Received: 15 May 2020 | Revised: 6 December 2020 | Accepted: 10 December 2020

DOI: 10.1002/sim.8862

RESEARCH ARTICLE

Statistics in Medicine WILEY

A comparison of parametric propensity score-based methods for causal inference with multiple treatments and a binary outcome

Youfei Yu[1] | **Min Zhang**[1] | **Xu Shi**[1] | **Megan E. V. Caram**[2,3,4] | **Roderick J. A. Little**[1] | **Bhramar Mukherjee**[1]

[1]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan

[2]Division of Hematology/Oncology, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan

[3]Veterans Affairs (VA) Health Services Research and Development, Center for Clinical Management and Research, VA Ann Arbor Healthcare System, Ann Arbor, Michigan

[4]Institute for Healthcare Policy and Innovation, University of Michigan Medical School, Ann Arbor, Michigan

Correspondence
Min Zhang, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA.
Email: mzhangst@umich.edu

Funding information
Division of Mathematical Sciences, Grant/Award Number: 1712933; National Cancer Institute, Grant/Award Number: 046591; Prostate Cancer Foundation, Grant/Award Number: Young Investigator Award; National Science Foundation

We consider comparative effectiveness research (CER) from observational data with two or more treatments. In observational studies, the estimation of causal effects is prone to bias due to confounders related to both treatment and outcome. Methods based on propensity scores are routinely used to correct for such confounding biases. A large fraction of propensity score methods in the current literature consider the case of either two treatments or continuous outcome. There has been extensive literature with multiple treatment and binary outcome, but interest often lies in the intersection, for which the literature is still evolving. The contribution of this article is to focus on this intersection and compare across methods, some of which are fairly recent. We describe propensity-based methods when more than two treatments are being compared, and the outcome is binary. We assess the relative performance of these methods through a set of simulation studies. The methods are applied to assess the effect of four common therapies for castration-resistant advanced-stage prostate cancer. The data consist of medical and pharmacy claims from a large national private health insurance network, with the adverse outcome being admission to the emergency room within a short time window of treatment initiation.

KEYWORDS
causal inference, comparative effectiveness research, electronic health records, multiple treatment comparison, propensity score

1 | INTRODUCTION

Comparative effectiveness research (CER) assesses alternative interventions for a particular clinical condition.[1] Randomized clinical trials are the gold standard for CER, but real-world evidence when drugs are released into the market is increasingly being used to make health care decisions.[2] CER for such observational data requires statistical methods for causal inference that control for confounding variables. The current literature on these methods largely focuses on two

treatments and continuous outcomes,[3,4] but often interest lies in comparing more than two treatments and outcomes are binary, for example, the occurrence of an event.[5] We compare here causal inference methods when the outcome is binary and there are more than two treatments.

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

In observational studies, the estimation of causal effects is prone to bias due to confounders related to both treatment and outcome. Methods to correct for this bias can be classified into two broad categories. The traditional approach is to model the multiple regression of the outcome on the treatment and measured potential confounders. This approach is vulnerable to misspecification of the regression model. An alternative approach is to model the propensity score, defined as the probability of being assigned to the treatment given a set of potential confounders. The treatment effect is then estimated by matching,[10,11] weighting,[3,12-14] stratification,[3,10,15] or regression[16,17] on the estimated propensity scores. This method was introduced by Rosenbaum and Rubin,[10] who showed that propensity scores have a balancing property, such that the conditional distribution of the potential confounders given the balancing scores are the same for treated and control. This property implies that propensity score methods provide some protection against misspecification of the outcome models. However, propensity score models are still required to be correctly specified.

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

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

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

Propensity score weighting methods are more easily extended to the multiple treatment setting.[37] The asymptotic distributions of the weighting-based estimators can be characterized using the theory of M-estimation,[38] which yields estimated standard errors that incorporate the uncertainty associated with the estimation of propensity scores. A common weighting scheme is to weight units in one group by their inverse probability of being in that group (IPW). Evaluations of IPW are mainly confined to the two treatment setting, and suggest that the estimator is sensitive to extreme weights and can have high variability.[3,24,39]

An important extension of IPW is the augmented inverse probability weighting (AIPW), where the IPW estimator is augmented using predictions from an outcome regression (OREG) model. To implement AIPW method in a multiple treatment setting, one can first obtain the estimated GPS, possibly from a multinomial logistic regression model, and then the predicted outcomes for each treatment group from outcome models that describe the conditional expectation of the outcome variable given measured covariates and treatment status. The resulting estimator is known as having a double robustness (DR) property such that the estimator remains consistent as long as either the propensity score model or the outcome model is correctly specified. AIPW estimator is asymptotically efficient within a broad class of estimators that includes the IPW estimator.[37] Lunceford and Davidian reviewed the theoretical properties of IPW, AIPW, and several other propensity score weighting estimators in the context of two treatments and continuous outcome.[3] Simulation studies

indicated that weighting-based methods with correct propensity score modeling produced approximately unbiased point estimates, and AIPW was more precise than IPW for sample sizes as small as 1000.

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

Studies of comparative effectiveness with continuous outcomes typically report an estimate of the average treatment effect (ATE), which is the difference in average outcome if individuals were all assigned the treatment and the average outcome if all the individuals were assigned the comparator treatment.[41] In this article, we measure treatment effectiveness by the risk difference,[42,43] which is a measure of the ATE for a binary outcome, where the average outcome is the proportion of successes.

In Section 2, we provide more detail on several of these methods. In Section 3, we describe simulation studies that compare the finite sample performance of these methods. In Section 4, we apply the methods to estimate comparative effectiveness of four common therapies for mCRPC patients, using claims data from the Optum Clinformatics Data Mart, with the outcome being admission to the ER within a short time window of treatment initiation. Conclusions and topics for future research are given in Section 5.

2 | NOTATION AND SETUP

2.1 | Estimands of interest

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

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

where the expectation is over the population of interest. When Y is binary, the ATE is the risk difference

$$\tau_{\text{ATE}}(z, z') = pr\{Y(z') = 1\} - pr\{Y(z) = 1\}.$$

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

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

A more general form of ATE is the weighted ATE[14,44]:

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

where $f(\boldsymbol{x})$ is the density function of the covariates $\boldsymbol{X}$ and $w(\boldsymbol{x})$ is a prespecified function of $\boldsymbol{x}$. Different choices of $w(\cdot)$ yield the ATE for different target populations, as discussed further in Section 4.2.

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

2.2 | Assumptions

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

Assumption 0. The individuals in the study are randomly sampled from the population.

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

Assumption 2. (strong unconfoundedness). Assignment to treatment Z is strongly unconfounded if $Z_i \perp\!\!\!\perp (Y_i(1), \cdots, Y_i(J)) \mid \boldsymbol{X}_i$, for all $z \in \{1, \cdots, J\}$.

Assumption 3. (overlap). For all values of z and $\boldsymbol{x}$, $0 < e_z(\boldsymbol{x}) < 1$, where $e_z(\boldsymbol{x}) \equiv pr(Z_i = z|\boldsymbol{x})$ is the GPS.[21]

SUTVA states that the potential outcomes of one unit are not affected by the treatments received by other units, and there are no hidden treatment versions.[45] Strong unconfoundedness and overlap are an extension of the strong ignorability assumption in Rosenbaum and Rubin[10] to the case of multiple treatments. In some cases, a weaker version of unconfoundedness is sufficient for identifying the causal effect,[21,24] namely

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

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

3 | GPS AND ITS ESTIMATION

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

$$\log \frac{pr(Z_i = z|\boldsymbol{X}_i)}{pr(Z_i = J|\boldsymbol{X}_i)} = \boldsymbol{X}_i^T \boldsymbol{\beta}_z, \tag{1}$$

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

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

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

Even moderate misspecification of the functional form for (1) may result in substantial bias in the estimates of treatment effects.[46] Imai and Ratkovic proposed the covariate balancing propensity score (CBPS) for the comparison of two groups and provided an extension to the multiple treatment case.[47] CBPS exploits the covariate balancing property of the GPS (ie, $\boldsymbol{X}_i \perp\!\!\!\perp D_i(z) \mid e_z(\boldsymbol{X}_i)$ for $z = 1, \cdots, J$[21]) by computing generalized method of moments estimates based on the covariate balancing moment conditions,

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

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

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

for $z = 1, \cdots, J$. The CBPS is called just-identified if the model only uses the covariate balancing conditions and overidentified if both conditions are used in the estimation step. These two types of CBPS have different asymptotic and finite sample properties, and the authors examined both types of scores in their simulation studies.[47] They showed that the use of CBPS, regardless of which conditions were involved, can improve the precision and reduce bias of some common weighting estimators (eg, IPW and AIPW) compared with using propensity score estimated by GLM when both propensity score and outcome models were misspecified. In our study, we only evaluate the just-identified CBPS, because of computational limitations. The CBPS method can be implemented through the R package CBPS.[48]

4 | METHODS FOR ESTIMATING THE ATE

4.1 | Matching methods based on the propensity scores

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

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

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

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

The SE can be computed using the delta method.

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

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

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

There are several differences between AI-type matching estimators and traditional matching estimators in applied research, such as nearest neighbor matching without replacement,[28] Traditional matching procedures address the issue of confounding by only including matches of high quality for the subsequent analysis. Normally each unit is only used once, as in a randomized control trial, and inferences on the matched dataset do not account for matching error. On the other hand, AI-type matching allows reuse of each unit, and does not ensure overlap of covariates unless combined with methods for dealing with limited overlap, such as trimming.[49] An advantage of the AI-type matching estimators is that their large-sample distributions can be characterized,[31,50] permitting calculation of variance estimates that take into account the uncertainty in the propensity score estimation and matching procedure. MCOV, MGPSV, and MGPSS estimate $\boldsymbol{\tau}_{\text{ATE}}$ while the estimand of traditional matching procedure may deviate from $\boldsymbol{\tau}_{\text{ATE}}$.

4.2 | Propensity score weighting-based methods

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

$$\sum_{i=1}^{n} \left\{ \frac{w(\boldsymbol{X}_i)D_i(z)(Y_i - \nu_z)}{\widehat{e}_z(\boldsymbol{X}_i)} \right\} = 0, \tag{2}$$

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

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

The ATE between treatment z and z' can then be estimated by $\widehat{\nu}_{z'} - \widehat{\nu}_z$. Different choices of $w(\boldsymbol{x})$ result in ATE with respect to different populations. In particular, $w(\boldsymbol{x}) = 1$ corresponds to the IPW estimator, whose target population is the combined population all sampled groups. The target population of ATT discussed in Section 2.1 is represented by units in a particular treatment group, say treatment J, and can be estimated by setting $w(\boldsymbol{x})$ to $e_J(\boldsymbol{x})$.

