Scientific Research Publishing

International Journal of Clinical Medicine, 2026, 17(1), 1-14
https://www.scirp.org/journal/ijcm
ISSN Online: 2158-2882
ISSN Print: 2158-284X

Comparative Performance of Propensity Score Methods for Clinical Multi-Group Data: Balancing Confounders and Estimating Treatment Effects

Xinlan Teng[1], Jiayu Chen[1], Ying Guan[1*], Kailiang Shen[1], Wenyi Lai[1], Emmanuel Nizeyimana[2], Teaway Angeline Patience[2], Eric Hu[3]

[1]Department of Biostatistics, School of Public Health, Southern Medical University, Guangzhou, China
[2]School of International Education, Southern Medical University, Guangzhou, China
[3]Guangdong Experimental High School, Guangzhou, China
Email: *guanying@smu.edu.cn

How to cite this paper: Teng, X.L., Chen, J.Y., Guan, Y., Shen, K.L., Lai, W.Y., Nizeyimana, E., Patience, T.A. and Hu, E. (2026) Comparative Performance of Propensity Score Methods for Clinical Multi-Group Data: Balancing Confounders and Estimating Treatment Effects. *International Journal of Clinical Medicine*, **17**, 1-14.
https://doi.org/10.4236/ijcm.2026.171001

Received: December 7, 2024
Accepted: January 2, 2026
Published: January 5, 2026

Abstract

Background: Propensity score methods have become a cornerstone of modern causal inference, enabling researchers to approximate the conditions of randomized experiments in observational studies. Despite their widespread adoption, most established propensity score approaches were originally developed for two-group comparisons, leaving a notable methodological gap for multi-group data commonly encountered in clinical trials, public health interventions, and comparative effectiveness research. **Methods:** We conducted a comparative evaluation of several propensity score methods in balancing confounders and estimating treatment effects using Monte Carlo simulation. Datasets of varying sample sizes were generated under two distinct hybrid data-generating structures. Propensity scores were estimated using both generalized linear models (GLM) and generalized boosting models (GBM), and were subsequently applied via inverse probability of treatment weighting (IPTW), overlap weighting (OW), and matching. Five specific method combinations were evaluated: GLM-IPTW, GLM-OW, GLM-matching, GBM-IPTW, and GBM-OW. Covariate balance was assessed using standardized mean differences (SMD), while treatment effect estimation performance was evaluated based on point estimate accuracy and root mean square error (RMSE). **Results:** Across simulation scenarios with both linear and non-linear underlying relationships, the GLM-matching approach generally outperformed other methods. GLM-OW and GBM-OW demonstrated superior performance in achieving covariate balance, while GLM-IPTW and GBM-IPTW yielded more accurate point

estimates of the treatment effect. **Conclusion:** When the relationship between covariates and outcome is relatively simple and treatment assignment follows a linear model, the GLM-matching method proved particularly advantageous. It produced estimates closer to the true value and exhibited a stronger ability to balance covariates compared to the other methods considered.

Keywords

Multi-Group Data, Generalized Boosting Model, Generalized Linear Model, Overlap Weighting

1. Background

Both observational studies and randomized controlled trials (RCTs) are essential methodologies in medical research. In RCTs, randomization is typically employed to assign participants to experimental or control groups, thereby minimizing confounding factors and ensuring balanced distribution of known and unknown confounders across groups in order to enable unbiased estimation of treatment effects. In contrast, observational studies are susceptible to selection bias due to the non-random allocation of subjects. To address the above issue, propensity score (PS) methods have emerged as a widely adopted approach for confounding adjustment in real-world studies [1]. By balancing the distributions of observed covariates across treatment groups, PS techniques reduce selection bias and strengthen the validity of effect estimates.

Propensity score methods have become a cornerstone in modern causal inference, enabling researchers to approximate randomized experiment conditions in observational settings. Several methods have been developed for propensity score estimation, such as generalized boosted models (GBM) and Generalized Linear Models (GLM) [2]. Once estimated, propensity scores can be incorporated into analyses through matching, weighting, or stratification to balance covariates across groups. Common implementations involve estimating propensity scores using either GBM or GLM, with the subsequent application of balancing techniques, such as Inverse Probability of Treatment Weighting (IPTW), Overlap Weighting (OW), or matching. Consequently, five distinct methodologies can be utilized: GBM with IPTW (GBM-IPTW), GBM with OW (GBM-OW), GLM with IPTW (GLM-IPTW), GLM with OW (GLM-OW), and GLM with matching (GLM-matching). Despite their widespread adoption, most established PS methodologies were tailored to two-group comparisons, leaving a significant gap for multi-group data commonly encountered in clinical trials, public health interventions, and comparative effectiveness research.

