



Penalized Spline of Propensity Methods for Treatment Comparison

Tingting Zhou, Michael R. Elliott & Roderick J. A. Little

To cite this article: Tingting Zhou, Michael R. Elliott & Roderick J. A. Little (2019) Penalized Spline of Propensity Methods for Treatment Comparison, Journal of the American Statistical Association, 114:525, 1-19, DOI: [10.1080/01621459.2018.1518234](https://doi.org/10.1080/01621459.2018.1518234)

To link to this article: <https://doi.org/10.1080/01621459.2018.1518234>



View supplementary material [↗](#)



Published online: 19 Apr 2019.



Submit your article to this journal [↗](#)



View Crossmark data [↗](#)



Penalized Spline of Propensity Methods for Treatment Comparison

Tingting Zhou^a, Michael R. Elliott^{a,b}, and Roderick J. A. Little^a

^aDepartment of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI; ^bSurvey Research Center, Institute for Social Research, Ann Arbor, MI

ABSTRACT

Valid causal inference from observational studies requires controlling for confounders. When time-dependent confounders are present that serve as mediators of treatment effects and affect future treatment assignment, standard regression methods for controlling for confounders fail. Similar issues also arise in trials with sequential randomization, when randomization at later time points is based on intermediate outcomes from earlier randomized assignments. 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 for missing data problems. PENCOMP estimates causal effects by imputing missing potential outcomes with flexible spline models and draws inference based on imputed and observed outcomes. Under the SUTVA, positivity, and ignorability assumptions, PENCOMP has a double robustness property for causal effects. Simulations suggest that it tends to outperform doubly robust marginal structural modeling when the weights are variable. We apply our method to the multicenter AIDS cohort study to estimate the effect of antiretroviral treatment on CD4 counts in HIV-infected patients. Supplementary materials for this article are available online. Code submitted with this article was checked by an Associate Editor for Reproducibility and is available as an online supplement.

ARTICLE HISTORY

Received January 2017
Revised June 2018

KEYWORDS



Causal inference; Double robustness; Time-dependent confounders; Rubin's causal model.

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 (PSP) method 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 model 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 ensure 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, impute the missing potential outcomes for all the subjects, and estimate the causal effects by combining the imputed datasets with Rubin's combining rule.

CONTACT Roderick J. A. Little  rlittle@umich.edu  Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

 These materials were reviewed for reproducibility.

© 2019 American Statistical Association

We extend PENCOMP to longitudinal treatments, which is not considered in Gutman and Rubin (2015).

As discussed in Section 2 and in the supplementary material (Appendix 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 mean 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 4954 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 (ART) 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 dataset 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 ART 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 article, we consider a longitudinal study with $T + 1$ 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, that is, 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 (IPTW) estimators, 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 potential outcome under each treatment sequence based on conditional distributions of covariates and outcomes estimated from the data, so consistently estimates 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 potential outcome under 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 datasets 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 article is structured as follows. In Section 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 3, we briefly describe IPTW, AIPTW, and g-computation. In Section 4, we compare PENCOMP with the MSM approaches in simulation studies, assessing empirical bias, root mean squared error, 95% CI coverage, and width of confidence intervals. In Section 5, we apply our method to the MACS dataset to evaluate the short term effect of ART on CD4 counts in HIV+ infected patients. In Section 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. Penalized Spline of Propensity Methods for Treatment Comparisons

2.1. PSPP for Missing Data

Zhang and Little (2009), refining earlier work by Little and An (2004), proposed the following 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 $X = (X_1, \dots, X_p)$ 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) showed 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) described 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. 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. Figure 1 thus emphasizes the fundamental problem of causal inference (Holland 1986): since Y^1 and Y^0 are never observed

Subjects	X_1	Z_1	Y^0	Y^1
1		0		?
2		0		?
...		0		?
n_0		0		?
$n_0 + 1$		1	?	
...		1	?	
$n = n_0 + n_1$		1	?	

Figure 1. Observed and missing outcomes for treatment at a single time point.

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 SUTVA assumption, 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, and 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 nonzero 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, that is, there are no unmeasured confounders. The plausibility of the 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 the supplementary material (Appendix 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 $d = 1, \dots, D$, 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

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

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 Equation (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 Equation (1), as are generalized linear mixed models for nonnormal outcomes Y^{z_1} . Note that a different spline function in Equation (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, \dots, 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 $C_1 = [1, \hat{P}_{z_1}^*, x_1]$, $C_2 = [(\hat{P}_{z_1}^* - K_1)_+, \dots, (\hat{P}_{z_1}^* - K_K)_+]$, and $C = [C_1, C_2]$. Then the spline model can be expressed as a linear mixed model (Wand 2003),

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

where $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ denote fixed effects, and $\theta = (\theta_{11}, \dots, \theta_{1K})$ 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 $\hat{y}^{z_1} = C(C^T C + \hat{\lambda} D)^{-1} C^T y$, 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 dataset 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 $\bar{\Delta}_D$ is $T_D = \bar{W}_D + (1 + 1/D)B_D$, where $\bar{W}_D = \sum_{d=1}^D W^{(d)}/D$, $B_D = \sum_{d=1}^D (\hat{\Delta}_d - \bar{\Delta}_D)^2 / (D - 1)$. The estimate Δ is t distributed with degree of freedom ν , $(\Delta - \bar{\Delta}_D) T_D^{-\frac{1}{2}} \sim t_\nu$, where $\nu = (D - 1)(1 + \bar{W}_D / ((D + 1) * B_D))^2$.

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

2.3. PENCOMP with Longitudinal Treatment Assignments

We now consider a longitudinal study with treatments assigned at multiple time points $t = 1, \dots, T$. 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 $\bar{Z}_t = (Z_1, \dots, Z_t)$. Let $Y^{\bar{Z}_T}$ denote the final potential outcome under the entire treatment regime $\bar{Z}_T = (Z_1, \dots, Z_T)$, 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 . Suppose 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, and 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

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

Figure 2. Observed and missing intermediate and final outcomes for treatment at two time points.

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} . Figure 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 2.1. 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 drawn 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 $d = 1, \dots, D$, 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

$$E(X_2^{z_1} | X_1, Z_1 = z_1, \theta_{z_1}, \beta_{z_1}) = s(\hat{P}_{z_1}^* | \theta_{z_1}) + g_{z_1}(\hat{P}_{z_1}^*, X_1; \beta_{z_1}), \quad (3)$$

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 Equation (3). Note that a distinct model of form (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 dataset 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 $\hat{P}_{z_2} = \hat{P}_{z_1}(X_1) \hat{P}_{z_2}(\bar{X}_2, Z_1)$, and define $\hat{P}_{z_2}^* = \log[\hat{P}_{z_2}/(1 - \hat{P}_{z_2})]$.

