An Ensemble Approach to CATE estimation with Super Learner in RCTs

Jong Soo Hong

The Graduate School
Yonsei University
Department of Biostatistics and Computing

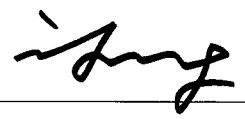
An Ensemble Approach to CATE estimation with Super Learner in RCTs

A Dissertation Submitted to the
Department of Biostatistics and Computing
and the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Biostatistics and Computing

Jong Soo Hong

December 2024

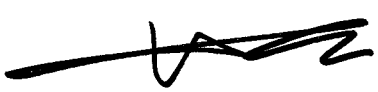
This certifies that the dissertation of ***Jong Soo Hong*** is approved.



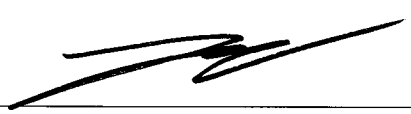
Inkyung Jung: Thesis Supervisor

Chung Mo Nam: Thesis Committee Member #1

Sohee Park: Thesis Committee Member #2



Min Jin Ha: Thesis Committee Member #3



Jeong Hoon Jang: Thesis Committee Member #4

The Graduate School

Yonsei University

December 2024

Contents

List of figures

List of tables

Appendix

ABSTRACT

An Ensemble Approach to CATE estimation
with Super Learner in RCTs

Previous studies have focused on average treatment effects rather than individual treatment effects in causal inference. Recently, with the growing interest in precision medicine, there has been a substantial increase in research on Conditional Average Treatment Effect (CATE) estimation and Individualized Treatment Rules (ITR). CATE estimation is a method for estimating the average treatment effect for individuals with identical feature attributes.

Various parametric and non-parametric methods have been proposed for CATE estimation, but recent studies indicate that no method has been found to be uniformly superior to the others across all criteria. To address this inconsistency, one study applied an ensemble method, such as causal stacking, to improve the consistency of CATE estimation. In this context, we proposed the Super Learner approach for CATE estimation to improve the performance metrics. Super Learner has the advantage of dividing data using cross-validation, which helps prevent overfitting and enables the generation of optimal results, compared to stacking or other methods.

Simulation results demonstrate that the proposed method has less MSE and outperformed in various performance metrics. These results indicate that, instead of relying on a single CATE estimation method for treatment decisions, utilizing the Super Learner to combine results from multiple methods provides a more robust and reliable framework for optimizing patient care. Furthermore, the Super Learner approach proves to be a practical and effective tool for developing individualized treatment rules, offering significant potential for optimizing patient care.

Key words: Causal inference, Heterogeneous treatment effect, Conditional average treatment effect, Individual treatment rules, Plug-in estimator

Chapter 1

Introduction

1.1 Background

Predicting individual treatment effects, beyond the average treatment effect (ATE), has become increasingly important. Clinical trials are primarily designed to estimate ATE. This is because the design and purpose of clinical trials focus on evaluating the average effect of a specific treatment across the entire population. However, the ATE does not capture differences in treatment effects at the individual or subgroup level. For instance, certain subgroups may experience significantly greater or smaller treatment effects, but ATE averages out such heterogeneity, potentially obscuring it. Therefore, the need for additional analysis, such as the conditional average treatment effect (CATE), has emerged to establish personalized treatment rules.

In clinical trials, CATE is typically estimated in two settings: the estimation of treatment effects in a relatively small number of predefined subgroups as per regulatory guidance, and data-driven assessments of treatment effect heterogeneity. This study focuses on the latter.

Over the past 15 years, advancements in machine learning and the growing interest in precision medicine within the field of causal inference have driven the development of various methods for estimating treatment effect heterogeneity. These methods primarily focus on evaluating the heterogeneity of treatment effects based on data-driven approaches. Methods for evaluating treatment effect heterogeneity which have been developed across various disciplines, can be broadly categorized into four primary approaches: (a) modeling the response surface, (b) direct estimation of CATE, (c) direct estimation of individualized

treatment rules (ITR), and (d) direct identification of subgroups (Lipkovich et al., 2024). The response surface modeling approach, initially introduced through the virtual twins method (Foster et al., 2011), has since been further expanded into several variations. In this study, we focus on approaches (a) and (b) as the foundation of our analysis to improve CATE estimation.

Several studies have reported challenges in selecting a model for evaluating HTE, as different models often perform better depending on the evaluation metrics used (Loh et al., 2019). Bouvier et al. (2024) demonstrated the issue of low agreement between CATE estimation methods when recommending treatments based on estimated individual treatment effects, with the methods showing weak correlations or inconsistent treatment recommendations.

