Paper RE01

Robust Causal Inference in Real-world Evidence Studies with Double Machine Learning

Jing Dai, Jazz Pharmaceuticals and Wayne State University, Detroit, MI, USA
Yilun Huang, Novartis, East Hanover, NJ, USA
Sarah Markt, Jazz Pharmaceuticals, Philadelphia, PA, USA
Adeniyi Togun, Jazz Pharmaceuticals, Philadelphia, PA, USA

ABSTRACT

Estimating causal treatment effects from Real-world Data (RWD) is frequently challenged by high-dimensional confounding, where traditional methods for adjustment may perform poorly with a large number of covariates and are subject to bias from model misspecification.

Double Machine Learning (DML) provides a rigorous framework for estimating causal effects in observational settings. It employs machine learning algorithms to model nuisance functions and uses orthogonalized regression to mitigate bias from regularization errors. It also leverages cross-fitting to produce the final causal estimate that is robust to modeling specifications.

This paper introduces the theory of DML and demonstrates its application through a simulated real-world dataset. We illustrate a comparison of DML against other methods, including multiple regression covariate adjustment and propensity score methods. The objective is to provide researchers with a clear understanding of DML's framework and its potential advantages for generating robust evidence from complex RWD.

INTRODUCTION

Real-world Evidence (RWE) from electronic health records, administrative claims, registries, and other routine data sources is increasingly used to support key evidence generation strategies, regulatory decision-making and clinical guidelines [1]. These data are attractive because they reflect heterogeneous patient populations and usual care settings; however, as they are not collected for research purposes, they also introduce challenges such as uncontrolled and unmeasured confounding, selection bias, measurement error, and missingness [2]. High dimensional confounding where the number of candidate covariates is large relative to the sample size and the presence of complex non-linear relations raises difficult problems for model specification and variable selection [3].

To minimize selection bias in observational studies, it is necessary to adjust for confounding variables that sufficiently capture the factors that influence treatment choice and clinical outcomes. Classical approaches such as multiple regression (e.g. adjustment for confounders in the exposure/outcome model), propensity score matching (PSM), and inverse probability of treatment weighting (IPTW) often rely on low dimensional models that may be mis-specified when the true treatment and outcome relationships are nonlinear or contain complex interactions [4]. Machine learning (ML) methods can flexibly capture these relationships, but naive use of ML for outcome regression or propensity score (PS) estimation can introduce regularization bias when the ML algorithms prioritize prediction over causal identification [5].

Double Machine Learning (DML), developed in econometrics and now increasingly used in biostatistics and pharmacoepidemiology [6], addresses these issues by constructing Neyman orthogonal score functions [7] and combining them with cross-fitting. The method uses ML only for nuisance components (i.e., outcome and treatment models), while the target causal parameter is estimated using a low-dimensional, orthogonalized regression that is robust to small errors in nuisance estimation [7]. This paper summarizes the DML framework, contrasts it with standard regression and propensity score (PS)-based approaches, and illustrates its performance in simulated RWE settings with high dimensionality and nonlinearity.

REGULARIZATION BIAS AND THE DML SOLUTION

High-dimensional confounding arises when the number of observed covariates (X) is large, such that traditional parametric models are unstable [8]. To mitigate overfitting, penalized regression (e.g., Lasso) or tree-based ensemble methods (e.g., Gradient Boosting) are often applied. While these algorithms are effective for prediction, they can distort causal estimates [9].

A typical ML model is used to predict the outcome using both the treatment and covariates. For high dimensional data, the algorithm applies regularization and shrinks coefficients to prevent overfitting. If the treatment variable is strongly correlated with confounders, the algorithm may credit the effect to the confounders rather than the treatment, shrinking the treatment coefficient toward zero. This phenomenon, known as regularization bias, prevents naive machine learning models from recovering the true causal effect [7, 10].

The DML method addresses the regularization bias by separating the 'prediction' tasks from the 'inference' task. Instead of fitting one complex model, DML proceeds in two stages (i.e., the prediction stage and the inference stage) using specific 'nuisance models to orthogonalize the data [7]. The target estimand for DML is the average treatment effect (ATE), defined as *E[Y(1) - Y(0)]*.

Step 1: Model the outcome using the outcome nuisance model.
We train an ML model to predict the outcome (Y) using only the baseline covariates (X).

$$Y = f(X) + residual_Y$$

The $residual_Y$ represents the variation in the outcome that cannot be explained by the observed covariates and is independent of the baseline covariates (X).