(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

$$E(Y^{\bar{Z}_2} | \bar{X}_2, Z_1 = z_1, Z_2 = z_2, \theta_{\bar{Z}_2}, \beta_{\bar{Z}_2}) = s(\hat{P}_{\bar{Z}_2}^* | \theta_{\bar{Z}_2}) + g_{\bar{Z}_2}(\hat{P}_{\bar{Z}_2}^*, \bar{X}_2, \bar{Z}_2; \beta_{\bar{Z}_2}) \quad (4)$$

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 (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 dataset 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 $T_D = \bar{W}_{jkD} + (1 + 1/D)B_{jkD}$, where $\bar{W}_{jkD} = \sum_{d=1}^D W_{jk}^{(d)}/D$, $B_{jkD} = \sum_{d=1}^D (\hat{\Delta}_{jk}^{(d)} - \bar{\Delta}_{jkD})^2 / (D - 1)$. As described in (e) of single treatment setting, draw inference about Δ_{jk} by assuming a t-distribution.

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 $E(X_{t+1}^{\bar{z}_t} | \bar{X}_t, \bar{Z}_t = \bar{z}_t, \theta_{\bar{z}_t}, \beta_{\bar{z}_t}) = s_{x_{t+1}}(\hat{P}_{\bar{z}_t}^*; \theta_{\bar{z}_t}) + g_{\bar{z}_t}[\hat{P}_{\bar{z}_t}^*, X_1, \dots, X_t; \beta_{\bar{z}_t}]$, 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 2.3, PENCOMP has a double robustness property for causal effects in a longitudinal study setting. The proof is outlined in the supplementary material (Appendix 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 $t = 1, \dots, T + 1$, 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 $\hat{P}_{\bar{z}_t}^*$ is correctly specified at each time point $t = 1, \dots, T + 1$. 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.

2.4. 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) suggested caliper matching when some units are dropped due to poor match quality. Rubin (1977) suggested 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) dropped 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) proposed 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) proposed 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 $\Delta_{\bar{Z}_T}$, relative to the null treatment regime 0_T , at a given time $1 \leq t \leq T$, we first obtain the set of observations $A_t = A_{\bar{Z}_t}$ such as

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

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

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

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 $A_1 \cap B_1 \cap \dots \cap A_T \cap B_T$. 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.

3. G-computation, IPTW, and AIPW

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 $Q_0 = P(Y | \bar{X}_T, \bar{Z}_T = \bar{z}_T) \prod_{t=1}^T P(X_t | \bar{X}_{t-1}, \bar{Z}_{t-1} = \bar{z}_{t-1})$ 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 $\bar{z}_T = (z_1, \dots, z_T)$, $E(Y^{\bar{z}_T}) = \sum_{X_1, \dots, X_T} E(Y | \bar{X}_T, \bar{Z}_T = \bar{z}_T) \times P(X_1) \times P(X_2 | X_1, Z_1 = z_1) \dots \times P(X_T | \bar{X}_{T-1}, \bar{Z}_{T-1} = \bar{z}_{T-1})$. For continuous X_s , 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.

3.2. IPTW 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:

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

where g_0 is defined in Section 3.1.

Under the assumptions stated in Section 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 MSM 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:

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

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

3.3. AIPTW Estimator

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

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

Under the assumptions stated in Section 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 supplementary material (Appendix 2) for more detailed descriptions of our implementations of IPTW and AIPTW.

4. Simulation Studies

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 versus nonlinear 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 datasets were created for a sample size of 500, but to reduce computation burden, only 500 simulated datasets 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% CI width, and 95% CI (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% CI width to the width of IPTW with the correct propensity model. In the main article, we present the results for RMSE and 95% noncoverage. The complete results are included in the supplementary material (Appendix 3).

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 dataset 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}

Table 1. Single time point treatment simulation scenarios: $\gamma = c(1.5, 1.5, 0.75)$, $c(1, 1, 0.5)$, and $c(0.1, 0.1, 0.05)$ corresponds to high, moderate, and low confounding, respectively.

<i>Linear outcome</i>		
True	$\text{logit}(P(Z_1 = 1 \bar{X}, \gamma)) = \gamma_1 X_{1a} + \gamma_2 X_{1b} + \gamma_3 X_{1a} X_{1b}$ $E(Y_1 \bar{X}, \beta_1) = \beta_{10} + \beta_{11} X_{1b} + \beta_{12} X_{1c}$ $E(Y_0 \bar{X}, \beta_0) = \beta_{00} X_{1b} + \beta_{01} X_{1c}$	$\gamma = c(\gamma_1, \gamma_2, \gamma_3) = c(1.5, 1.5, 0.75)$, $c(1, 1, 0.5)$, 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	$\text{logit}(P(Z_1 = 1 \bar{X}, \lambda)) = \lambda_0 + \lambda_1 X_{1a}$ $E(Y_1 \bar{X}, \alpha_1) = \alpha_{10} + \alpha_{11} X_{1c}$ $E(Y_0 \bar{X}, \alpha_0) = \alpha_{00} + \alpha_{01} X_{1c}$	
<i>Nonlinear outcome</i>		
True	$\text{logit}(P(Z_1 = 1 \bar{X}, \gamma)) = \gamma_1 X_{1a} + \gamma_2 X_{1b} + \gamma_3 X_{1a} X_{1b}$ $E(Y_1 \bar{X}, \beta_1) = \beta_{10} + \beta_{11} X_{1b} + \beta_{12} X_{1c} + \beta_{13} X_{1b}^2 + \beta_{14} X_{1c}^2$ $E(Y_0 \bar{X}, \beta_0) = \beta_{00} X_{1b} + \beta_{01} X_{1c}$	$\gamma = c(\gamma_1, \gamma_2, \gamma_3) = c(1.5, 1.5, 0.75)$, $c(1, 1, 0.5)$, 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	$\text{logit}(P(Z_1 = 1 \bar{X}, \lambda)) = \lambda_1 X_{1a}$ $E(Y_1 \bar{X}, \alpha_1) = \alpha_{10} + \alpha_{11} X_{1c}$ $E(Y_0 \bar{X}, \alpha_0) = \alpha_{00} + \alpha_{01} X_{1c}$	

NOTE: The true coefficients associated with each model are listed next to each model.

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 1. The data were generated based on the true models shown in Table 1. The treatment effects under linear and nonlinear outcome models were 5 and 9, respectively.