Meanwhile, attempts have been made to improve the accuracy of CATE estimation methods using ensemble methods, which combine the predictions of multiple models to enhance overall performance. In ensemble models, since the true value of CATE is unknown, it must be replaced with an alternative value. Specifically, in clinical trials, a mathematically derived unbiased estimator can be used as a substitute for CATE. We demonstrated the unbiased estimator of true CATE in randomized clinical trials in Chapter 3. In observational studies, it is not possible to obtain an unbiased estimator to substitute for the true CATE when applying ensemble models. When attempting to obtain an unbiased estimator of the substitute value in observational studies, the propensity score must be estimated. As the propensity score appears in the denominator of the estimation formula, the estimator can become unstable, particularly when the propensity score is close to zero. Therefore, this study limited the analysis to clinical trial settings.

1.2 Objective and outline

This paper aims to improve CATE estimation by applying the Super Learner approach to estimate the weights of CATE estimation methods. A key advantage of this approach is its ability to prevent overfitting by training model weights using cross-validation, which distinguishes it from the stacking method. The CATE estimation methods used in the Super Learner demonstrate superior performance only in specific scenarios. Therefore, we evaluate the proposed method based on various performance metrics to determine whether it performs well under diverse conditions. We conducted simulation studies to compare our proposed method with existing methods. Simulation studies were designed under conditions where the true CATE is known to verify whether the proposed method outperforms existing methods.

In Chapter 2, we introduce the notations and definitions related to the potential outcome framework. In Chapter 3, we provide a brief review of CATE estimation and ensemble methods. In Chapter 4, we propose the Super Learner based CATE estimation. The simulations and their results are summarized in Chapter 5. Finally, Chapter 6 concludes and provides a discussion on the proposed method.

Chapter 2

Potential outcome framework

2.1 Notations and definitions

We adopt potential outcome framework introduced by Neyman and Rubin (2005). This framework serves as a theoretical foundation for causal inference and is widely used for quantifying and analyzing causal effects. Also known as the Rubin Causal Model (RCM), this approach was formalized by Rubin and defines causal effects through assumptions and comparisons not only of "what actually happened" but also of "what could have happened."

Potential outcomes are defined as the possible outcomes under the different treatment conditions: $Y(1)$ represents the potential outcome if the individual receives the treatment and $Y(0)$ represents the potential outcome if the individual does not receive the treatment (or receive the control).

For each individual, only one of the two potential outcomes is observed, depending on the treatment assignment. This issue is referred to as the counterfactual problem, as it is impossible to observe both outcomes for the same individual. The observed outcome (Y_{obs}) is determined as:

$$Y_{\text{obs}} = \begin{cases} Y(1), & \text{if } T = 1, \\ Y(0), & \text{if } T = 0. \end{cases} \tag{2.1}$$

To evaluate heterogeneous treatment effects (HTE), we first define the individual treatment effect (causal effect). The individual treatment effect for a binary or continuous outcome Y is represented in terms of potential outcomes as:

$$\tau_i = Y_i(1) - Y_i(0), \tag{2.2}$$

where $Y_i(t), t \in \{0,1\}$, represents a potential outcome that could have been observed. However, since both potential outcomes cannot be observed simultaneously for the same individual, the causal effect cannot be directly measured.

To enable causal inference within the potential outcome framework, the following assumptions are required:

First, we assume general Stable Unit Treatment Value Assumption (SUTVA):

$$Y_i = Y_i(T_i) = Y_i(1)T_i + Y_i(0)(1 - T_i), \tag{2.3}$$

where Y_i is the observed outcome and T_i is the treatment received for the i-th subject. From the perspective of precision medicine, we are interested in modeling the heterogeneity of individual treatment effect (ITE) as a function of observed subject characteristics, leading to the conditional average treatment effect (CATE), defined as:

$$\tau(x_i) = E(Y_i(1) - Y_i(0)|X = x_i), \tag{2.4}$$

where $x_i = (x_{1i}, \dots, x_{pi})$ is a vector of p covariates, denoted by $X_1, \dots, X_p$, for the i-th subject.

Removing the patient index i, let $\mu(t,x) = E(Y(t)|X = x), t \in \{0,1\}$, and define $\tau(x) = \mu(1,x) - \mu(0,x)$. Note that under strong treatment ignorability, ensured by randomization in RCTs and assumed in observational studies conditional on the covariates (Rosenbaum & Rubin, 1983), we can replace the potential outcomes with the conditional expectations of the observable random variables:

$$\mu(t,x) = E(Y|T = t, X = x). \tag{2.5}$$

The response surface can be represented without loss of generality as:

$$\mu(t,x) = h(x) + \frac{1}{2}\tau(x)(2t-1), t \in \{0,1\}, \tag{2.6}$$

where $h(x)$ is the main covariate effect, that is,

$$h(x) = \frac{1}{2}\{\mu(1,x) + \mu(0,x)\}. \tag{2.7}$$

In observational data, estimating causal effects, such as ATE and CATE, requires additional assumptions. First, we assume treatment ignorability conditional on the observed covariates that is,

$$T \perp \{Y(1), Y(0)\}|X \tag{2.8}$$

Second, we often estimate the propensity score function $\pi(x) = \Pr(T = 1|X = x)$ from the observed data. To make valid inferences, we assume positivity, $0 < \pi(x) < 1$.

