BMC Bioinformatics

RESEARCH

Open Access

Robust double machine learning model with application to omics data

Xuqing Wang[1†], Yahang Liu[1†], Guoyou Qin[1,2*] and Yongfu Yu[1,2*]

†Xuqing Wang and Yahang Liu have contributed equally to this work.

*Correspondence: gyqin@fudan.edu.cn; yu@fudan.edu.cn

[1] Department of Biostatistics, Key Laboratory of Public Health Safety of Ministry of Education, Key Laboratory for Health Technology Assessment, National Commission of Health, School of Public Health, Fudan University, Shanghai, China
[2] Shanghai Institute of Infectious Disease and Biosecurity, Shanghai, China

Abstract

***Background*:** Recently, there has been a growing interest in combining causal inference with machine learning algorithms. Double machine learning model (DML), as an implementation of this combination, has received widespread attention for their expertise in estimating causal effects within high-dimensional complex data. However, the DML model is sensitive to the presence of outliers and heavy-tailed noise in the outcome variable. In this paper, we propose the robust double machine learning (RDML) model to achieve a robust estimation of causal effects when the distribution of the outcome is contaminated by outliers or exhibits symmetrically heavy-tailed characteristics.

***Results*:** In the modelling of RDML model, we employed median machine learning algorithms to achieve robust predictions for the treatment and outcome variables. Subsequently, we established a median regression model for the prediction residuals. These two steps ensure robust causal effect estimation. Simulation study show that the RDML model is comparable to the existing DML model when the data follow normal distribution, while the RDML model has obvious superiority when the data follow mixed normal distribution and t-distribution, which is manifested by having a smaller RMSE. Meanwhile, we also apply the RDML model to the deoxyribonucleic acid methylation dataset from the Alzheimer's disease (AD) neuroimaging initiative database with the aim of investigating the impact of Cerebrospinal Fluid Amyloid β42 (CSF Aβ42) on AD severity.

***Conclusion*:** These findings illustrate that the RDML model is capable of robustly estimating causal effect, even when the outcome distribution is affected by outliers or displays symmetrically heavy-tailed properties.

Keywords: Causal inference, Observational study, Double machine learning, Outlier, Heavy-tailed, Robustness

Background

Causal inference holds paramount importance in biomedical research. The gold standard for causal inference is typically the Randomized Controlled Trial (RCT) [3, 32]. However, the implementation of RCTs frequently encounters challenges related to ethical considerations, participant compliance, and time cost [32]. Consequently, researchers often explore causality through observational studies; nevertheless, these studies are

subject to confounding bias due to the lack of randomization. The propensity score (PS) method is widely employed in observational studies to adjust for measured confounders. Yet this method heavily depends on the correct specification of the model; any misspecification may lead to biased estimates. Furthermore, when dealing with datasets with high-dimensional covariates-such as omics data-the curse of dimensionality poses significant issues, making PS estimation exceedingly complex. Therefore, addressing problems related to model misspecification and variable dimensionality while performing reliable causal inference on high-dimensional, complex biological information datasets remains a substantial challenge in observational studies.

Currently, there is a growing body of literature that explores the integration of machine learning techniques into causal inference to address the above issues [2, 23, 24]. Notably, the double machine learning (DML) model proposed by Chernozhukov et al. has garnered widespread attention [2]. Within the framework of a partially linear model, DML allows for the estimation of the average treatment effect. The estimation process can be decomposed into two stages: in the first stage, machine learning algorithms are employed to predict both treatment and outcome variables; in the second stage, a least squares regression model is constructed to estimate the treatment effect. The machine learning techniques introduced in the DML model are not only capable of handling complex function forms for variables such as quadratic terms and interaction terms, but also of relaxing the constraint on variables dimensionality. This effectively solves the problem of model misspecification caused by complex function forms and the curse of dimensionality associated with high-dimensional data. Furthermore, the cross-fitting technique adopted in the two-stage estimation process of DML eliminates the regularization bias introduced by machine learning techniques. At present, DML has been extended to more complex causal models to identify more intricate causal effects [6, 7, 21, 22]. For instance, Farbmacher et al. [6] combined causal mediation analysis with DML to propose estimation methods for natural direct effect, natural indirect effect, and controlled direct effect in contexts with numerous potential confounding variables. Bodory et al. [7] integrated dynamic analysis with DML method to measure the causal effects of multiple treatment variables over different periods and employed weighted estimation to assess dynamic treatment effects for specific subgroups, thus enhancing the dynamic quantification extension of DML models.

However, the existing research based on DML model mostly use traditional machine learning techniques for prediction, which can achieve satisfactory prediction results when the data is free from outliers and follows normal distribution. In the field of bioinformatics, the observational data collected frequently contains outliers [27–29]. An outlier, or outlying observation, is one that appears to deviate markedly from other members of the sample in which it occurs [25, 26]. In our real data application, we used the Alzheimer's disease neuroimaging initiative (ADNI) dataset to explore the impact of Cerebrospinal Fluid Amyloid β42 (CSF Aβ42) on AD severity, with AD severity serving as the outcome variable and being measured by 11-item Alzheimer's Disease Assessment Scale (ADAS-11) cognitive scores. Nevertheless, we have detected some potential outliers in the ADAS-11 cognitive scores through residual analysis, as shown in Fig. 3. In such a case, the sensitivity of traditional machine learning models to outliers in the outcome variable can hinder achieving satisfactory prediction results within the DML

framework. Fuhr et al. [1] conducted simulation experiments comparing multiple DML models based on different machine learning techniques, finding that smaller prediction errors in the first stage correlated with superior estimation performance in the second stage. Therefore, poor prediction performance caused by outliers impacts the final causal effect estimation of DML. In addition, many bioinformatics data such as genomic data exhibit heavy-tailed distribution [34]. Heavy-tailed distribution has thicker tails compared to normal distribution, which can accommodate outliers in the data [35]. Thus, applying DML model to datasets with heavy-tailed noise in the outcome variable may also lead to unreliable estimate of treatment effect. Owing to the fact that the median regression model demonstrates a certain degree of robustness to outliers and heavy-tailed noise [30], combining machine learning models with median regression can enhance the robustness of machine learning models to outliers heavy-tailed noise, improve prediction accuracy, and further obtain more reliable causal effect estimate. In this paper, to reduce the impact of outliers and heavy-tailed noise in the outcome variable on the causal effect estimation of DML model, we propose a robust double machine learning (RDML) model within the framework of partially linear regression model. Our proposal differs from the DML proposed by Chernozhukov et al. [2] in that it employs median machine learning methods instead of traditional machine learning methods for predicting treatment and outcome variables in the first stage, and in the second stage, it uses linear median regression model instead of an ordinary least squares regression model to estimate treatment effect. Simulation results indicate that when the outcome variable follows a standard normal distribution, our proposed model is comparable to general DML model. However, when the outcome variable follows a mixture of normal distributions or t-distribution, our proposal demonstrates significant advantages over general DML model, as evidenced by a smaller mean squared error.

This paper is organized as follows. First, we begin with a brief introduction to the DML model and propose the new model of RDML. Next, several numerical simulations are implemented to illustrate the superior performance of our proposal. Then, we apply the RDML model to the deoxyribonucleic acid methylation dataset from the Alzheimer's disease (AD) neuroimaging initiative database to investigating the impact of Cerebrospinal Fluid Amyloid β42 (CSF Aβ42) on AD severity. Finally, we conduct some discussion.