Results for sample size 500 are shown in Figures 3 and 4 and Tables 9–12, supplementary material (Appendix 3). The RMSEs of the methods are shown in Figure 3, 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% CI noncoverage rates are shown in Figure 4. 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 9, supplementary material (Appendix 3) 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 12, supplementary material (Appendix 3). 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 4. 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 to 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 Tables 5–8 and 13–16, supplementary material (Appendix 3). 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. PENCOMPs 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

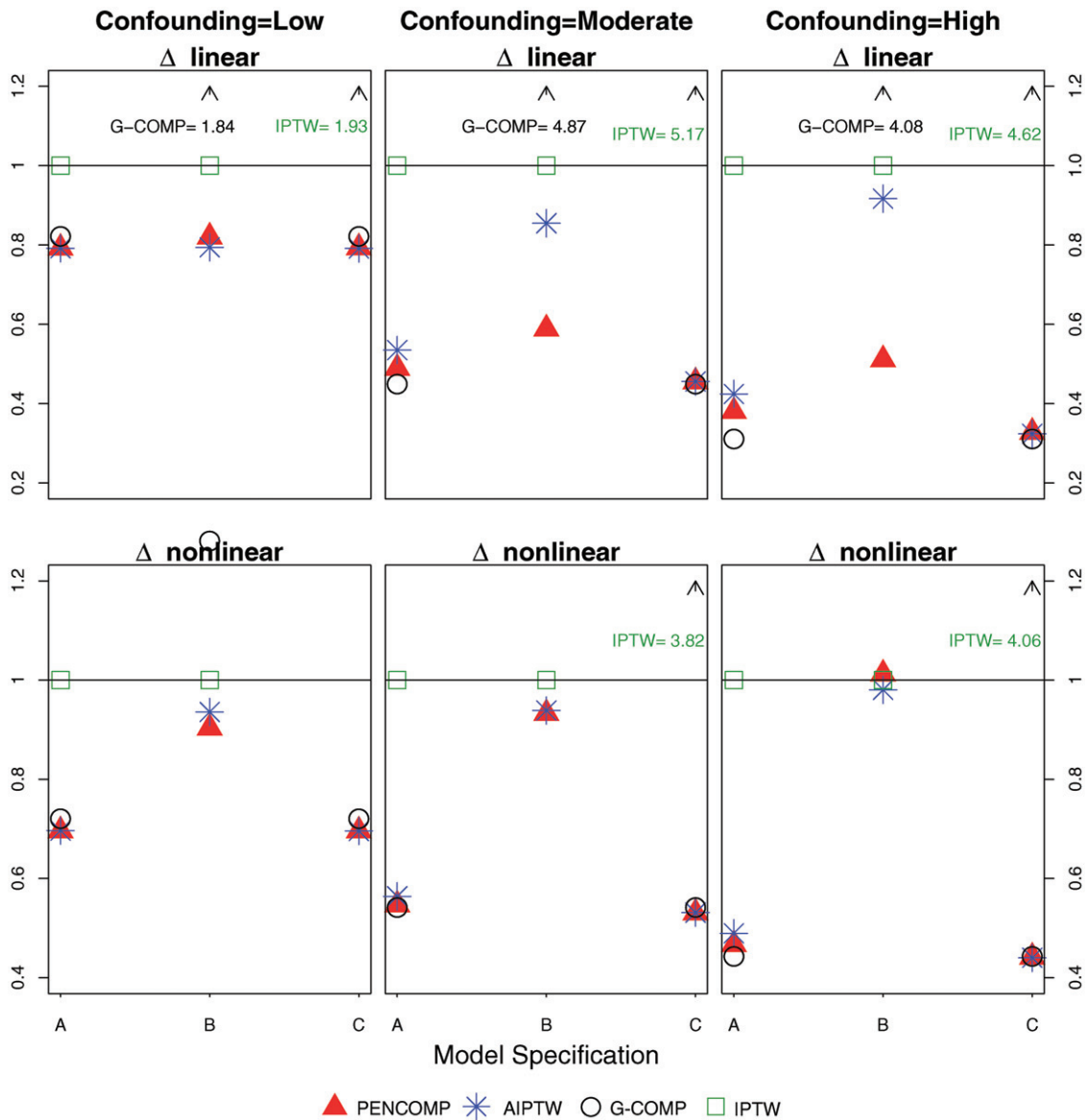


Figure 3. Ratio of RMSE over RMSE of IPTW (A) with correct propensity score model across four methods—PENCOMP, AIPTW, IPTW, and g-computation for treatment effect Δ in a linear and nonlinear outcome model. (A) Correctly specified propensity and prediction models, (B) a correctly specified propensity 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.

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 15, supplementary material, Appendix 3).

4.3. Simulations for Treatments Assigned at Two Time Points

In the two time-point treatment scenario, each simulated dataset 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 and variance 1. The first treatment Z_1 is Bernoulli

