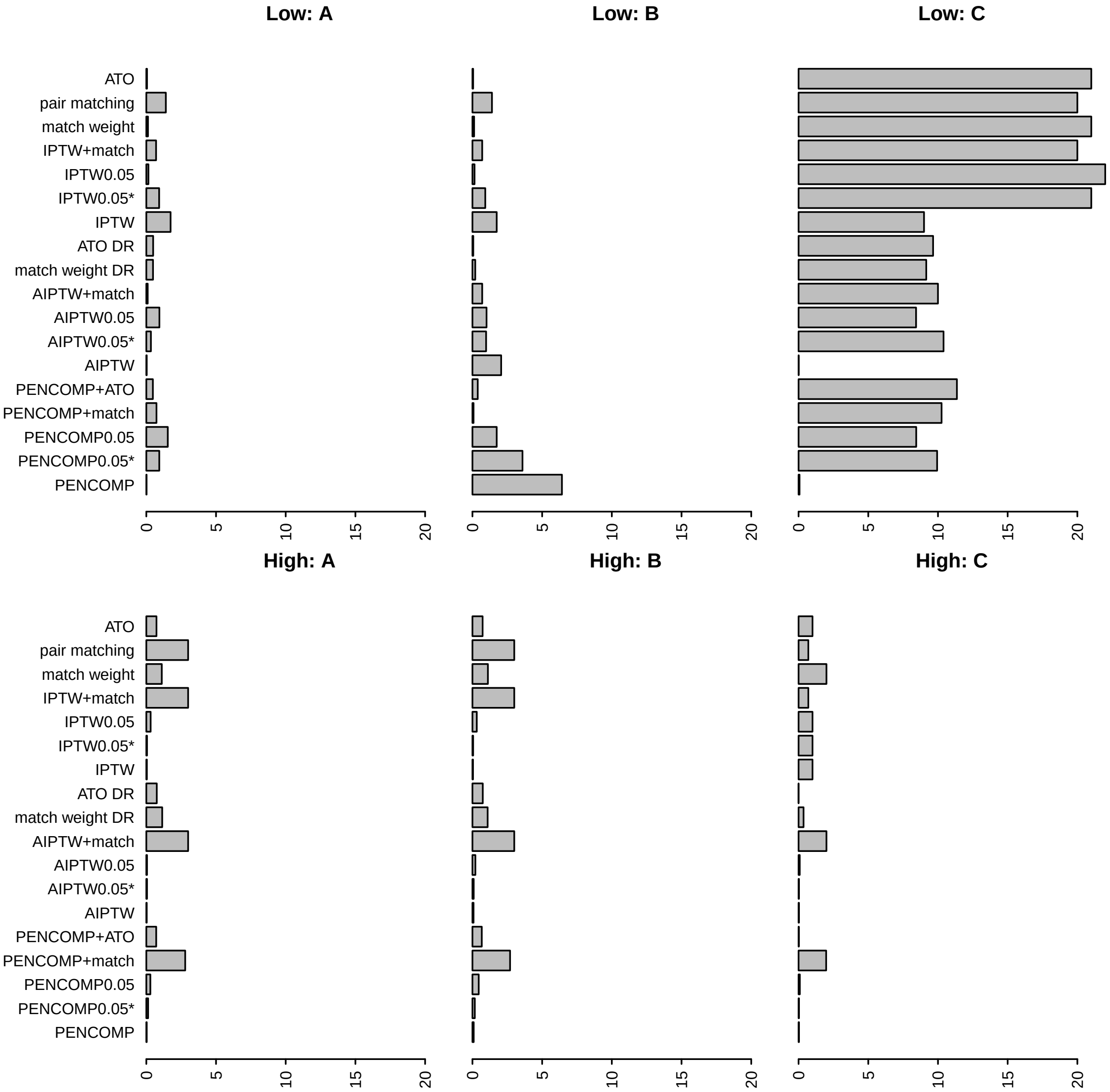

Figure B.4: Nonparallel surface and Aligned: absolute bias in percentage, sample size of 200. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

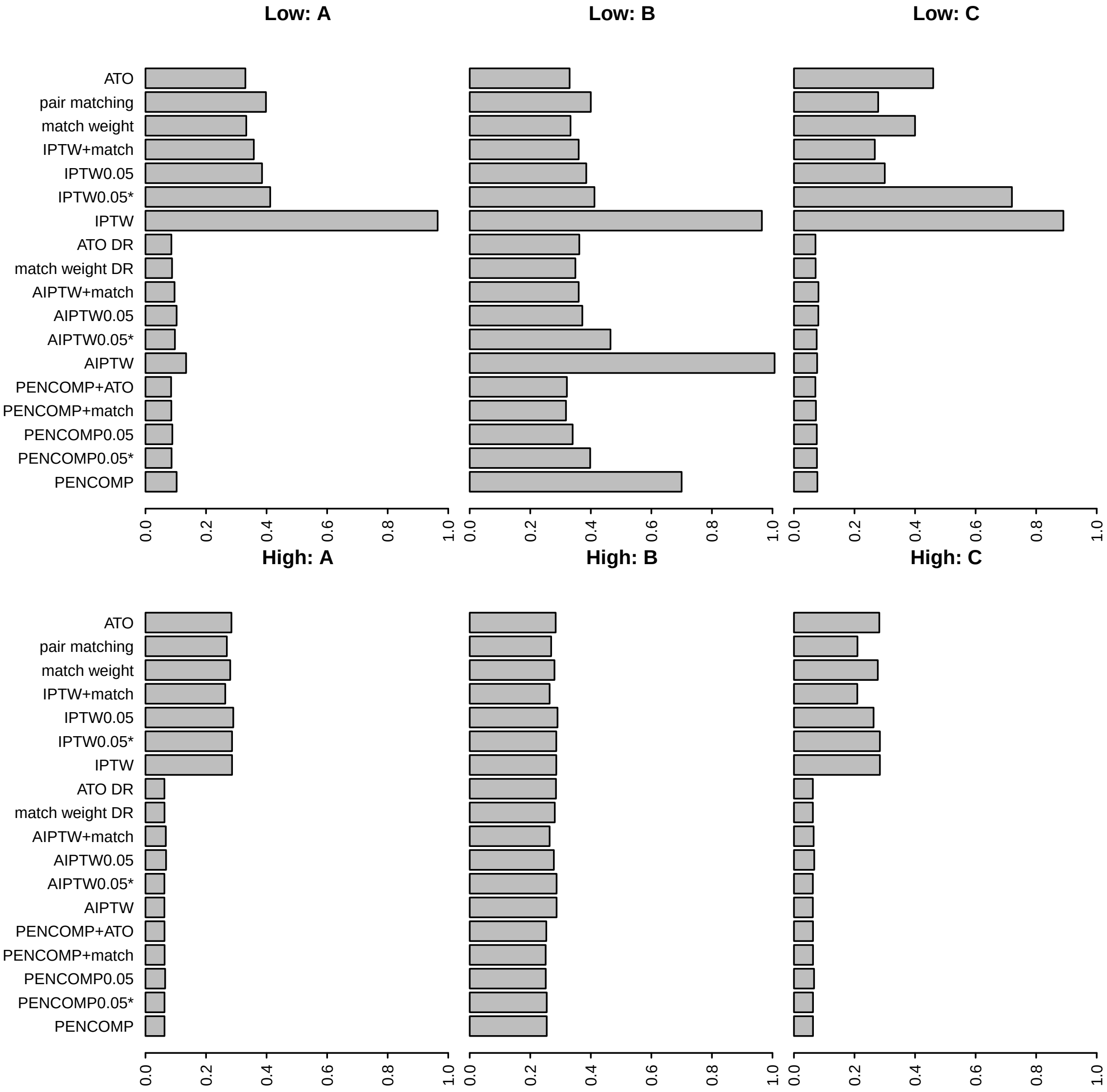

Figure B.5: Parallel surface and Misaligned: Empirical RMSE, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