Methods

Double machine learning

In this section, we review the double machine learning model [2]. Considering the following partial linear regression (PLR) model,

$$Y = D\theta_0 + g_0(X) + U, E[U|X, D] = 0, \tag{1}$$

$$D = m_0(X) + V, E[V|X] = 0 \tag{2}$$

where D represents the continuous treatment variable and Y denotes the continuous outcome variable. The vector $(X_1, \ldots, X_p)^T$ comprises p-dimensional observable confounding variables, while U and V are disturbance terms and follow symmetric distribution. We assume that the functional forms $m(\cdot)$ and $g(\cdot)$, which describe the effects of X on D and Y, respectively, are unknown. Additionally, we posit that the effect of D on Y is linear.

Our main interest is to estimate the average treatment effect of D on Y, i.e., θ_0, where θ_0 is a scalar. The DML model proposed by Chernozhukov et al. [2] serves as a novel method for estimating θ_0. This approach leverages machine learning algorithms to relax assumptions about functional forms and constraints on variable dimensionality, thereby enabling the flexible handling of high-dimensional complex datasets. Additionally, it incorporates the cross-fitting technique to eliminate regularization bias introduced by machine learning algorithms. The specific estimation steps of this model are as follows: Firstly, randomly divide the dataset into K subsamples. Secondly, for each K-th subsample, use the remaining K-1 samples to train two machine learning models for D and Y with respect to X. Thirdly, apply the trained machine learning models to predict the conditional expectation of D given X, $E(D|X)$ and the conditional expectation of Y given X, $E(Y|X)$ for the K-th subsample. Next, obtain the residuals of D and Y by subtracting the predicted values from their actual values. Then, fit a ordinary least squares model of the residuals of Y on the residuals of D, i.e., $Y - \hat{E}(Y|X) = [D - \hat{E}(D|X)]\theta_0 + U$, define score function

$$\begin{aligned}\psi(W;\theta,\eta) &:= -\frac{1}{2}\frac{d\left\{Y - \hat{E}(Y|X) - \left[D - \hat{E}(D|X)\right]\theta\right\}^2}{d\theta} \\ &= \left\{Y - \hat{E}(Y|X) - \left[D - \hat{E}(D|X)\right]\theta\right\}\left[D - \hat{E}(D|X)\right] \\ &= \left\{Y - g(X) - [D - m(X)]\theta\right\}[D - m(X)], \eta = (g, m).\end{aligned} \tag{3}$$

It has been proved in Chernozhukov et al. [2] that θ_0 satisfies both the moment condition $E\psi(W;\theta_0,\eta_0) = 0$, and the orthogonality condition $\partial_\eta E\psi(W;\theta_0,\eta_0)[\eta - \eta_0] = 0$ where $\eta_0 = (g_0, m_0)$, $g_0(X) = E(Y|X)$ and $m_0(X) = E(D|X)$. Based on the moment condition, we can solve for θ_0 using the K-th subsample. Finally, average these K estimates to obtain the final causal effect estimate.

Robust double machine learning

Existing DML model employs traditional machine learning algorithms to predict $E(D|X)$ and $E(Y|X)$ in the first stage. When the data follows a normal distribution and free from outliers, machine learning algorithms can flexibly handle high-dimensional complex data structures, achieving satisfactory predictive performance. However, if the data is contaminated by outliers or heavy-tailed noise, the predictive accuracy of traditional machine learning models is compromised. Therefore, in order to avoid the impact of inaccurate predictions of D and Y in the first stage on the estimation of regression coefficient in the second stage, we propose a robust double machine learning model.

The first phase

Considering the robustness of median regression in machine learning models against outliers and heavy-tailed noise, we employs machine learning median regression models such as Median Regression Forests model [36] and eXtreme Gradient Boosting Median Regression model [37] instead of traditional machine learning models in the first phase of our RDML framework. In this phase, we make predictions for conditional medians of D and Y with respect to X, obtaining prediction results denoted as $\widehat{Q}_{0.5}(D|X)$ and $\widehat{Q}_{0.5}(Y|X)$ respectively.

The second phase

For the prediction results $\widehat{Q}_{0.5}(D|X)$ and $\widehat{Q}_{0.5}(Y|X)$ obtained in the first phase, subtracting the predicted values from the true values yields residuals $\widehat{V}_D = D - \widehat{Q}_{0.5}(D|X)$ and $\widehat{V}_Y = Y - \widehat{Q}_{0.5}(Y|X)$, respectively. Subsequently, fit a median regression model of $\widehat{V}_Y$ on $\widehat{V}_D$. The causal effect estimate, denoted as $\widehat{\theta}_0$, is then derived by minimizing the following equation:

$$\widehat{\theta}_0 = \arg\min_{\theta_0} \|\widehat{V}_Y - \theta_0 \widehat{V}_D\|_1. \tag{4}$$

The detailed steps of the RDML model can be found in Algorithm 1.

---

1. Input: dataset $W = \{X, D, Y\}$.

2. Data Partitioning: Partition the dataset W into K equal parts, denote the kth sub-sample as W_k and its complement as W_k^C.

3. Cross-Fitting:
For $k = 1$ to K do
On dataset W_k^C, construct two median machine learning models for D based on X and Y based on X, respectively.
On dataset W_k, predict D and Y using two constructed models, yielding $\widehat{Q}^k_{0.5}(D|X)$ and $\widehat{Q}^k_{0.5}(Y|X)$.
Calculate residuals: $\widehat{V}^k_D = D_k - \widehat{Q}^k_{0.5}(D|X)$ and $\widehat{V}^k_Y = Y_k - \widehat{Q}^k_{0.5}(Y|X)$.
Fit a median regression model of $\widehat{V}^k_Y$ on $\widehat{V}^k_D$ to obtain a coefficient estimate $\widehat{\theta}^k_0$.
end

4. Output: Return $\widehat{\theta}_0 = 1/K \sum_{k=1}^{K} \widehat{\theta}^k_0$.

Simulations

In this section, to evaluate the performance of the proposed RDML model, we conducted the following simulation study.

Data generation process 1

We first consider the following data generation process (DGP):

$$D = \alpha_1 + \delta_1 X_1 + \delta_2 X_2^2 + \delta_3 X_1 X_2 + \delta_4 |X_3| + \delta_5 X_4^3 + U \tag{5}$$

$$Y = \alpha_2 + D\theta + \gamma_1 X_1 + \gamma_2 X_2^2 + \gamma_3 X_1 X_2 + \gamma_4 |X_3| + \gamma_5 X_4^3 + V \tag{6}$$

where $X = (X_1, X_2, X_3, X_4)$ are randomly generated from a multivariate normal distribution with mean 0 and covariance Σ, denoted as $N(0, \Sigma)$, $\Sigma = AA^T$, A is a four-dimensional column vector, with each element randomly generated from the normal distribution $N(0, 0.5)$. In addition, we set $\alpha_1 = \alpha_2 = 0.5$, $\delta_j = \gamma_j = t, j = 1, 2, 3, 4, 5$ represent the magnitude of the confounding coefficient, and the sample size is denoted as n.

For t and n, we consider the following three combinations: $(t, n) = (0.3, 200)$, (0.6, 200), and (0.3, 500).

In order to fully explore the robustness of the proposed model to the presence of outliers and symmetrically heavy-tailed noise in the outcome variable, we set the noise term U to follow a standard normal distribution, but for the noise term V, consider the following three scenarios:

Scenario 1

In Scenario 1, we postulate that the noise term V follows a standard normal distribution, specifically $N(0, 1)$. Consequently, both the treatment variable D and the outcome variable Y exhibit symmetric and light-tailed distributions, resulting in a low probability of outlier generation.