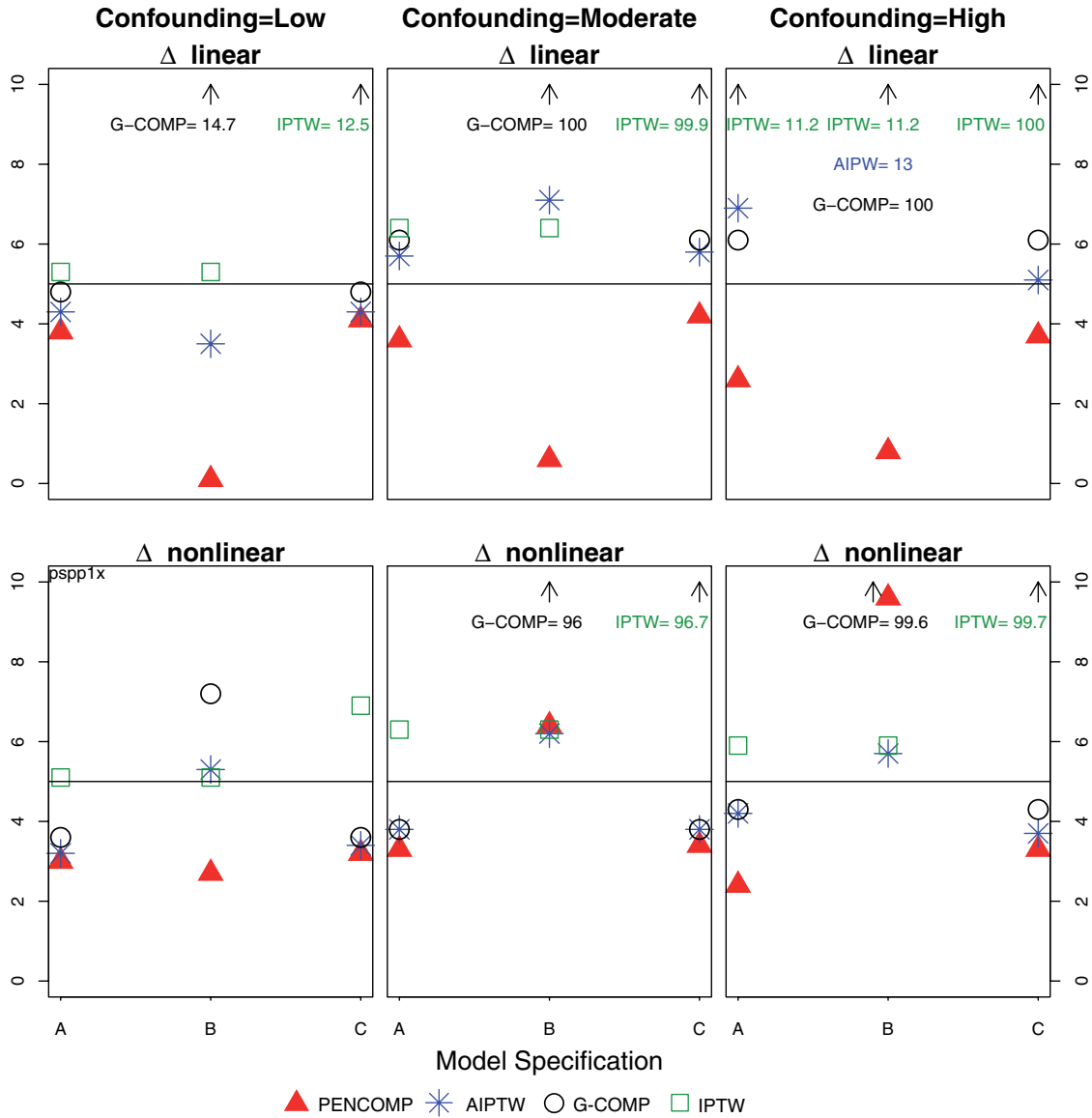


Figure 4. About 95% noncoverage rate across four methods PENCOMP, AIPTW, IPTW, and g-computation for treatment effect Δ in a linear and nonlinear outcome model. (A) Correctly specified propensity and prediction models; (B) a correctly specified propensity model only; (C) a correctly specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

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. 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. The data were generated based on the true models in Table 2. 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.

Results for RMSE and 95% CI noncoverage for sample size 500 are shown in Figures 5–8; other results are given in Tables 21–24, supplementary material (Appendix 3). The RMSEs of the methods are presented in Figure 5 for the linear outcome model and in Figure 6 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.

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

Noncoverage rates of the 95% intervals are shown in Figures 7 and 8. 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 21 and supplementary material (Appendix 3) 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% CIs widths are shown in Table 24, supplementary material (Appendix 3). 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 confounding. Finally,

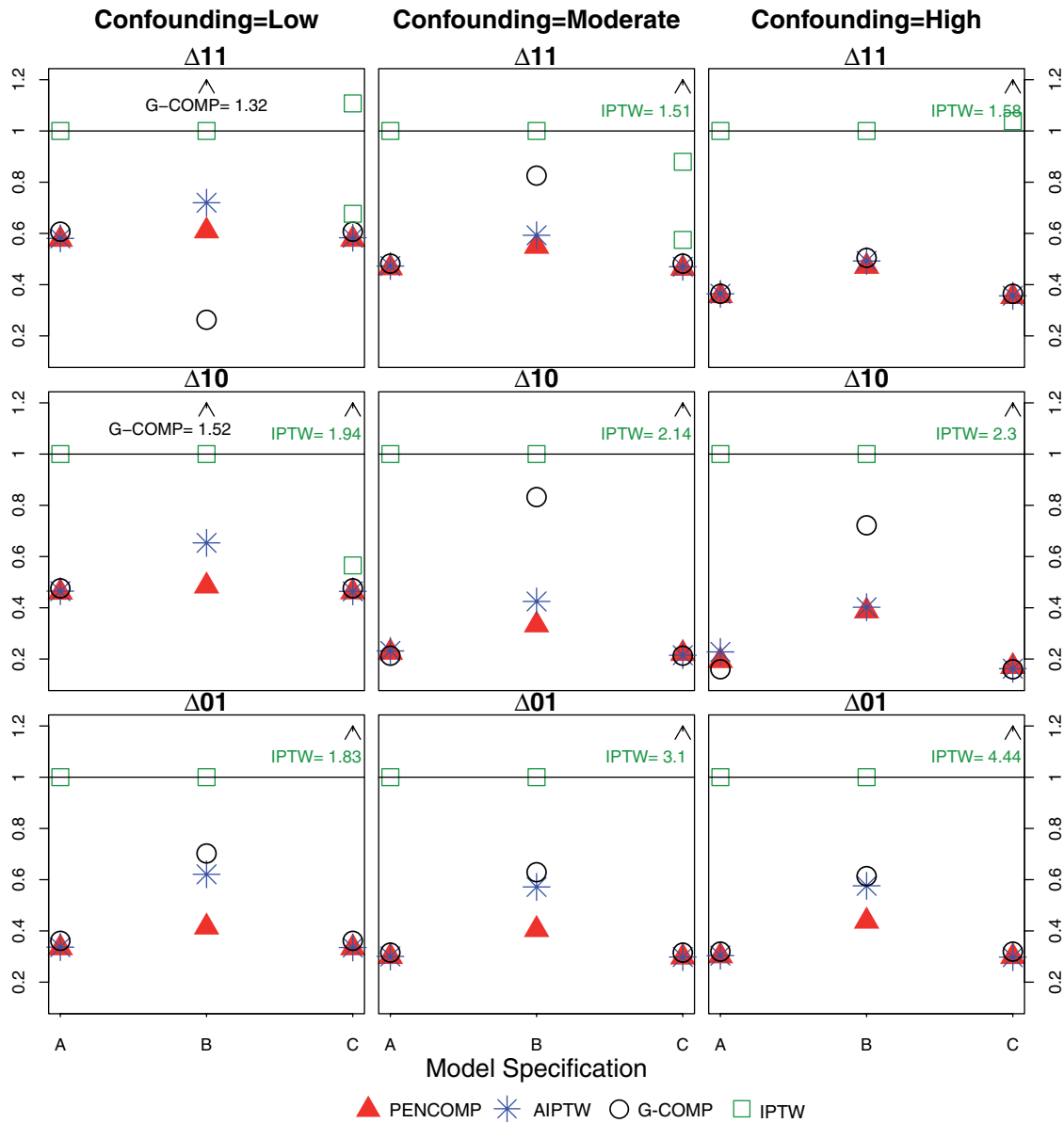


Figure 5. Ratio of RMSE over RMSE of IPTW (A) with correct propensity score model across four methods PENCOMP, AIPTW, IPTW, and g-computation for three treatment effects Δ_{11} , Δ_{10} , and Δ_{01} in a linear outcome model. (A) Correctly specified propensity and prediction models; (B) a correctly specified 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.

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 Tables 17–20 and 25–28, supplementary material (Appendix 3). 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.