Occasionally we use a general treatment assignment function $\pi(t,x) = \Pr(T = t|X = x)$ allowing us to simplify certain expressions. In this context, the propensity score is defined as $\pi(x) \equiv \pi(1,x)$. In this paper, we focus on randomized clinical trial settings and use the treatment assignment probability p instead of estimating the propensity score $\pi(x)$.

Let us assume we obtained a good estimate of CATE, $\widehat{\Delta}(x)$. we define a subgroup as a set of all subjects with a positive treatment effect, that is,

$$\hat{S}(x) = \{x : \widehat{\Delta}(x) > \delta\}, \tag{2.9}$$

where δ is a predefined. In this paper, we only consider cases where $\delta = 0$. This implies that subgroup includes all individuals with a treatment effect that is even slightly positive. This approach is closely related to developing individualized treatment assignment rules or regimens that, given a subject's covariate profile $X = x$, select the optimal treatment $D(x) \in \{0,1\}$.

Chapter 3

Reviews of methods for CATE estimation

3.1 Estimating CATE via meta-learner

The meta-learner emerged as a powerful framework to address the complexity of CATE estimation by combining machine learning techniques with causal inference theory. It estimates nuisance parameters through outcome modeling and propensity score modeling, leveraging these results to estimate CATE. There are no restrictions on the methods used to estimate the model, allowing for the application of various machine learning techniques such as extreme gradient boosting (XGBoost) or random forests (RF).

3.1.1 S-learner

The S-learner estimates the treatment effect within a single regression model, where the treatment is included as a feature and where interactions between the treatment and relevant covariates are introduced in the parametric settings. First, a model is used to estimate the response function $\mu(t,x)$:

$$\mu(t,x) = E[Y|T = t, X = x]. \tag{3.1}$$

Then, the individual treatment effect τ is estimated as:

$$\hat{\tau}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x). \tag{3.2}$$

Since the S-learner trains only one outcome model by using machine learning models such as XGBoost and RF, it is simple and computationally efficient. However, if the

treatment effect is highly heterogeneous, the S-learner may introduce bias due to the single-model approach, making it difficult to capture interaction effects between treatment groups effectively.

3.1.2 T-learner

In the T-learner algorithm, two models are built, one for the treatment group and one for the control group. These models are used to calculate the response functions:

$$\mu_1(x) = E[Y|T = 1, X = x],$$
$$\mu_0(x) = E[Y|T = 0, X = x].$$

The ITE is estimated as the difference between the two predicted risks:

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x).$$

The T-learner offer the advantage of simplicity of implementation and flexibility for integration with various machine learning models. However, a potential drawback is the risk of introducing bias in estimated. T-learners may be prone to bias arising from inconsistent estimations across independently trained models and challenges related to data imbalance.

3.1.3 X-learner

Künzel et al. (2019) proposed a method called X-learner, which is a hybrid estimator of CATE formed as a weighted average of two estimators $\hat{\tau}_1(x)$ and $\hat{\tau}_0(x)$ constructed using a multi-step procedure:

1. Estimate the response function as in the T-learner:

$$\mu_0(x) = E[Y|X = x, T = 0],$$

$$\mu_1(x) = E[Y|X = x, T = 1].$$

2. Imputed the treatment effects for the individuals in the treated group based on the control-outcome estimator and the treatment effects for the individuals in the control group based on the treatment-outcome estimator and estimate $\hat{\tau}_1(x)$ and $\hat{\tau}_0(x)$:

$$\widetilde{D}^1 = Y_i - \hat{\mu}_0(X_i), i \in \{i : T_i = 1\},$$
$$\widetilde{D}^0 = \hat{\mu}_1(X_i) - Y_i, i \in \{i : T_i = 0\},$$
$$\hat{\tau}_1(x) = E\left[\widetilde{D}^1 \middle| X = x\right],$$
$$\hat{\tau}_0(x) = E\left[\widetilde{D}^0 \middle| X = x\right].$$

3. Define the ITE by a weighted average of the two estimates:

$$\hat{\tau}(x) = w(x)\,\hat{\tau}_0(x) + \big(1 - w(x)\big)\,\hat{\tau}_1(x),$$

where the weight function is often taken as the estimated propensity score, $w(x) = \hat{\pi}(x) = \widehat{\Pr}(T = 1|X = x)$ or constant probability of treatment assignment.