Li and Greene[12] proposed to specify $w(\boldsymbol{x})$ as the minimum of the probabilities of receiving treatment and control in the binary case, which they call matching weights (MW). MW can be extended to the case with more than two treatments[13] with weights

$$w_{\text{MW}}(\boldsymbol{x}) = \min(e_1(\boldsymbol{x}), \cdots, e_J(\boldsymbol{x})).$$

For the three treatment case, the MW estimator uses weights to mimic the 1:1:1 matching procedure without replacement and yields more efficient estimation of τ^*_{ATE}.[12,13] The MW estimator and the estimator from 1:1:1 matching without replacement have asymptotically the same estimand,[13] and therefore the corresponding target population of the MW estimator is the "matched" population of units that can be matched in 1:1:1 matching.

Li et al[14] and Li and Li[26] proposed weighting by the OW:

$$w_{\text{OW}}(\boldsymbol{x}) = \left(1 / \sum_{j=1}^{J} e_j(\boldsymbol{x}) \right)^{-1}.$$

We refer to the corresponding population as the overlap population. Both MW and OW upweight the units whose GPS is in the middle range, which have approximately equal chances of being assigned to any of the candidate treatments.

Inversely weighted estimators have a number of issues. The first is that their variance may be inflated if the weights are highly variable. The second issue is that they rely heavily on the correct specification of the propensity score model for valid inference. In addition, the inference for treatment group z is made only based on individuals with $D_i(z) = 1$, with individuals in other treatment groups not contributing. To improve the robustness to model misspecification and make more effective use of the available data, augmented versions of these estimators have been proposed.[12,39] The estimating equation (2) is augmented by an extra term that involves a function of $\boldsymbol{x}$. The resulting estimating equation is

$$\sum_{i=1}^{N}\left\{\frac{w(\boldsymbol{X}_i)D_i(z)(Y_i-\nu_z)}{e_z(\boldsymbol{X}_i)}-\frac{w(\boldsymbol{X}_i)(e_z(\boldsymbol{X}_i)-D_i(z))}{e_z(\boldsymbol{X}_i)}h(\boldsymbol{X}_i)\right\}=0.$$

The resulting estimator $\widehat{\nu}_z$ achieves the smallest asymptotic variance when $h(\boldsymbol{X}_i) = E(Y_i - \boldsymbol{\nu}_z|Z_i = z, \boldsymbol{X}_i)$.[37] We label the augmented versions of IPW, MW, and OW estimators as AIPW, AMW, and AOW, respectively. Besides asymptotic efficiency, as shown in the original set of articles,[12,14] for any scalar outcome, the corresponding estimator has the property of double robustness, which means that only one of the propensity score and outcome models need to be correctly specified to obtain a consistent estimator for $\boldsymbol{\nu}_z$. Semiparametric theory shows that these estimators are asymptotically normal, and variances can be estimated using sandwich-type estimators or the bootstrap.[3,39]

4.3 | OREG model methods

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

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

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

$$\widehat{\tau}_{\mathrm{OREG}}(z,z')=\widehat{\mu}_{z'}-\widehat{\mu}_z,$$

where

$$\widehat{\mu}_z=\frac{1}{n}\sum_{i=1}^{n}\mathrm{expit}(\widehat{\gamma}+\widehat{\theta}_z+\boldsymbol{X}_i^T\widehat{\boldsymbol{\alpha}})$$

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

$$\widehat{\mu}_z=\frac{1}{n}\sum_{i=1}^{n}\mathrm{expit}(\widehat{\gamma}+\boldsymbol{X}_i^T\widehat{\boldsymbol{\alpha}})$$

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

Utilizing this idea, Zhou et al[17] proposed PENCOMP, which estimates causal effects comparing two treatments for a continuous outcome by imputing unobserved potential outcomes from the corresponding predictive distributions.

PENCOMP incorporates splines of propensity scores as predictors in the outcome model, which gives it a double robustness property for a continuous outcome such that the estimator for the marginal mean is consistent if (a) the prediction models are correctly specified or (b) the propensity model and the relationship between the outcome and the splines are correctly specified. We extend PENCOMP at a single time point to more than two treatments and a binary outcome, calling the method PEN-GAM. The double robustness property for PEN-GAM has not yet been theoretically established. However, our simulation studies shed light on its finite sample performance. The steps for PEN-GAM can be summarized as follows:

(a) Generate a bootstrap sample $S^{(b)}$ for $b = 1, \cdots, B$, stratified on treatment groups, from the original dataset. For each $S^{(b)}$, repeat steps (b)-(d).
(b) Estimate the GPS, possibly from a multinomial logistic regression model. Denote the estimated values as $\widehat{\boldsymbol{e}}_i = \{\widehat{e}_1(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}_1^{(b)}), \cdots, \widehat{e}_{J-1}(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}_{J-1}^{(b)})\}^T$, where $\widehat{e}_z(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}_z^{(b)}) = pr(Z = z|X; \widehat{\boldsymbol{\beta}}_z^{(b)})$ and $\widehat{\boldsymbol{\beta}}_z^{(b)}$ is the maximum likelihood estimate of $\boldsymbol{\beta}_z$ for sample $S^{(b)}$. Define $\widehat{\boldsymbol{e}}_i^* = \{\widehat{e}_{i1}^*, \cdots, \widehat{e}_{i(J-1)}^*\}^T$, where $\widehat{e}_{iz}^* = \log\{\widehat{e}_z(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}_z^{(b)})/[1 - \widehat{e}_z(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}}_z^{(b)})]\}$.
(c) For $z = 1, \cdots, J$, fit a generalized linear regression model

$$\log \frac{pr(Y_i(z) = 1|Z_i = z, \boldsymbol{X}_i, \boldsymbol{\theta}_z, \boldsymbol{\alpha}_z)}{pr(Y_i(z) = 0|Z_i = z, \boldsymbol{X}_i, \boldsymbol{\theta}_z, \boldsymbol{\alpha}_z)} = s(\widehat{\boldsymbol{e}}_i^*|\boldsymbol{\theta}_z) + g(\boldsymbol{X}_i, \widehat{\boldsymbol{e}}_i^*; \boldsymbol{\alpha}_z) \tag{3}$$

where $s(\widehat{\boldsymbol{e}}_i^*|\boldsymbol{\theta}_z)$ denotes a penalized spline with fixed knots, and $g(\cdot)$ denotes a parametric function of the covariates and propensity scores and has to be constrained to ensure identifiability. In this case, we assume truncated linear basis, namely, $s(\widehat{\boldsymbol{e}}_i^*|\boldsymbol{\theta}_z) = \sum_{z=1}^{J-1} \left\{ \theta_{0z} + \theta_{1j}\widehat{e}_{iz}^* + \sum_{k=1}^{K} \theta_{1zk}(\widehat{e}_{iz}^* - Q_k)_+ \right\}$, where $Q_1, \cdots, Q_K$ are fixed knots, and $(\widehat{e}_{iz}^* - Q_k)_+ = \widehat{e}_{iz}^* - Q_k$ if $\widehat{e}_{iz}^* > Q_k$, and $(\widehat{e}_{iz}^* - Q_k)_+ = 0$ otherwise. Note that following Zhou et al,[17] we fit different spline functions in (3) for each treatment level z. For linear regression of $Y_i(j)$, the coefficients in the spline model can be estimated in a linear mixed model framework[51] and implemented using standard statistical software, as was done in Zhou et al[17] In principal, the coefficients of a generalized linear model with penalized spline terms as (3) can be obtained by fitting a generalized linear mixed models (GLMM). However, to the best of our knowledge, current GLMM implementation in R either does not allow the specification of the structure of the covariance matrices or will take unreasonable running time. Therefore, we instead fit a generalized additive model (GAM) using the gam function in the *mgcv* package in R.[52]

(d) For $z = 1, \cdots, J$, impute the values of $Y(z)$ for subjects with $D(z) = 0$ in the original dataset with draws from the Bernoulli distribution with predictive probability $pr(Y_i(z) = 1|Z_i = z, \boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_z^{(b)}, \widehat{\boldsymbol{\alpha}}_z^{(b)})$, where $\widehat{\boldsymbol{\theta}}_z^{(b)}$ and $\widehat{\boldsymbol{\alpha}}_z^{(b)}$ are estimates for the coefficients $\boldsymbol{\theta}_z^{(b)}$ and $\boldsymbol{\alpha}_z^{(b)}$, respectively, for the bth bootstrap replicate. For subjects with $D_i(z) = 1$, $Y_i(z) = Y_i$. Denote the estimates of treatment effects and associated pooled variances as $\widehat{\tau}^{(b)}$ and $\widehat{v}^{(b)}$, respectively.
(e) Derive the estimated treatment effects and associated SE using Rubin's Rules.[53]

For all methods discussed in this section, we refer the readers to the corresponding R packages developed by the authors (Table 1). In the cases where there are no R packages available, we provide accessible code for easier implementation at https://github.com/youfeiyu/multiTreatment.

5 | SIMULATION STUDIES

We conducted simulation studies to assess the finite sample properties of the 12 estimators listed in Table 1 combined with the two GPS estimation methods (GLMPS and CBPS) discussed in Section 3. We used direct comparison of the proportions of each group as a benchmark, which is referred to as the naive estimator. We considered two levels of covariate overlap (good and poor), two functional forms for the true propensity score model (linear and nonlinear in covariates), two levels of associations for the outcome model (strong and weak), two levels of overall marginal outcome prevalence (common [0.3] and rare [0.1]) and two sample sizes (300 and 1500). Simulation results are presented in terms of bias from ATE, empirical SD, average SE, root mean squared error (RMSE), average width of 95% confidence intervals (CIs), and 95% coverage rate.

TABLE 1 Causal inference methods under comparison and their corresponding R implementation

Method	Reference	R package/author generated code	Which unconfoundedness assumption is made[a]
NAIVE[b]	N/A	https://github.com/youfeiyu/multiTreatment	N/A
OREG	N/A	https://github.com/youfeiyu/multiTreatment	Assumption 2*
PENCOMP[c]	Zhou et al[17]	https://github.com/youfeiyu/multiTreatment	Assumption 2
Propensity score matching			
MCOV	Abadie and Imbens[31]	Matching,[54] Matchit[55,56]	Assumption 2
MGPSV	Yang et al[24]	https://github.com/youfeiyu/multiTrea	Assumption 2
MGPSS	Yang et al[24]	Multilevelmatching (https://github.com/shuyang1987/multilevelMatching/)	Assumption 2*
Propensity score weighting			
IPW, AIPW	Lunceford and Davidian,[3] among others	https://github.com/youfeiyu/multiTreatment	Assumption 2*
MW, AMW	Li and Greene,[12] Yoshida et al[13]	https://github.com/youfeiyu/multiTrea	Assumption 2*
OW, AOW	Li and Li[26]	PSweight,[57] or https://github.com/youfeiyu/multiTreatment	Assumption 2*

Abbreviations: AIPW, augmented inverse probability weighting; IPW, inverse probability weighting; MW, matching weights; OREG, outcome regression; PENCOMP, propensity methods for treatment comparison.
[a] Assumptions 2 and 2* are the strong and weak unconfoundedness assumption, respectively.
[b] NAIVE estimator refers to the direct comparison of the proportions of each treatment group.
[c] The authors developed PENCOMP in the context of binary treatment and continuous outcome. We extend it to the case of multiple treatment and binary outcome.