This study aims to address these methodological gaps by conducting a comprehensive comparison of five PS approaches in multi-group contexts. We will evaluate their relative performance in achieving covariate balance and producing accurate treatment effect estimates under varying conditions, including different

sample sizes, functional forms of covariate-outcome relationships, and model specifications. These findings will contribute to the growing methodological literature on causal inference and provide evidence-based recommendations for applied researchers.

2. Methodology

2.1. Generalized Augmented Model (GBM)

The GBM estimates the propensity score $p(x)$ by iteratively forming a collection of simple regression tree models and then combining them into a robust model [3].

$$g(x) = \log\left(p(x)/\left(1-p(x)\right)\right) \tag{1}$$

In the GBM model (Formula (1)), initial propensity scores are estimated using a logistic regression model to estimate the initial propensity scores $g(x)$, and these initial propensity scores are used as the starting point for the GBM. Then, the adjustment function $h(x)$ is found to add to $g(x)$. In each iteration, the GBM model trains a new $h(x)$ in this regression tree based on the current propensity score, observes the data, and adds it to the model. This process continues several times until a predetermined number of iterations is reached or a convergence condition is met [4]. In each iteration, the GBM weights the propensity score to update it based on the performance of the previous model and other factors. This process can help the model better adapt to the characteristics of the data and improve the accuracy of the estimation. The GBM is the algorithm for estimating the greatest likelihood of the logarithmic dominance of the propensity score values $g(x)$ [5].

2.2. Generalized Linear Model (GLM)

The Generalized Linear Model (GLM) can be modeled in various ways to estimate propensity scores, and we used a polynomial logistic regression model in this study. When modeling the GLM, appropriate covariates need to be selected to predict the probability that an individual will be assigned to a treatment group, conditional on the specified covariates:

$$e_j(x) = Pr(j \mid x) \tag{2}$$

where x is the specified covariate and $j = \{1, 2, \ldots, J\}$ represents the different treatment groups ($J > 2$) [6]. A GLM is used to fit the data, with the fitting process determining the model parameters by maximizing the likelihood function or minimizing the loss function to obtain optimal parameter estimates [7]. Ordinary least squares (OLS) is generally used, or a machine algorithm is employed to find the optimal form of a weighted combination of several prediction algorithms through data-driven cross-validation [8], where the prediction algorithms include both parametric and nonparametric forms.

2.3. Weighted Methods

Weighting is the use of propensity scores to calculate weighting coefficients to weight the study population through different weighting methods in the study. The objects are assigned weights to make the confounders of the two groups converge. The weights are calculated using the formula [9].

$$W = \frac{h(x)}{A \times e(x) + (1-A) \times (1-e(x))} \tag{3}$$

The inverse probability weighted target population includes all study subjects across different treatment groups, with the inverse probability weight being the inverse of the propensity score of the respective treatment groups, using $h(x) = 1$ in the weighting formula. The overlap weighting method, proposed by Li [10], defines the overlap weighting target population as a group with similar covariate characteristics. After identifying the overlap region between treatment groups, the density or probability of the propensity scores within this overlap region is used as the weight to adjust for individual subjects. Overlap weighting increases the relative weight (*i.e.*, the ratio of individual sample weights to the overall weight) of the overlapping portion of the propensity score distribution [11].

2.4. Matching Method

Matching methods include nearest neighbor matching, caliper matching, and Mahalanobis distance matching. We used caliper matching in our study. In caliper matching, the difference in propensity scores of research subjects within each treatment group is matched within a predefined range, with this predefined range referred to as the caliper value [12]. Starting with the first subject in the first group, the subject from the second group with the closest propensity score is selected. If there is more than one subject in the second group with a propensity score identical to that of the first group's subject, one is randomly chosen for matching. After matching the first two groups, the differences between the propensity score values of the third group's research subjects and those of the first two groups are calculated separately, and the subject with the smallest sum of these differences is selected for matching [13]. The caliper value is set during the matching process to control the maximum allowable matching distance, thereby ensuring the accuracy of the matching.

2.5. Balance Analysis

Propensity score balance analysis is designed to assess whether the distributions of baseline characteristics are similar between the matched treatment and control groups, thereby ensuring the reliability and validity of subsequent comparisons. Statistically, this involves testing whether the differences in baseline characteristics between groups are significant. A common metric for this assessment is the Standardized Mean Difference (SMD) [14], which quantifies the magnitude of the difference between two group means, standardized by the pooled standard deviation. This standardization allows for the comparison of balance across variables

measured on different scales. The SMD is calculated using the following formula:

$$\mathrm{SMD} = \frac{\bar{X}_1 - \bar{X}}{\mathrm{SD}_{\mathrm{pooled}}} \tag{4}$$

where $\bar{X}_1$ and $\bar{X}$ correspond to the mean of group 1 and the overall mean, respectively, and $\mathrm{SD}_{\mathrm{pooled}}$ is the pooled standard deviation. A reduction in the SMD value reflects improved covariate balance. When extending this evaluation to multi-group data, the maximum absolute value of the SMD from all possible group pairs for a given covariate is used as the summary metric. This strategy is grounded in the logic that demonstrating balance between the most dissimilar groups provides sufficient evidence to infer balance among all groups, thereby justifying the pooling of data for subsequent analysis [15].