Scenario 2

In Scenario 2, to assess the robustness of the proposed model in the presence of outliers in the outcome variable, we specify that the noise term V follows a mixture of normal distributions. This mixture is constructed by combining several distinct normal distributions with specific weights. In our simulation study, we focus on a mixture consisting of two normal distributions, both with a mean of 0, but with variances of 1 and 100, respectively. The incorporation of a component with a significantly larger variance alongside the standard normal distribution introduces a long-tail behavior, which is more susceptible to outlier generation. Recognizing that higher mixing proportions elevate the likelihood of outlier occurrence, we systematically vary the mixing weights to 0.1, 0.2, and 0.3, aiming to explore their influence on the estimation of causal effect within our modeling framework.

Scenario 3

In Scenario 3, we assume that V follows a t-distribution. When the degree of freedom (df) at a low level, the t-distribution exhibits heavier tail compared to standard normal distribution, thereby increasing the probability of extreme values occurring. As the df increase, the t-distribution gradually approaches the standard normal distribution. In our simulations, we set df of t-distribution to be 1.5 and 3.

Regarding the simulated data generated, we have constructed two robust DML models utilizing Median Regression Forest (MRF) and eXtreme Gradient Boosting Median Regression (MXGBoost) as the respective machine learning algorithms, which aim to predict the treatment variable D and outcome variables Y. To facilitate comparison with robust DML models, we have further developed two corresponding general DML models, employing Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) Regression as predictors. The chosen predictors, namely RF and XGBoost are ensemble learning algorithms that are widely recognized in the field of machine learning. They operate by generating multiple weak learners and subsequently combining them to enhance the model's accuracy and generalization capability. All four DML models adopt two-fold cross-fitting technique to mitigate the regularization bias inherent in machine learning algorithms. Each machine learning algorithm utilized default hyperparameters, and the DGP is replicated 100 times for each model. For the purpose of comparing and

analyzing the performance of these models, our study used relative bias and root mean square error (RMSE) as the primary evaluation metrics.

Data generation process 2

To further investigate the impact of the number of confounding variables on model performance, we consider the following second DGP.

$$D = \alpha_1 + \sum_{j=1}^{p} \delta_j X_j^2 + U \tag{7}$$

$$Y = \alpha_2 + D\theta + \sum_{j=1}^{p} \gamma_j X_j^2 + V \tag{8}$$

where $\delta_j = \gamma_j = t/j, j = 1, 2, ..., p$ represent the confounding coefficients, whose value decrease as j increase. Two combinations are considered for the sample size (n) and variable dimensions (p): $(n, p) = (200, 20)$ and $(500, 50)$. For the generation of the noise term V, in order to investigate the robustness of models to outliers in the outcome variable, we also consider three different generating distributions for V: normal, mixture of normal and, t-distribution. Additionally, the generation of other variables and the setting of parameters remain consistent with those described in the previous subsection.

Scenario 4

In Scenario 4, we assume that V follows a standard normal distribution, implying that the distribution of the outcome variable Y exhibits symmetry and light tails, thereby resulting in a low likelihood of generating outliers.

Scenario 5

In Scenario 5, with the aim of investigating the robustness of different models to the presence of outliers in the outcome variable Y, we model V as a mixture of normal distributions, with the mixture ratios set to 0.1, 0.2, and 0.3.

Scenario 6

In Scenario 6, to examine the robustness of various models against heavy-tailed noise, we model the noise term V as a t-distribution while considering two cases for the degrees of freedom parameter: 1.5 and 3.

Simulation results

We have summarized all computational results of the four models under different settings in Tables 1, 2, 3, 4, 5, 6 and 7, and plotted box plots of the causal effect estimates from 100 simulations for some settings in Figs. 1 and 2.

Regarding DGP1, all results are shown in Tables 1, 2 and 3. From Table 1, which shows the results for six data distributions with n=200 and t=0.3, when both D and Y follow a standard normal distribution, the four models of Dml_RF, Dml_XGBoost, RDml_RF, and RDml_XGBoost all exhibit relatively small biases and RMSE values.

Table 1 Simulation results from DGP 1 with n=200 and t=0.3

Distributions	Methods	Estimate	Bias(%)	RMSE
$N(0, 1)$	Dml_RF	1.0483	4.8291	0.1556
	Dml_XGBoost	1.0659	6.5887	0.1898
	RDml_MRF	1.0230	2.3019	**0.1431**
	RDml_MXGBoost	1.0020	**0.2038**	0.1636
$0.9N(0, 1) + 0.1N(0, 10^2)$	Dml_RF	1.0221	2.2104	0.3338
	Dml_XGBoost	1.0580	5.8013	0.3250
	RDml_MRF	1.0086	0.8597	**0.1524**
	RDml_MXGBoost	0.9980	**−0.1971**	0.1684
$0.8N(0, 1) + 0.2N(0, 10^2)$	Dml_RF	1.0189	1.8889	0.4229
	Dml_XGBoost	1.0768	7.6789	0.4251
	RDml_MRF	0.9937	**−0.6287**	**0.1830**
	RDml_MXGBoost	0.9826	−1.7397	0.1849
$0.7N(0, 1) + 0.3N(0, 10^2)$	Dml_RF	0.9819	−1.8113	0.4860
	Dml_XGBoost	1.0254	2.5400	0.4734
	RDml_MRF	0.9931	**−0.6901**	**0.2360**
	RDml_MXGBoost	0.9899	−1.0102	0.2459
$t(3)$	Dml_RF	1.0616	6.1591	0.2070
	Dml_XGBoost	1.0701	7.0124	0.2237
	RDml_MRF	1.0420	4.1956	**0.1855**
	RDml_MXGBoost	1.0302	**3.0224**	0.1875
$t(1.5)$	Dml_RF	1.0770	7.6975	0.5492
	Dml_XGBoost	1.1257	12.5715	0.5843
	RDml_MRF	1.0331	**3.3131**	0.2392
	RDml_MXGBoost	1.0361	3.6095	**0.2167**

Notations: The minimum Bias and RMSE have been indicated in bold. RMSE is calculated as $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\theta}_0^m - \theta_0)^2}$, where M denotes the number of simulations; Bias(%) refers to Relative Bais is calculated as $\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_0^m - \theta_0\right)\Big/\theta_0 \times 100\%$

When the noise term V is drawn from a mixed normal distribution, the RMSE values of two general DML models increase significantly compared to those obtained under the standard normal distribution. Furthermore, as the mixing proportion gets larger, the extent of RMSE value increment becomes greater. Specifically, at a mixing ratio of 0.3, compare with a mixing ratio of 0, the RMSE values increase from 0.1556 and 0.1898 to 0.4860 and 0.4734, respectively for the Dml_RF and Dml_XGBoost models. This indicates that general DML models are not robust to outliers in the outcome variable. While the RMSE values of the two robust DML models also increase to some extent, the increment is slight. At a mixing ratio of 0.3, the RMSE values of two robust DML models are close to half of those of the corresponding general DML models. When the noise term V follows a t-distribution, as the degree of freedom decrease, the RMSE values of the two general DML models increase by 165.31% and 161.20%, respectively. However, the two robust DML models maintain relatively robust performance, exhibiting no significant changes in both bias and RMSE metrics. To provide a more intuitive presentation of the estimation results from different models, we have plotted box plots for each scenario under DGP1 with n=200 and t=0.3 in Fig. 1. It can