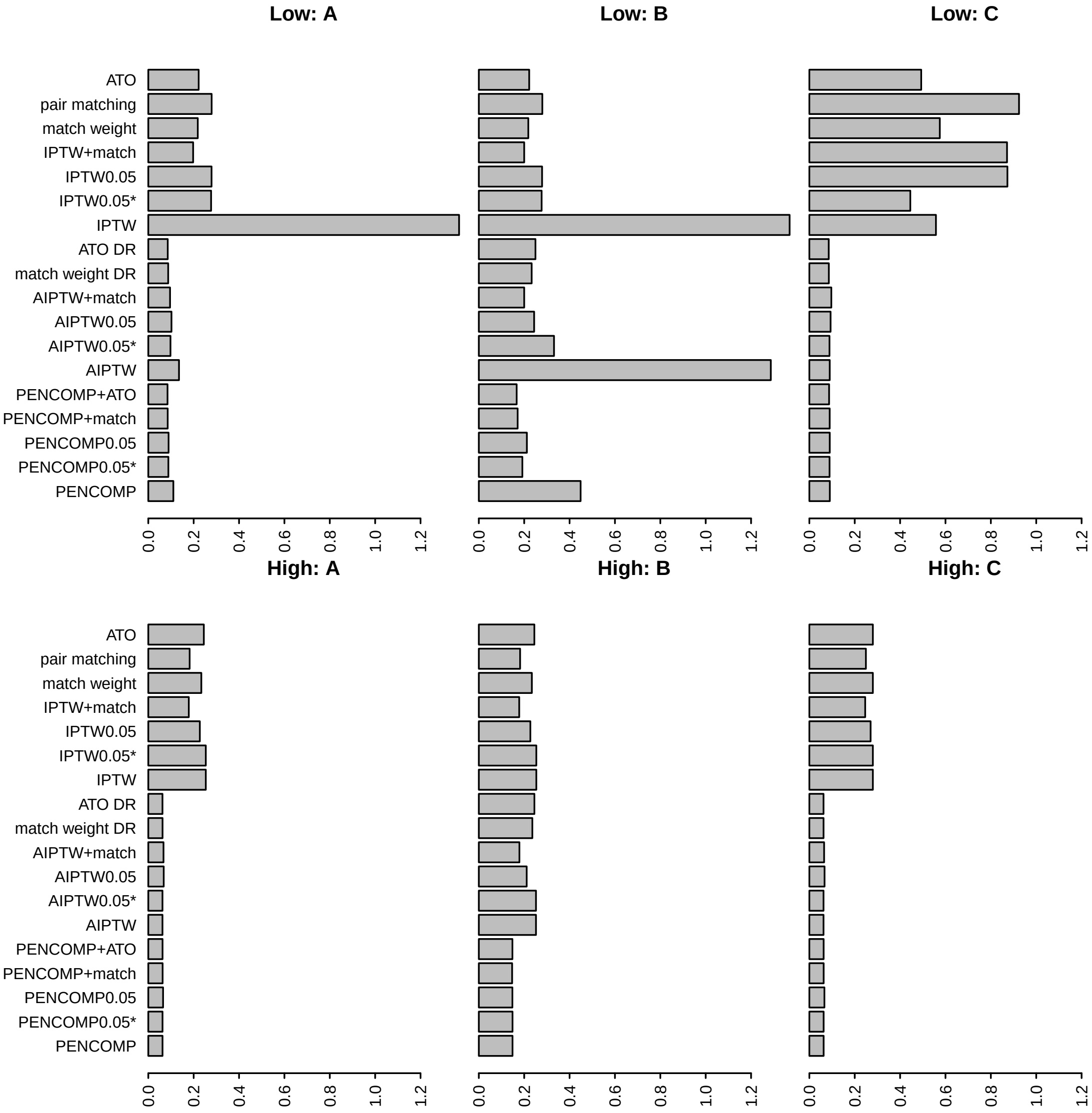

Figure B.6: Parallel surface and Aligned: Empirical RMSE, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

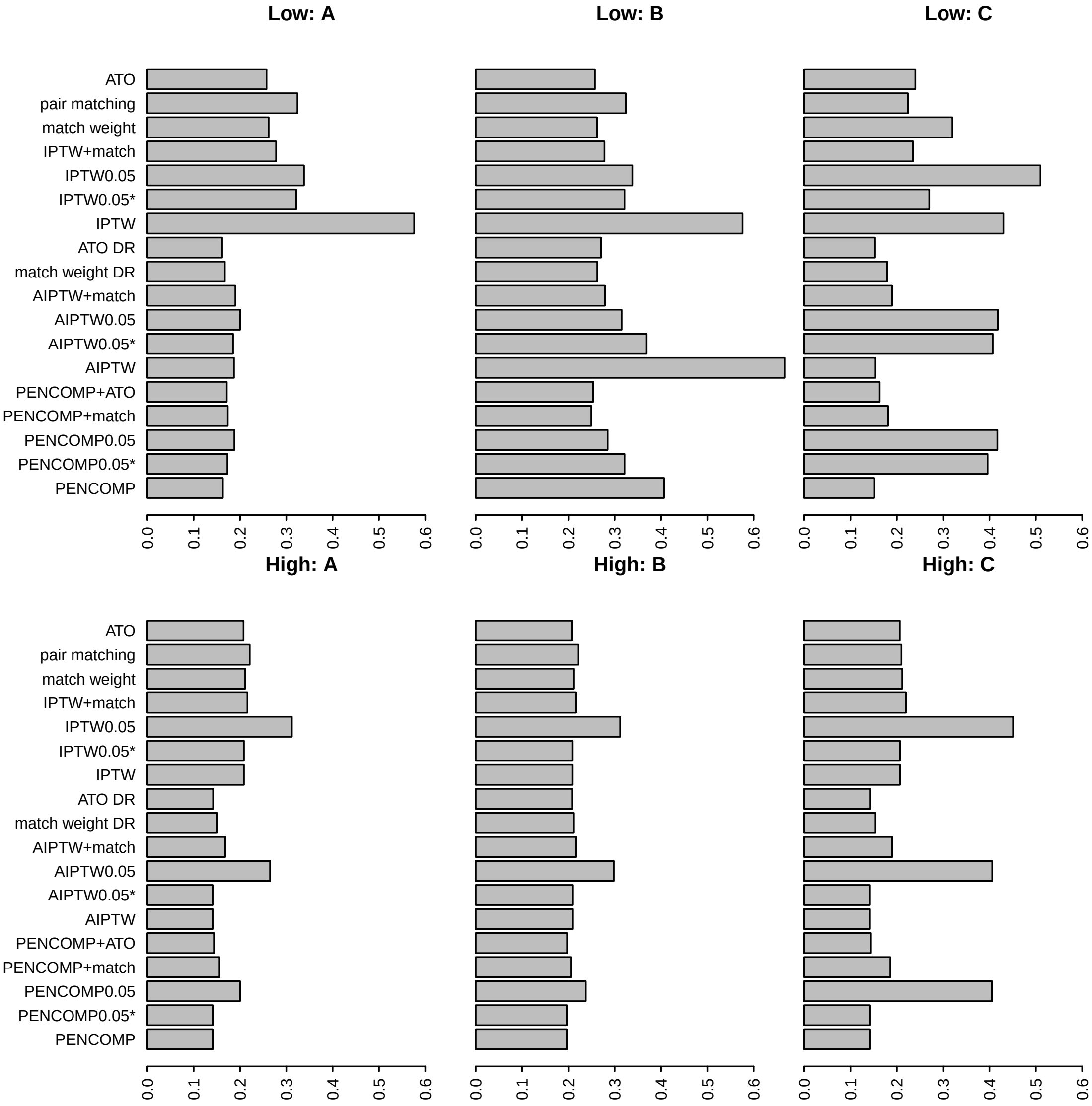

Figure B.7: Nonparallel surface and Misaligned: Empirical RMSE, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