3. Simulation Study Design

A comprehensive simulation study was designed to evaluate and compare the performance of the five propensity score methodologies—GBM-IPTW, GBM-OW, GLM-IPTW, GLM-OW, and GLM-matching—under a range of conditions typically encountered in clinical research. The study was structured around the following key components:

3.1. Data Generation Mechanism

Multi-group clinical trial data were simulated with varying sample sizes (e.g., from 500 to 5000) to assess performance in both small-sample and large-sample settings. Baseline covariates were generated to include both continuous and binary variables. Crucially, the relationship between these covariates and the hypothetical outcome variable was manipulated to span two primary scenarios: a simple linear relationship and a more complex nonlinear relationship.

Firstly, six total sample sizes were established for the simulation: 500, 1000, 2000, 3000, 4000, and 5000. For each sample, an unordered trichotomous grouping variable T_j (j = 1, 2, 3) was randomly generated. The sample allocation ratios for Groups 1, 2, and 3 were set as *pertr1*, *pertr2*, and 1-*pertr1*-*pertr2*, respectively. In this study, these ratios were fixed at 1:2:7. Corresponding dummy variables for the three groups were then included in the subsequent models.

$$j = \begin{cases} 1 & x_1 = 1, x_2 = 0 \\ 2 & x_1 = 0, x_2 = 1 \\ 3 & x_1 = 0, x_2 = 0 \end{cases}$$

Second, to emulate a clinical setting where baseline characteristics influence treatment assignment [16], we simulated eight covariates with imbalanced distributions across the treatment groups. This set comprised four continuous variables generated from normal distributions and four binary variables from binomial distributions. Furthermore, the continuous outcome variable was generated to depend on both these covariates and the treatment assignments. To comprehensively evaluate methodological performance, we specifically designed two distinct

simulation scenarios.

Scene 1:

$$y = 1.6 + c_1 \times 1.5 + c_2 \times 2 + c_3 \times 3 + c_4 \times 4 + b_1 \times 4 + b_2 \times 3 + b_3 \times 2 + b_4 \times 1.5 + x_1 + x_2 + \text{ƀ}$$

Scene 2:

$$y = 0.5 \times c_1 \times c_2 + c_2 \times c_4 + 1.8 \times c_3^2 + c_4 \times c_3 + b_1 \times 4 + b_2 \times 3 + b_3 \times 2 + b_4 \times 1.5 + x_1 + x_2 + \mathcal{E}$$

The error term $\mathcal{E}$ follows a normal distribution with a mean of 0 and a standard deviation of 2. Two distinct simulation scenarios were defined based on the relationship between the outcome and covariates: Scenario 1 assumed a simple linear relationship, whereas Scenario 2 incorporated a complex nonlinear relationship. Using the simulated dataset, the grouping variable was treated as the response, with all eight covariates included as independent variables in the propensity score model. Confounder adjustment was performed using each of the five propensity score methods. For the GLM-matching approach, 1:1:1 caliper matching without replacement was implemented, with a caliper width set to 0.1 [17]. The performance of alternative caliper widths was not explored and could be considered in future work. All simulations and analyses were conducted in the R software environment. To ensure robust performance evaluation, the entire data generation and analysis process was repeated 3000 times for each scenario, enabling a precise comparison of the five methods in handling multi-group propensity scores.

3.2. Performance Metrics

Following successful covariate balance via propensity score adjustment, the analysis proceeds analogously to a randomized controlled trial, wherein treatment effects are estimated by comparing the weighted means of the outcome across groups. To evaluate the performance of the five methods, we employed two primary classes of metrics. The first assessed covariate balance, quantified by the maximum absolute Standardized Mean Difference (SMD) across all pairwise group comparisons for each covariate. An SMD below the threshold of 0.1 was considered indicative of adequate balance [18]. The second assessed estimation accuracy, primarily measured using the Root Mean Square Error (RMSE) of the treatment effect estimate [19], as provided by the R software output. The RMSE, in conjunction with the point estimate, was used to determine each method's ability to recover the true, pre-specified parameter value and to evaluate their relative strengths and weaknesses.