5. Application

We applied our method to the MACS to analyze the effect of ART on CD4 counts. We restricted 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 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 estimated the short-term (1 year) effects of using any ART for HIV+ subjects. Treatment was coded to 1 if the patient reported taking any of the four mentioned 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

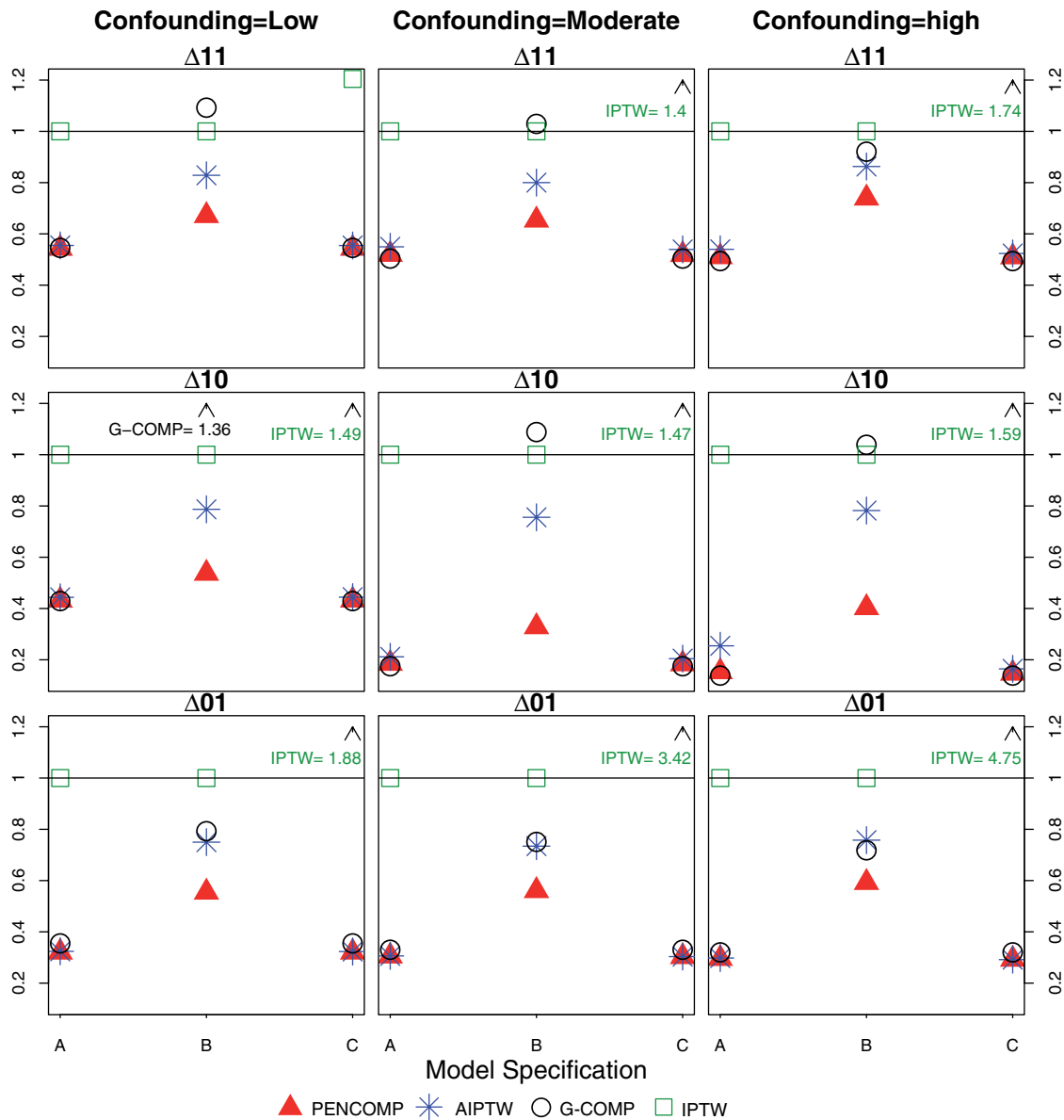


Figure 6. Ratio of RMSE over RMSE of IPTW (A) with correct propensity score model across four methods—PENCOMP, AIPTW, IPTW, and g-computation for three treatment effects Δ_{11} , Δ_{10} , and Δ_{01} in a nonlinear outcome model. (A) Correctly specified propensity and prediction models; (B) a correctly 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.

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 ART 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, that is, 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