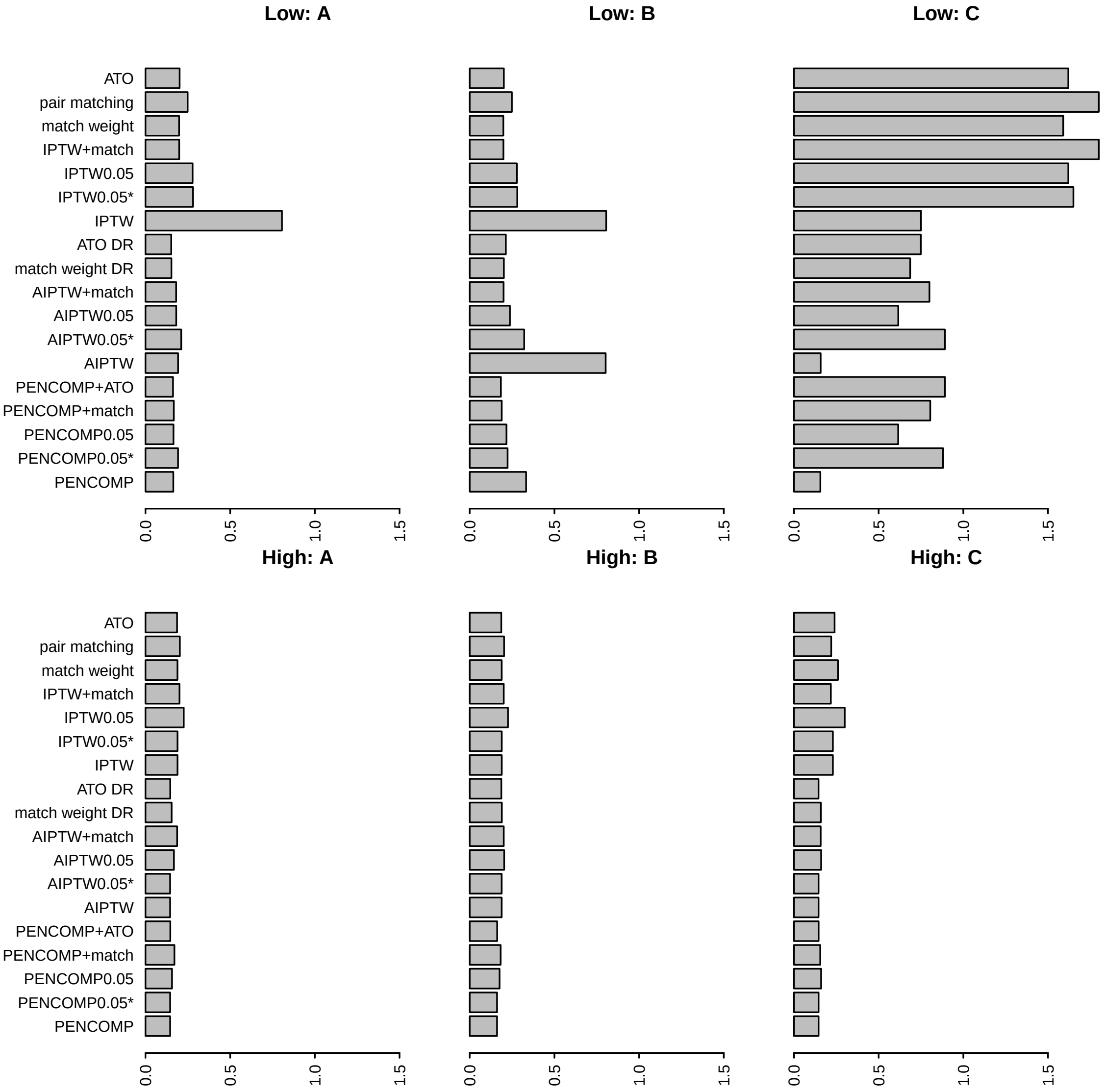

Figure B.8: Nonparallel surface and Aligned: Empirical RMSE, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