5.1 | Simulation design

Each simulated dataset contains six covariates. $(X_{i1}, X_{i2}, X_{i3})^T$ follows a multivariate normal distribution with mean $(0,\ 0,\ 0)^T$ and covariance matrix $[(2,\ 1, -1)^T,\ (1,\ 1, -0.5)^T,\ (-1, -0.5,\ 1)^T]$, $X_{i4}\sim$Bernoulli(0.5), $X_{i5}\sim$Bernoulli$(0.75X_{i4} + 0.25(1 - X_{i4}))$, and X_{i6} follows a chi-squared distribution with 1° of freedom. Let $\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})^T$. Three treatment groups are compared, and the true GPS model is given by

$$Z_i \sim \text{Multinomial}(e_1(\widetilde{\boldsymbol{X}}_i), e_2(\widetilde{\boldsymbol{X}}_i), e_3(\widetilde{\boldsymbol{X}}_i)),$$

where $\widetilde{\boldsymbol{X}}_i$ is a function of $\boldsymbol{X}_i$ that corresponds to a model specification and $e_z(\widetilde{\boldsymbol{X}}_i) = \exp(\widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta}_z) / \sum_{j=1}^{3} \exp(\widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta}_j)$. The potential outcome $Y_i(z)$ was sampled from a binomial distribution with probability $pr\{Y_i(z)|\boldsymbol{X}_i\} = \text{expit}(\boldsymbol{X}_i^T \boldsymbol{\alpha}_z)$. We considered five scenarios (Supplemental Table 1) with different specifications of GPS and outcome models:

(1) $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i$, good covariate overlap, weak outcome-covariate associations, and common outcome.
(2) $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i$, poor covariate overlap, weak outcome-covariate associations, and common outcome.
(3) $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i$, poor covariate overlap, strong outcome-covariate associations, and common outcome.
(4) $\widetilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i, X_{i2}^2, X_{i1} \times X_{i3})^T$, poor covariate overlap, weak outcome-covariate associations, and common outcome.
(5) $\widetilde{\boldsymbol{X}}_i = \boldsymbol{X}_i$, poor covariate overlap, weak outcome-covariate associations, and rare outcome.

The GPS were estimated in two ways, the first using a multinomial logistic regression and the second using the CBPS framework that incorporates covariate balancing conditions.[47] Since PENCOMP is computationally intensive, we only

implemented GLMPS (not CBPS) for this method. We used 10 equally spaced knots on the logit scale for each GPS component. We used 200 imputed datasets to estimate treatment effects and the associated standard errors and CIs.

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

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

(1) both GPS and outcome models are correctly specified denoted by (c, c),
(2) the GPS model is correct while the outcome model is incorrect denoted by (c, m),
(3) the outcome model is correct while the GPS model is incorrect denoted by (m, c),
(4) both models are misspecified denoted by (m, m).

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

The 95% CIs were calculated using: (1) bootstrapped standard errors from 200 bootstrap samples for OREG, IPW, AIPW, MW, AMW, OW, AOW, and CBPS-based MGPSS; (2) Wald-type CI based on original data for NAIVE; (3) Abadie and Imbens[24,31] CI for MCOV and both GLMPS- and CBPS-based MGPSV; (4) Abadie and Imbens[24,50] CI for GLMPS-based MGPSS; (5) Rubin's[17] imputation rule for PEN-GAM.

5.2 | Simulation results

The main results of the simulation studies for sample size 1500 are summarized in Figures 1 to 7. The complete results are presented in Supplemental Tables 2 to 8 for sample size 1500, and Supplemental Figures 1 to 7 and Supplemental Tables 9 to 15 for sample size 300. In all scenarios, all estimators for $\boldsymbol{\tau}_{\text{ATE}}$ with at least one model correctly specified yielded smaller empirical bias compared with the naive estimator.

Three key takeaways from the simulation studies are summarized below:

1. The improvement in precision was limited for AIPW and PEN-GAM compared with IPW when (a) there was sufficient covariate overlap or (b) the prevalence of the outcome was low.
2. With moderate prevalence of the outcome (0.3 in our simulation setting) or relatively poor covariate overlap, AIPW and PEN-GAM outperformed IPW and AI-type matching algorithms considered in this study in terms of RMSE across the scenarios, as AIPW and PEN-GAM incorporate the outcome information, which tended to provide efficiency gains over IPW and AI-type matching.
3. For a relatively small sample size, PEN-GAM with at least one model being correctly specified were noted to be slightly biased away from the true risk difference. Moreover, PEN-GAM tended to show overcoverage and produce wider confidence width than IPW when the outcome is sparse. One reason is that the fitting of spline models in PEN-GAM is more unstable with low outcome prevalence and small sample size. The empirical bias and overcoverage tended to disappear as the outcome prevalence and sample size increased.

Results of RMSE for each of the treatment comparisons averaged over 2000 datasets for sample size 1500 across all methods that estimate $\boldsymbol{\tau}_{\text{ATE}}$ are presented in Figures 1 to 3. Note that the corresponding estimands for MW, AMW, OW, and AOW were in general different from $\boldsymbol{\tau}_{\text{ATE}}$, and the RMSE for these estimators are shown in Supplemental Tables 2 and 9. We report the ratio of RMSE to the RMSE of the GLMPS-based IPW estimator with correctly specified GPS model. When both models were correctly specified and the overlap in covariate distributions was good (Figure 1, scenario 1), OREG, IPW, AIPW, and PEN-GAM had similar RMSE. Matching methods had larger RMSE than GLMPS-based IPW,

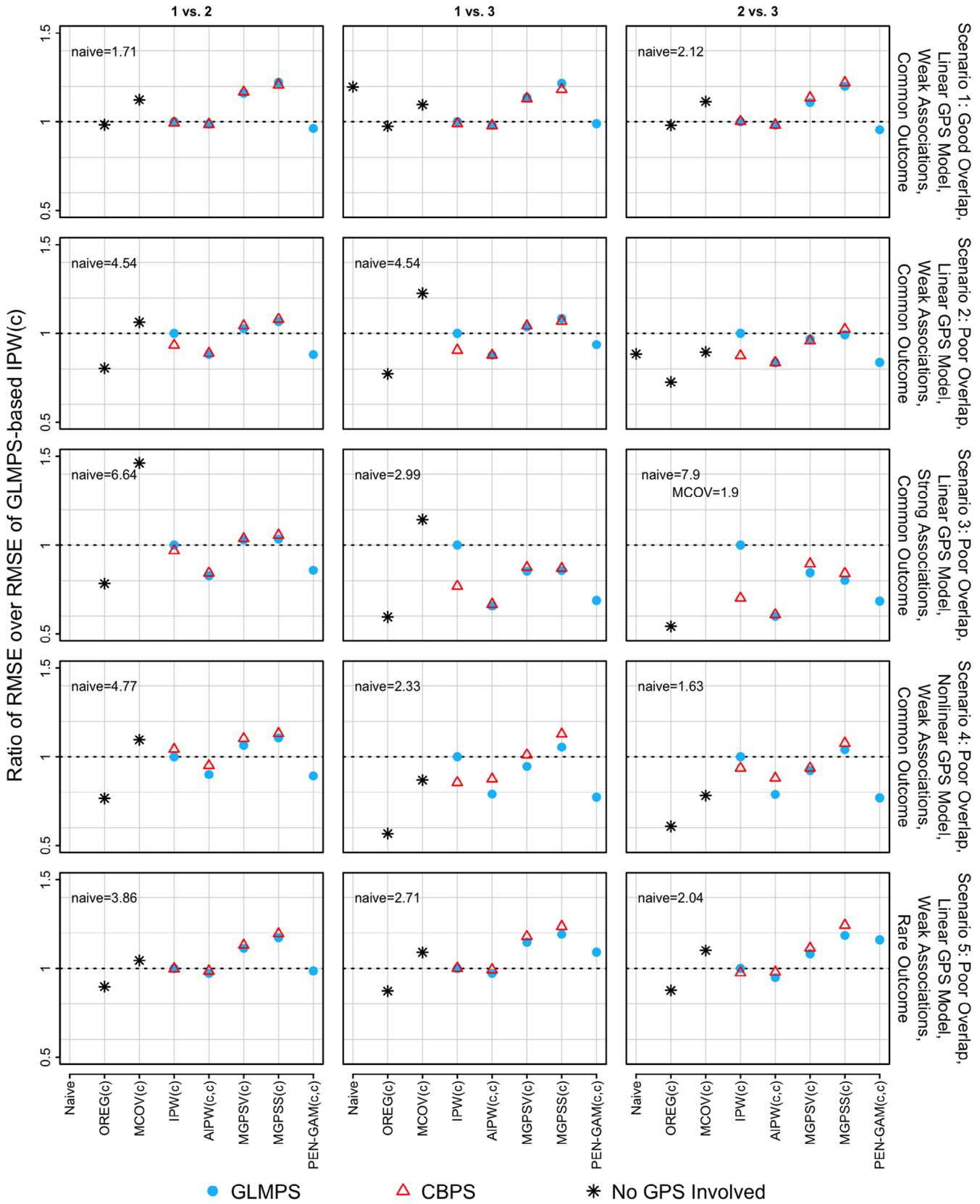

FIGURE 1 Ratio of RMSE over RMSE of GLMPS-based IPW(c) for sample size 1500 across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets. IPW, inverse probability weighting; RMSE, root mean squared error [Colour figure can be viewed at wileyonlinelibrary.com]