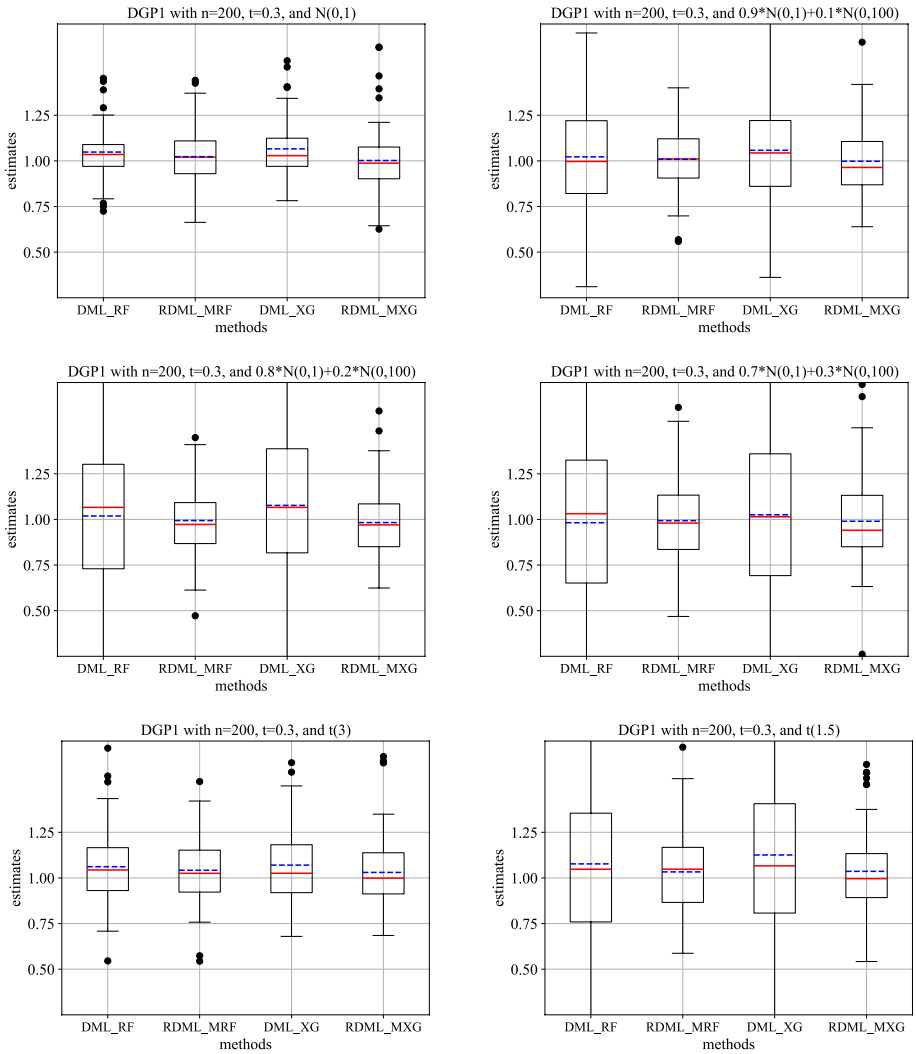

Fig. 1 Box plots of the estimates of θ for 100 simulations under DGP 1 with n=200 and t=0.3. The solid red line and dashed blue line represent the median and mean of 100 estimates, respectively, while the black dots signify outliers

be clearly observed that, except for cases where the data follow a standard normal distribution or a t-distribution with 3 degrees of freedom, the general DML models exhibit unstable estimation results characterized by high variance. In contrast, RDML models demonstrate small variance across all scenarios. These findings are consistent with those presented in Table 1. For Table 2, the simulation results are shown with all other settings kept the same as in Table 1, but with the confounding coefficients increased from 0.3 to 0.6. From these results, we can draw similar conclusions as those from Table 1. Additionally, a comparison between Tables 1 and 2 reveals that as the confounding coefficient increases, both bias and RMSE for the four models generally exhibit an upward trend in most settings. Table 3 shows the simulation results when the sample size is increased to 500 while keeping the confounding coefficient at 0.3. Comparing Tables 1 and 3 indicates that for robust DML models, an increase in sample size leads to a reduction in RMSE values across all settings.

Table 2 Simulation results from DGP 1 with n=200 and t=0.6

Distributions	Methods	Estimate	Bias(%)	RMSE
$N(0, 1)$	Dml_RF	1.1098	10.9843	0.2268
	Dml_XGBoost	1.1553	15.5257	0.2926
	RDml_MRF	1.0524	**5.2364**	**0.1681**
	RDml_MXGBoost	1.0731	7.3140	0.2388
$0.9N(0, 1) + 0.1N(0, 10^2)$	Dml_RF	1.0941	9.4077	0.3642
	Dml_XGBoost	1.1435	14.3508	0.3858
	RDml_MRF	1.0470	**4.7039**	**0.1860**
	RDml_MXGBoost	1.0581	5.8104	0.2169
$0.8N(0, 1) + 0.2N(0, 10^2)$	Dml_RF	1.0923	9.2261	0.4405
	Dml_XGBoost	1.1616	16.1625	0.4716
	RDml_MRF	1.0346	3.4556	**0.2105**
	RDml_MXGBoost	1.0291	**2.9127**	0.2367
$0.7N(0, 1) + 0.3N(0, 10^2)$	Dml_RF	1.0562	5.6168	0.4921
	Dml_XGBoost	1.1132	11.3185	0.5184
	RDml_MRF	1.0350	3.5049	**0.2476**
	RDml_MXGBoost	1.0331	**3.3094**	0.2635
$t(3)$	Dml_RF	1.1298	12.9794	0.2674
	Dml_XGBoost	1.1661	16.6110	0.3225
	RDml_MRF	1.0808	**8.0813**	**0.219**1
	RDml_MXGBoost	1.0885	8.8548	0.2643
$t(1.5)$	Dml_RF	1.1404	14.0365	0.5486
	Dml_XGBoost	1.1984	19.8380	0.6100
	RDml_MRF	1.0745	**7.4501**	**0.2755**
	RDml_MXGBoost	1.0983	9.8318	0.2881

Notations: The minimum Bias and RMSE have been indicated in bold. RMSE is calculated as $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\theta}_0^m - \theta_0)^2}$, where M denotes the number of simulations; Bias(%) refers to Relative Bais is calculated as $\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_0^m - \theta_0\right)\Big/\theta_0 \times 100\%$

Concerning DGP 2, we vary the data generation process while augmenting the number of confounding variables. The simulation results are summarized in Tables 4, 5, 6 and 7, and Fig. 2 corresponds to the estimated results of Table 4. Tables 4 and 5 are related to the settings of $(n, p) = (200, 20)$, from Table 4, we observe that while the biases in the causal effect estimates of Dml_RF and Dml_XGBoost remain low, there is a considerable increase in RMSE compared to a standard normal distribution, particularly when the mixture ratio set to 0.3. As the coefficient of confounding increases, Table 5 indicates that the general DML models exhibit poor performance in both bias and RMSE metrics. Furthermore, when the noise term V adheres to a t-distribution and the degree of freedom vary from 3 to 1.5, regardless of whether the confounding coefficient is set to 0.3 or 0.6, the bias and RMSE values of the general DML models increase to some extent, particularly in terms of RMSE. Above analyses reaffirm that the general DML models are not robust to outliers in the outcome variables. In contrast, the robust DML models, when the noise term V follows a mixed normal distribution or a t-distribution, generally demonstrate smaller biases and RMSE values compared to the general DML models. Similar conclusions can be drawn for the setting of $(n, p) = (500, 50)$ from the results presented in Tables 6 through 7.

