# Comparative Analysis of Propensity Score Methods and Double Machine Learning in Estimating Treatment Effects

Jiazheng Li

May 23, 2024

## 1 Introduction

The estimation of treatment effects is a cornerstone of empirical research across various domains, including healthcare, education and social sciences. Accurately determining the impact of interventions is crucial for decision-making and policy formulation. Traditionally, randomized controlled trials (RCTs) are considered the gold standard for estimating treatment effects due to their ability to mitigate confounding biases. However, RCTs are often impractical or unethical to implement in real-world scenarios, leading researchers to rely on observational data. In observational studies, the treated subjects often differ systematically from untreated subjects, thus an unbiased estimate of the average treatment effect cannot be obtained by directly comparing outcomes between the two treatment groups, making the task of accurate effect estimation particularly challenging and susceptible to biases.

In response to these challenges, advanced causal inference methods have been developed to enhance the accuracy of treatment effect estimations from observational data. Propensity Score and Double Machine Learning represent two sophisticated approaches that attempt to address the limitations inherent in traditional statistical methods. The comparison of these methods is not merely academic; it has profound implications for the robustness of policy recommendations and the efficacy of interventions across numerous fields. Thus, by reviewing these methods' theoretical foundation as well as examining their strengths and weaknesses through a practical application, this paper seeks to contribute valuable insights into the ongoing development of causal inference methodologies and their applicability in empirical research.

## 2 Methodology Review

### 2.1 Propensity Score Methods

The propensity score is defined as the conditional probability of receiving a treatment given a vector of observed covariates, it allows one to design and analyze an observational (nonrandomized) study so that it mimics some of the particular characteristics of a randomized controlled trial. Rosenbaum and Rubin [RR84] demonstrated that "subclassification on the propensity score will balance all observed covariates", meaning within subclasses defined by the propensity score, the distribution of covariates is similar between treated and control groups, thereby mimicking randomization.

Rosenbaum & Rubin [RR84] define treatment assignment as strongly ignorable if two conditions are met: (a) treatment assignment is independent of potential outcomes given the observed covariates, i.e. $(Y(1), Y(0)) \perp Z \mid X$, also known as the "no unmeasured confounders" assumption, meaning all variables affecting treatment assignment and outcomes have been measured, and (b) each subject has a nonzero probability of receiving either treatment, i.e. $0 < P(Z = 1 \mid X) < 1$. These conditions ensure that conditioning on the propensity score allows for unbiased estimates of average treatment effects. To address the crucial nature of this assumption, Rosenbaum and Rubin proposed sensitivity analyses to assess the impact of unmeasured confounders. Additionally, they suggested using a second control group to verify that adjustments for measured covariates eliminate bias in estimating treatment effects.

The core idea is by adjusting for the propensity score, researchers can control for confounding variables. The balancing property is formally stated as: within strata of units with the same propensity score, the distribution of covariates will be the same for treated and control units. Mathematically, this is expressed as:

$$\Pr(x, z \mid e(x)) = \Pr(x \mid e(x)) \Pr(z \mid e(x))$$

where $x$ represents covariates, $z$ represents treatment assignment, and $e(x)$ is the propensity score.

Several propensity score methods are used for removing the effects of confounding when estimating the effects of treatment on outcomes: matching on the propensity score, stratification on the propensity score and inverse probability of treatment weighting using the propensity score [Aus11].