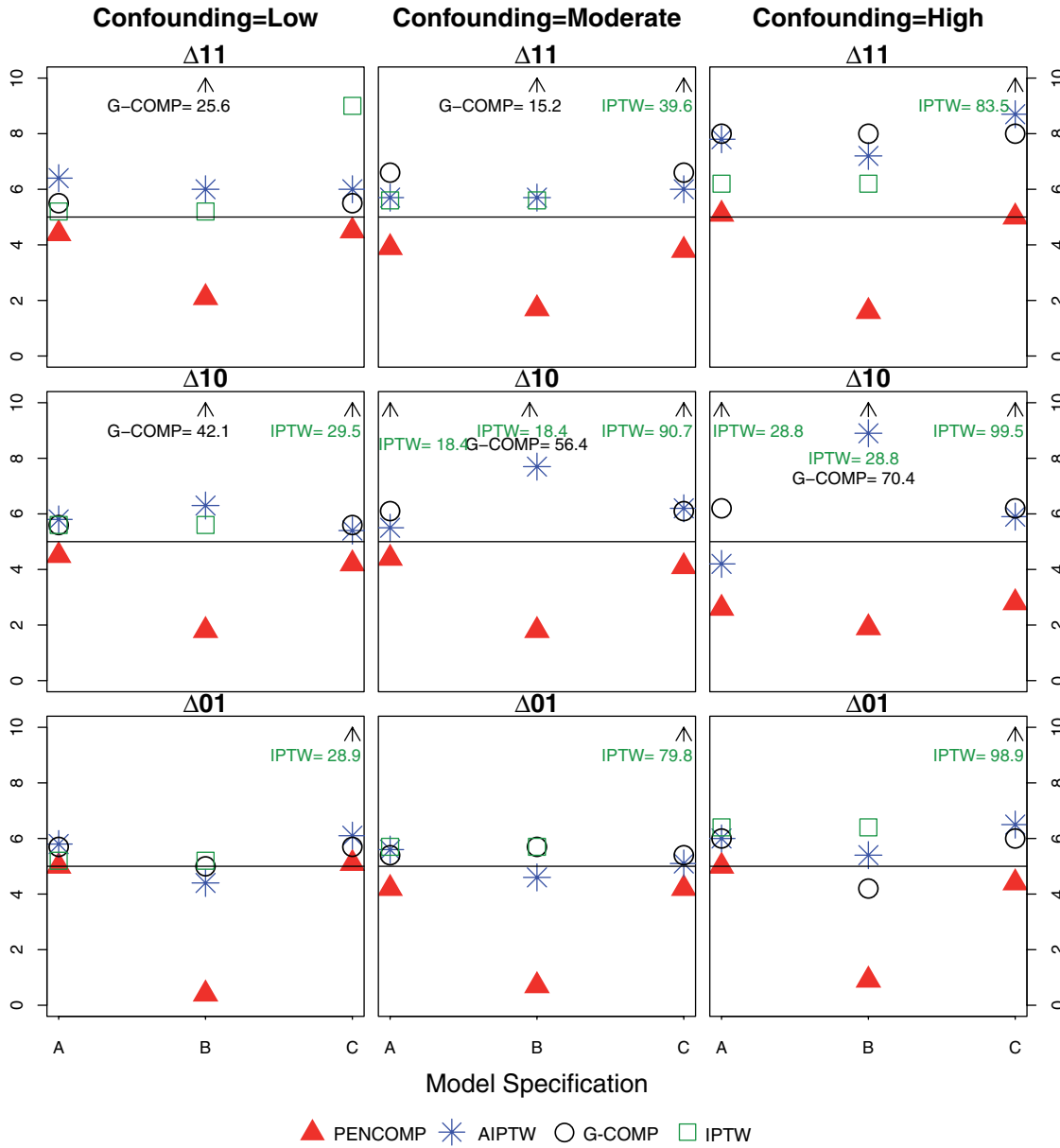


Figure 7. About 95% noncoverage rate across four methods PENCOMP, AIPTW, IPTW, and g-computation for three treatment effects Δ_{11} , Δ_{10} , and Δ_{01} in a linear outcome model. (A) Correctly specified propensity and prediction models; (B) a correctly specified propensity model only; (C) a correctly specified prediction model only; based on 1000 simulations with sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

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 29, supplementary material (Appendix 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 30, 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 11, supplementary material (Appendix 4). The results without truncation are in Figure 10 (see Zubizarreta (2015) for an alternative that minimizes weight variance while retaining covariate balance). For PENCOMP, we chose a minimum of 35 and $1/4$ of unique data points as the number of knots. Equally spaced knots and a truncated linear basis were used. In addition, for estimating outcomes for a particular treatment regimen \bar{Z}_t , we excluded cases where the propensity of \bar{Z}_t lay outside the observed ranges of the propensity of \bar{Z}_t as described in Section 2.4, to avoid extrapolating

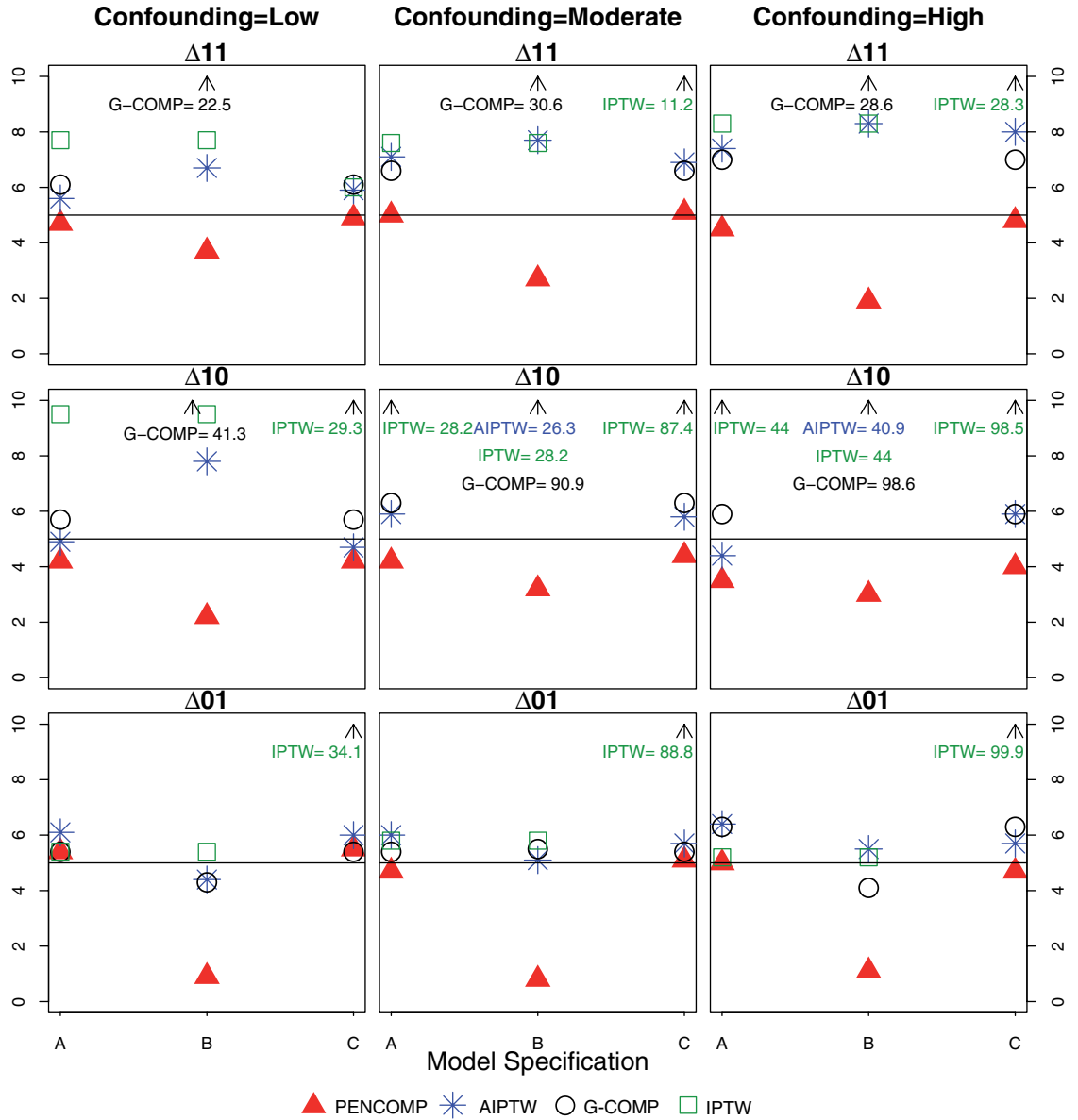


Figure 8. About 95% noncoverage rate across four methods PENCOMP, AIPTW, IPTW, and g-computation for three treatment effects Δ_{11} , Δ_{10} , and Δ_{01} in a nonlinear outcome model. (A) Correctly specified propensity and prediction models; (B) a correctly specified propensity model only; (C) a correctly specified prediction model only; based on 1000 simulations with a sample size of 500 and 500 bootstraps, and 200 complete datasets for PENCOMP.