The X-learner is expected to outperform the T-learner in scenarios where the control arm is significantly larger than the treated arm. This is because the two estimators, $\hat{\tau}_1(x)$ and $\hat{\tau}_0(x)$, in the X-learner rely on comparing observed outcomes and counterfactual outcomes predicted from the alternative arm, rather than comparing predicted potential outcomes generated by models fitted to different arms. Additionally, with the X-learner, differences in model complexity across the two arms are effectively "smoothed out" within each estimator, $\hat{\tau}_1(x)$ and $\hat{\tau}_0(x)$.

3.2 Direct modeling of CATE

A common approach to addressing the challenges associated with tuning complexity parameters for predictive and prognostic effects is to redefine the problem in a way that eliminates the need to estimate prognostic effects. Over the past 10 years, several methods have been developed to model CATE directly, bypassing the need to incorporate the prognostic component of the outcome model. The key benefit of these methods is that they reduce the risk of errors resulting from misspecifying the prognostic effects.

3.2.1 R-learner

One proposal for outcome transformation is based on the so-called Robinson's transformation (Kennedy, 2023) of an outcome variable that involves simultaneously centering the response and treatment indicator around their estimated expected values. Specifically, consider

$$Y_i^* = \frac{Y_i - \mu(X_i)}{T_i - \pi(X_i)}, \tag{3.3}$$

where $\mu(x) = E(Y|X = x)$ is the overall response function, capturing the main effect of covariates on the outcomes in the pooled data. It is easy to show that $E(Y^*|X = x) = \tau(x)$, therefore a simple approach similar to the modified outcome is to estimate CATE by regressing Y^* on the covariates. The residualization of marginal outcomes and treatment effects has recently been promoted in the literature as part of efforts to estimate overall treatment effects from observational data, particularly under the framework of double/debiased machine learning. Additionally, it has been a focus in research on HTE, as demonstrated by Athey et al. (2019) in their work on generalized random forests (GRF).

The transformation leads to the following data representation:

$$Y_i - \mu(X_i) = \big(T_i - \pi(X_i)\big)\tau(X_i) + \epsilon_i, \tag{3.4}$$

where the plug-in estimates of nuisance parameters $\mu(x)$ and $\pi(x)$ are obtained from some off-the-shelf machine learning methods with a cross-fitting step following the estimation of $\tau(x)$. These ideas were first introduced in the proposal by Zhao and Panigrahi (2019) and further generalized in R-learning of Nie and Wager (2021).

The R-learner estimates the ITEs in two steps:

1. Fit the response function $\hat{\mu}^{-i}(x)$ and the propensity score $\hat{\pi}^{-i}(x)$ with a base learner.

2. Estimate ITEs by minimizing the R-loss, which uses Robinson's decomposition:

$$\hat{\tau}(\cdot) = \arg\min(\hat{L}_n\{\tau(\cdot)\} + \Lambda_n\{\tau(\cdot)\})$$
$$\hat{L}_n\{\tau(\cdot)\} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}^{-i}(X_i) - \left(T_i - \hat{\pi}^{-i}(X_i)\right)\tau(X_i)\right)^2.$$

Here $\Lambda_n\{\tau(\cdot)\}$ is a regularizer on the complexity of the $\tau(\cdot)$ to a given machine learning method. It was shown that the cross-fitting procedure controls the convergence rates of the target CATE estimator independently of learning the nuisance parameters, as if those were provided by an approximate oracle.

3.2.2 Causal Forest

The causal forests algorithm is a special case of generalized random forests (GRF), a flexible and general framework to estimate the ITEs (Athey et al., 2019). Causal forests extend the original random forest algorithm by borrowing ideas from kernel-based methods and the R-learner. If we know that $\tau(x)$ were constant over some neighbourhood $N(x)$, we could solve partially linear model over $N(x)$ using the residual-on-residual approach: first, we estimate $\pi(x) = E(T_i|X_i = x)$ and second, $\mu(x) = E(Y_i|X_i = x)$. We can any non-parametric method like the lasso, RF, boosting methods and others. The final step is to estimate $\tau(x)$ over the neighbourhood $N(x)$:

$$\hat{\tau}(x) = \frac{\sum_{\{i:X_i \in N(x)\}}\{Y_i - \hat{\mu}(X_i)\}\{T_i - \hat{\pi}(X_i)\}}{\sum_{\{i:X_i \in N(x)\}}\{T_i - \hat{\pi}(X_i)\}}.$$