### 2.1.1 Propensity Score Matching

Propensity score matching involves creating matched sets of treated and untreated subjects with similar propensity scores. The propensity score $e(X)$ is the probability of receiving treatment given observed covariates $X$, and matching on it helps estimate the average treatment effect on the treated (ATT) [Imb04]:

$$e(X) = \Pr(T = 1 \mid X)$$

The most common method is one-to-one matching, where each treated subject is matched with an untreated subject with a similar propensity score. After forming matched pairs, treatment effects can be estimated by directly comparing outcomes between treated and untreated subjects. For continuous outcomes, the effect is the difference in mean outcomes:

$$\hat{\Delta} = \bar{Y}_{T=1} - \bar{Y}_{T=0}$$

where $\bar{Y}_{T=1}$ and $\bar{Y}_{T=0}$ are the mean outcomes for treated and untreated subjects in the matched sample, respectively. For dichotomous outcomes, the effect can be estimated as the difference in proportions:

$$\hat{\Delta} = \frac{1}{n_1} \sum_{i \in \text{Treated}} Y_i - \frac{1}{n_0} \sum_{j \in \text{Untreated}} Y_j$$

or using relative risk:

$$\text{RR} = \frac{\Pr(Y = 1 \mid T = 1)}{\Pr(Y = 1 \mid T = 0)}$$

Estimating the variance of treatment effects in matched samples accounts for the lack of independence between matched subjects. Schafer and Kang [SK08] suggest treated and untreated subjects in the matched sample should be regarded as independent, while Imbens[Imb04] recommends using variance estimation methods for paired experiments.

Alternative matching methods include:

**Matching with Replacement:** matching with replacement allows an untreated subject to be matched to multiple treated subjects, which requires special variance estimation [HR06]

**Greedy vs. Optimal Matching:**

- **Greedy Matching:** Selects the nearest untreated subject iteratively [GR93]). For a treated subject $i$, it finds the untreated subject $j$ such that:

$$j = \arg \min_{k \in \text{Untreated}} |e(X_i) - e(X_k)|$$

- **Optimal Matching:** Minimizes the total within-pair difference in propensity scores. The objective is to minimize:

$$\sum_{(i,j) \in \text{Matched Pairs}} |e(X_i) - e(X_j)|$$

**Caliper Matching:**

Matches treated and untreated subjects within a specified propensity score range (caliper distance), enhancing balance and reducing bias [RR85]. For a treated subject $i$, an untreated subject $j$ is selected if:

$$|e(X_i) - e(X_j)| \leq \delta$$

where $\delta$ is the caliper distance.

**Many-to-One Matching:**

Involves matching multiple untreated subjects to each treated subject, which can improve bias reduction compared to fixed one-to-one matching [MR00]. The treated subject $i$ is matched with $M$ untreated subjects:

$$\{j_1, j_2, \ldots, j_M\} = \arg \min_{j \in \text{Untreated}} |e(X_i) - e(X_j)|$$

**Full Matching:** Forms matched sets with either one treated subject and at least one untreated subject or one untreated subject and at least one treated subject, offering a flexible approach to matching. This method aims to use all available data by creating the most balanced groups possible.

In summary, propensity score matching is a versatile method that can be adapted using various approaches to improve balance and reduce bias in observational studies. By carefully selecting the matching method and adjusting for remaining imbalances, researchers can effectively estimate causal treatment effects from non-randomized data.

### 2.1.2 Stratification on the Propensity Score

Stratification on the propensity score involves dividing subjects into mutually exclusive subsets based on their estimated propensity scores. Subjects are ranked according to their estimated propensity scores and then stratified into subsets based on predefined thresholds. A common approach is to divide subjects into five equal-sized groups using the quintiles of the estimated propensity score. Cochran demonstrated that stratifying on the quintiles of a continuous confounding variable can eliminate approximately 90% of the bias due to that variable [Coc68]. Rosenbaum and Rubin extended it to stratification on propensity score, showing that it similarly eliminates about 90% of the bias due to measured confounders when estimating a linear treatment effect. While increasing the number of strata can further reduce bias, the marginal reduction decreases as the number of strata increases. Within each stratum, treated and untreated subjects will have roughly similar propensity scores, ensuring that the distribution of measured baseline covariates is approximately similar between the two groups.