Step 2: Model the exposure/treatment using the treatment nuisance model.
We also train an ML model to predict the treatment (T) using the same baseline covariates (X).

$$T = g(X) + residual_T$$

Similarly, the $residual_T$ represents the variation in the treatment assignment that is independent of the observed covariates.

Step 3: Estimate the causal effect. We estimate the causal effect (θ) by regressing the outcome residuals on the treatment residuals using ordinary least squares (OLS) linear regression. By construction, OLS minimizes the mean squared error by projecting the outcome residuals onto the treatment residuals. This operation ensures that the resulting structural error term ε is mathematically orthogonal to the treatment variation.

$$residual_Y = \theta \times residual_T + \varepsilon$$

By using residuals, we have mathematically 'subtracted out' the confounding influence of the covariates from both the outcome and the treatment. The resulting coefficient θ represents the causal effect of the treatment on the outcome, robust to high-dimensional confounding [7]. For binary outcomes, this linear approximation yields the ATE as risk difference.

It is critical to distinguish between the prediction residuals derived in the nuisance stages and the structural error term in the inference stage. The outcome residual $residual_Y$ and the treatment residual $residual_T$ represent variations in the observed data that are independent of the baseline covariates (X), effectively removing the confounding signal [7]. In contrast, the structural error ε captures the remaining unexplained variation in the outcome residuals. While ε is intrinsically part of the outcome variation, the final OLS stage enforces that ε is independent of the treatment residual $residual_T$, satisfying the condition required for orthogonality and unbiased parameter estimation [7, 11].

To prevent overfitting, this process is performed using cross-fitting [7, 12]. The data is split into K folds. The nuisance models ($f(X)$ and $g(X)$) are trained on one partition of the data, but the residuals are calculated on the held-out partition. This ensures that the nuisance errors are independent of the estimation noise, allowing for valid statistical inference (p-values and confidence intervals) even when complex ML algorithms are used.

PS-BASED METHODS AND COMPARISON TO DML

The PS-based methods remain standard tools in RWE because they can reduce the dimensionality of multiple confounding variables into a probability that can then be applied in a model [13, 14], as weights or matched to create two populations that are similar to one another on the potential measured confounding variables to emulate a randomized controlled trial (RCT). [4]. The PS is defined as the probability that a specific patient receives the treatment given their baseline characteristics (e.g., age, comorbidities). By modeling this probability, usually via logistic regression, we can adjust the data so that the treated/exposed and control/unexposed groups look comparable [9].

The PSM approach uses the 'matching' strategy. It attempts to find 'twins' (for every treated or exposed patient, it searches for an untreated or unexposed patient with a similar PS) [15].

- **Mechanism**: If a treated or exposed patient has a score of 0.8, we pair them with an untreated or unexposed patient who also has a score of 0.8.
- **Analysis**: We discard unmatched patients and calculate the difference in outcomes between the matched pairs.
- **Limitation**: In small datasets (common in rare diseases), finding exact matches is difficult. We often lose a significant amount of data (unmatched subjects), reducing statistical power [16]. Furthermore, this exclusion can shift the

demographic and clinical characteristics of the cohort, meaning the final matched sample may no longer represent the original target population (limited generalizability).

The IPTW approach uses the 'weighting' strategy. Instead of discarding data, it creates a synthetic population where the two groups are balanced [17].

- **Mechanism**: A weight is assigned to each patient based on the inverse of their PS. A treated/exposed patient who looks like an untreated/unexposed patient is given a very high weight, while a treated patient with a high probability of receiving treatment (high PS) is assigned a lower weight.
- **Analysis**: We calculate the weighted average difference in outcomes.
- **Limitation**: This method is highly sensitive to extreme weights. If the PS model is overfit (common in high dimensions), some observations may get massive weights, destabilizing the final estimate [17, 18].

While both PS methods and DML aim to control for confounding, they differ fundamentally in their mechanism [19]:

PS methods focus almost exclusively on the treatment model and usually ignore the outcome relationship until the final step. This might lead to the risk of balancing the wrong variables. These methods may prioritize variables that strongly predict treatment but have no effect on the outcome (adding noise), while failing to adequately adjust for strong prognostic factors (drivers of the outcome) that are only weakly associated with treatment [20]. The DML focuses equally on the treatment and outcome models. Instead of reweighting patients, DML mathematically 'subtracts' the variation explained by covariates from both the treatment and the outcome using residuals. This ensures that strong prognostic factors are explicitly modeled and controlled for, yielding more precise causal estimates.