In contrast to the standard random forest algorithm in which a prediction for a new observation is obtained by averaging predictions of each tree, here, the trees are used to compute a weighting scheme similar to kernel-based methods. The trees act as weights between training points and any new observations:

$$\alpha_{bi}(x) = \frac{1\{X_i \in L_b(x)\}}{|L_b(x)|}, \alpha_i(x) = \mathrm{B}^{-1}\sum_{b=1}^{B}\alpha_{bi}(x),$$

$$\hat{\mu}(x) = \sum_{i=1}^{n} Y_i \alpha_i(x).$$

where X_i corresponds to the covariates of individual i in the training dataset and $L_b(x)$ corresponds to the set of observations in the training set that fall in the same leaf as x for tree b.

Then, the prediction for a new observation is obtained using the adaptive weights by minimizing the R-loss described above.

Another characteristic of causal forests (and more generally of GRF) is the notion of

honesty where the training data is split into two parts: one for constructing the tree and the other (the estimation sample) for estimating leaf values for each tree. In doing so, the estimates are less prone to bias and more consistent. The notion of honesty is similar to employing the cross-fitting in non-parametric meta-learners.

3.2.3 W-learning

Directly estimating CATE using W-learning is achieved through the following procedure. Based on an inverse probability weighted (IPW) transformation, a continuous or binary outcome Y can be transformed.

$$Y_i^* = Y_i \frac{T_i}{\pi(X_i)} - Y_i \frac{1 - T_i}{1 - \pi(X_i)} = Y_i \frac{T_i - \pi(X_i)}{\pi(X_i)(1 - \pi(X_i)}.$$

It is instructive, as will be seen from the following, to express the modified outcome via the treatment indicator $A_i = 2T_i - 1 \in \{-1, 1\}$.

$$Y_i^* = Y_i \frac{A_i}{A_i \pi(X_i) + (1 - A_i)/2}.$$

An alternative way of obtaining a CATE estimator is by using a weighted squared loss with the outcome multiplied by $2A$ and subject weights

$$W_i = \frac{1}{A_i \hat{\pi}(X_i) + (1 - A_i)/2},$$

$$\begin{aligned}
&\underset{g(x)}{\operatorname{argmin}} E\left(W_i\left(2A_i Y_i - g(X_i)\right)^2 \middle| X_i = x\right) \\
&= \underset{g(x)}{\operatorname{argmin}} E\left(4W_i\left(Y_i - \frac{A_i}{2} g(X_i)\right)^2 \middle| X_i = x\right) \\
&= \tau(x).
\end{aligned}$$

W-learning with augmentation method can be obtained by transforming continuous responses into doubly robust augmented inverse probability weighted (AIPW) scores.

$$\hat{\tau}_{AIPW}(X_i) = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i\big(Y_i - \hat{\mu}_1(X_i)\big)}{\hat{\pi}(X_i)} - \frac{(1-T_i)\big(Y_i - \hat{\mu}_0(X_i)\big)}{1-\hat{\pi}(X_i)}$$
$$= \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i - \hat{\pi}(X_i)}{\hat{\pi}(X_i)\big(1-\hat{\pi}(X_i)\big)}\big(Y_i - \hat{\mu}(T_i, X_i)\big),$$

where $\hat{\mu}_1(x) \equiv \hat{\mu}(1,x)$ and $\hat{\mu}_0(x) \equiv \hat{\mu}(0,x)$ are based on the outcome regression.

3.2.4 A-learning

Chen et al. (2017) proposed A-learning based on minimizing the modified loss $E\left(Y - \big(T-\pi(X)\big)g(X)\right)^2$ is the method for directly estimating CATE. Its population minimizer for the squared loss is $g(x) = \tau(x)$. This property holds because the centered interaction $\big(T-\pi(X)\big)g(X)$ is orthogonal to the main effect $h(X)$ in (2.7).

It is important to note that direct estimation methods are prone to high variability resulting in poor efficiency, which can be improved by augmenting the estimating equations with an additional zero-expectation term.

$$E\left(\frac{1}{\pi(X)-(1-A)/2}\left(Y - \frac{A}{2}g(X)\right)\middle| X = x\right) = 0.$$

It can be shown that the solution of the equation does not change if we add a tern $\frac{b(X)}{\pi(X)-(1-A)/2}$ for an arbitrary function of covariates $b(X)$. Now the goal is to choose the function $b(X)$ that minimizes the variance of the estimating equations.

A-learning with augmentation method can be implemented by transforming outcome variable based on Robinson's transformation (3.3) to the outcome variable. Through

augmentation, the estimation has doubly robust property if either the outcome model or the propensity score model is correctly specified.