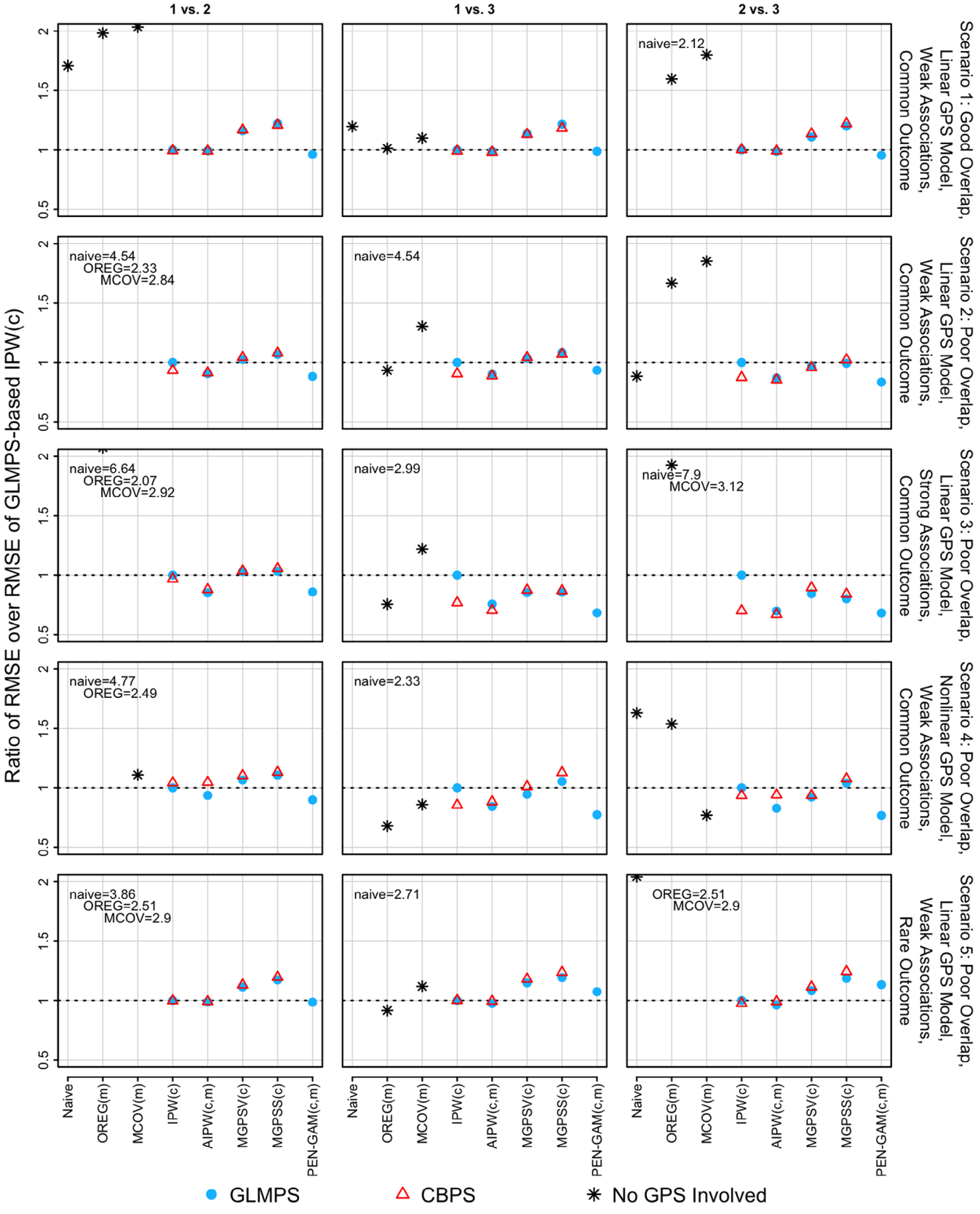

FIGURE 2 Ratio of RMSE over RMSE of GLMPS-based IPW(c) for sample size 1500 across methods based on a correctly specified propensity model only. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets. IPW, inverse probability weighting; RMSE, root mean squared error [Colour figure can be viewed at wileyonlinelibrary.com]

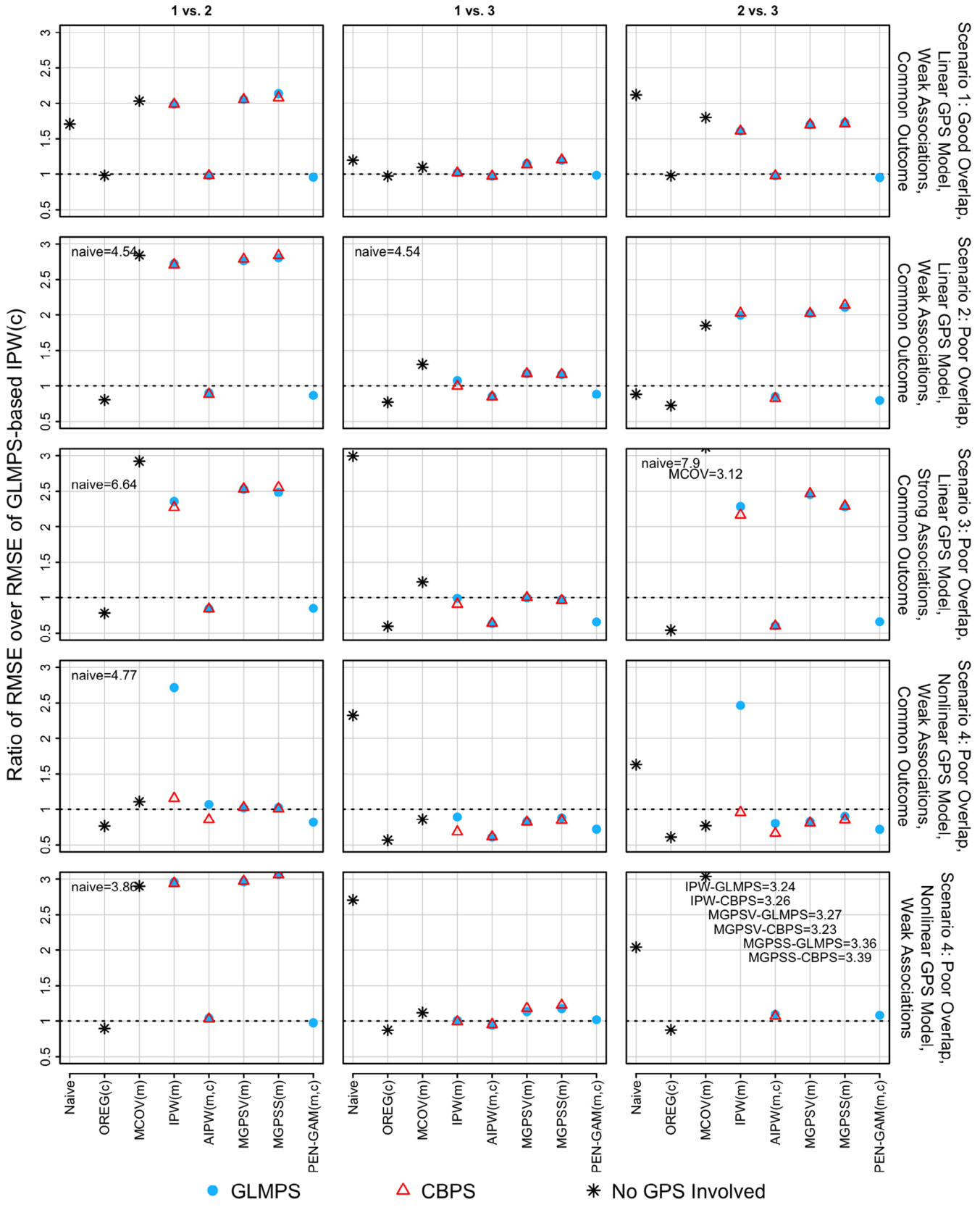

FIGURE 3 Ratio of RMSE over RMSE of GLMPS-based IPW(c) for sample size 1500 across methods based on a correctly specified outcome model only. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets. IPW, inverse probability weighting; RMSE, root mean squared error [Colour figure can be viewed at wileyonlinelibrary.com]

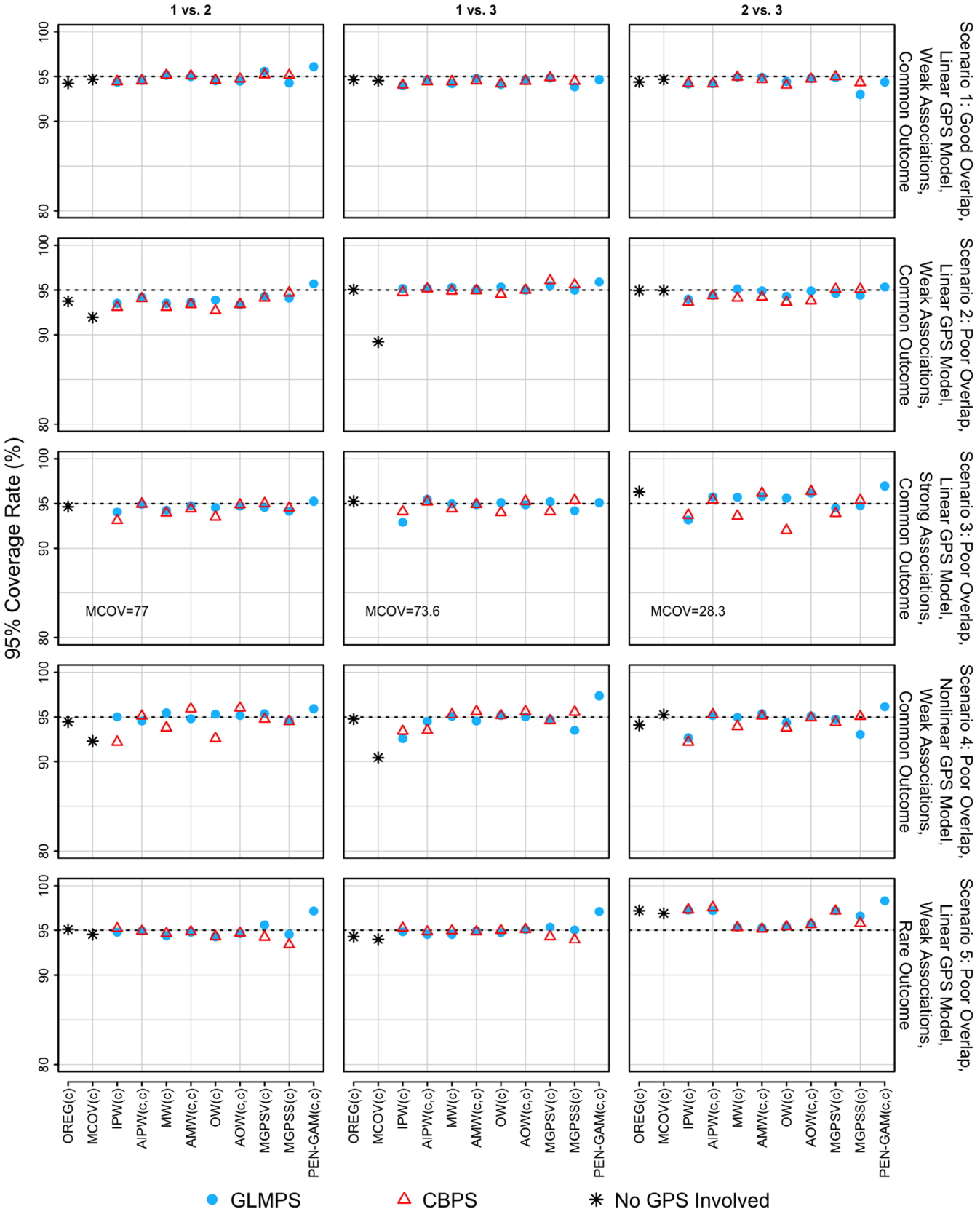

FIGURE 4 95% Coverage probability for sample size 1500 across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets [Colour figure can be viewed at wileyonlinelibrary.com]