The DML also offers more flexibility in modeling approach by allowing for ML models of choice. Unlike traditional PS modeling, which relies almost exclusively on logistic regression [21], DML offers a model-agnostic approach. Researchers can deploy any ML algorithm and can use different ML models for the treatment and outcome models, to capture non-linearity and address high dimensionality [7, 19].

SIMULATION STUDY

To evaluate the performance of DML vs traditional methods in a real-world setting, we conducted a simulation study designed to mimic a rare disease cohort. This design reflects common real-world scenarios involving both continuous and binary outcomes. Specifically, we simulated a continuous outcome to mimic metrics like systolic blood pressure or total healthcare costs, and a binary outcome to represent event-driven measures such as treatment response, remission status, or the incidence of adverse events.

1. Simulation design

We generated a synthetic dataset with the following characteristics:

- **Sample size (*N*):** 500 subjects.
- **High dimensionality (*P*):** 100 baseline covariates (e.g., demographics, labs, and comorbidities).
- **Sparsity:** While 100 covariates were observed, only 5 were 'true' confounders affecting both the treatment and the outcome. The remaining 95 were noise variables.
- **Complexity:** The true confounders influenced the outcome via non-linear relationships (e.g., quadratic terms) and interactions between variables, rather than simple linear additivity.
- **True effect:** For the continuous outcome simulation, the true causal effect of the treatment (e.g., low adherence relative to high adherence) on the outcome (e.g., total healthcare costs) was set to $1,000 increase annually. For the binary outcome simulation, the true causal effect of the treatment (e.g., a combo therapy relative to monotherapy) on the outcome (e.g., adverse event risk) was set to an absolute risk increase of approximately 15% (0.15) for a target adverse event.
- **Repetitions:** We performed 1,000 Monte Carlo simulation runs for both continuous and binary outcome scenarios.
- **Results presentation:** The estimated effect represents the mean and standard deviation (SD) across 1,000 Monte Carlo runs, with the 95% confidence interval (CI) defined by the 2.5th and 97.5th percentiles of the simulation results.
- **Code availability:** The Python codes used to generate these simulations can be downloaded at: https://github.com/JingDai777/PHUSE-USCONNECT-2026-RE01-DML

2. Methods compared

We applied five estimators to this synthetic dataset:

1. **Naive (unadjusted):** Unadjusted regression of the outcome on the treatment variable, ignoring covariates.
2. **Regression adjustment:** A standard multivariable regression (Continuous: linear; Binary: logistic) including all 100 covariates.
3. **IPTW:** Using a logistic regression model to estimate PS and weight the population.

4. **PSM:** 1:1 Nearest Neighbor matching based on logistic regression generated PS.
5. **DML:** Using Gradient Boosting estimates for both nuisance models (treatment and outcome), combined with 3-fold cross-fitting.

3. Results

We summarize the estimated treatment effects and their bias (absolute deviation from the true effect) in the tables below.

Table 1: Simulation results for the continuous outcome summarized over 1,000 runs.

Estimator	Estimated Effect (SD)	Bias (95% CI)
True Effect	$1,000 (NA)	NA
Unadjusted	$1,569.7 ($73.44)	$569.7 ($429.20, $709.11)
Regression Adjusted	$1,463.4 ($78.34)	$463.4 ($309.28, $618.87)
IPTW	$1,469.7 ($86.49)	$469.7 ($300.96, $642.66)
PSM	$1,463.1 ($92.09)	$463.1 ($287.58, $645.99)
DML	$1,109.6 ($66.84)	$112.4 ($6.38, $241.34)

Table 2: Simulation results for the binary outcome summarized over 1,000 runs.

Estimator	Estimated Effect (SD)	Bias (95% CI)
True Effect	~0.15 (NA)	NA (NA)
Unadjusted	0.27 (0.041)	0.11 (0.032, 0.191)
Regression adjusted	0.25 (0.048)	0.10 (0.010, 0.192)
IPTW	0.25 (0.052)	0.09 (0.005, 0.200)
PSM	0.25 (0.054)	0.09 (0.006, 0.198)
DML	0.19 (0.052)	0.05 (0.002, 0.137)

4. Findings

The simulation illustrates the challenges of causal inference in small cohorts characterized by high-dimensional data.