Hence, A-learning is well-suited for simpler data and models, providing an efficient approach when quick results are prioritized. In contrast, A-learning with augmentation is more appropriate for complex data settings, offering robust estimation, particularly in scenarios where model misspecification is a concern.

3.3 Ensemble method

3.3.1 Causal stacking

Han and Wu (2022) proposed causal stacking method for CATE function estimation in RCTs (Bernoulli design or completely randomized design). The analysis procedure is as follows:

1. Choose *K* CATE estimation algorithms

2. Partition the data into train/validation set and calculate the proportion of treated units in validation set.

3. Estimate τ_k using training data and predict $\hat{\tau}_k(X_i), i \in S_{\text{valid}}$

4. Optimize the weights

$$\widehat{w} = \underset{w \geq 0, ||w||_1 = 1}{\operatorname{argmin}} \frac{1}{S_{valid}} \sum_{i \in S_{\text{valid}}} \left(\hat{\tau}_i - w^T \hat{\tau}_{1:K}(X_i)\right)^2,$$

where $\hat{\tau}_i$ is defined as:

$$\hat{\tau}_i = [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)] + \frac{[Y_i - \hat{\mu}_1(X_i)]T_i}{p} - \frac{[Y_i - \hat{\mu}_0(X_i)](1 - T_i)}{1 - p},$$

within the validation set.

5. Predict $\hat{\tau}_s(\cdot) = \widehat{w}^T \hat{\tau}_{1:K}(\cdot)$.

First, researcher selects K methods for estimating CATE as candidate methods to apply the ensemble method. Next, the dataset is divided into a training set and validation set. The K CATE estimation methods are trained on the training set and the proportion of treated units p is calculated using the validation set.

Using the K CATE models, predicted value $\hat{\tau}_k(X_i)$ are calculated on the validation set and weights that minimize the loss function are optimized. Since the true CATE is unknown in the loss function used for $\widehat{w}$, an unbiased plug-in estimator substitutes true CATE as the surrogate CATE. The plug-in estimator is one of the methods proposed by many researchers to approximate the true CATE (Saito & Yasui, 2020). Aronow and Middleton (2013) demonstrated that an unbiased for $\tau(X_i)$ under any choice of the regression functions $\hat{\mu}_0$ and $\hat{\mu}_1$ when using validation set. An estimator that substitutes the true CATE using this idea is referred to as a plug-in estimator. The predicted values and weights from each method are linearly combined to estimate the CATE. Previous studies have shown that causal stacking outperforms individual methods in terms of mean squared error.

3.3.2 Surrogate CATE

As introduced in Section 3.3.1, implementing ensemble methods typically requires a surrogate value for the true Conditional Average Treatment Effect (CATE) during the estimation process. In most practical settings, however, it is not feasible to observe the exact CATE for each subject, thereby necessitating a suitable surrogate. This surrogate effectively serves as a “true” label in the training procedure and critically affects both the bias and variance of the resulting estimators.

In the context of a RCT, the treatment assignment probability p is known a priori, and randomization ensures that the treatment assignment is independent of the potential outcomes. Under these conditions, the plug-in estimator is unbiased, providing a surrogate for the true CATE. This property is especially advantageous when applying ensemble methods, as it allows the algorithm to rely on a well-founded surrogate for the underlying CATE, ultimately improving estimation accuracy and robustness.

Chapter 4

Proposed method

4.1 Super Learner CATE estimation method

We propose Super Learner-based method for CATE estimation. This idea is derived from Super Learner (Van der Laan et al., 2007). Super Learner identifies the optimal weights that minimize loss based on cross-validation results. Unlike other ensemble methods that use predefined combination rules, the Super Learner learns the weights in a data-driven manner. We expect our proposed method to prevent overfitting and ensure estimation stability, improving upon the causal stacking method suggested in previous studies. The proposed algorithm is outlined as follows:

Algorithm 1. Super Learner-based method for CATE estimation

Input: Dataset $S = \{(X_i, Y_i, T_i)\}_{i=1}^{n}$, candidate CATE algorithms $\{\mathcal{A}_k\}_{k=1}^{K}$.

1: Split the dataset into V-fold datasets $\{S_1, S_2, \dots, S_V\}$.

2: **for** $v \in \{1,2, \dots, V\}$ **do**

3: Define the dataset excluding the v-th fold as $S_{-v} = S \setminus S_v$.

4: For each CATE algorithm $\mathcal{A}_k$, fit $\mathcal{A}_k$ using S_{-v} to estimate $\hat{\tau}_{k,v}$, the predicted CATE for the data in S_v, based on the model trained on S_{-v}.