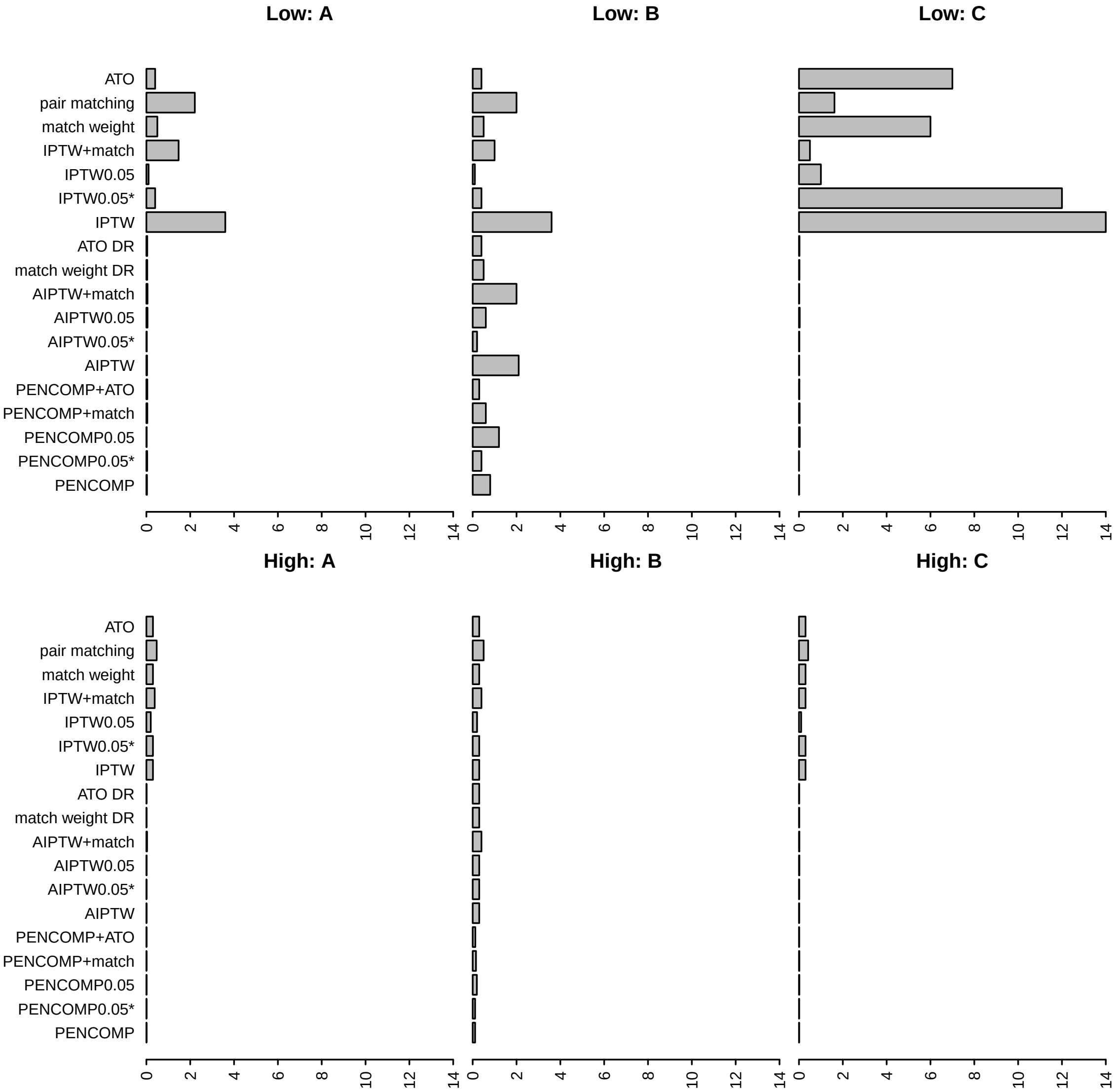

Figure B.9: Parallel surface and Misaligned: absolute bias in percentage, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

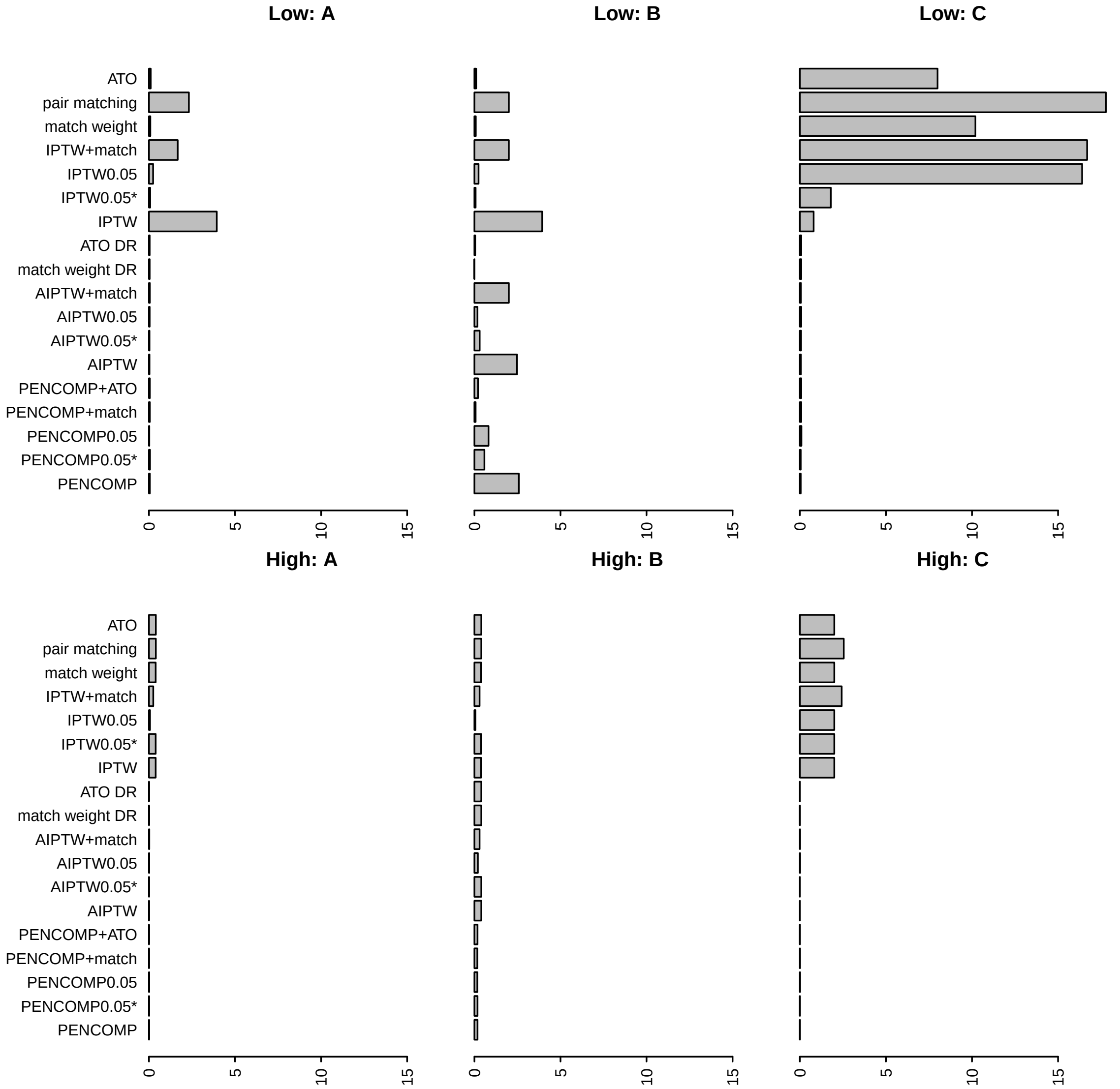

Figure B.10: Parallel surface and Aligned: absolute bias in percentage, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