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 $\hat{P}_{z_2} = \hat{P}(Z_1 = z_1 | \bar{X}_1) * \hat{P}(Z_2 = z_2 | Z_1 = z_1, \bar{X}_1)$. At $t = 1$, subjects were divided into two groups using indicators $I(Z_1^{\text{obs}} = z_1)$ and $I(Z_1^{\text{obs}} \neq z_1)$. 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 $I\{(Z_1^{\text{obs}}, Z_2^{\text{obs}}) = (z_1, z_2)\}$ and $I\{(Z_1^{\text{obs}}, Z_2^{\text{obs}}) \neq (z_1, z_2)\}$. Again, we removed subjects whose estimated propensity scores \hat{P}_{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 9 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 29, supplementary material. 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 $I(Z_1^{\text{obs}} = z_1)$ and $I(Z_1^{\text{obs}} \neq z_1)$, we checked

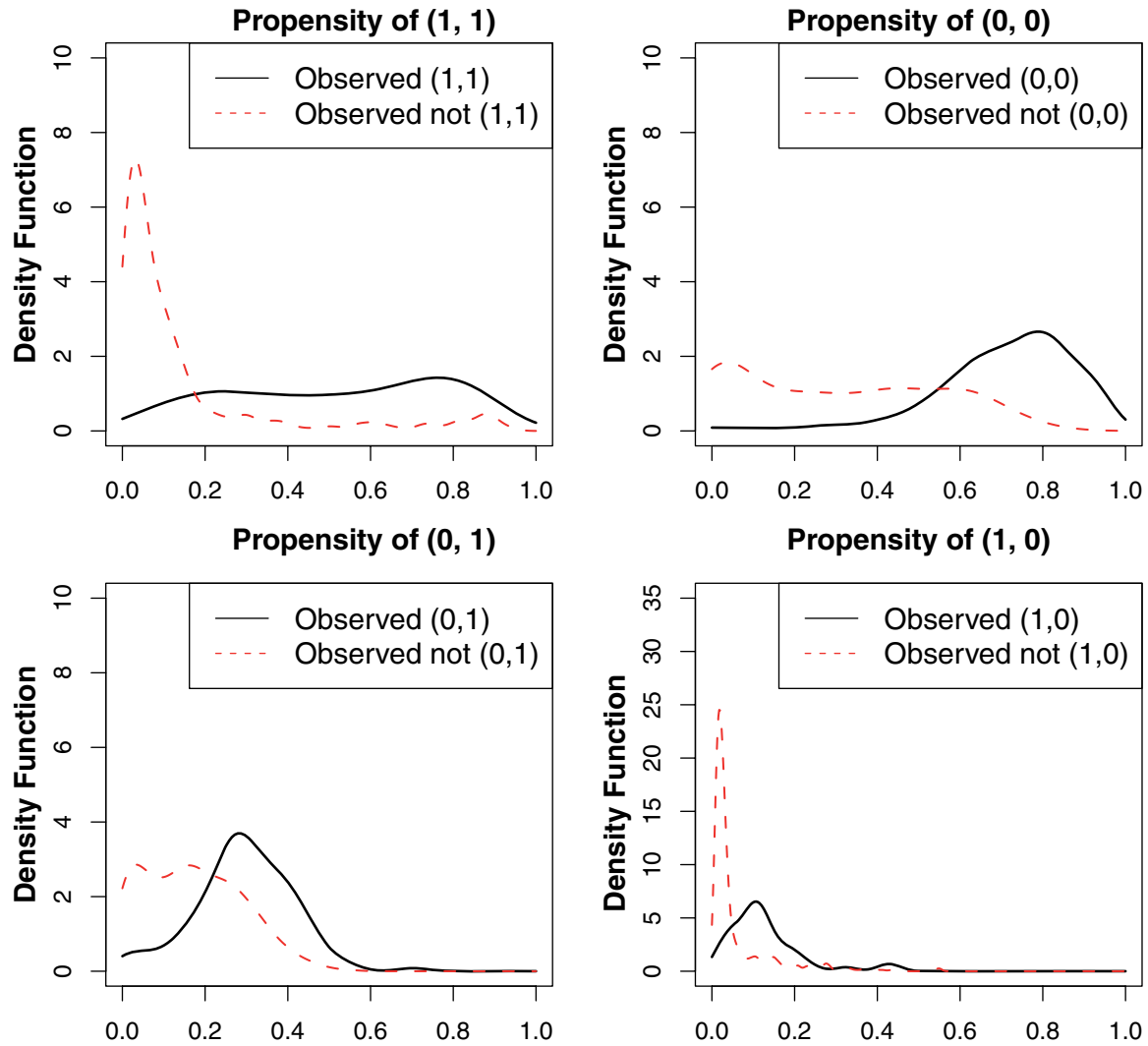


Figure 9. Distributions of the propensity scores in subjects whose observed treatment sequence is (z_1, z_2) and subjects whose observed treatment sequence is not (z_1, z_2) for window 4.

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 $I\{(Z_1^{\text{obs}}, Z_2^{\text{obs}}) = (z_1, z_2)\}$ and $I\{(Z_1^{\text{obs}}, Z_2^{\text{obs}}) \neq (z_1, z_2)\}$. 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:

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

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 the t -test. Table 3 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

Table 3. 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.

Covariate	Before adjusting		After adjusting	
	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

NOTE: 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.

*significant at 0.05 level.

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 $\pi_{(z_1, z_2)}^{1-\alpha} = F_{\neq(z_1, z_2)}(F_{(z_1, z_2)}^{-1}(1 - \alpha)) -$