5: **end for**

6: Construct the predicted CATE matrix $C = \begin{pmatrix} \hat{\tau}_{1,1} & \cdots & \hat{\tau}_{K,1} \\ \vdots & \cdots & \vdots \\ \hat{\tau}_{1,V} & \cdots & \hat{\tau}_{K,V} \end{pmatrix}$.

7: Estimate the *plug-in estimator* for true CATE $\hat{\tau}^* = (\hat{\tau}_1^*, \hat{\tau}_2^* \ldots, \hat{\tau}_n^*)^t$, where each individual estimator

$$\hat{\tau}_i^* = [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)] + \frac{[Y_i - \hat{\mu}_1(X_i)]T_i}{p} - \frac{[Y_i - \hat{\mu}_0(X_i)](1 - T_i)}{1 - p},$$

Using the data in S_{train} fit the regression model $\hat{\mu}_1, \hat{\mu}_0$ that Y_i using X_i and fraction of treated units in the S_v should be p.

8: Estimate the weight w by solving the problem: $\underset{w>0\,,||w||=1}{\text{minimize}} \|\hat{\tau}^* - Cw\|^2$.

9: fit CATE $\hat{\tau}_{k,S} \leftarrow \mathcal{A}_k(S_{\text{train}}), k \in \{1,2, \ldots, K\}$

Output: predict test dataset $\sum_{k=1}^{K} \widehat{w}_k \hat{\tau}_{k,S}(X_i)$ $for\ i \in S_{\text{test}}$

First, researcher selects K methods for estimating CATE as candidate methods to apply the Super Learner-based method. Then, the data is divided into V-fold datasets. The K CATE estimation methods are trained using the dataset excluding the v-fold, and the predicted CATE for the v-fold dataset using prediction models. This process is repeated V times to create the predicted CATE matrix C.

Next, a plug-in estimator is calculated to utilize the surrogate CATE as substitute for the true CATE. To optimize the weight w, following loss function is minimized:

$$\underset{w>0\,,||w||=1}{\text{minimize}} \|\hat{\tau}^* - Cw\|^2$$

To obtain the K CATE estimations for the entire dataset, each model is refitted to generate predictions (Polley & Van der Laan, 2010). These predictions, along with the estimated weights, are used to estimate the CATE for the test dataset.

4.2 Unbiased plug-in estimator in RCTs

In Algorithm 1, the plug-in estimator plays a crucial role, serving as a benchmark for estimating the true conditional average treatment effect. As mentioned in Section 3.3.2, identifying an appropriate substitute for the true CATE is essential for estimating weights in ensemble methods. To ensure reliable CATE estimation, an unbiased estimator must be used. However, in observational studies, even if the propensity score is carefully modeled, the property of unbiasedness is not guaranteed. Consequently, this study focuses on randomized clinical trials to secure an unbiased plug-in estimator for the true CATE.

In causal inference and machine learning, many studies utilize plug-in estimators as substitutes for the true CATE, employing outcome models based on meta-learner framework such as S-learner, T-learner, and X-learner (Mahajan et al., 2022). In this study, we adopt plug-in estimators derived from S-learner and T-learner to evaluate their impact on the performance of ensemble methods.

Finally, using Algorithm 1, we compared the Super Learner-based method with existing methods to examine whether it prevents overfitting and produces consistent results. The results were validated through simulations, as detailed in Chapter 5.

Chapter 5

Simulation studies

5.1 Simulation scheme

We conducted a simulation study with two scenarios reflecting different challenges encountered in HTE assessments. This is done to demonstrate that, if certain key assumptions are violated, it would no longer be feasible to accurately estimate CATE even with very sophisticated methods.

The simulations used in this paper partially adopted the settings proposed by Lipkovich (2024). Data generating process $D1$ and $D2$ represent data from an RCT setting, where subjects are randomized to treatment than to the control groups at ratios of 3:1, 2:1 or 1:1. The outcome Y is continuous, with large values indicating treatment benefits. The train dataset contains N=1,000 observations and the data-generating process is defined as follows:

$$Y = 100 - (X_1 + 5X_2) + T \times \big(g_1(X_3) + g_2(X_4)\big) + \epsilon,$$

where $X_1, X_3, X_4 \sim N(0.5, 1), X_2 \sim Categorical\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \epsilon \sim N(0,1), T \in \{0,1\}$.