4. Results

4.1. Comparison of Effect Estimates of PS Methods under Scenario 1

Under a simple linear confounding structure (Scenario 1), all five methods pro-

duced similar Root Mean Square Error (RMSE) values. In terms of RMSE, GLM-matching exhibited the smallest variability in effect estimates across sample sizes, followed by GLM-OW and GBM-OW, while GLM-IPTW and GBM-IPTW showed notably higher variability (Table 1). Similarly, with respect to point estimation accuracy, GLM-matching provided the most precise estimates, with GLM-OW and GBM-OW also performing reasonably well, and GLM-IPTW and GBM-IPTW demonstrating larger deviations from the true value (Table 2). As sample size increased, the RMSE consistently decreased across all methods, and their point estimates progressively converged toward the true parameter value.

Table 1. RMSE of the five propensity scores under Scenario 1 (linear relationship only).

Sample size	500	1000	2000	3000	4000	5000
GLM-OW	0.08949	0.06322	0.04474	0.03654	0.03163	0.02830
GLM-IPTW	0.08980	0.06336	0.04479	0.03657	0.03165	0.02831
GLM-matching	0.08860	0.06293	0.04464	0.03649	0.03159	0.02828
GBM-OW	0.08958	0.06324	0.04474	0.03654	0.03163	0.02830
GBM-IPTW	0.08981	0.06335	0.04478	0.03656	0.03164	0.02831

Table 2. Point estimation of the five propensity scores under Scenario 1 (linear relationship only).

Sample size	500	1000	2000	3000	4000	5000
GLM-OW	1.1494	1.1051	1.0884	1.0733	1.0722	1.0753
GLM-IPTW	1.2178	1.1420	1.1170	1.0993	1.0957	1.1012
GLM-matching	1.1464	1.1066	1.0899	1.0752	1.0736	1.0766
GBM-OW	1.1579	1.1075	1.0902	1.0748	1.0739	1.0765
GBM-IPTW	1.2076	1.1378	1.1111	1.0965	1.0934	1.0971

Both evaluation metrics lead to a consistent conclusion: under a simple linear confounding structure, GLM-matching performed best, closely followed by GLM-OW and GBM-OW, whereas GLM-IPTW and GBM-IPTW were less accurate. All methods benefited from increased sample size, showing improved estimation accuracy as sample size grew (see Table 1 and Table 2).

4.2. Comparison of Effect Estimates of PS Methods under Scenario 2

Under a nonlinear relationship between covariates and the outcome variable (Scenario 2), GLM-matching, GLM-IPTW, and GBM-IPTW exhibited similar variability in effect estimates across all sample sizes, which was consistently lower than that of GLM-OW and GBM-OW (Table 3). As the sample size increased, the RMSE values of all five methods progressively declined. In terms of point estimation accuracy, GLM-matching outperformed the other methods under small sam-

ple sizes, followed by GLM-OW and GBM-OW, while GLM-IPTW and GBM-IPTW showed considerable deviation from the true treatment effect (Table 4). However, as the sample size grew, the accuracy of GLM-IPTW and GBM-IPTW improved substantially. At a sample size of 5000, GLM-IPTW achieved the highest precision, followed by GLM-OW and GLM-matching, whereas GBM-IPTW and GBM-OW displayed the largest deviations. Overall, increasing the sample size enhanced the estimation accuracy across all five methods (Table 3 and Table 4).

Table 3. RMSE of the five propensity scores under Scenario 2 (nonlinear relationships).

Sample size	500	1000	2000	3000	4000	5000
GLM-OW	0.17094	0.12072	0.08538	0.06965	0.06029	0.05394
GLM-IPTW	0.16409	0.11603	0.08195	0.06679	0.05780	0.05168
GLM-matching	0.16062	0.11452	0.08134	0.06644	0.05756	0.05151
GBM-OW	0.16927	0.11969	0.08472	0.06918	0.05989	0.05359
GBM-IPTW	0.16365	0.11571	0.08180	0.06671	0.05775	0.05165

Table 4. Point estimation of the five propensity scores under Scenario 2 (nonlinear relationships).

Sample size	500	1000	2000	3000	4000	5000
GLM-OW	1.1152	1.0028	0.9500	0.9300	0.9219	0.9221
GLM-IPTW	1.2724	1.0277	0.9265	0.8899	0.8688	0.8718
GLM-matching	1.0988	0.9996	0.9452	0.9234	0.9156	0.9166
GBM-OW	1.2349	1.0982	1.0432	1.0180	1.0068	1.0020
GBM-IPTW	1.3382	1.1341	1.0539	1.0172	0.9997	0.9947

4.3. Comparison of Covariate Equilibrium of PS Methods under Scenario 1

In large-sample studies, covariates are more likely to be balanced across groups. In contrast, with relatively small samples, achieving covariate balance—even after randomization—becomes challenging, making balance testing necessary to ensure comparability between groups. Standardized mean differences (SMD) were used to assess covariate balance in small-sample settings. This study specifically examined the performance of five covariate balancing methods under small-sample conditions.