Mathematically, let $e(X)$ be the estimated propensity score for a subject with covariates $X$. Subjects are stratified into $K$ strata based on the quantiles of $e(X)$. For example, for quintiles, $K = 5$, and the strata are defined by the quintile thresholds $Q_1, Q_2, \ldots, Q_{K-1}$.

$$S_i = \begin{cases} 1 & \text{if } e(X_i) \leq Q_1 \\ 2 & \text{if } Q_1 < e(X_i) \leq Q_2 \\ \vdots & \\ K & \text{if } Q_{K-1} < e(X_i) \end{cases}$$

Within each stratum $S_k$, the treatment effect on outcomes $Y$ can be estimated by directly comparing treated ($T = 1$) and untreated ($T = 0$) subjects:

$$\hat{\Delta}_k = \bar{Y}_{T=1,S_k} - \bar{Y}_{T=0,S_k}$$

where $\bar{Y}_{T=1}$ and $\bar{Y}_{T=0}$ are the mean outcomes for treated and untreated subjects in the matched sample, respectively.

Stratification on the propensity score can be conceptualized as conducting a meta-analysis of a set of quasi-randomized controlled trials (quasi-RCTs). Within each stratum, the treatment effect on outcomes can be estimated by directly comparing treated and untreated subjects. These stratum-specific estimates of treatment effect can then be pooled to estimate an overall treatment effect. Stratum-specific differences in means or risk differences can be averaged to produce an overall difference in means or risk difference. Typically, stratum-specific estimates of effect are weighted by the proportion of subjects within each stratum. When using equal-size strata, each stratum is given equal weight. This method allows for the estimation of the Average Treatment Effect (ATE). Alternatively, using weights proportional to the treated subjects within each stratum allows for the estimation of the Average Treatment Effect on the Treated (ATT)[Imb04]. The overall treatment effect can be estimated as:

$$\hat{\Delta} = \sum_{k=1}^{K} w_k \hat{\Delta}_k$$

where $w_k$ is the weight for stratum $k$. For estimating the ATE, $w_k = \frac{n_k}{n}$, where $n_k$ is the number of subjects in stratum $k$ and $n$ is the total number of subjects. For estimating the ATT, $w_k = \frac{n_{T=1,S_k}}{n_T}$, where $n_{T=1,S_k}$ is the number of treated subjects in stratum $k$ and $n_T$ is the total number of treated subjects.

A pooled estimate of the variance of the estimated treatment effect can be obtained by pooling the variances of the stratum-specific treatment effects:

$$\widehat{\text{Var}}(\hat{\Delta}) = \sum_{k=1}^{K} w_k^2 \widehat{\text{Var}}(\hat{\Delta}_k)$$

As with matching, within-stratum regression adjustment may be used to account for residual differences between treated and untreated subjects

### 2.1.3 Inverse Probability of Treatment Weighting Using the Propensity Score

Inverse probability of treatment weighting (IPTW) uses the propensity score to create a synthetic sample in which the distribution of measured baseline covariates is independent of treatment assignment. This approach is similar to using survey sampling weights to ensure that samples are representative of specific populations [MT08].

The weight for each subject is defined as follows:

$$w_i = \frac{Z_i}{e_i} + \frac{(1 - Z_i)}{1 - e_i}$$

where $Z_i$ is an indicator variable denoting whether the $i$th subject was treated, and $e_i$ is the propensity score for the $i$th subject. A subject's weight is the inverse of the probability of receiving the treatment that the subject actually received. This method is a form of model-based direct standardization.

Lunceford and Davidian [LD04] review various estimators for treatment effects based on IPTW. Assuming $Y_i$ denotes the outcome variable measured on the $i$th subject, an estimate of the Average Treatment Effect (ATE) is:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{e_i} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1 - Z_i)Y_i}{1 - e_i}$$

where $n$ is the number of subjects. They discuss the theoretical properties of this estimator and compare it to stratification.