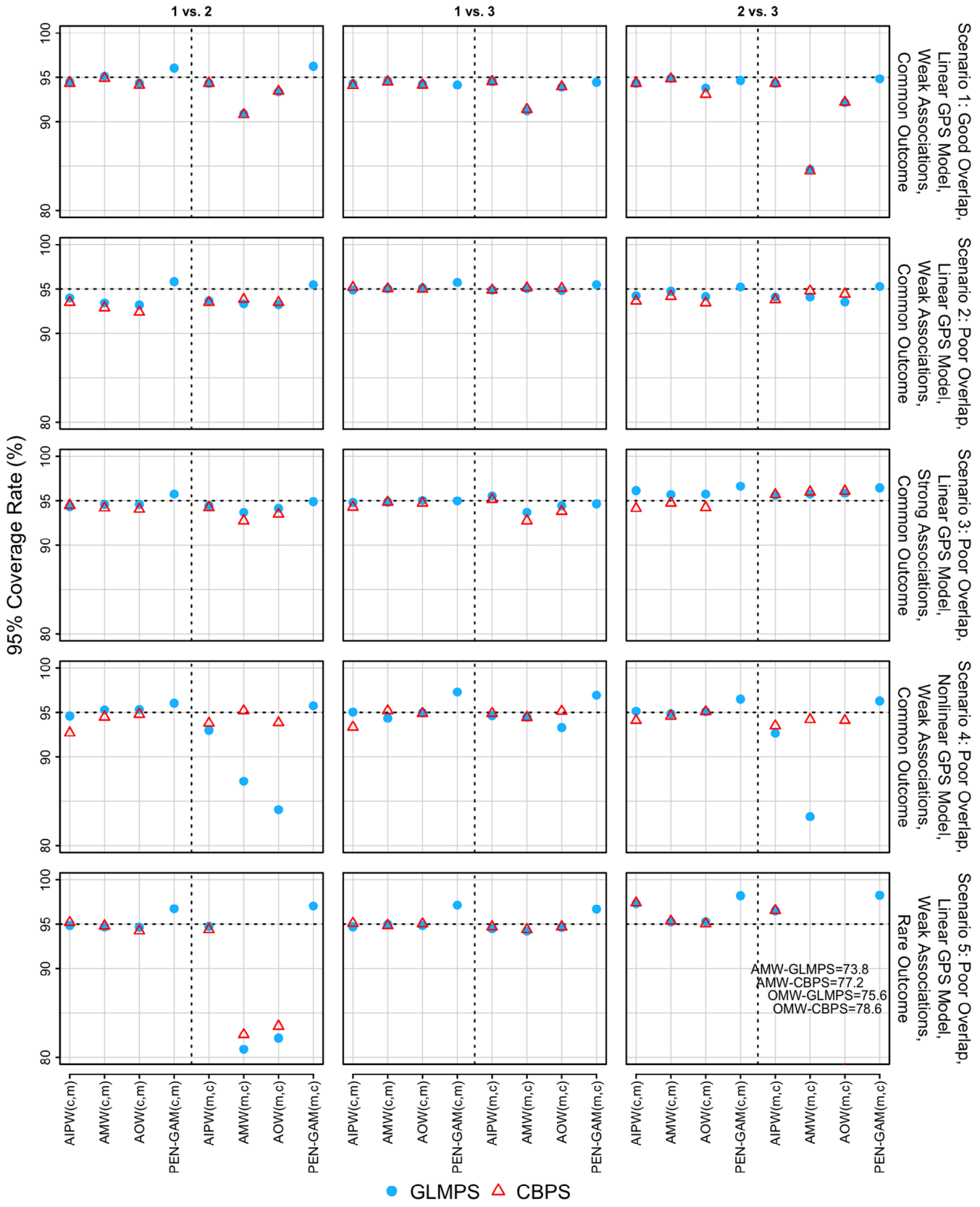

FIGURE 5 95% Coverage probability for sample size 1500 across methods based on a correctly specified propensity score or outcome model. For methods that involve both models, the first and second letter in the parentheses correspond to the propensity and outcome model, respectively. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets [Colour figure can be viewed at wileyonlinelibrary.com]

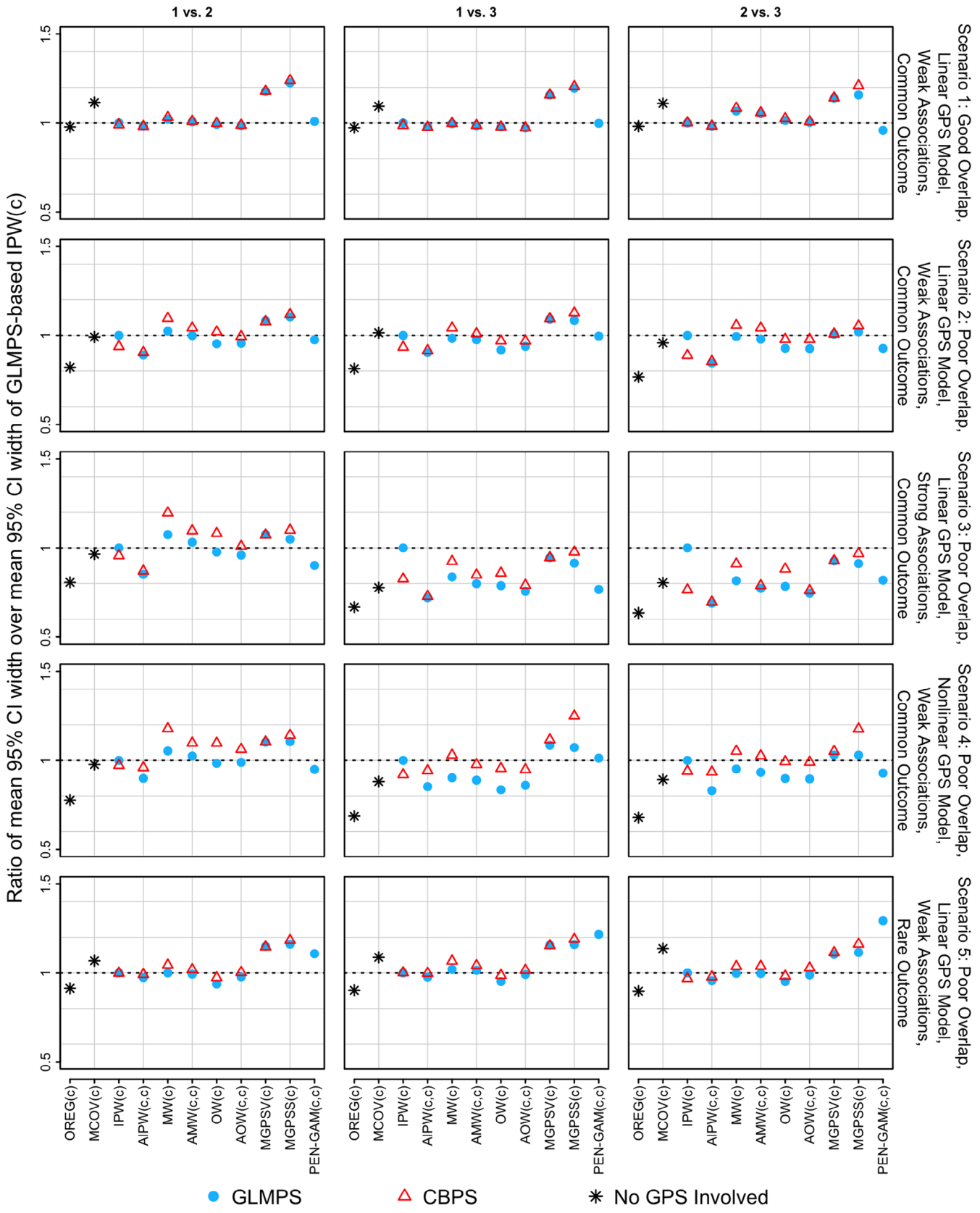

FIGURE 6 Ratio of mean 95% CI width over mean 95% CI width of GLMPS-based IPW(c) for sample size 1500 across methods based on correctly specified outcome and propensity models. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets. CI, confidence interval; IPW, inverse probability weighting [Colour figure can be viewed at wileyonlinelibrary.com]

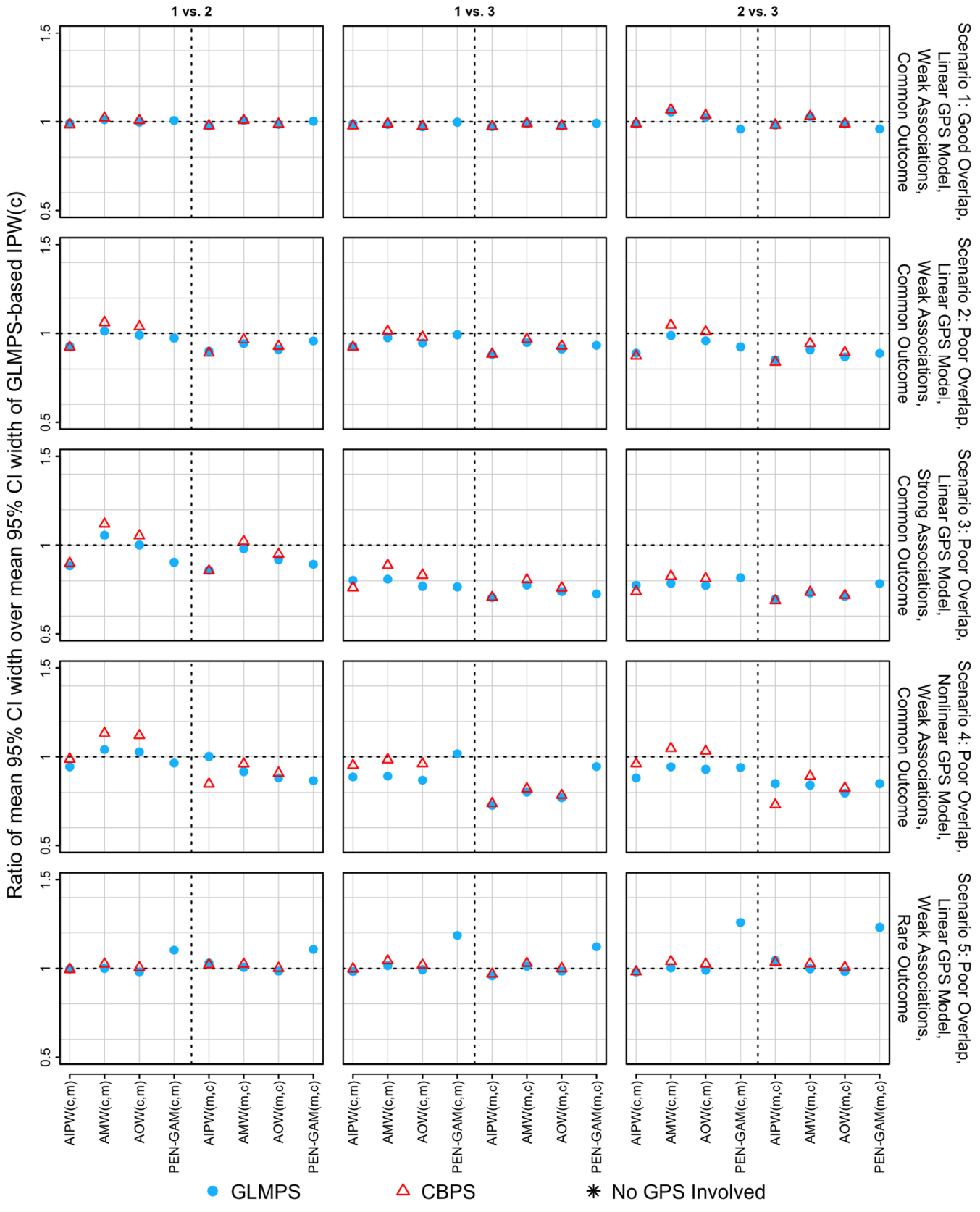

FIGURE 7 Ratio of mean 95% CI width over mean 95% CI width of GLMPS-based IPW(c) for sample size 1500 across methods based on a correctly specified propensity score or outcome model. For methods that involve both models, the first and second letter in the parentheses correspond to the propensity and outcome model, respectively. The rows represent scenarios and columns represent pairs of comparison. Results were obtained using 2000 simulated datasets. CI, confidence interval; IPW, inverse probability weighting [Colour figure can be viewed at wileyonlinelibrary.com]

with the ratios ranging from 1.1 to 1.2. In this case, AIPW and PEN-GAM had similar empirical SD (and therefore RMSE) to IPW (Supplemental Table 3). A study conducted by Austin showed similar results that AIPW provided little efficiency gain over IPW.[58]