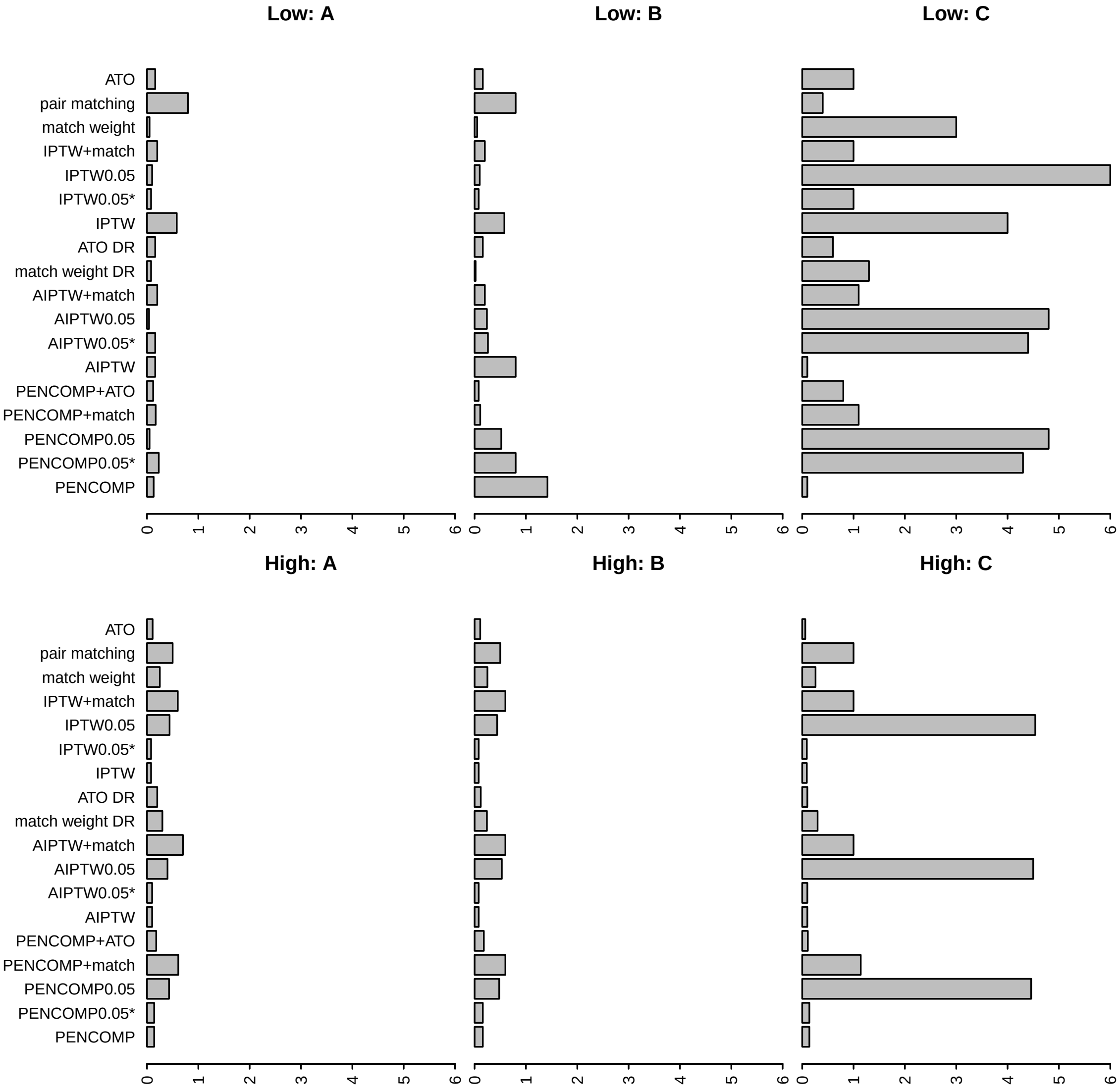

Figure B.11: Nonparallel surface and Misaligned: absolute bias in percentage, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

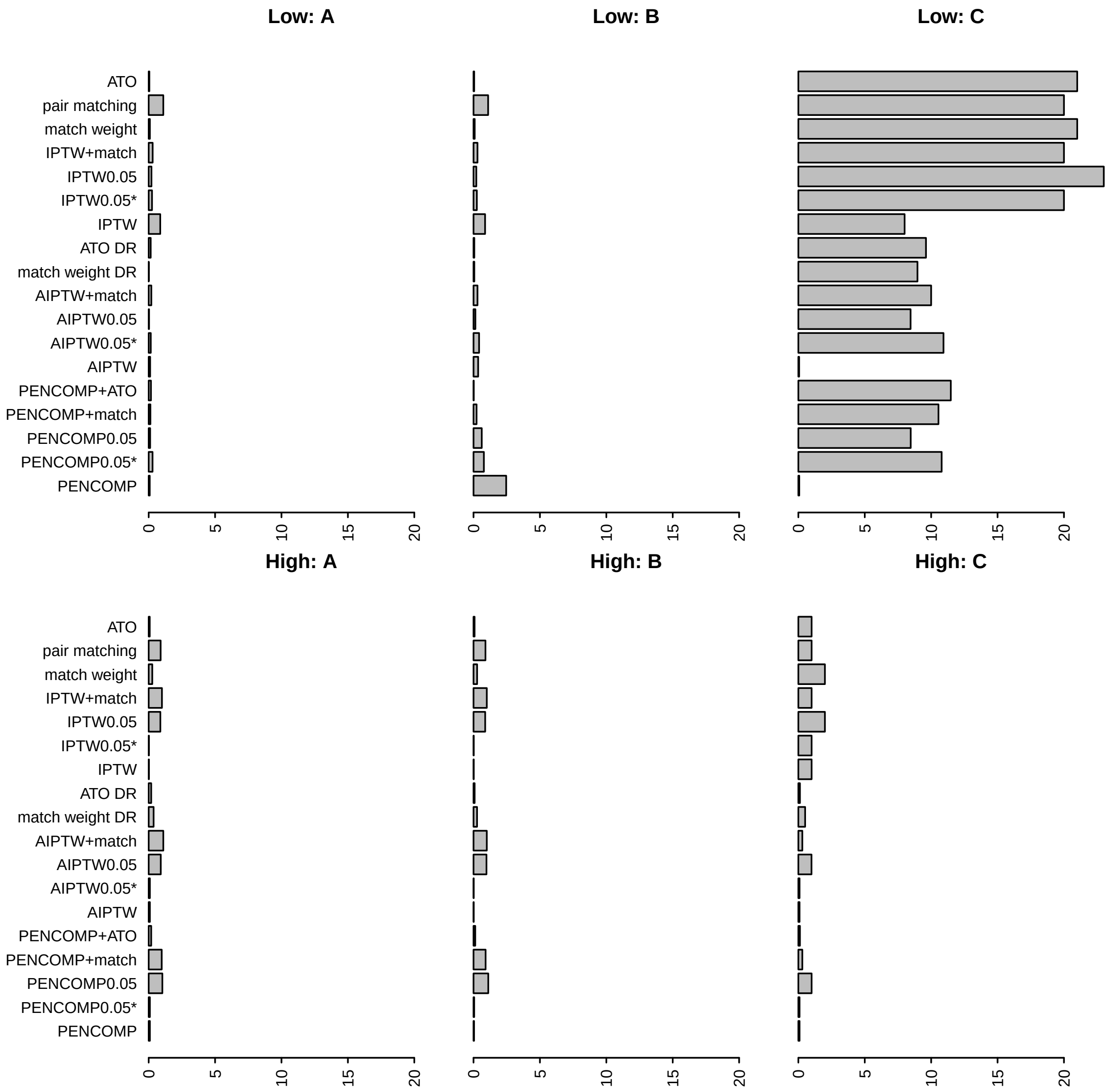

Figure B.12: Nonparallel surface and Aligned: absolute bias in percentage, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