Weights can be unstable for subjects with a very low probability of receiving treatment. Stabilizing weights [HBR00] address this issue. Using stabilized weights, the ATE can be estimated.
For the ATT, the weights are:

$$w_{i,\text{ATT}} = Z_i + (1 - Z_i)\frac{e_i}{1 - e_i}$$

For the average effect of treatment in the controls (ATC), the weights are:

$$w_{i,\text{ATC}} = Z_i \frac{1 - e_i}{e_i} + (1 - Z_i)$$

## 2.2 Double Machine Learning

Double or Debiased Machine Learning (DML) represents a significant advancement over traditional methods like Propensity Score Methods(PSM) by relaxing stringent assumptions about model specification and the functional forms of covariates. While effective, PSM relies on correctly specifying the propensity score model and can struggle with high-dimensional data where the number of covariates is large relative to the sample size. DML not only accommodates high-dimensional settings but also leverages modern mach One central purpose of DML is to address the limitations of traditional parametric regression models, which can lead to biased estimates of the treatment effect if confounding variables are included with incorrect functional forms. [IW09]. Unlike PSM, which primarily focuses on balancing covariates through propensity scores, DML uses a combination of sample splitting and orthogonalization to account for and mitigate the effects of confounding variables, leading to more accurate and unbiased estimates of treatment effects even in complex scenarios.

Classical semi-parametric methods, such as those discussed by Athey [AI17], attempt to mitigate this issue by making functional form assumptions solely about the treatment parameter while remaining agnostic about the covariates. However, these methods often suffer from slow convergence rates and require larger sample sizes compared to their parametric counterparts [Pow86]. This challenge becomes particularly pronounced in high-dimensional settings, where the number of parameters to estimate is large relative to the sample size. High dimensional settings can arise either from having more variables/features than observations (p¿n) or from including various transformations and interactions of a relatively low-dimensional set of covariates to safeguard against model misspecification. Estimating parameters for each transformation can quickly escalate into a high-dimensional problem [BCH14].

One of the pioneering semi-parametric regression methods is Robinson (1988), who employed kernel regression to model the relationship between covariates and both the outcome and treatment. The model is typically specified as:

$$Y_i = D_i\tau + g(X_i) + \epsilon_i$$

where $Y_i$ is the outcome variable, $D_i$ is the treatment indicator, $X_i$ represents covariates, $g(X_i)$ is an unknown function of covariates, $\tau$ is the treatment effect, and $\epsilon_i$ is the error term.

In parallel developments within biostatistics, Laan and Rubin[LR06] introduced Targeted Maximum Likelihood Estimation (TMLE), another semi-parametric estimation technique akin to DML that also incorporates ML methods. TMLE and DML share the common goal of enhancing the robustness and efficiency of causal inference models.

A more direct predecessor of DML is the "double selection" procedure [BCH14]. This method involves three steps:

1. **Outcome Model Selection**: Use lasso regression to select covariates predictive of the outcome $Y$.
2. **Treatment Model Selection**: Use lasso regression to select covariates predictive of the treatment $D$.
3. **Final Estimation**: Apply Ordinary Least Squares (OLS) to regress the outcome on the treatment and the union of all selected covariates.

Mathematically, this can be expressed as:

$$\hat{\beta} = \arg\min_{\beta} \left(Y - D\beta - X_{\hat{S}_Y \cup \hat{S}_D}\gamma\right)^2$$

where $\hat{S}_Y$ and $\hat{S}_D$ are the sets of covariates selected by lasso for the outcome and treatment models, respectively.

DML builds upon this foundation by providing a framework that allows these kernel regressions to be replaced with modern machine learning (ML) methods [CCD$^+$17]. This adaptation enables the application of DML in high-dimensional contexts, where traditional methods struggle.

The DML framework typically involves the following steps:

1. **Sample Splitting**: Divide the sample into $K$ folds to mitigate overfitting.
2. **Nuisance Parameter Estimation**: Use ML methods to estimate the nuisance parameters $\hat{m}(X_i)$ and $\hat{g}(X_i)$ in each fold.
3. **Orthogonalization**: Construct orthogonalized scores to remove the bias due to nuisance parameters. For instance, the orthogonalized score for the treatment effect can be written as:

   $$\psi(W_i, \hat{\eta}) = \left(\hat{g}(X_i) - \hat{g}(X_i^{(-k)})\right)\left(Y_i - \hat{m}(X_i^{(-k)})\right)$$

   where $\hat{g}(X_i^{(-k)})$ and $\hat{m}(X_i^{(-k)})$ are estimates of $g(X_i)$ and $m(X_i)$ obtained without using the $k$-th fold.