When sample sizes were small ($n \leq 700$), regardless of whether the outcome variable had a simple linear relationship or a complex nonlinear relationship with the covariate, both GLM-OW and GLM-matching outperformed the other three methods in balancing covariate effects, with all SMD values below 0.1—indicating that each covariate achieved adequate balance (Table 5 and Table 6).

Table 5. Equilibrium analysis of the five propensity scores under sample Scenario 1 (linear relationship only).

n	SMD	c_1	c_2	c_3	c_4	b_1	b_2	b_3	b_4
500	prematch	0.145	0.139	0.177	0.232	0.206	0.145	0.312	0.124
	GLM-OW	0.086	0.030	0.002	0.008	0.062	0.056	0.016	0.012
	GLM-IPTW	0.188	0.005	0.085	0.065	0.225	0.195	0.024	0.130
	GLM-matching	0.039	0.036	0.021	0.034	0.049	0.021	0.011	0.027
	GBM-OW	0.011	0.013	0.036	0.026	0.100	0.079	0.119	0.009
	GBM-IPTW	0.083	0.026	0.248	0.158	0.178	0.186	0.030	0.092
700	prematch	0.163	0.02	0.157	0.189	0.121	0.118	0.374	0.153
	GLM-OW	0.059	0.022	0.002	0.008	0.066	0.037	0.013	0.010
	GLM-IPTW	0.170	0.108	0.101	0.090	0.259	0.095	0.047	0.118
	GLM-matching	0.039	0.014	0.020	0.015	0.035	0.030	0.017	0.009
	GBM-OW	0.015	0.107	0.050	0.045	0.100	0.039	0.088	0.007
	GBM-IPTW	0.095	0.149	0.155	0.080	0.183	0.148	0.091	0.196

Table 6. Equilibrium analysis of the five propensity scores under sample Scenario 2 (non-linear relationships).

n	SMD	c_1	c_2	c_3	c_4	b_1	b_2	b_3	b_4
500	prematch	0.174	0.031	0.377	0.162	0.158	0.114	0.225	0.060
	GLM-OW	0.032	0.011	0.020	0.019	0.074	0.083	0.005	0.028
	GLM-IPTW	0.018	0.031	0.012	0.100	0.057	0.029	0.075	0.127
	GLM-matching	0.034	0.011	0.019	0.035	0.041	0.040	0.014	0.074
	GBM-OW	0.110	0.106	0.014	0.157	0.118	0.139	0.082	0.018
	GBM-IPTW	0.058	0.090	0.051	0.085	0.264	0.025	0.221	0.037
700	prematch	0.135	0.026	0.124	0.071	0.011	0.130	0.306	0.039
	GLM-OW	0.023	0.002	0.028	0.005	0.021	0.064	0.045	0.017
	GLM-IPTW	0.014	0.075	0.010	0.114	0.069	0.000	0.162	0.048
	GLM-matching	0.043	0.007	0.073	0.036	0.017	0.037	0.032	0.003
	GBM-OW	0.119	0.059	0.056	0.005	0.005	0.031	0.100	0.010
	GBM-IPTW	0.127	0.002	0.072	0.026	0.088	0.004	0.183	0.016

In cases where only a linear relationship existed between the covariates and the outcome, GBM-OW demonstrated better covariate balance compared to GBM-IPTW and GLM-IPTW (Table 5). However, under more complex scenarios involving nonlinear relationships—particularly with smaller sample sizes—the performance of GBM-OW in balancing covariates was inferior to that of GBM-IPTW and GLM-IPTW (Table 6).

As sample size increased, the ability of GBM-OW, GBM-IPTW, and GLM-IPTW to balance covariates improved. Notably, the performance of GBM-OW gradually surpassed that of GBM-IPTW and GLM-IPTW. When the sample size exceeded 1000, all five methods achieved SMD values below 0.1 for each covariate, indicating satisfactory balance. For example, taking C_1 as an illustration, the SMD decreased with increasing sample size, and GBM-OW, GBM-IPTW, and GLM-IPTW performed similarly to GLM-OW and GLM-matching in balancing this covariate (Figure 1 and Figure 2).