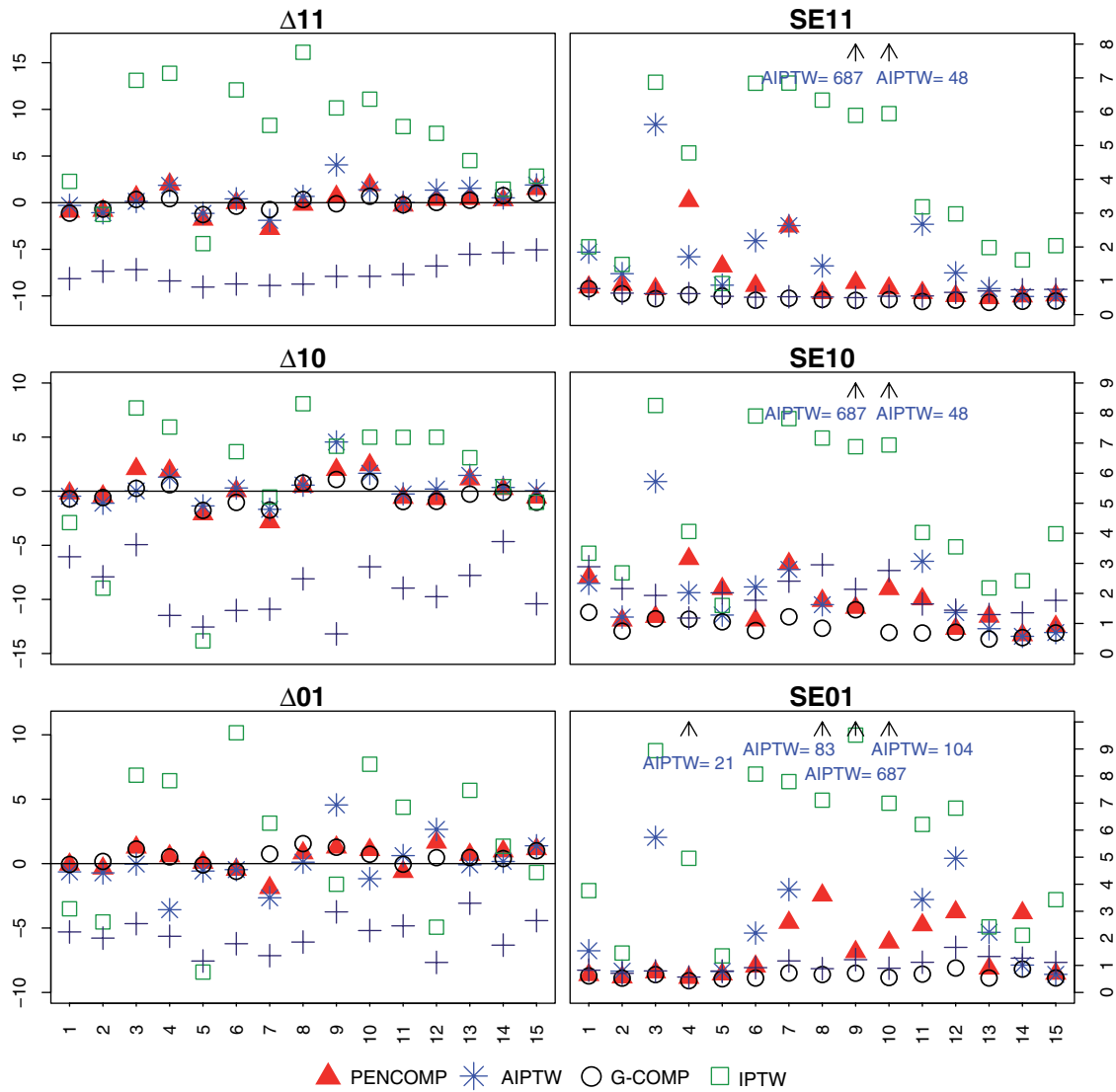


Figure 10. For each of the three-visit windows 1, . . . , 15, the estimates and standard errors (SE) of the treatment effects Δ_{11} , Δ_{10} , and Δ_{01} of the four methods: PENCOMP, AIPTW, IPTW, and Naive. For some windows, AIPTW had very large bootstrap standard errors because of a few extreme bootstrap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

$F_{\neq(z_1, z_2)}(F_{(z_1, z_2)}^{-1}(\alpha))$, 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 9 and Table 31, supplementary material (Appendix 4).

We estimated the short-term effect of ART on CD4 count using four methods: naive crude estimate, g-computation, IPTW, AIPTW, and PENCOMP. The results are summarized in Figure 10. 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 ART 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 suggested 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

windows 1–3, 15, and 16 and the means of the stabilized 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. Furthermore, 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.

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 misspecification 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 bases 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 article. Specifically, if 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, Ruppert, and Wand 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 tended 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 datasets 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 dataset 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 be 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, Frangakis, and Griswold (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 fit a proportional odds logistic regression models based on the propensity quintiles for the transition probabilities between strata. PENCOMP is similar to Achy Brou et al.’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) considered 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, Genovese, and Kass 2001). In addition, we have focused on estimating causal effects for a continuous and normally distributed outcome. Extensions to nonnormal 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) examined the performance of a similar spline method for binary outcome in one time point treatment. However, the performance of such extensions to nonnormal 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 semiparametric 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.

Supplementary Material

Appendix 1 outlines a proof of the double robustness property of PENCOMP. Appendix 2 provides more details on our implementations of IPTW and AIPTW. Appendix 3 provides more detailed results for the simulation studies. Appendix 4 contains more detailed results for the Application in Section 5.

References

- 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. [18]
- 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. [3,4]
- 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. [6]
- Coull, B. A., Ruppert, D., and Wand, M. P. (2001), "Simple Incorporation of Interactions into Additive Models," *Biometrics*, 57, 539–545. [18]
- 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. [6]
- Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062. [6]
- Di Matteo, I., Genovese, C. R., and Kass, R. E. (2001), "Bayesian Curve-Fitting With Free-Knot Splines," *Biometrika*, 88, 1055–1071. [18]
- Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties," *Statistical Science*, 11, 89–121. [4]
- Elliott, M. R., and Little, R. J. A. (2015), "Discussion of 'on Bayesian Estimation of Marginal Structural Models,'" *Biometrics*, 71, 288–291. [2,3,5]
- Glynn, N. A., and Quinn, M. K. (2010), "An Introduction to the Augmented Inverse Propensity Weighted Estimator," *Political Analysis*, 18, 36–56. [7]
- Gutman, R., and Rubin, D. B. (2012), "Robust Estimation of Causal Effects of Binary Treatments in Unconfounded Studies With Dichotomous Outcomes," *Statistics in Medicine*, 32, 1795–1814. [19]
- (2015), "Estimation of Causal Effects of Binary Treatments in Unconfounded Studies," *Statistics in Medicine*, 34, 3381–3398. [1,2,6]
- Heitjan, D. F., and Little, R. J. A. (1991), "Multiple Imputation for the Fatal Accident Reporting System," *Applied Statistics*, 40, 13–29. [3]
- Holland, P. W. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960. [3]
- Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press. [16]
- 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. [18]
- Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R., Jr. (1987), "The Multicenter AIDS Cohort Study: Rationale, Organization, and Selected Characteristics of the Participants," *American Journal of Epidemiology*, 126, 310–318. [2]
- Koo, J.-Y. (1997), "Spline Estimation of Discontinuous Regression Functions," *Journal of Computational and Graphical Statistics*, 6, 266–284. [18]
- Little, R. J. A., and An, H. (2004), "Robust Likelihood-Based Analysis of Multivariate Data With Missing Values," *Statistica Sinica*, 14, 949–968. [1,3]
- Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data*. New York: Wiley. [3]
- Murphy, S. A. (2005), "An Experimental Design for the Development of Adaptive Treatment Strategies," *Statistics in Medicine*, 24, 455–481. [19]
- 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. [19]
- Ngo, L., and Wand, M. P. (2004), "Smoothing With Mixed Model Software," *Journal of Statistical Software*, 9, 1–54. [4]
- 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. [2,6]
- 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. [2]
- (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.
- Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701. [1,3,5]
- (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592. [3]
- (1977), "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics*, 2, 1–26. [6]
- (1980), "Discussion of 'Randomization Analysis of Experimental Data: The Fisher Randomization Test,' by D. Basu," *Journal of the American Statistical Association*, 75, 591–593. [3,4]
- (1987), *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley. [3]
- 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. [2]
- 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. [2,7]
- Wand, M. P. (2003), "Smoothing and Mixed Models," *Computational Statistics*, 18, 223–249. [4]
- 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. [3]
- 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. [2,7]
- Zhang, G., and Little, R. J. A. (2009), "Extensions of the Penalized Spline of Propensity Prediction Method of Imputation," *Biometrics*, 65, 911–918. [1,3,17]
- Zubizarreta, J. R. (2015), "Stable Weights That Balance Covariates for Estimation With Incomplete Outcome Data," *Journal of the American Statistical Association*, 110, 910–922. [14]