In the presence of poor covariate overlap (Figure 1, scenarios 2-4), OREG had the smallest RMSE, followed by PEN-GAM and AIPW. We observed 6.3% to 16.5% reduction in RMSE for AIPW and PEN-GAM compared with GLMPS-based IPW when the associations between the outcome and covariates was weak (scenario 2). Greater reduction (14.1%-40.1%) was noted as the associations became stronger (scenario 3). When the prevalence of the outcome was low (scenario 5), AIPW barely reduced RMSE compared with IPW, and PEN-GAM had larger RMSE than IPW. The increased RMSE for PEN-GAM may result from the instability of model fitting with low prevalence. MGPSS had larger RMSE than MGPSV, which was also observed for the scenario with good covariate overlap. For all scenarios considered, RMSEs of GLMPS-based estimators were close to those of their CBPS-based counterparts (Figure 1 and Supplemental Table 2). One exception is that for IPW, the use of CBPS tended to reduce RMSE compared with GLMPS when the covariate overlap was poor.

When only the GPS model was correctly specified (Figure 2), PEN-GAM and AIPW in general had the lowest RMSEs across the scenarios with moderate prevalence, and the RMSEs for PEN-GAM were close to or lower than those for AIPW. When only the outcome was modeled correctly (Figure 3), the RMSEs for AIPW and PEN-GAM remained similar to or lower than those for IPW with correctly specified GPS model. In scenario 4 where the misspecification of the GPS model was caused by incorrect functional form, the use of GLMPS may lead to substantial RMSE for IPW (Figure 3) and AIPW with misspecified outcome model (Supplemental Table 6) due to large empirical bias, which is consistent with previous findings.[46,47] The bias was greatly reduced and became close to zero when GLMPS were replaced by CBPS with misspecified functional form, which leaded to smaller RMSEs. The RMSEs of the AI-type matching methods (MGPSS and MGPSV) were noted to be smaller than those of GLMPS-based IPW in scenario 4, since the matching methods yielded approximately unbiased estimates of ATE (Supplemental Table 6) even when the GPS model was incorrect but adjusted for the whole set of confounders, which indicates that matching methods are more robust to the omission of higher order and interaction terms in the GPS model than IPW.

The empirical coverage rates of 95% CI for sample sizes 1500 with both models correctly specified and either one of the models misspecified are shown in Figures 4 and 5, respectively. The true values for MW, AMW, OW, and AOW were determined using the true GPS based on 10^6 sample units and used to evaluate the corresponding coverage rates. In general, when both models were correctly specified (Figure 4), all methods except MCOV had close to nominal coverage of 95% for moderate prevalence. Coverage for MCOV was far below nominal in scenarios 2 and 3 with moderate and strong confounding, respectively. This undercoverage was primarily the result of empirical bias (Supplemental Tables 4 and 5).

With the outcome model being misspecified (Figure 5), all of the augmented estimators showed reasonable coverage. Note that the corresponding estimands of MW, OW, and their augmented versions depend on the actual values of GPS. Therefore, different specifications of GPS model lead to different estimands, while the estimands based on the true GPS model were used for evaluating the coverage rates, which explains the undercoverage of AMW and AOW in some scenarios when the GPS model was misspecified (Figure 5). For a small sample size ($n = 300$) or sparse outcome (scenario 5), we consistently observed overcoverage for PEN-GAM methods across all scenarios regardless of the specifications of the models, with some of the CIs achieving 99% coverage (Supplemental Figures 4 and 5, and scenario 5 in Figures 4 and 5). This finding agrees with the overestimation of the standard errors for PEN-GAM observed in Supplemental Tables 7 and 10-14. The undercoverage for GLMPS-based MGPSS in scenario 3 (Supplemental Figure 4) was caused by the underestimation of the standard errors using the asymptotic formula provided in Yang et al.[24] Such undercoverage was remedied as the sample size increased.

The average 95% CI widths for sample size 1500 are shown in Figures 6 and 7. When both models were correctly specified (Figure 6), the average widths of OREG, AIPW, and PEN-GAM were close to or smaller than those of GLMPS-based IPW for common outcome. MGPSS and MGPSV tended to have wider CIs than IPW across all scenarios. The average widths of CBPS-based estimators tended to be larger than those of their corresponding GLMPS-based ones. Figure 7 displays the results for the augmented estimators with either one of the models being misspecified. The relative relationships among IPW, AIPW, and PEN-GAM were similar to the ones in Figure 6 where both models were correct. In general, for all estimators considered in Figure 7, the CIs were wider when the outcome model was misspecified compared with the case with a misspecified GPS model only. For $n = 300$, the CIs for PEN-GAM were in general wider than those of IPW (Supplemental Figures 6-7). The average SEs of PEN-GAM were greater than their corresponding Monte Carlo standard deviations for all scenarios (Supplemental Tables 10-14), suggesting that PEN-GAM tends to be more sensitive to small sample size in terms of SE estimation compared with IPW and AIPW.

MW, OW estimators and their augmented version provide stable estimates of τ^*_{ATE}, regardless of the overlap status in the covariate distribution of the original population (Supplemental Tables 3-7 and 10-14). This is as expected since MW and OW artificially downweight the units with extreme GPS and upweight the units whose GPS for each treatment are similar, the latter of which tend to have a common support in their covariate distribution.

6 | DATA ANALYSIS

6.1 | Data analysis methods

We applied the methods in Table 1 to claims data of patients with mCRPC, which was obtained from a large national private health insurance network (Optum Clinformatic Data Mart). Our data consisted of a subset of a previously identified cohort,[59-61] which included patients who had at least one diagnosis of prostate cancer from January 1, 2010 to September 30, 2016 and used at least one of the six focus drugs (docetaxel, abiraterone, enzalutamide, sipuleucel-T, cabazitaxel, and radium-233) after the diagnosis. Since radium-233 were approved by FDA and released to the market later than the other five drugs, we restricted our cohort to patients who initiated treatment after January 1, 2014 to give them a fair comparison and make the results more generalizable to the current mCRPC population. We observed that the cabazitaxel and radium-233 groups had much fewer samples ($n_{\text{cabazitaxel}} = 11$ and $n_{\text{radium}} = 57$) than the other four groups, and therefore we further dropped those patients who received the two drugs as their first-lines therapy from our analysis. We assessed the safety of the four remaining drugs for mCRPC with the outcome being the occurrence of postprescription ER visits during a fixed period of time. Specifically, we evaluated the risk difference of ER visits among the four drugs within 180-day time window of the initiation of each therapy.

Medical and pharmacy claims pertaining to ER visits were identified by procedure code and type of service variables in the database. In this study, we did not consider treatment sequence and hence were only interested in ER visits associated with the first drug used. Patients who switched treatment or dropped out of the insurance plan within 180 days of the first prescription with no events (ie, ER visits) occurring during the follow-up period were regarded as being censored. Censored patients exhibited similar demographic and baseline clinical characteristics (Supplemental Table 16) to uncensored ones and were dropped from the analysis. We first calculated the crude risks of at least one ER visit for 180-day follow-up for each of the four focus drugs, and compared the risk among the four treatment groups using causal inference methods described in the previous section.

The GPS for each subject was estimated from a multinomial logistic regression model adjusting for age, race, education level, household income, geographic region, insurance product type, whether the insurance plan is administrative services only, metastatic status of cancer, year of first prescription, comorbid conditions, and provider type. All covariates were binary or categorical, and the categorization was summarized in Supplemental Table 17. We observed insufficient overlap among the four treatment groups in terms of the logit propensity of receiving docetaxel, especially at the left end of the distribution (Supplemental Figure 8A), which indicates that we may not be able to find a good match in docetaxel users for some patients receiving abiraterone, enzalutamide, or sipuleucel-T. Similar patterns occurred for the logit propensity of receiving the other three drugs (Supplemental Figure 8C,E,G). One can use trimming methods that discard the tails of propensity score distributions to remedy the lack of overlap. Several trimming criteria for three or more treatment groups are discussed in the literature.[24,25,62] In our case, we trimmed the data using the criteria described in.[25] In brief, for each treatment $z \in \{1, 2, 3, 4\}$, let $l_z = \max_j\{\min_i(pr(Z_i = z|Z_i = j, \boldsymbol{X}_i))\}$ and $u_z = \min_j\{\max_i(pr(Z_i = z|Z_i = j, \boldsymbol{X}_i))\}$, where $pr(Z = z|Z = j, \boldsymbol{X})$ is the treatment assignment probability for z among those receiving treatment j. Subjects with $e_z(\boldsymbol{x}) \notin [l_z, u_z]$ for any z were discarded. GPS were recalculated using the remaining subjects. One important step in propensity score modeling is balance checking. Ways to check for balance in covariates and their corresponding results for the methods considered are described in supplemental section 1. The log odds of the outcome was modeled as a linear combination of the same set of covariates adjusted in the GPS model for each treatment group. The CIs for each method were obtained in the same way as described in the simulation studies. Specifically, 200 bootstrap replicates were used for OREG, PEN-GAM, and all weighting-based methods.

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

TABLE 2 Emergency room visits following the first prescription (N = 2628)

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

Abbreviation: ER, emergency room.

[a] Percentage was calculated using the number uncensored patients as the denominator.

6.2 | Data analysis results

A total of 2628 mCRPC patients with at least 180 days of continuous enrollment prior to the receipt of the first focus drug were identified. The average and median length of the enrollment period that covers January 1, 2014 is 6.16 and 4.75 years, respectively. Among the 2628 patients, 670 (25.5%) were censored and 4 (0.2%) had incomplete covariates. We further excluded these patients from the analysis. The demographic and baseline clinical data of the remaining 1955 patients are presented in Supplemental Table 17. Table 2 presents the crude risks of at least one ER visit during 180-day follow-up among uncensored patients for each of the four treatment groups. The unadjusted risk was the highest in the docetaxel group (51.5%), followed by Sipuleucel-T group (44.3%). Enzalutamide users had the lowest risk (25.5%) of at least one ER visit within 180 days.