- Direction of bias: It is notable that most methods consistently overestimated the true effect (positive bias). This reflects the underlying 'positive confounding' structure of the simulations, where specific comorbidities were generated to simultaneously increase the likelihood of treatment and the magnitude of adverse outcomes (higher costs/risk).
- Performance of traditional estimators: Traditional methods, including regression adjustment and PS-based methods, showed higher levels of bias in this setting. With 100 variables and only 500 patients, standard models likely faced difficulties in effectively distinguishing the 5 true confounders from the 95 noise variables. This potential overfitting to noise may have contributed to the overestimation of costs observed in the regression and PS-based models. In the binary setting, we found that standard logistic regression can fail to converge in extreme conditions where we specified more variables than patients. The high dimensionality relative to sample size likely led to issues of perfect separation, causing coefficient estimates to become unstable or undefined, highlighting a practical limitation of traditional approaches in high-dimensional, data-scarce settings.
- DML stability: DML yielded the estimate closest to the truth (Continuous: $1,109.6; Binary: 0.19). The DML framework appears to be better equipped to handle internal feature selection by focusing on relevant non-linear confounders while filtering out noise. In the continuous case, this resulted in a substantial reduction in bias: approximately 75% lower bias compared to PSM ($112.4 vs $463.1). Similarly, in the binary setting, DML reduced bias by approximately 45% compared to PSM (0.05 vs 0.09). While the limited sample size precluded a perfect recovery of the true effect, these findings suggest that DML offers advantages in minimizing bias within data-scarce, complex environments.

REFERENCES

1. Girman C, Ritchey M, and Lo Re V III. Real-world data: assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. *Pharmacoepidemiology and Drug Safety* 31.7 (2022): 717.
2. Sherman RE, et al. Real-world evidence—what is it and what can it tell us. *N Engl J Med* 375.23 (2016): 2293-2297.

3. Nørgaard M, Ehrenstein V, and Vandenbroucke J. Confounding in observational studies based on large health care databases: problems and potential solutions–a primer for the clinician. *Clinical epidemiology* (2017): 185-193.
4. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46.3 (2011): 399-424.
5. Wyss R, et al. Machine learning for improving high-dimensional proxy confounder adjustment in healthcare database studies: An overview of the current literature. *Pharmacoepidemiology and drug safety* 31.9 (2022): 932-943.
6. Jiang, Cong, et al. "A double machine learning approach for the evaluation of COVID-19 vaccine effectiveness under the test-negative design: Analysis of Québec administrative data." *Statistics in medicine* 44.5 (2025): e70025.
7. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*. (2018);21(1):C1–C68.
8. Bühlmann, Peter, and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.
9. Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. "High-dimensional methods and inference on structural and treatment effects." *Journal of Economic Perspectives* 28.2 (2014): 29-50.
10. Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*. (2014);81(2):608-650.
11. Robinson, Peter M. "Root-N-consistent semiparametric regression." *Econometrica: journal of the Econometric Society* (1988): 931-954.
12. Zivich PN, Breskin A. Machine learning for causal inference: on the use of cross-fitting. *Epidemiology*. (2021);32(3):393-401.
13. Fitzke, Heather, et al. "Real-world evidence: state-of-the-art and future perspectives." *Journal of comparative effectiveness research* 14.4 (2025): e240130.
14. Mahendraratnam, Nirosha, et al. "Understanding use of real-world data and real-world evidence to support regulatory decisions on medical product effectiveness." *Clinical Pharmacology & Therapeutics* 111.1 (2022): 150-154.
15. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* (1983);70(1):41–55.
16. Pirracchio R, Resche-Rigon M, and Chevret S. Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. *BMC medical research methodology* 12.1 (2012): 70.
17. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. (2000);11(5):550-560.
18. Austin PC, and Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine* 34.28 (2015): 3661-3679.
19. Bach P, et al. DoubleML-an object-oriented implementation of double machine learning in python. *Journal of Machine Learning Research* 23.53 (2022): 1-6.
20. Brookhart MA, et al. Variable selection for propensity score models. *American journal of epidemiology* 163.12 (2006): 1149-1156.
21. Westreich D, Lessler J, and Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *Journal of clinical epidemiology* 63.8 (2010): 826.