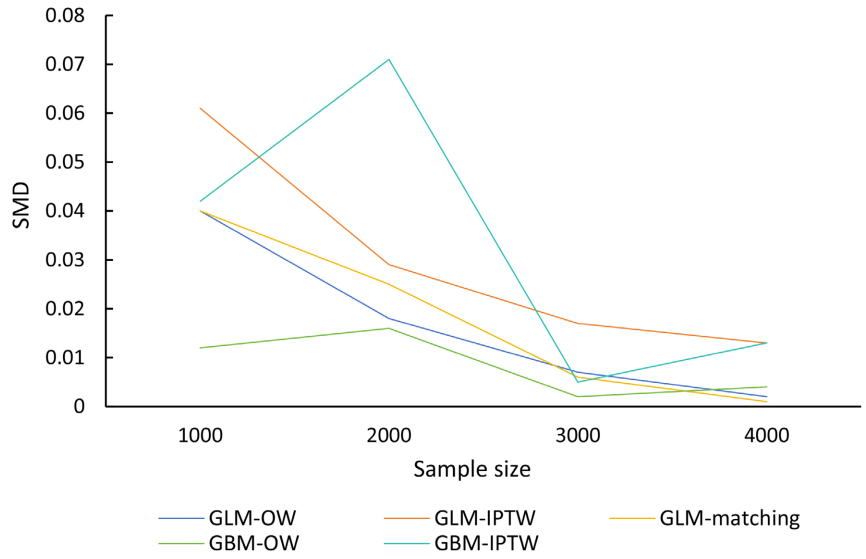

Figure 1. Comparison of C1's MD of five propensity score methods under Scenario 1.

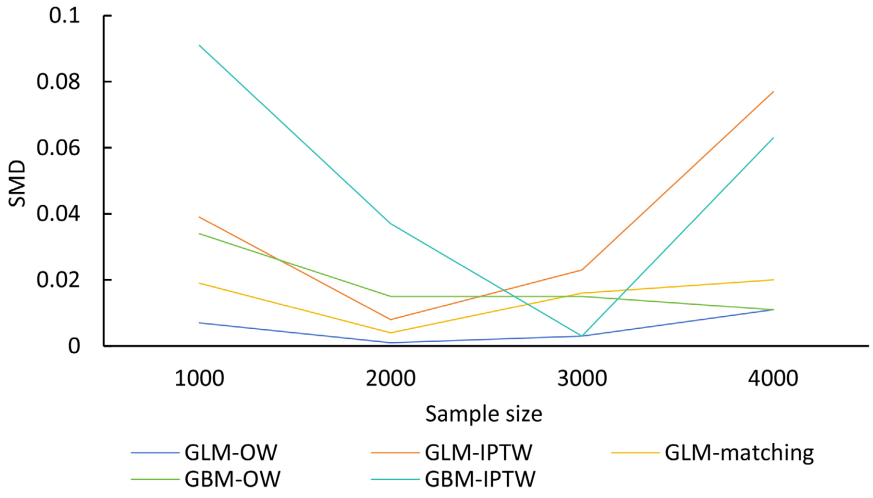

Figure 2. Comparison of C1'SMD of five propensity score methods under Scenario 2.

5. Discussions

In recent years, propensity score methods have gained increasing popularity in clinical trial design for their ability to emulate randomized experiments and improve comparability between study groups. By addressing imbalances in the dis-

tribution of baseline covariates between treatment and control groups, these methods help reduce confounding and enhance causal inference. However, while the majority of existing applications focus on two-group comparisons, many real-world clinical trials involve multi-arm interventions, raising important methodological questions regarding the appropriate use of propensity score approaches in multi-group settings.

Existing PS estimation methods each present distinct advantages and limitations. Parametric approaches such as GLM rely on distributional assumptions that, when violated, could lead to model misspecification and biased estimates [20]. Machine learning alternatives like GBM offer flexibility in capturing complex relationships and handling data complexities, yet might introduce estimation instability [21] [22]. Recent innovations such as overlap weighting demonstrate superior theoretical properties in achieving exact covariate balance, though their practical performance across diverse multi-group settings warrants further investigation.

To address this limitation, data-driven approaches have been proposed, including the use of cross-validation to optimize weighted combinations of multiple prediction algorithms. Another alternative, the generalized boosted model (GBM), employed an iterative ensemble of regression trees. This method automatically selected covariates, captured complex nonlinear relationships, handled missing data directly, and resisted overfitting. However, in practice, GBM might still exhibit instability in weight estimation and yield imprecise effect estimates. In simpler clinical scenarios, both GLM and GBM demonstrated comparable performance in covariate balancing and estimation accuracy. With the methodological advance introduced in the overlap weighting (OW) approach, it could ensure exact balance of all covariate moments across groups, driving standardized mean differences toward zero. Moreover, OW was robust to extreme propensity scores and consistently achieved superior covariate balance compared to other weighting schemes, such as inverse probability of treatment weighting (IPTW) [23]. Our findings aligned with this theoretical advantage: both GBM-OW and GLM-OW yielded better covariate balance than their IPTW counterparts, reinforcing the value of overlap weighting in clinical trial settings where balance is prioritized. The inferior performance of IPTW in small samples or under nonlinear data-generating mechanisms could be attributed to two primary factors. First, model misspecification in the treatment model (e.g., using a GLM for a nonlinear relationship) lead to biased propensity scores and, consequently, biased effect estimates. Second, IPTW was particularly susceptible to extreme weights in these scenarios, which increased variance and can destabilize estimates [24].