We observed imbalance in some of the covariates (Supplemental Tables 1.1 and 17). For example, patients who received abiraterone or enzalutamide tend to be older than those receiving docetaxel. Sipuleucel-T users tend to have more pre-treatment osteoporosis (16.0%) than patients receiving the other three drugs (5.3% for docetaxel, 8.4% for abiraterone, and 9.0% for enzalutamide).

To improve the covariate overlap among the treatment groups, we applied data trimming[25] with criteria discussed previously, which left us with 1777 subjects. Results of data analysis are presented in Figure 8 and Supplemental Table 18. Direct comparison of the four groups (naive method) revealed that docetaxel users had significantly higher risk of at least one ER visits within 180 days of follow-up than users of abiraterone (risk difference = 0.130 [0.073, 0.186]), enzalutamide (risk difference = 0.177 [0.115, 0.239]), and sipuleucel-T (risk difference = 0.099 [0.001, 0.197]). The directions of the average effects between docetaxel and the other drugs were preserved for the other methods, though the effect sizes varied. The 95% CIs for the average causal effects between docetaxel and enzalutamide consistently excluded 0 for all methods. However, for the Sipuleucel-T-docetaxel comparison, only MCOV showed a significant difference. For the enzalutamide-abiraterone comparison, all methods considered indicated a higher risk for enzalutamide, while none of these estimated risk differences were significant. For the sipuleucel-T-abiraterone comparison, PEN-GAM yielded negative point estimates (indicating higher risk for abiraterone), while the other methods indicated a reversed relationship. Again, none of the corresponding CIs excluded 0. In general, there was a larger uncertainty in regard to the direction and magnitude of the risk differences that involve the Sipuleucel-T group due to its smaller sample size. Notably, PEN-GAM tended to have wider CIs than the other methods, which was consistent with the simulation results for small sample size. The results of MW, AMW, OW, and AOW were close to one another in terms of point estimates as well as standard errors for all pairwise comparisons, possibly because their corresponding target populations were similar. This finding aligns with what was observed in the simulation studies. The results of our data analysis agree well with the clinical evidence in current literature.[59-61] The naive method yielded results that were highly consistent with those of the methods that adjust for potential confounding, suggesting that the treatment effects were relatively strong compared with the confounding effects.

7 | DISCUSSION

This article has reviewed and compared a set of causal inference strategies that account for confounding for multiple treatment comparison with a binary outcome variable. Some of these methods, for example, MGPSS[24] and PENCOMP,[17]

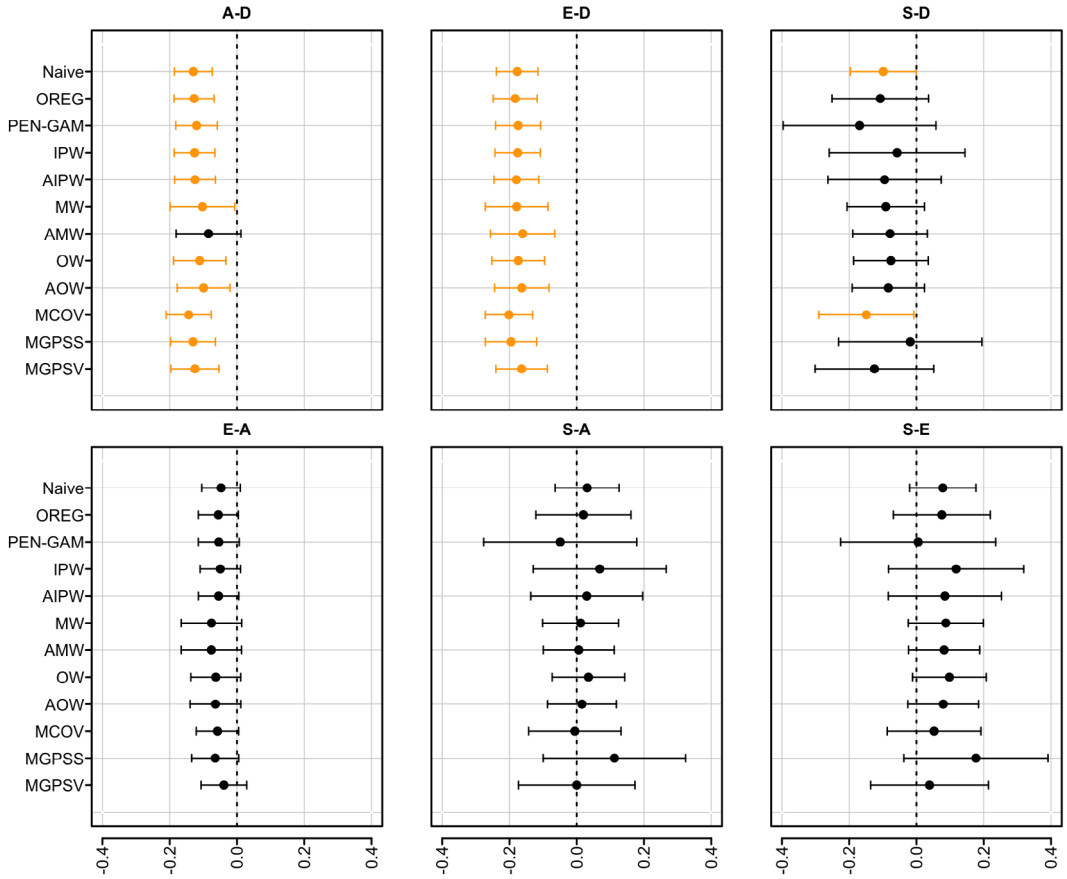

FIGURE 8 Results for treatment effects of the four focus drugs and associated 95% confidence intervals. Data were obtained from Optum Clinformative Data Mart, with the outcome interest being the occurrence of emergency room visit within 180 days of treatment initiation. Total sample size is N = 1777 (NA = 699, ND = 519, NE = 438, NS = 121). Confidence intervals that exclude zero are highlighted in orange. Abbreviations: A, abiraterone; D, docetaxel; E, enzalutamide; S, sipuleucel-T [Colour figure can be viewed at wileyonlinelibrary.com]

were recently proposed and less explored under the setting of binary outcome in current literature. Our simulation studies show that when there is sufficient overlap in covariate distributions, MGPSS, and in general all AI-type matching methods, are less efficient than the conventional IPW estimator. The gain in precision of AIPW over IPW that has been observed for continuous outcomes[3,63] was less evident in our simulations for a binary outcome and good covariate overlap. Thus, while augmentation was still useful for the robustness of estimating the causal effect, it was less useful for improving efficiency. When there was lack of common support, PEN-GAM and AIPW provided more precise estimation than IPW. The improvement in precision increased as the associations of the outcome with baseline covariates became stronger. With moderate outcome prevalence, PEN-GAM tended to perform better than AIPW in terms of RMSE when only the propensity model was correctly specified. One possible reason was that when the covariate overlap is poor, the weights tend to have large variations and some individuals may receive extreme weights, which results in highly variable estimates. PEN-GAM avoids weights by adjusting for the splines of propensity scores (in logit scale) in the outcome model. When the outcome model was misspecified, the estimates relied more on the use of propensity scores. On the other hand, when the outcome was sparse, the fitting of the spline models tended to be unstable, which leads to larger RMSE for PEN-GAM than AIPW.

For propensity score-based methods, correctly modeling the propensity scores is the key to yielding valid inference. The generalized linear model based on maximum likelihood (GLMPS) is sensitive to both unmeasured confounders and misspecified functional form, which tend to lead to large bias in ATE estimation. Efforts have been made to improve the robustness of propensity score estimation and the CBPS, which utilizes the covariate balancing property of the propensity scores and achieves robustness in the presence of incorrect functional forms, in one of the examples.[47] In particular, when the GPS model has misspecified functional form but adjusts for the whole set of confounders, the use of CBPS can reduce

the bias of the ATE estimates compared with using GLMPS. In addition to CBPS, methods based on machine learning technique have also been proposed for propensity score estimation.[64]

Our focus in this article has remained on simple parametric models. There is extensive literature on using machine learning methods[65-67] to capture potential nonlinearities and higher-order interactions. The relative gain by using such flexible methods depends on the sample size, the number of predictors, and the true structure of the underlying models (the propensity model or the outcome model).

The computational time for each of the methods considered in the simulation studies for a sample size of 1500 and three treatment groups is reported in Supplemental Table 19. All simulations were run on an Intel Xeon Gold 6138 Processor (2.00 GHz). The average run time of overidentified CBPS was almost twice as much as that of just-identified CBPS. The average run time of PEN-GAM for one bootstrap replicate was around 2 seconds. The projected computational time for 200 bootstrap replicates is approximately 7 minutes.

The methods examined in this study only accounts for the selection bias associated with differences in the covariates. However, the outcome of the data we used is also subject to censoring, which may introduce another layer of selection bias. In particular, approximately 30% of the patients in our dataset were censored due to treatment switch or dropout within 180 days of treatment initiation. Weighting-based methods have been proposed to achieve unbiased estimation of average causal effect in the presence of right-censored observations under certain assumptions.[68-70]

ACKNOWLEDGEMENTS

The research of BM was supported by NSF DMS 1712933 and NCI grant CA 046591. The research of MC was funded by a Cancer Center Population Science Career Development Award and a Prostate Cancer Foundation Young Investigator Award.

DATA AVAILABILITY STATEMENT

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

ORCID

Youfei Yu https://orcid.org/0000-0002-4986-848X
Min Zhang https://orcid.org/0000-0003-3331-3583
Xu Shi https://orcid.org/0000-0001-8566-9552
Bhramar Mukherjee https://orcid.org/0000-0003-0118-4561

REFERENCES

1. Sox HC. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med*. 2009;151(3):203-205. https://doi.org/10.7326/0003-4819-151-3-200908040-00125.
2. Office of the Commissioner. *Real-World Evidence*. Silver Spring, MD: FDA; 2019. http://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Accessed September 28, 2019.
3. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*. 2004;23(19):2937-2960. https://doi.org/10.1002/sim.1903.
4. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biom J*. 2009;51(1):171-184. https://doi.org/10.1002/bimj.200810488.
5. Kim SY, Solomon DH. Use of administrative claims data for comparative effectiveness research of rheumatoid arthritis treatments. *Arthritis Res Ther*. 2011;13(5):129. https://doi.org/10.1186/ar3472.
6. Tannock IF, de Wit R, Berry WR, et al. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *N Engl J Med*. 2004;351(15):1502-1512. https://doi.org/10.1056/NEJMoa040720.
7. Kantoff PW, Higano CS, Shore ND, et al. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. *N Engl J Med*. 2010;363(5):411-422. https://doi.org/10.1056/NEJMoa1001294.
8. de Bono JS, Logothetis CJ, Molina A, et al. Abiraterone and increased survival in metastatic prostate cancer. *N Engl J Med*. 2011;364(21):1995-2005. https://doi.org/10.1056/NEJMoa1014618.
9. Scher HI, Fizazi K, Saad F, et al. Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med*. 2012;367(13):1187-1197. https://doi.org/10.1056/NEJMoa1207506.
10. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. https://doi.org/10.1093/biomet/70.1.41.

11. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*. 1985;39(1):33-38. https://doi.org/10.1080/00031305.1985.10479383.
12. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*. 2013;9(2):215-234. https://doi.org/10.1515/ijb-2012-0030.
13. Yoshida K, Hernández-Díaz S, Solomon DH, et al. Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. *Epidemiol Camb Mass*. 2017;28(3):387-395. https://doi.org/10.1097/EDE.0000000000000627.
14. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2018;113(521):390-400. https://doi.org/10.1080/01621459.2016.1260466.
15. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79(387):516-524. https://doi.org/10.2307/2288398.
16. Vansteelandt S, Daniel RM. On regression adjustment for the propensity score. *Stat Med*. 2014;33(23):4053-4072. https://doi.org/10.1002/sim.6207.
17. Zhou T, Elliott MR, Little RJA. Penalized spline of propensity methods for treatment comparison. *J Am Stat Assoc*. 2019;114(525):1-19. https://doi.org/10.1080/01621459.2018.1518234.
18. Borah BJ, Moriarty JP, Crown WH, Doshi JA. Applications of propensity score methods in observational comparative effectiveness and safety research: where have we come and where should we go? *J Comp Eff Res*. 2013;3(1):63-78. https://doi.org/10.2217/cer.13.89.
19. Laliberté F, Cloutier M, Nelson WW, et al. Real-world comparative effectiveness and safety of rivaroxaban and warfarin in nonvalvular atrial fibrillation patients. *Curr Med Res Opin*. 2014;30(7):1317-1325. https://doi.org/10.1185/03007995.2014.907140.
20. Sooriakumaran P, Nyberg T, Akre O, et al. Comparative effectiveness of radical prostatectomy and radiotherapy in prostate cancer: observational study of mortality outcomes. *BMJ*. 2014;g1502:348. https://doi.org/10.1136/bmj.g1502.
21. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika*. 2000;87(3):706-710. https://doi.org/10.1093/biomet/87.3.706.
22. Imai K, van Dyk DA. Causal inference with general treatment regimes. *J Am Stat Assoc*. 2004;99(467):854-866. https://doi.org/10.1198/016214504000001187.
23. Feng P, Zhou X-H, Zou Q-M, Fan M-Y, Li X-S. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat Med*. 2012;31(7):681-697. https://doi.org/10.1002/sim.4168.
24. Yang S, Imbens GW, Cui Z, Faries DE, Kadziola Z. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*. 2016;72(4):1055-1065. https://doi.org/10.1111/biom.12505.
25. Lopez MJ, Gutman R. Estimation of causal effects with multiple treatments: a review and new ideas. *Stat Sci*. 2017;32(3):432-454. https://doi.org/10.1214/17-STS612.
26. Li F, Li F. Propensity score weighting for causal inference with multiple treatments. *Ann Appl Stat*. 2019;13(4):2389-2415. https://doi.org/10.1214/19-AOAS1282.
27. Pearl J. The foundations of causal inference. *Sociol Methodol*. 2010;40(1):75-149. https://doi.org/10.1111/j.1467-9531.2010.01228.x.
28. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci Rev J Inst Math Stat*. 2010;25(1):1-21. https://doi.org/10.1214/09-STS313.
29. Rassen JA, Solomon DH, Glynn RJ, Schneeweiss S. Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design.". *Pharmacoepidemiol Drug Saf*. 2011;20(7):675-683. https://doi.org/10.1002/pds.2121.
30. Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. *Epidemiol Camb Mass*. 2013;24(3):401-409. https://doi.org/10.1097/EDE.0b013e318289dedf.
31. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica*. 2006;74(1):235-267. https://doi.org/10.1111/j.1468-0262.2006.00655.x.
32. Cui ZL, Hess LM, Goodloe R, Faries D. Application and comparison of generalized propensity score matching versus pairwise propensity score matching. *J Comp Eff Res*. 2018;7(9):923-934. https://doi.org/10.2217/cer-2018-0030.
33. He X, Wang Y, Cong H, Lu C, Wu J. Impact of optimal medical therapy at discharge on 1-year direct medical costs in patients with acute coronary syndromes: a retrospective, observational database analysis in China. *Clin Ther*. 2019;41:456-465.e2. https://doi.org/10.1016/j.clinthera.2019.01.005.
34. Gupta K, Trocio J, Keshishian A, et al. Real-world comparative effectiveness, safety, and health care costs of oral anticoagulants in nonvalvular atrial fibrillation patients in the U.S. Department of Defense Population. *J Manag Care Spec Pharm*. 2018;24(11):1116-1127. https://doi.org/10.18553/jmcp.2018.17488.
35. Shirvani SM, Jiang J, Chang JY, et al. Comparative effectiveness of 5 treatment strategies for early-stage non-small cell lung cancer in the elderly. *Int J Radiat Oncol*. 2012;84(5):1060-1070. https://doi.org/10.1016/j.ijrobp.2012.07.2354.
36. Mauri L, Silbaugh TS, Garg P, et al. Drug-eluting or bare-metal stents for acute myocardial infarction. *N Engl J Med*. 2008;359(13):1330-1342. https://doi.org/10.1056/NEJMoa0801485.
37. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866. https://doi.org/10.1080/01621459.1994.10476818.
38. Stefanski LA, Boos DD. The calculus of M-estimation. *Am Stat*. 2002;56(1):29-38. https://doi.org/10.1198/000313002753631330.
39. Mao H, Li L, Greene T. Propensity score weighting analysis and treatment effect discovery. *Stat Methods Med Res*. 2018;28:2439-2454. https://doi.org/10.1177/0962280218781171.

40. Joffe MM, Ten Have TR, Feldman HI, Kimmel SE. Model selection, confounder control, and marginal structural models. *Am Stat*. 2004;58(4):272-279. https://doi.org/10.1198/000313004X5824.
41. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701. https://doi.org/10.1037/h0037350.
42. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310(6977):452-454.
43. North D. Number needed to treat. Absolute risk reduction may be easier for patients to understand. *BMJ*. 1995;310(6989):1269.
44. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*. 2003;71(4):1161-1189. https://doi.org/10.1111/1468-0262.00442.
45. Rubin DB. Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591-593. https://doi.org/10.2307/2287653.
46. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4):523-539. https://doi.org/10.1214/07-STS227.
47. Imai K, Ratkovic M. Covariate balancing propensity score. *J R Stat Soc Ser B Stat Methodol*. 2014;76(1):243-263. https://doi.org/10.1111/rssb.12027.
48. Fong C, Ratkovic M, Imai K, Hazlett C, Yang X, Peng S. CBPS*:* Covariate Balancing Propensity Score; 2019. https://CRAN.R-project.org/package=CBPS. Accessed April 1, 2019.
49. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199. https://doi.org/10.1093/biomet/asn055.
50. Abadie A, Imbens GW. Matching on the estimated propensity score. *Econometrica*. 2016;84(2):781-807. https://doi.org/10.3982/ECTA11293.
51. Wand MP. Smoothing and mixed models. *Comput Stat*. 2003;18(2):223-249. https://doi.org/10.1007/s001800300142.
52. Wood S. Mgcv: mixed GAM computation vehicle with automatic smoothness estimation; 2019. https://CRAN.R-project.org/package=mgcv. Accessed January 27, 2020.
53. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. 1st ed. New York, NY: John Wiley & Sons, Ltd, 1987. doi:https://doi.org/10.1002/9780470316696
54. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J Stat Softw*. 2011;42(1):1-52. https://doi.org/10.18637/jss.v042.i07.
55. Ho D, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42(1):1-28. https://doi.org/10.18637/jss.v042.i08.
56. Ho D, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal*. 2007;15:199-236.
57. Zhou T, Tong G, Li F, Thomas L, Li F. *PSweight:* propensity score weighting for causal inference with observational studies and randomized trials; 2020. https://CRAN.R-project.org/package=PSweight. Accessed September 20, 2020.
58. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*. 2010;29(20):2137-2148. https://doi.org/10.1002/sim.3854.
59. Caram MEV, Wang S, Tsao P, et al. Patient and provider variables associated with variation in the systemic treatment of advanced prostate cancer. *Urol Pract*. 2019;6:234-242. https://www.auajournals.org/doi/abs/10.1097/UPJ.0000000000000020.
60. Caram MEV, Estes JP, Griggs JJ, Lin P, Mukherjee B. Temporal and geographic variation in the systemic treatment of advanced prostate cancer. *BMC Cancer*. 2018;18(1):258. https://doi.org/10.1186/s12885-018-4166-3.
61. Caram MEV, Ross R, Lin P, Mukherjee B. Factors associated with use of Sipuleucel-T to treat patients with advanced prostate cancer. *JAMA Netw Open*. 2019;2(4):e192589. https://doi.org/10.1001/jamanetworkopen.2019.2589.
62. Yoshida K, Solomon DH, Haneuse S, et al. Multinomial extension of propensity score trimming methods: a simulation study. *Am J Epidemiol*. 2019;188(3):609-616. https://doi.org/10.1093/aje/kwy263.
63. Robins J, Sued M, Lei-Gomez Q, Rotnitzky A. Comment: performance of double-robust estimators when "inverse probability" weights are highly variable. *Stat Sci*. 2007;22(4):544-559. https://doi.org/10.1214/07-STS227D.
64. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med*. 2013;32(19):3388-3414. https://doi.org/10.1002/sim.5753.
65. Hu L, Gu C, Lopez M, Ji J, Wisnivesky J. Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Stat Methods Med Res*. May 2020;29:3218-3234. https://doi.org/10.1177/0962280220921909.
66. McConnell KJ, Lindner S. Estimating treatment effects with machine learning. *Health Serv Res*. 2019;54(6):1273-1282. https://doi.org/10.1111/1475-6773.13212.
67. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185(1):65-73. https://doi.org/10.1093/aje/kww165.
68. Wang X, Beste LA, Maier MM, Zhou X-H. Double robust estimator of average causal treatment effect for censored medical cost data. *Stat Med*. 2016;35(18):3101-3116. https://doi.org/10.1002/sim.6876.
69. Anstrom KJ, Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. *Biometrics*. 2001;57(4):1207-1218. https://doi.org/10.1111/j.0006-341X.2001.01207.x.

70. Zhang M, Schaubel DE. Estimating differences in restricted mean lifetime using observational data subject to dependent censoring. *Biometrics*. 2011;67(3):740-749. https://doi.org/10.1111/j.1541-0420.2010.01503.x.

SUPPORTING INFORMATION

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

How to cite this article: Yu Y, Zhang M, Shi X, Caram MEV, Little RJA, Mukherjee B. A comparison of parametric propensity score-based methods for causal inference with multiple treatments and a binary outcome. *Statistics in Medicine*. 2021;40:1653–1677. https://doi.org/10.1002/sim.8862