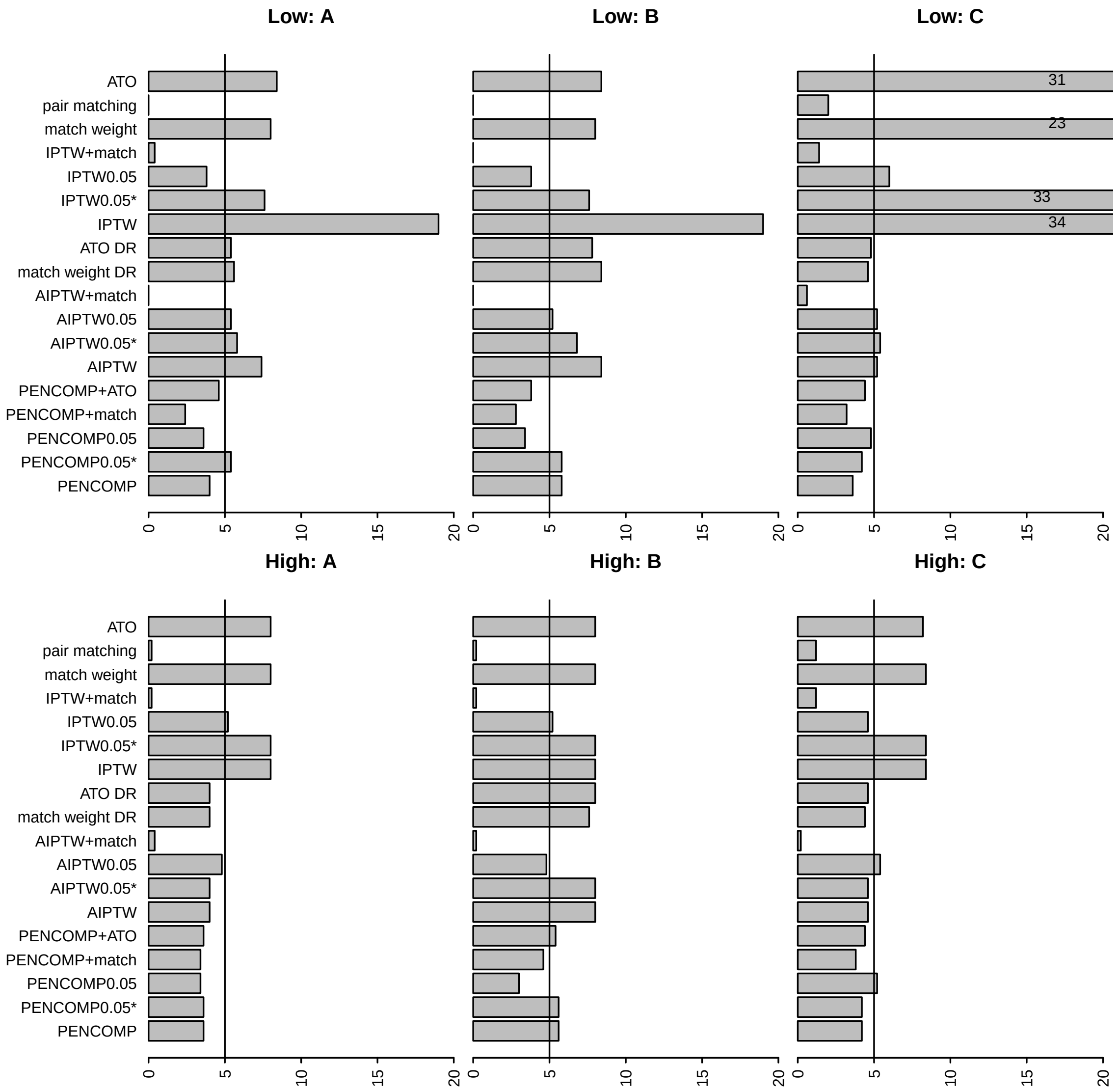

Figure B.13: Parallel surface and Misaligned: 100 * 95% non coverage rate, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

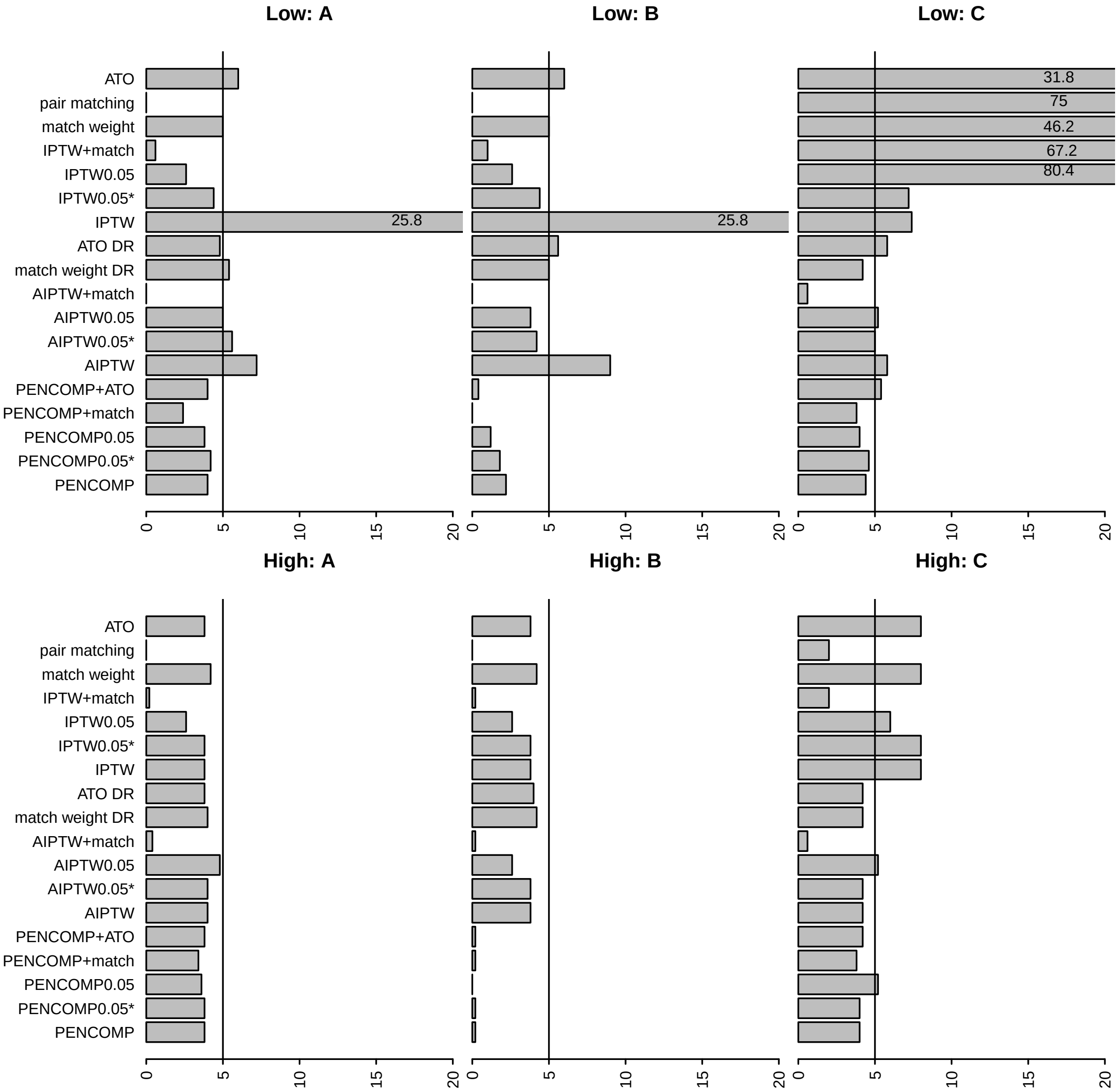

Figure B.14: Parallel surface and Aligned: 100 * 95% non coverage rate, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