When the relationship between covariates and the outcome was linear or mildly nonlinear, GLM-based matching emerged as a favorable option, offering optimal balance and estimation accuracy combined with analytical simplicity. In such cases, GLM-matching not only produced estimates closer to the true treatment effect but also maintained strong balancing properties. Meanwhile, GLM-OW and

GBM-OW provided competitive balance performance, whereas GLM-IPTW and GBM-IPTW tended to yield more accurate estimates under correct model specification. Thus, in conventional clinical trial contexts with straightforward covariate-outcome relationships and well-specified treatment models, GLM-matching represented a straightforward and effective strategy for robust inference.

6. Limitations

This study had several limitations. First, the simulations considered only two confounders (related to both treatment assignment and the outcome) and a linear treatment-covariate relationship—idealized conditions under which GLM and GBM performed similarly. Real-world clinical data often involve more complex structures, where their performance might differ. Second, the degree of propensity score overlap was not varied, limiting insight into its influence, particularly in small samples. Third, only continuous outcomes were examined; performance with binary, time-to-event, or other outcome types remains to be evaluated. Third, the 1:2:7 allocation ratio was chosen to mimic a clinical trial with one smaller treatment arm. Different allocation ratios, especially those with more extreme imbalance, may affect the relative performance of the methods, particularly those reliant on weighting, and represent an important area for future research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

Funding

Research reported in this publication was supported by College Students' Innovation and Entrepreneurship Training Program (2023, Southern Medical University, S202312121136).

References

[1] Kendall, M.A., Zander, T., Wolansky, R.L., Teixeira, L. and Kuo, P.C. (2025) Propensity Score Matching: A Step-by-Step Guide to Coding in R and Application in Observational Research Studies. *The American Surgeon™*, **91**, 1949-1955. https://doi.org/10.1177/00031348251331293

[2] Shurrab, M., Ko, D.T., Jackevicius, C.A., Tu, K., Middleton, A., Michael, F., *et al.* (2023) A Review of the Use of Propensity Score Methods with Multiple Treatment Groups in the General Internal Medicine Literature. *Pharmacoepidemiology and Drug Safety*, **32**, 817-831. https://doi.org/10.1002/pds.5635

[3] McCaffrey, D.F., Griffin, B.A., Almirall, D., Slaughter, M.E., Ramchand, R. and Burgette, L.F. (2013) A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. *Statistics in Medicine*, **32**, 3388-3414. https://doi.org/10.1002/sim.5753

[4] Wang, J. and Marion-Gallois, R. (2022) Propensity Score Matching and Stratification Using Multiparty Data without Pooling. *Pharmaceutical Statistics*, **22**, 4-19. https://doi.org/10.1002/pst.2250

[5] Yang, S., Zhou, R., Li, F. and Thomas, L.E. (2023) Propensity Score Weighting Methods for Causal Subgroup Analysis with Time-to-Event Outcomes. *Statistical Methods in Medical Research*, **32**, 1919-1935. https://doi.org/10.1177/09622802231188517

[6] Woo, M., Reiter, J.P. and Karr, A.F. (2008) Estimation of Propensity Scores Using Generalized Additive Models. *Statistics in Medicine*, **27**, 3805-3816. https://doi.org/10.1002/sim.3278

[7] Gabriel, E.E., Sachs, M.C., Martinussen, T., Waernbaum, I., Goetghebeur, E., Vansteelandt, S., *et al.* (2023) Inverse Probability of Treatment Weighting with Generalized Linear Outcome Models for Doubly Robust Estimation. *Statistics in Medicine*, **43**, 534-547. https://doi.org/10.1002/sim.9969

[8] Judkins, D.R. and Porter, K.E. (2015) Robustness of Ordinary Least Squares in Randomized Clinical Trials. *Statistics in Medicine*, **35**, 1763-1773. https://doi.org/10.1002/sim.6839

[9] Tang, T., Austin, P.C., Lawson, K.A., Finelli, A. and Saarela, O. (2020) Constructing Inverse Probability Weights for Institutional Comparisons in Healthcare. *Statistics in Medicine*, **39**, 3156-3172. https://doi.org/10.1002/sim.8657