Table 3 Simulation results from DGP 1 with n=500 and t=0.3

Distributions	Methods	Estimate	Bias(%)	RMSE
$N(0, 1)$	Dml_RF	1.0395	3.9460	0.1119
	Dml_XGBoost	1.0556	5.5612	0.1575
	RDml_MRF	1.0166	1.6626	**0.0860**
	RDml_MXGBoost	1.0094	**0.9420**	0.1041
$0.9N(0, 1) + 0.1N(0, 10^2)$	Dml_RF	1.0219	2.1853	0.2169
	Dml_XGBoost	1.0254	2.5400	0.2186
	RDml_MRF	1.0187	1.8686	**0.0976**
	RDml_MXGBoost	1.0025	**0.2485**	0.1081
$0.8N(0, 1) + 0.2N(0, 10^2)$	Dml_RF	1.0375	3.7542	0.2763
	Dml_XGBoost	1.0245	2.4462	0.2693
	RDml_MRF	1.0205	**2.0479**	**0.1172**
	RDml_MXGBoost	1.0291	2.9127	0.2367
$0.7N(0, 1) + 0.3N(0, 10^2)$	Dml_RF	1.0562	5.6168	0.4921
	Dml_XGBoost	1.1132	11.3185	0.5184
	RDml_MRF	1.0350	3.5049	0.2476
	RDml_MXGBoost	0.9979	**-0.2124**	**0.1214**
$t(3)$	Dml_RF	1.0242	2.4212	0.1481
	Dml_XGBoost	1.0314	3.1431	0.1625
	RDml_MRF	1.0082	0.8238	**0.0980**
	RDml_MXGBoost	0.9978	**-0.2162**	0.1208
$t(1.5)$	Dml_RF	1.0421	4.2147	0.5865
	Dml_XGBoost	1.0417	4.1695	0.4773
	RDml_MRF	1.0126	1.2557	0.1232
	RDml_MXGBoost	0.9998	**-0.0203**	**0.1020**

Notations: The minimum Bias and RMSE have been indicated in bold. RMSE is calculated as $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\theta}_0^m - \theta_0)^2}$, where M denotes the number of simulations; Bias(%) refers to Relative Bais is calculated as $\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_0^m - \theta_0\right) \Big/ \theta_0 \times 100\%$

Real data application

In this section, we apply both the proposed robust double machine learning model and the general double machine learning model to the ADNI dataset with the aim of evaluating the impact of Cerebrospinal Fluid Amyloid β42 (CSF Aβ42) on AD severity. It has been shown that one of the central neuropathological features of Alzheimer's disease (AD) is the accumulation of β-amyloid (Aβ) containing neuriticplaques [10]. And the impaired clearance of Aβ is believed to accounts for 99% of sporadic AD [8] [9]. As a type of Aβ, CSF Aβ42 containing 42 amino acid residues and present in cerebrospinal fluid, is an important biomarker of AD. As such, exploring the intrinsic connection between CSF Aβ42 and AD severity, observing changes in the level of CSF Aβ42 can not only help in the early diagnosis of AD, but also monitor the progress of AD patients.

We considered CSF Aβ42 level at baseline as an exposure variable and explored its causal effect on the outcome variable, AD severity at month 24. For the measurement of outcome variable, we utilized the well-recognized 11-item Alzheimer's Disease Assessment Scale (ADAS-11) cognitive score to assess the severity of AD. The ADAS-11 scores span from 0 to 70, with higher scores indicating more severe symptoms. In addition, to reduce confounding bias, we selected age, gender, educational

Table 4 Simulation results from DGP 2 with n=200, p=20, and t=0.3

Distributions	Methods	Estimate	Bias(%)	RMSE
$N(0, 1)$	Dml_RF	1.0306	3.0586	0.1058
	Dml_XGBoost	1.0396	3.9575	0.1139
	RDml_MRF	1.0172	**1.7187**	0.1220
	RDml_MXGBoost	0.9759	−2.4059	**0.1046**
$0.9N(0, 1) + 0.1N(0, 10^2)$	Dml_RF	1.0152	1.5211	0.2850
	Dml_XGBoost	1.0383	3.8332	0.2895
	RDml_MRF	1.0128	1.2791	0.1440
	RDml_MXGBoost	0.9893	**−1.0707**	**0.1382**
$0.8N(0, 1) + 0.2N(0, 10^2)$	Dml_RF	0.9797	−2.0260	0.4523
	Dml_XGBoost	1.0261	2.6105	0.4148
	RDml_MRF	1.0027	**0.2701**	0.1828
	RDml_MXGBoost	0.9944	−0.5585	**0.1681**
$0.7N(0, 1) + 0.3N(0, 10^2)$	Dml_RF	0.9644	−3.5630	0.5770
	Dml_XGBoost	1.0112	1.1245	0.5008
	RDml_MRF	0.9915	**−0.8470**	0.2543
	RDml_MXGBoost	0.9879	−1.2094	**0.2256**
$t(3)$	Dml_RF	1.0447	4.4656	0.1891
	Dml_XGBoost	1.0542	5.4169	0.1873
	RDml_MRF	1.0294	2.9362	0.1570
	RDml_MXGBoost	0.9865	**−1.3482**	**0.1312**
$t(1.5)$	Dml_RF	1.0590	5.8990	0.6889
	Dml_XGBoost	1.0799	7.9932	0.5932
	RDml_MRF	1.0013	**0.1277**	0.1988
	RDml_MXGBoost	0.9864	−1.3567	**0.1702**

Notations: The minimum Bias and RMSE have been indicated in bold. RMSE is calculated as $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\theta}_0^m - \theta_0)^2}$, where M denotes the number of simulations; Bias(%) refers to Relative Bais is calculated as $\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_0^m - \theta_0\right) \Big/ \theta_0 \times 100\%$

attainment, and genome-wide CpG sites as candidate covariates for control. Zhang et al. found that the epigenetic changes linked to various pathological processes differ between cognitively normal individuals (some of whom may later develop Alzheimer's disease) and patients with Alzheimer's disease [11]. Furthermore, they identified a variety of novel associations between DNAm at multiple CpG sites in blood and CSF biomarkers, including CSF Aβ42, suggesting that changes in various pathological processes in the CSF are reflected in the blood epigenome [11]. In total, we collected data from 321 participants containing the above variables, with all covariates measured at baseline.

For data preprocessing, we first performed following operations on the DNAm data: (i) discarding probes with a P-value greater than 0.05; (ii) excluding probes associated with gender; (iii) removing probes containing SNPs at CpG sites; (iv) eliminating cross-reactive probes, and (v) averaging the DNAm levels for samples that were measured multiple times [12]. Following this pre-processing procedure, we retained 865,859 CpG sites as candidate covariates for further analysis. Due to the excessive dimensionality of the variables, we then performed epigenome-wide association study (EWAS) analyses and chose the leading 100 CpG sites according to the Bonferroni-adjusted P values for each site. In addition, controlling for all confounders in