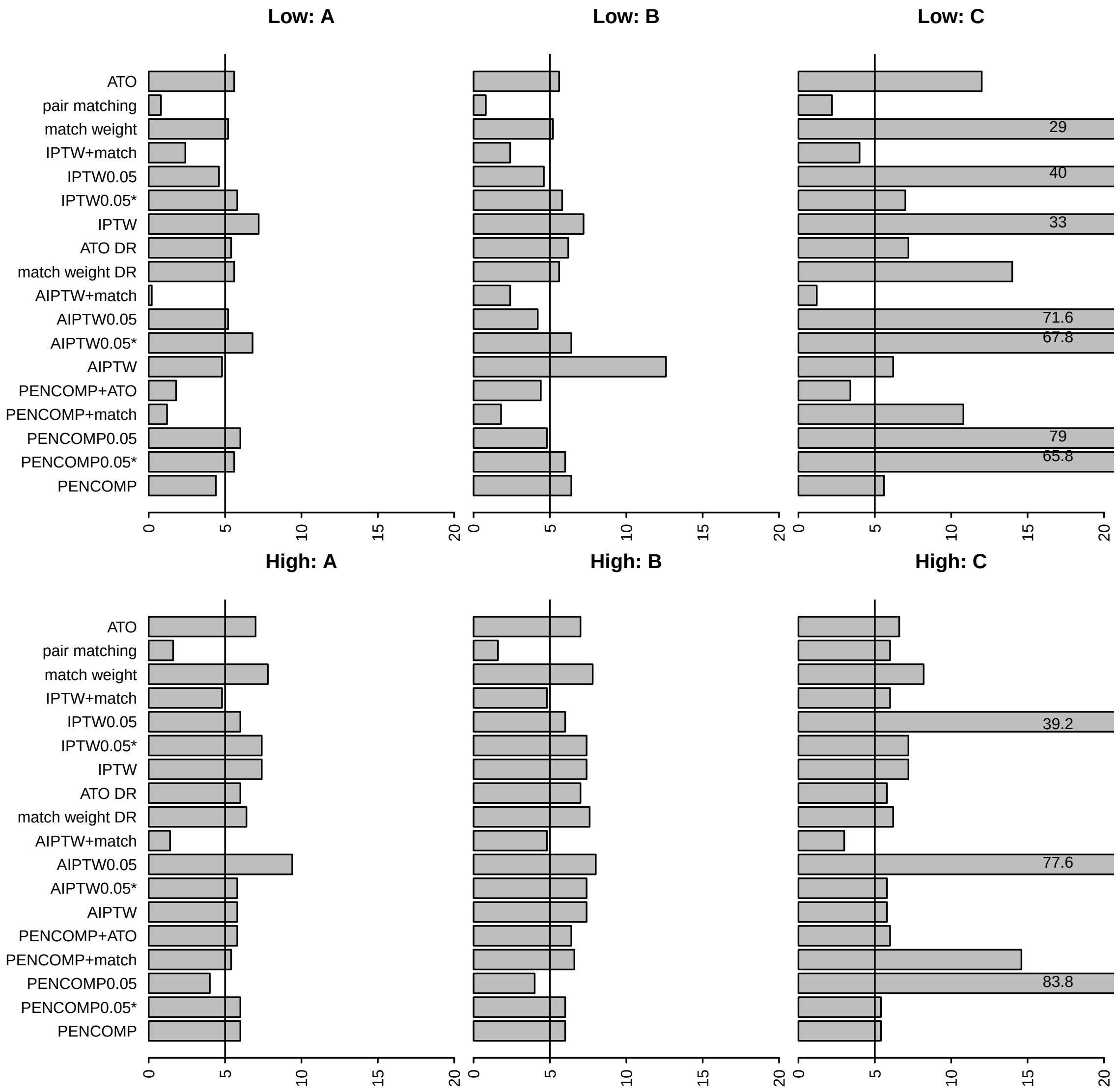

Figure B.15: Nonparallel surface and Misaligned: 100 * 95% non coverage rate, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

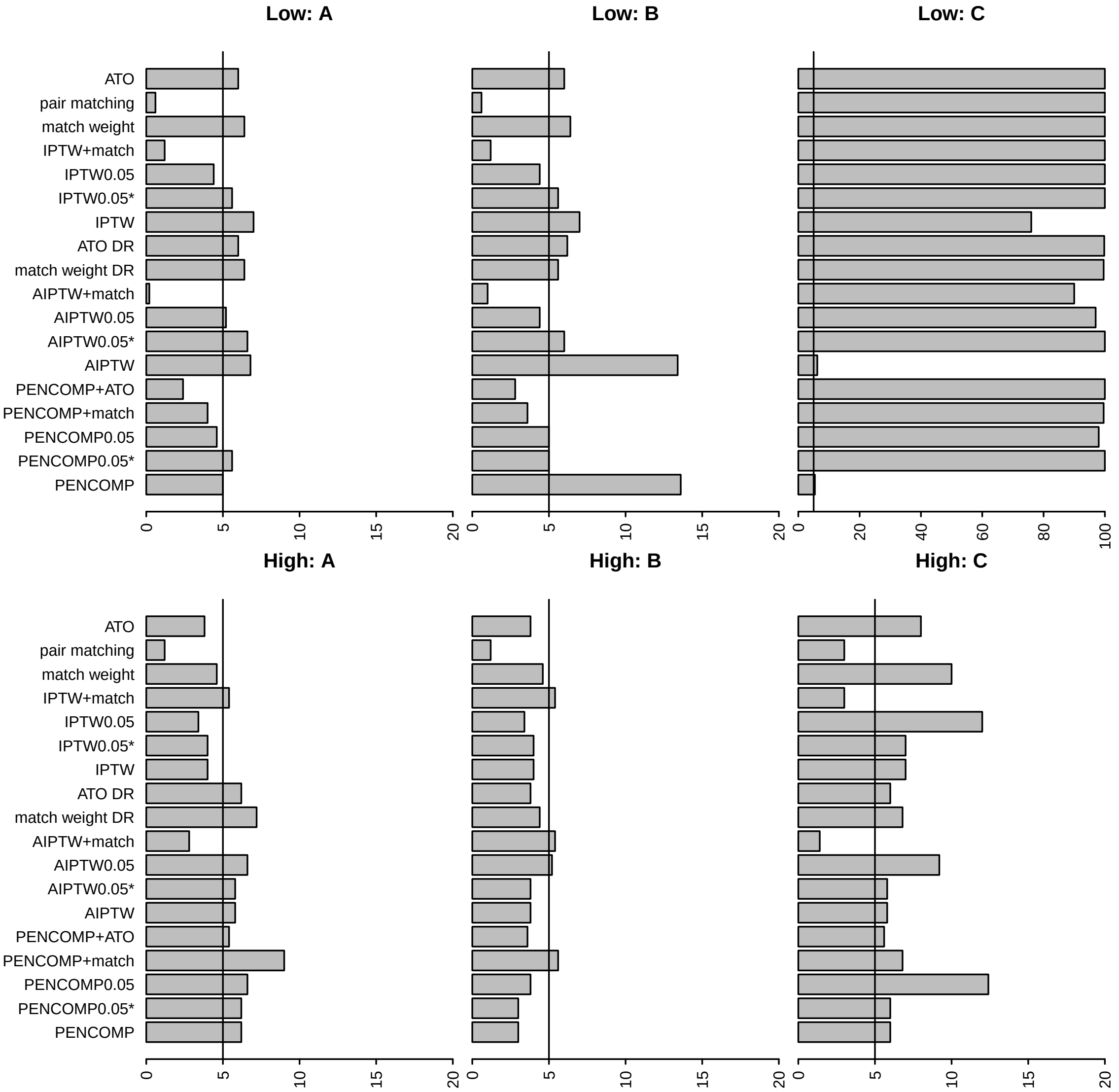

Figure B.16: Nonparallel surface and Aligned: 100 * 95% non coverage rate, sample size of 1000. (A)-Both propensity and prediction models are correct; (B) Prediction models are incorrect; (C) Propensity models are incorrect. Top Panel-Low overlap in the propensity distributions; Bottom Panel-high overlap in the propensity distributions.

BIBLIOGRAPHY

BIBLIOGRAPHY

Achy-Brou, A., Frangakis, C., and Griswold, M. (2010). Estimating Treatment Effects of Longitudinal Designs using Regression Models on Propensity Scores. *Biometrics, 66*, 824-833.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association, 91*, 444-455.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J. and Strmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*, 1149-1156.

Cochran, W.G. and Rubin, D.B. (1973). Controlling Bias in Observational Studies: A Review. *Sankhya: The Indian Journal of Statistics, Series A, 35*, 417-446.