[10] Li, L. and Greene, T. (2013) A Weighting Analogue to Pair Matching in Propensity Score Analysis. *The International Journal of Biostatistics*, **9**, 215-234. https://doi.org/10.1515/ijb-2012-0030

[11] Mlcoch, T., Hrnciarova, T., Tuzil, J., Zadak, J., Marian, M. and Dolezal, T. (2019) Propensity Score Weighting Using Overlap Weights: A New Method Applied to Regorafenib Clinical Data and a Cost-Effectiveness Analysis. *Value in Health*, **22**, 1370-1377. https://doi.org/10.1016/j.jval.2019.06.010

[12] Komen, J.J., Belitser, S.V., Wyss, R., Schneeweiss, S., Taams, A.C., Pajouheshnia, R., *et al.* (2021) Greedy Caliper Propensity Score Matching Can Yield Variable Estimates of the Treatment-Outcome Association—A Simulation Study. *Pharmacoepidemiology and Drug Safety*, **30**, 934-951. https://doi.org/10.1002/pds.5232

[13] Yoshida, K., Hernández-Díaz, S., Solomon, D.H., Jackson, J.W., Gagne, J.J., Glynn, R.J., *et al.* (2017) Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-Way Matching. *Epidemiology*, **28**, 387-395. https://doi.org/10.1097/ede.0000000000000627

[14] Austin, P.C. (2009) Balance Diagnostics for Comparing the Distribution of Baseline Covariates between Treatment Groups in Propensity-Score Matched Samples. *Statistics in Medicine*, **28**, 3083-3107. https://doi.org/10.1002/sim.3697

[15] Huang, F.Q., Xu, J. and An, S.L. (2018) Study on the Evaluation Method of Covariate Balance among Multiple Groups. *Chinese Journal of Health Statistics*, **35**, 172-176. (In Chinese)

[16] Li, M., Liu, J., Zheng, J., Liu, K., Wang, J., Miner Ross, A., *et al.* (2019) The Relationship of Workplace Violence and Nurse Outcomes: Gender Difference Study on a Propensity Score Matched Sample. *Journal of Advanced Nursing*, **76**, 600-610. https://doi.org/10.1111/jan.14268

[17] Wang, Y., Cai, H., Li, C., Jiang, Z., Wang, L., Song, J., *et al.* (2013) Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study. *PLOS ONE*, **8**, e81045. https://doi.org/10.1371/journal.pone.0081045

[18] Austin, P.C. (2011) An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, **46**, 399-424. https://doi.org/10.1080/00273171.2011.568786

[19] Alzahrani, B., Abd El-Aty, A., Elatriby, S.A., Sobh, A.S., Bhlol, M.A., Elfar, A.A., *et al.*

(2024) A Modified Johnson Cook Model-Based Kalman Filter Method to Determine the Hot Flow Behavior of Sustainable AA6082 Al Alloy. *Materials*, **17**, Article No. 5169. https://doi.org/10.3390/ma17215169

[20] Camirand Lemyre, F., Lévesque, S., Domingue, M., Herrmann, K. and Ethier, J. (2024) Distributed Statistical Analyses: A Scoping Review and Examples of Operational Frameworks Adapted to Health Analytics. *JMIR Medical Informatics*, **12**, e53622. https://doi.org/10.2196/53622

[21] Setodji, C.M., McCaffrey, D.F., Burgette, L.F., Almirall, D. and Griffin, B.A. (2017) The Right Tool for the Job: Choosing Between Covariate-Balancing and Generalized Boosted Model Propensity Scores. *Epidemiology* (*Cambridge*, *Mass.*), **28**, 802-811. https://doi.org/10.1097/ede.0000000000000734

[22] Fuentes, A., Lüdtke, O. and Robitzsch, A. (2021) Causal Inference with Multilevel Data: A Comparison of Different Propensity Score Weighting Approaches. *Multivariate Behavioral Research*, **57**, 916-939. https://doi.org/10.1080/00273171.2021.1925521

[23] Stürmer, T., Webster-Clark, M., Lund, J.L., Wyss, R., Ellis, A.R., Lunt, M., *et al.* (2021) Propensity Score Weighting and Trimming Strategies for Reducing Variance and Bias of Treatment Effect Estimates: A Simulation Study. *American Journal of Epidemiology*, **190**, 1659-1670. https://doi.org/10.1093/aje/kwab041

[24] Lane, S.T.W. and Stuart, E.A. (2020) Propensity Score Analysis: Fundamentals and New Developments. In: van de Schoot, R. and Miočević, M., Eds., *Small Sample Size Solutions*: *A Guide for Applied Researchers and Data Scientists*, Routledge, 37-49.