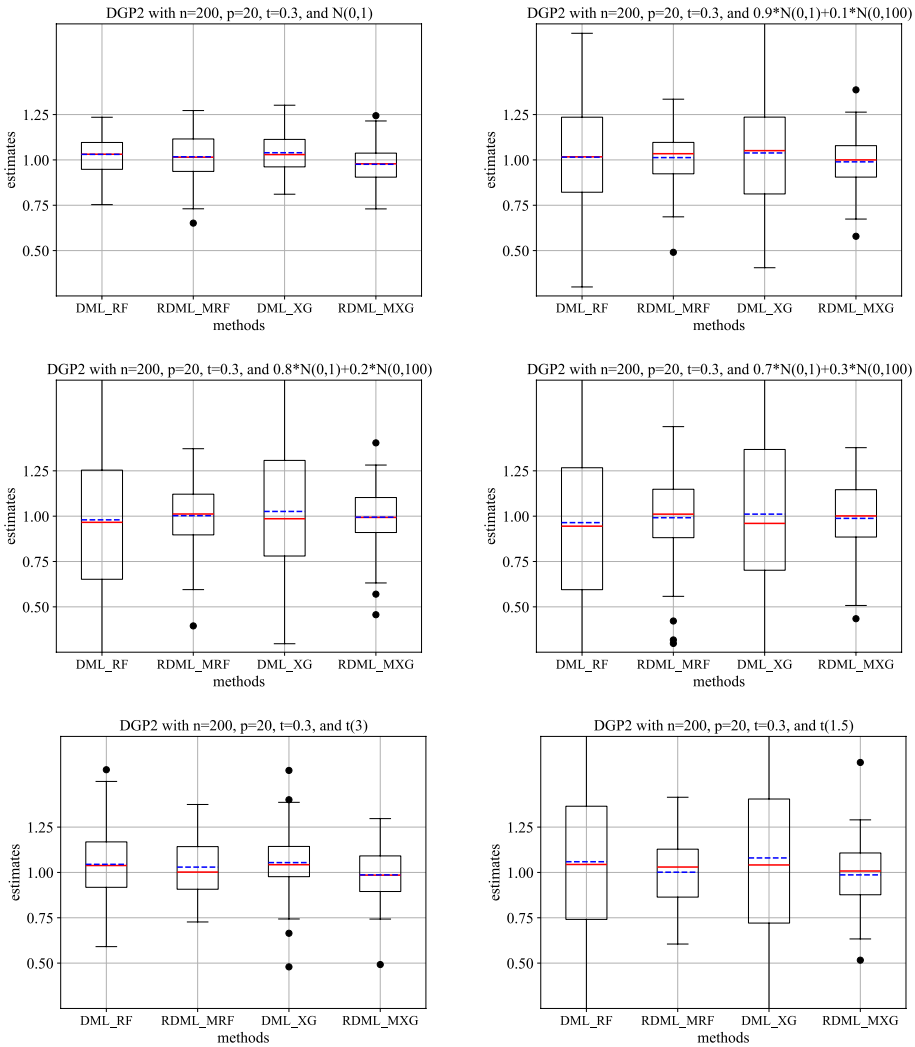

Fig. 2 Box plots of the estimates of θ for 100 simulations under DGP 2 with n=200, p=20, and t=0.3. The solid red line and dashed blue line represent the median and mean of 100 estimates, respectively, while the black dots signify outliers

observational studies is crucial to obtaining unbiased estimate of treatment effect, but adding instrumental covariates except for confounders might reduce the efficiency of the estimation and incorporating prognostic covariates can enhance estimation efficiency [17–20]. Therefore, we subsequently applied the generalized median adaptive lasso (GMAL) method to perform variable selection on the 100 CpG sites after dimensionality reduction, along with three covariates: age, gender, and education level [17]. Ultimately, we selected 44 CpG sites, as shown in Table 8.

Before establishing models to estimate the treatment effect, we first identify potential outliers in the outcome variable ADAS-11. In regression analysis, an outlier is typically identified when the standardized residual exceeds a given threshold c [32]. Therefore, we begin by constructing an ordinary least squares regression model of the outcome variable ADAS-11 on the treatment variable CSF Aβ42 and covariates selected via the GMAL method. Subsequently, we calculate the standardized residuals for each sample

Table 5 Simulation results from DGP 2 with n=200, p=20, and t=0.6

Distributions	Methods	Estimate	Bias(%)	RMSE
$N(0, 1)$	Dml_RF	1.0974	9.7404	0.1606
	Dml_XGBoost	1.1489	14.8916	0.2222
	RDml_MRF	1.0472	**4.7156**	**0.1299**
	RDml_MXGBoost	1.0538	5.3753	0.1447
$0.9N(0, 1) + 0.1N(0, 10^2)$	Dml_RF	1.0796	7.9598	0.2958
	Dml_XGBoost	1.1395	13.9486	0.3243
	RDml_MRF	1.0458	**4.5779**	0.1649
	RDml_MXGBoost	1.0514	5.1447	**0.1632**
$0.8N(0, 1) + 0.2N(0, 10^2)$	Dml_RF	1.0484	4.8408	0.4542
	Dml_XGBoost	1.1330	13.3049	0.4282
	RDml_MRF	1.0333	**3.3320**	**0.2023**
	RDml_MXGBoost	1.0596	5.9632	0.2037
$0.7N(0, 1) + 0.3N(0, 10^2)$	Dml_RF	1.0292	**2.9197**	0.5611
	Dml_XGBoost	1.1148	11.4752	0.4997
	RDml_MRF	1.0307	3.0722	0.2668
	RDml_MXGBoost	1.0579	5.7886	**0.2483**
$t(3)$	Dml_RF	1.1145	11.4548	0.2218
	Dml_XGBoost	1.1604	16.0408	0.2521
	RDml_MRF	1.0592	**5.9213**	**0.1651**
	RDml_MXGBoost	1.0692	6.9174	0.1816
$t(1.5)$	Dml_RF	1.1214	12.1413	0.6805
	Dml_XGBoost	1.1850	18.5015	0.5720
	RDml_MRF	1.0310	**3.1003**	0.2008
	RDml_MXGBoost	1.0648	6.4836	**0.1899**

Notations: The minimum Bias and RMSE have been indicated in bold. RMSE is calculated as $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\theta}_0^m - \theta_0)^2}$, where M denotes the number of simulations; Bias(%) refers to Relative Bais is calculated as $\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_0^m - \theta_0\right)\Big/\theta_0 \times 100\%$

observation, as illustrated in Fig. 3. In our study, we set c = 2.5 [33]; six observations with standardized residuals exceeding this threshold are marked in red and identified as outliers. Consequently, it is justifiable to investigate the causal effect of CSF Aβ42 on AD severity using our proposed RDML model due to its robustness against outliers. For the selection of the machine learning model, the tuning of the parameters and the number of folds for cross-fitting, we are consistent with the simulation section.

The causal effect estimates and the corresponding 95% confidence intervals (95% CIs) for the different models are presented in Table 9. These CIs were calculated using bootstrapping with 200 replications. As shown in Table 9, there is a negative correlation between CSF Aβ42 levels and AD scores. This relationship can be attributed to the tendency of Aβ42 to form plaques in the brains of patients with Alzheimer's disease, resulting in lower concentrations of Aβ42 in the CSF. When comparing the robust double machine learning model to the general double machine learning model, it is evident that the estimate of causal effect from two robust double machine learning models are smaller.