4. **Final Estimation**: Combine the orthogonalized scores across all folds to obtain the final estimate of the treatment effect $\hat{\tau}$.

DML generalizes the double selection approach by introducing a sample splitting procedure, which enhances its applicability to a wide range of modern ML methods beyond lasso regression. The sample splitting technique helps to mitigate overfitting and provides more reliable inference by ensuring that the model's training and testing phases are separated, thus enabling the effective use of ML algorithms for causal estimation [CCD$^{+}$17].

In summary, DML represents a robust and flexible framework for causal inference, particularly in high-dimensional settings. By leveraging modern ML methods and incorporating sample splitting, DML addresses many limitations of traditional parametric and semi-parametric methods, providing more accurate and unbiased estimates of treatment effects.

# 3 Experimental Setup

## 3.1 Data

The dataset utilized in this study originates from the Infant Health and Development Program (IHDP), which was initially designed as a randomized controlled trial (RCT). The primary objective of the IHDP was to assess the impact of home visits by specialist doctors on the cognitive test scores of premature infants. This intervention was aimed at enhancing the developmental outcomes of these at-risk children by providing timely medical and educational support through specialist interventions.

For the purposes of benchmarking causal inference methodologies, the dataset has been adapted to mimic the conditions typically found in observational studies. This adaptation was achieved by introducing selection bias into the originally randomized dataset. Specifically, a non-random subset of treated individuals—those with certain predefined characteristics—was systematically removed. This manipulation creates an artificial scenario where the treatment assignment mimics the biases often inherent in observational data, thereby providing a robust platform for testing the effectiveness of causal inference methods under realistic, non-experimental conditions.

The IHDP dataset comprises data on 747 subjects, each represented by 25 variables. These variables include both the treatment assignment—whether the infant received home visits by specialist doctors—and a range of covariates considered relevant to the study's focus on cognitive development. The outcome variable of interest, the cognitive test score, is generated based on these covariates and the treatment status, providing a comprehensive dataset for analyzing the causal impact of the intervention.

## 3.2 Simulation Design

In order to compare the effectiveness of Propensity Score Matching (PSM) and Double Machine Learning (DML) in estimating treatment effects, I will conduct a simplified simulation study using the Infant Health and Development Program (IHDP) dataset. This dataset is well-suited for such a comparison as it mimics observational study conditions through induced selection bias, providing a robust platform for testing causal inference methodologies.

To simulate an observational study, I will introduce selection bias by systematically removing a non-random subset of treated individuals, thereby creating an artificial scenario that mimics the biases often present in real-world observational data. This manipulation allows us to test the effectiveness of the causal inference methods under realistic, non-experimental conditions. The dataset is subsequently split into a training set (70%) and a test set (30%) to facilitate model training and evaluation.

For the Propensity Score Matching (PSM) method, I begin by estimating the propensity scores using logistic regression. The propensity score represents the conditional probability of receiving treatment given the observed covariates. This step ensures that subjects with similar covariate profiles are assigned comparable propensity scores. Once the propensity scores are estimated, I perform one-to-one matching, where each treated subject is paired with an untreated subject with the closest propensity score. This matching process aims to balance the distribution of covariates between treated and untreated groups, thereby reducing bias in the estimation of treatment effects. The treatment effect for continuous outcomes is estimated by calculating the difference in mean outcomes between treated and untreated subjects within the matched sample, while for dichotomous outcomes, the effect is estimated as the difference in proportions or using relative risk. Variance estimation is performed using methods such as bootstrap or robust standard errors to account for the lack of independence between matched subjects.