Coull, B. A., Ruppert, D. and Wand, M. P. (2001). Simple Incorporation of Interactions into Additive Models. *Biometrics, 57*, 539-545.

Crump, R.K., Hotz, V.J, Imbens, G.W. and Mitnik, O.A. (2009). Dealing with Limited Overlap in Estimation of Average Treatment Effects. *Biometrika, 96*, 187-199.

de Luna, X., Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika, 98*, 861-875.

Dehejia, R.H. and Wahba, Sadek (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evalutation of Training Programs. *Journal of the American Statistical Association, 94*, 1053-1062.

Di Matteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian Curve-Fitting with Free-Knot Splines. *Biometrika, 88*, 1055-1071.

Efron, B. (2014). Estimation and Accuracy after Model Selection (with discussion). Journal of the American Statistical Association, 109, 991-1007.

Elliott, M. R. and Little, R. J. A. (2015). Discussion of “on Bayesian Estimation of Marginal Structural Models. *Biometrics, 71*, 288-291.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with b-splines and Penalties. *Statistical Science, 11*, 89-121.

Frangakis, C.E. and Rubin, D.B. (2002). Principal Stratification in Causal Inference *Biometrics, 58*, 21-29.

Gelman, A., and Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation or "p-hacking and the research hypothesis was posited ahead of time. Retrieved from http://www.stat.columbia.edu/ gelman/research/unpublished/p_hacking.pdf

Glymour M.M, Weuve J, Chen J.T. (2008). Methodological challenges in causal research on racial and ethnic patterns of cognitive trajectories: Measurement, selection, and bias. *Neuropsychology Review, 18*, 194-213.

Glynn, N. A, and Quinn, M. K. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis, 18*, 36-56.

Gutman, R. and Rubin, D.B. (2012). Robust Estimation of Causal Effects of Binary Treatments in Unconfounded Studies with Dichotomous Uutcomes. *Statistics in Medicine, 32*, 1795-1814.

Gutman, R. and Rubin, D.B. (2015). Estimation of Causal Effects of Binary Treatments in Unconfounded Studies. *Statistics in Medicine, 34*, 3381-3398.

Hansen, Ben B. (2008). The prognostic analogue of the propensity score. *Biometrika, 95*, 481-488.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple Imputation for the Fatal Accident Reporting System. *Applied Statistics, 40*, 13-29.

Hill, J. and Su, Y. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *the Annals of Applied Statistics, 7*, 1386-1420.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007). Matching As Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political analysis, 15(3)*, 199-236.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*, 945-960.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York, NY: Cambridge University Press.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data (with discussion). *Statistical Science, 22*, 523-539.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, CR. Jr. (1987). The Multicenter AIDS Cohort Study: Rationale, Organization, and Selected Characteristics of the Participants. *American Journal Epidemiology, 126*, 310-318.

Koo, J.-Y. (1997). Spline Estimation of Discontinuous Regression Functions. *Journal of Computational and Graphical Statistics, 6*, 266-284.

Lechnr, M. (2008). A Note on the Common Support Problem in Applied Evaluation Studies. *Econometric Evaluation of Public Policies: Methods and Applications*, 217-235.

Li, F, Morgan, K.L. and Zaslavsky, A.M (2017). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association, 113:521*, 390-400.

Li, L. and Greene, T. (2013), A Weighting Analogue to Pair Matching in Propensity Score Analysis, The International Journal of Biostatistics 9(2), 215-234.

Little, R. J. A. and An, H. (2004). Robust Likelihood-Based Analysis of Multivariate Data with Missing Values. *Statistica Sinica, 14*, 949-968.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York, NY: John Wiley & Sons.

Mao,H., Li, L., and Greene, T. (2018). Propensity score weighting analysis and treatment effect discovery. *Statistical Methods in Medical Research*, 1-16.

Murphy, S. A. (2005). An Experimental Design for the Development of Adaptive Treatment Strategies. *Statistics in Medicine, 24*, 455-481.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., and Murphy, S. A. (2012). Q-learning: A Data Analysis Method for Constructing Adaptive Interventions. *Psychological Methods, 17*, 478-494.

Neyman, Jerzy. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. Master's Thesis (1923). *Excerpts reprinted in English, Statistical Science, 5*, 463-472. (D. M. Dabrowska, and T. P. Speed, Translators.)

Ngo, L. and Wand, M. P. (2004). Smoothing with Mixed Model Software. *Journal of Statistical Software, 9*, 1-54.

Robins, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease, 40*, 139-161.

Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika, 70*, 41-55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score *Journal of the American Statistical Association, 79*, 516-524.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score. *The American Statistician, 39*, 33-38.

Rosenbaum, P.R. (2012). Optimal Matching of an Optimally Chosen Subset in Observational Studies. *Journal of Computational and Graphical Statistics, 21*, 57-71.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology, 66*, 688-701.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika, 63*, 581-592.

Rubin, D.B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics, 2*, 1-26.

Rubin, D. B. (1980). Discussion of "Randomization Analysis of Experimental Data: The Fisher Randomization Test, by D. Basu. *Journal of the American Statistical Association, 75*, 591-593.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D.B. (2006). Causal Inference through Potential Outcomes and Principal Stratification: Application to Studies with "Censoring" Due to Death *Statistical Science, 21*, 299-309.

Rubin, D.B. (2007). The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials. *Statistics in Medicine, 26*, 20-36.

Saarela, O., Stephens, D. A., Moodie, E. E. M., and Klein, M. B. (2015). On Bayesian Estimation of Marginal Structural Models. *Biometrics, 71*, 379-388.

Samuels, L. R. Aspects of Causal Inference within the Evenly Matchable Population: The Average Treatment Effect on the Evenly Matchable Units, Visually Guided Cohort Selection, and Bagged One-to-One Matching. Dissertation (2017). http://etd.library.vanderbilt.edu/available/etd-12122016-113901/unrestricted/Samuels.pdf

Scharfstein, D., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models. *Journal of the American Statistical Association, 94*, 1096-1120.

Shortreed, S.M and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics, 73(4)*, 1111-1122.

VanderWeele, T.J and Shpitser, I. (2011). A New Criterion for Confounder Selection. *Biometrics, 67*, 1406-13.

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics, 18*, 223-249.

Yang, Y. and Little, R. J. A. (2015). A Comparison of Doubly Robust Estimators of the Mean with Missing Data. *Journal of Statistical Computation and Simulation, 85*, 3383-3403.

Yu, Z. and van der Laan, M. J. (2006). Double Robust Estimation in Longitudinal Marginal Structural Models. *Journal of Statistical Planning and Inference, 136*, 1061-1089.

Zhang, G. and Little, R. J. A. (2009). Extensions of the Penalized Spline of Propensity Prediction Method of Imputation. *Biometrics, 65*, 911-918.

Zhang, J.L. and Rubin, D.B. (2003). Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death". *Journal of Educational and Behavioral Statistics, 28*, 353-368.

Zhang, J.L. and Rubin, D.B. (2009). Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification. *Journal of the American Statistical Association, 104*, 166-176.

Zhou, T., Elliott, M. R. and Little, R. J. A. (2018). Penalized Spline of Propensity Score for Treatment Comparison. *Journal of the American Statistical Association*, Accepted.

Zigler, c. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *Journal of American Statistical Association, 109*, 95-107.

Zou H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*, 1418-1429.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association, 110*, 910-922.