Table 6 Simulation results from DGP 2 with n=500, p=50, and t=0.3

Distributions	Methods	Estimate	Bias(%)	RMSE
$N(0, 1)$	Dml_RF	1.0160	1.6031	0.0665
	Dml_XGBoost	1.0215	2.1506	0.0705
	RDml_MRF	1.0146	1.4619	0.0775
	RDml_MXGBoost	0.9955	**-0.4511**	**0.0589**
$0.9N(0, 1) + 0.1N(0, 10^2)$	Dml_RF	1.0350	3.4966	0.2187
	Dml_XGBoost	1.0187	1.8669	0.2040
	RDml_MRF	1.0013	**0.1328**	0.0888
	RDml_MXGBoost	0.9947	−0.5319	**0.0793**
$0.8N(0, 1) + 0.2N(0, 10^2)$	Dml_RF	1.0489	4.8871	0.2846
	Dml_XGBoost	1.0288	2.8781	0.2509
	RDml_MRF	1.0040	0.4037	0.1044
	RDml_MXGBoost	0.9976	**−0.2408**	**0.0938**
$0.7N(0, 1) + 0.3N(0, 10^2)$	Dml_RF	1.0757	7.5656	0.3328
	Dml_XGBoost	1.0548	5.4756	0.2953
	RDml_MRF	1.0092	0.9164	0.1339
	RDml_MXGBoost	1.0053	**0.5279**	**0.1249**
$t(3)$	Dml_RF	1.0218	2.1848	0.1147
	Dml_XGBoost	1.0331	3.3080	0.1146
	RDml_MRF	1.0096	0.9633	0.1017
	RDml_MXGBoost	1.0045	**0.4471**	**0.0818**
$t(1.5)$	Dml_RF	0.9095	−9.0519	0.6393
	Dml_XGBoost	0.9142	−8.5839	0.5011
	RDml_MRF	1.0047	**0.4681**	0.1099
	RDml_MXGBoost	0.9809	−1.9064	**0.0980**

Notations: The minimum Bias and RMSE have been indicated in bold. RMSE is calculated as $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\theta}_0^m - \theta_0)^2}$, where M denotes the number of simulations; Bias(%) refers to Relative Bais is calculated as $\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_0^m - \theta_0\right)\Big/\theta_0 \times 100\%$

Discussion

In our article, we have improved the general double machine learning model and proposed a robust double machine learning model. Different from the general DML model, our proposal employs the median machine learning algorithms in the first stage to predict the treatment variable D and the outcome variable Y, and then utilizes the median regression model in the second stage to estimate the causal effect. Our proposed model achieves a robust causal effect estimation for outcome variable with outliers and heavy-tailed noise. Simulation results show that when the data follows a mixed normal distribution or a t-distribution, our proposed RDML model always performs better than the general DML model. As the mixing proportion in the mixed normal distribution increases or the degree of freedom parameter in the t-distribution decreases, indicating a longer-tailed or thicker-tailed distribution, the probability of generating outliers increases. In such cases, the general DML model provides unsatisfactory estimates of causal effect, manifesting as a significant increase in RMSE values. However, our proposed model exhibits greater robustness. One possible explanation is that traditional machine learning algorithms tend to have poor generalization performance when outliers exist in the data, which further impacts

Table 7 Simulation results from DGP 2 with n=500, p=50, and t=0.6

Distributions	Methods	Estimate	Bias(%)	RMSE
$N(0, 1)$	Dml_RF	1.0666	6.6558	0.1107
	Dml_XGBoost	1.1202	12.0234	0.1684
	RDml_MRF	1.0277	**2.7734**	**0.0811**
	RDml_MXGBoost	1.0510	5.1043	0.0967
$0.9N(0, 1) + 0.1N(0, 10^2)$	Dml_RF	1.0866	8.6573	0.2359
	Dml_XGBoost	1.1157	11.5734	0.2462
	RDml_MRF	1.0181	**1.8117**	**0.0932**
	RDml_MXGBoost	1.0450	4.4992	0.1003
$0.8N(0, 1) + 0.2N(0, 10^2)$	Dml_RF	1.1020	10.1969	0.2934
	Dml_XGBoost	1.1312	13.1221	0.2842
	RDml_MRF	1.0184	**1.8399**	**0.1086**
	RDml_MXGBoost	1.0427	4.2716	0.1099
$0.7N(0, 1) + 0.3N(0, 10^2)$	Dml_RF	1.1310	13.0978	0.3405
	Dml_XGBoost	1.1549	15.4864	0.3247
	RDml_MRF	1.0256	**2.5594**	0.1336
	RDml_MXGBoost	1.0477	4.7657	**0.1181**
$t(3)$	Dml_RF	1.0723	7.2263	0.1500
	Dml_XGBoost	1.1306	13.0592	0.2001
	RDml_MRF	1.0268	**2.6771**	**0.1056**
	RDml_MXGBoost	1.0588	5.8769	0.1176
$t(1.5)$	Dml_RF	0.9605	−3.9480	0.6331
	Dml_XGBoost	1.0024	**0.2350**	0.4992
	RDml_MRF	1.0241	2.4114	**0.1133**
	RDml_MXGBoost	1.0489	4.8950	0.1350

Notations: The minimum Bias and RMSE have been indicated in bold. RMSE is calculated as $\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\widehat{\theta}_0^m - \theta_0)^2}$, where M denotes the number of simulations; Bias(%) refers to Relative Bais is calculated as $\left(\frac{1}{M}\sum_{m=1}^{M}\widehat{\theta}_0^m - \theta_0\right)\Big/\theta_0 \times 100\%$

Table 8 The CpG sites selected by GMAL method

cg21464891	cg06201680	cg25250374	cg03159926
cg17332016	cg11944093	cg10493585	cg08967134
cg07542043	cg04426031	cg10455672	cg20208633
cg03778029	cg19772847	cg14143728	cg03801758
cg08155817	cg21904271	cg04933176	cg02830555
cg10260572	cg05880455	cg13109095	cg01195881
cg02564761	cg15664161	cg20210263	cg20901246
cg26372998	cg17318529	cg12974637	cg00382930
cg14825116	cg13296755	cg08339744	cg00267207
cg16390418	cg05499127	cg10263181	cg18302890
cg09850632	cg05214708	cg12551751	cg25181751

the estimation of causal effects in the second stage. Additionally, the ordinary least squares regression model in the second stage is also susceptible to the influence of outliers. In contrast, our proposed model employs both the median machine learning

Table 9 Causal estimator and corresponding 95% CIs using four models

Methods	Estimate	95%CI
Dml_RF	−0.0470	[−0.0600, −0.0313]
Dml_XGBoost	−0.0403	[−0.0575, −0.0230]
RDml_MRF	−0.0323	[−0.0454, −0.0127]
RDml_MXGBoost	−0.0274	[−0.0436, −0.0136]

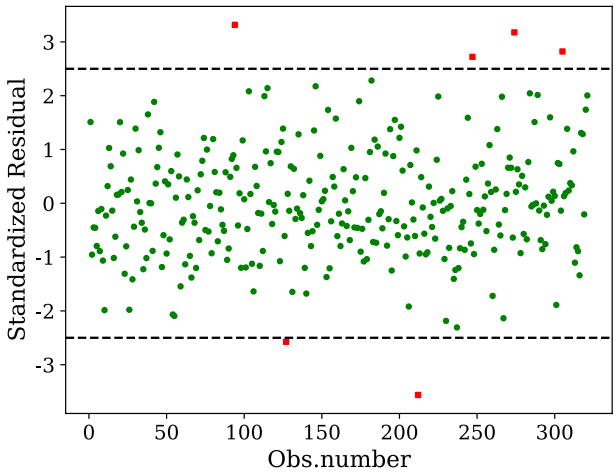

Fig. 3 The standardized residuals for 321 observations. Points with standardized residuals exceeding the threshold c=2.5, marked in red, are considered as outliers

algorithms and the median regression model, which perform more robustly against outliers.