For the Double Machine Learning (DML) method, I employ the DoubleML library, which offers a sophisticated approach that accommodates high-dimensional settings and leverages modern machine learning techniques. Using the DoubleML library, I start by defining the DML model. This library allows for efficient implementation of DML methods by integrating machine learning models to estimate nuisance parameters and then using orthogonalization techniques to debias the treatment effect estimates.

The DML process involves several steps. First, I split the dataset into multiple folds to prevent overfitting and facilitate cross-validation, specifically, here I set the number of folds equals to five. This step ensures that the model's performance is evaluated on unseen data, enhancing its generalizability. In each fold, I estimate the nuisance parameters, including the expected outcome given covariates and the propensity score model, using machine learning methods such as Random Forest, Gradient Boosting and Neural Network. For Random Forest, I set the number of estimators to 100, the maximum depth of each tree to 10 and a fixed random state for reproducibility. For Gradient Boosting (XGBoost), the number of estimators is also set to 100, with a maximum depth of 10 for each tree. For Neural Network, I use a multi-layer perceptron with a hidden layer size of 100 neurons and a maximum of 500 iterations.These estimates are crucial for constructing orthogonalized scores. I then proceed with orthogonalization, which involves calculating residuals from the nuisance parameter models and using these residuals to adjust the treatment effect estimates, thereby removing bias due to nuisance parameters. The final estimate of the treatment effect is obtained by combining the orthogonalized scores from all folds, leveraging the entire dataset while mitigating overfitting.

## 4 Results and Discussions

The table 1 presents the results of the simulation comparing Propensity Score Matching (PSM) and Double Machine Learning (DML) using different machine learning models (Random Forest, XGBoost, and Neural Network). The metrics reported include the mean estimate of the treatment effect, the average treatment effect on the treated (ATT), the variance estimate, and the 95% confidence interval for each method.

| **Method** | **Mean Estimate** | **ATT** | **Variance Estimate** | **95% CI** |
|---|---|---|---|---|
| PSM | 3.852 | 3.794 | 0.082 | [3.189, 4.342] |
| DML Random Forest | 3.820 | 3.850 | 0.028 | [3.481, 4.113] |
| DML XGBoost | 3.395 | 3.500 | 0.071 | [2.940, 3.922] |
| DML Neural Network | 3.828 | 3.827 | 0.061 | [3.360, 4.320] |

Table 1: Comparison of Methods

The mean estimates and ATT values indicate the central tendency of the estimated treatment effects across different methods. Both PSM and DML with Random Forest show mean estimates close to 3.8, suggesting that these methods provide similar central estimates. DML with XGBoost has a lower mean estimate (3.3950), while DML with Neural Network also shows a similar central tendency to PSM and Random Forest.

The variance estimates highlight the precision of the treatment effect estimates. DML with Random Forest has the lowest variance (0.0278), indicating high precision and stability in the estimates. PSM has the highest variance (0.0816), suggesting more variability in the estimates. DML with XGBoost and Neural Network have intermediate variances, with Neural Network being more precise than XGBoost.

The 95% confidence intervals provide a range within which the true treatment effect is expected to lie with 95% confidence. PSM has the widest confidence interval, reflecting its higher variance. DML with Random Forest has the narrowest confidence interval, indicating greater confidence in the estimates. The confidence intervals for XGBoost and Neural Network are broader than Random Forest but narrower than PSM, aligning with their variance estimates, this may due to their complex architecture, making the estimation easier to overfit using a relatively small dataset provided.

DML with Random Forest demonstrates the highest precision and stability, as indicated by its low variance and narrow confidence interval. This suggests that, for IDHP dataset, Random Forest is particularly effective in capturing the treatment effect with minimal variability. The mean estimates from PSM, Random Forest, and Neural Network are similar, indicating these methods are consistent in estimating the central tendency of the treatment effect. XGBoost, while slightly lower, still provides a reasonably close estimate. The wider confidence intervals and higher variance in PSM suggest that this method may be more sensitive to sample variations, making it less stable compared to DML methods. Among the DML methods, Random Forest offers the most confidence in the estimates, followed by Neural Network and XGBoost.

# 5 Conclusions

The simulation results indicate that Double Machine Learning (DML) methods, particularly those using Random Forest, offer more precise and stable estimates of treatment effects compared to Propensity Score Matching (PSM). This conclusion is supported by the lower variance and narrower confidence intervals observed in the DML estimates.

The theoretical foundation underlying this conclusion stems from the strengths of DML in addressing high-dimensional settings and model misspecification. DML leverages modern machine learning techniques to flexibly model complex relationships between covariates and outcomes. By orthogonalizing the estimation process, DML effectively mitigates biases arising from nuisance parameter estimation, leading to more robust causal inferences. Specifically, the use of machine learning models like Random Forest in DML provides an advantage due to their ability to capture non-linear relationships and interactions between covariates, thereby improving the accuracy and stability of the treatment effect estimates. The superior performance of Random Forest in estimating the ATE compared to Neural Networks and XGBoost in this simulation can be attributed to the dataset size and complexity. Random Forest, known for its robustness and ability to handle small to medium-sized datasets effectively, reduces overfitting and captures essential patterns without requiring large datasets. In contrast, Neural Networks and XGBoost, while more advanced, require larger datasets to optimize their complex architectures and fully leverage their capabilities. The IHDP dataset, with its limited size, is insufficient to support these complex models, leading to higher variance and less stable estimates. Therefore, Random Forest's simplicity and robustness make it better suited for smaller dataset, providing more precise and stable ATE estimates.

PSM, on the other hand, relies on the assumption that all confounding variables are adequately controlled through the propensity score model. This method can be sensitive to model misspecification and may result in higher variability when the propensity score model does not fully capture the underlying data structure. This sensitivity is reflected in the higher variance and wider confidence intervals observed in the PSM estimates in our simulation.

It is important to note that these results were drawn using the Infant Health and Development Program (IHDP) dataset. While the IHDP dataset is well-suited for benchmarking causal inference methods, it has limitations that must be considered. The dataset size and the specific covariate structures may influence the generalizability of the findings. In real-world applications, datasets may vary in size and complexity, potentially affecting the performance of both PSM and DML methods. Therefore, while the conclusions from this study are informative, further validation with other datasets and in different contexts is necessary to fully understand the robustness and applicability of these methods.

# 6 Appendix

For the code of the simulation in this study, please see the following GitHub repository: Causal Inference Repository.

# References

[AI17] Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, May 2017.

[Aus11] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[BCH14] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, May 2014.

[CCD$^+$17] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters, 2017.

[Coc68] William G Cochran. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, pages 295–313, 1968.

[GR93] Xing Sam Gu and Paul R Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

[HBR00] Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men, 2000.

[HR06] Jennifer Hill and Jerome P Reiter. Interval estimation for treatment effects using propensity score matching. *Statistics in medicine*, 25(13):2230–2256, 2006.

[Imb04] Guido W. Imbens. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29, February 2004. _eprint: https://direct.mit.edu/rest/article-pdf/86/1/4/1613802/003465304323023651.pdf.

[IW09] Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, March 2009.

[LD04] Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19):2937–2960, 2004.

[LR06] Mark Laan and Daniel Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2:1043–1043, 02 2006.

[MR00] Kewei Ming and Paul R Rosenbaum. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56(1):118–124, 2000.

[MT08] Stephen L Morgan and Jennifer J Todd. A diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, 38(1):231–281, 2008.

[Pow86] James L. Powell. Estimation of semiparametric models. In R. F. Engle and D. McFadden, editors, *Handbook of Econometrics*, volume 4 of *Handbook of Econometrics*, chapter 41, pages 2443–2521. Elsevier, 1986.

[RR84] Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[RR85] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[SK08] Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4):279, 2008.