However, in our simulation study, we did not consider the robustness of the model to skewed distributions. Given this limitation, it would be worthwhile to investigate the robustness of the model to the presentation of skewed distributions for the outcome variable in future studies. Furthermore, in our analysis of the ADNI study, we encountered a challenge: numerous CpG sites were still present after pre-processing. To address this, we first performed an EWAS analysis to reduce the dimensionality of the variables. Subsequently, we used the GWAL method for variable selection and finally built both general DML and robust DML models. Among the various methods we considered, we acknowledge that there may be other more suitable dimension reduction methods worth exploring. Additionally, within the framework of our research method, it is possible to extend median regression to more flexible quantile regression. This extension allows for the estimation of quantile treatment effects and enables us to capture the impact of the treatment variable on the entire conditional distribution of the outcome variable.

Abbreviations

DML	Double machine learning
RDML	Robust double machine learning
AD	Alzheimer's disease
CSF Aβ42	Cerebrospinal Fluid Amyloid β42
RCT	Randomized Controlled Trial
PS	Propensity score
ADNI	Alzheimer's disease neuroimaging initiative

ADAS-11	11-Item Alzheimer's Disease Assessment Scale
PLR	Partial linear regression
DGP	Data generation process
df	Degree of freedom
MRF	Median Regression Forest
MXGBoost	EXtreme Gradient Boosting Median Regression
RF	Random Forest
XGBoost	EXtreme Gradient Boosting
RMSE	Root mean square error
GMAL	Generalized median adaptive lasso
CI	Confidence intervals

Acknowledgements
Not applicable.

Author contributions
G.Q. and Y.Y. designed the study. X.W. and Y.L. conducted the analysis and drafted the manuscript, including figures and tables. All authors have offered essential feedback on the draft, as well as reviewed and approved the final version of the manuscript. X.W. and Y.L.'s contributions to this work were equal.

Funding
This work was supported by National Natural Science Foundation of China (82473724 to GYQ, No. 82273730 to YFY), Shanghai Rising-Star Program (21QA1401300 to YFY), Shanghai Municipal Natural Science Foundation (22ZR1414900 to YFY), the Three-Year Public Health Action Plan of Shanghai (GWVI—11.2-XD10 to YFY), and Shanghai Municipal Science and Technology Major Project (ZD2021CY001 to GYQ).

Data availability
The publicly available datasets analysed during the current study can be found at: https://adni.loni.usc.edu.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
No Conflict of interest is declared.

Received: 21 July 2024 Accepted: 4 November 2024
Published online: 14 November 2024

References

1. Fuhr J, Berens P, Papies D. Estimating Causal Effects with Double Machine Learning–A Method Evaluation. Preprint at https://arxiv.org/abs/2403.14385 2024.
2. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018;21(1):C1-68.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
5. Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology. 2000;11(5):561–70.
6. Farbmacher H, Huber M, Lafférs L, Langen H, Spindler M. Causal mediation analysis with double machine learning. Econom J. 2022;25(2):277–300.
7. Bodory H, Huber M, Lafférs L. Evaluating (weighted) dynamic treatment effects by double machine learning. Econom J. 2022;25(3):628–48.
8. Selkoe DJ, Hardy J. The amyloid hypothesis of Alzheimer's disease at 25 years. EMBO Mol Med. 2016;8(6):595–608.
9. Wang J, Fan DY, Li HY, He CY, Shen YY, Zeng GH, Chen DW, Yi X, Ma YH, Yu JT, Wang YJ. Dynamic changes of CSF sPDGFRβ during ageing and AD progression and associations with CSF ATN biomarkers. Mol Neurodegener. 2022;17(1):9.
10. Janelidze S, Zetterberg H, Mattsson N, Palmqvist S, Vanderstichele H, Lindberg O, van Westen D, Stomrud E, Minthon L, Blennow K. Swedish BioFinder Study Group. CSF Aβ42/Aβ40 and Aβ42/Aβ38 ratios: better diagnostic markers of Alzheimer disease. Ann Clin Transl Neurol. 2016;3(3):154–65.
11. Zhang W, Young JI, Gomez L, et al. Distinct CSF biomarkerassociated DNA methylation in Alzheimer's disease and cognitively normal subjects. Alzheimers Res Ther. 2023;15(1):78.
12. Shireby GL, Davies JP, Francis PT, et al. Recalibrating the epigenetic clock: implications for assessing biological age in the human cortex. Brain. 2020;143(12):3763–75.

13. Polimanti R, Peterson RE, Ong JS, et al. Evidence of causal effect of major depression on alcohol dependence: findings from the psychiatric genomics consortium. Psychol Med. 2019;49(7):1218–26.
14. Sanderson E, Richardson TG, Morris TT, et al. Estimation of causal effects of a time-varying exposure at multiple time points through multivariable mendelian randomization. PLoS Genet. 2022;18(7):e1010290.
15. Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Stat Sci. 1999;14(1):29–46.
16. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med. 2017;167(4):268–74.
17. Liu Y, Gao Q, Wei K, Huang C, Wang C, Yu Y, Qin G, Wang T. High-dimensional generalized median adaptive lasso with application to omics data. Brief Bioinform. 2024;25(2):bbae059.
18. Ertefaie A, Asgharian M, Stephens DA. Variable selection in causal inference using a simultaneous penalization method. J Causal Inference. 2018;6(1):20170010.
19. Koch B, Vock DM, Wolfson J. Covariate selection with group lasso and doubly robust estimation of causal effects. Biometrics. 2018;74(1):8–17.
20. Wilson A, Reich BJ. Confounder selection via penalized credible regions. Biometrics. 2014;70(4):852–61.
21. Liu M, Zhang Y, Zhou D. Double/debiased machine learning for logistic partially linear model. Econom J. 2021;24(3):559–88.
22. Chang NC. Double/debiased machine learning for difference-in-differences models. Econom J. 2020;23(2):177–91.
23. Athey S, Imbens GW. Machine learning methods for estimating heterogeneous causal effects. stat. 2015;1050(5):1–26.
24. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat Mach Intell. 2020;2(7):369–75.
25. Grubbs FE. Procedures for detecting outlying observations in samples. Technometrics. 1969;11(1):1–21.
26. Hodge V, Austin J. A survey of outlier detection methodologies. Artif Intell Rev. 2004;22:85–126.
27. Zhou X, Lindsay H, Robinson MD. Robustly detecting differential expression in RNA sequencing data using observation weights. Nucleic Acids Res. 2014;42(11):e91–e91.
28. Chen X, Zhang B, Wang T, et al. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinform. 2020;21:1–20.
29. Jensch A, Lopes MB, Vinga S, et al. ROSIE: robust sparse ensemble for outlIEr detection and gene selection in cancer omics data. Stat Methods Med Res. 2022;31(5):947–58.
30. Yuan Y, MacKinnon DP. Robust mediation analysis based on median regression. Psychol Methods. 2014;19(1):1.
31. Sibbald B, Roland M. Understanding controlled trials. Why are randomised controlled trials important? BMJ. 1998;316(7126):201.
32. Sullivan JH, Warkentin M, Wallace L. So many ways for assessing outliers: What really works and does it matter? J Bus Res. 2021;132:530–43.
33. Kalina J, Tichavský J. On robust estimation of error variance in (highly) robust regression. Meas Sci Rev. 2020;20(1):6–14.
34. Fang EX, Li Y, Zhang H, Wang J, Chen L. Mining massive amounts of genomic data: a semiparametric topic modeling approach. J Am Stat Assoc. 2017;112(519):921–32.
35. Li G, et al. Robust differential abundance analysis of microbiome sequencing data. Genes. 2023;14(11):2000.
36. Meinshausen N, Ridgeway G. Quantile regression forests. J Mach Learn. 2006;7(6):983–99.
37. Smelyakov K, Klochko O, Dudar Z. Building Quantile Regression Models for Predicting Traffic Flow. COLINS (1). 2023:117-132.

Publisher's Note