Here, X_1, X_2 form the prognostic component and X_3, X_4 the predictive part with CATE given as $\tau(x) = g_1(x_3) + g_2(x_4)$. Nonlinearity is induced in CATE via a $g_1(\cdot)$

$$g_1(x) = \begin{cases} a - b \cdot 0.25 & \text{if } x < 0 \\ a - b(x - 0.5)^2 & \text{if } 0 \le x \le 1, \\ a - b \cdot 0.25 & \text{if } x > 1 \end{cases}$$

and a monotone $g_2(\cdot)$

$$g_2(x) = \begin{cases} 0 & \text{if } x < 0 \\ \dfrac{c}{1 + \exp\left(-d(x - 0.5)\right)} & \text{if } 0 \le x \le 1, \\ c & \text{if } x > 1 \end{cases}$$

The constant a = 0.625, b = 5, c = 0.625, d = 20 are calibrated so as to make the overall treatment effect slightly positive, $\mathrm{E}[\tau(X)] = 0.0119$, the true signature $S_{true} = \{\tau(X) > 0\}$ has the proportion of subjects $E[I(X \in S_{true})] = 0.330$ and the true mean treatment effect in S_{true} is $\mathrm{E}[\tau(X)|X \in S_{true}] = 0.665$ and the true subgroup's utility index η is 0.22.

Data under D_2 are simulated similarly to D_1, except the prognostic part is more complex:

$$Y = 100 - (X_1 + 5X_2) + 2(X_5 + X_6 + X_7 + X_8 + X_9) + T \times \left(g_1(X_3) + g_2(X_4)\right) + \epsilon,$$
$$X_5, X_6, \dots, X_9 \sim N(0.5, 1),$$

To make analysis more challenging each dataset includes an additional 10 noise covariates independently drawn from the standard normal distribution.

Unequal sample sizes in treatment groups, such as when the treatment group is three times larger than the control group, seem to be an advantage since the larger sample size is in the arm where the true outcome model is more complex. Therefore, additional simulations were conducted with treatment and control ratios of 2:1 and 1:1.

We selected 11 methods to estimate CATE estimation in each scenario. Selected methods are T-, S-, R-, X-, Causal Forest, A-learning, A-learning with augmentation, W-learning, W-learning with augmentation, causal stacking, and our proposed Super Learner-based CATE method. The CATE estimation methods included in the comparison were determined based on the availability of R packages or publicly available codes from other papers. The outcome model was estimated using XGBoost.

We utilized a modified version of meta-learning code by (Nie & Wager, 2021), which

includes cvboost3, tboost3, sboost3, rboost3, and xboost3 to estimate T-learner, S-learner, R-learner, and X-learner (Lipkovich et al., 2023). This modified version focuses on the three key tuning parameters (size, depth, and eta) out of the seven tuning parameters in the original cvboost. Hyperparameters for all models were estimated using 5-fold cross-validation combined with grid search, and the ranges of hyperparameters explored during the search process are summarized in detail in the Appendix Table A1. The estimation of Causal Forest was performed using the 'grf' package, while the estimation of A-learning and W-learning was performed using the 'personalized' package.

In addition to the previous simulations, we conducted additional simulations for cases where the CATE is absent, referred to as a null model. Through these simulations, we aimed to examine the relative performance of the Super Learner compared to other methods in the absence of CATE. We constructed the null model by partially modifying $D1$ and $D2$ to remove the interaction term ($X \times T$). The data generating process is as follows.

Null model:

($D1$ null) $Y = 100 - (X_1 + 5X_2) + \big(g_1(X_3) + g_2(X_4)\big) + \epsilon$

($D2$ null) $Y = 100 - (X_1 + 5X_2) + 2(X_5 + X_6 + X_7 + X_8 + X_9) + \big(g_1(X_3) + g_2(X_4)\big) + \epsilon$

To make analysis more challenging each dataset includes an additional 10 noise covariates independently drawn from the standard normal distribution.

The results of the additional simulations are summarized in the Appendix Table A2 and Figure A3-A14.

Table 1. Simulation scenarios for the data generating process $D1$: 2 prognostic and 2 predictive covariates; $D2$: 7 prognostic and 2 predictive covriates

Scenario	Data generating process	Treatment-to-control ratio	Surrogate CATE
1-1-1	$D1$	3:1	S-learner
1-2-1	$D1$	2:1	S-learner
1-3-1	$D1$	1:1	S-learner
1-1-2	$D1$	3:1	T-learner
1-2-2	$D1$	2:1	T-learner
1-3-2	$D1$	1:1	T-learner
2-1-1	$D2$	3:1	S-learner
2-2-1	$D2$	2:1	S-learner
2-3-1	$D2$	1:1	S-learner
2-1-2	$D2$	3:1	T-learner
2-2-2	$D2$	2:1	T-learner
2-3-2	$D2$	1:1	T-learner

5.2 Evaluation metrics

We used eight evaluation metrics to summarize the performance of various methods for estimating CATE on the simulated data (Table 2). We mainly focus on mean squared error and subgroup utility index, as they are critical for assessing both estimation accuracy and subgroup identification performance.

Table 2. Evaluation metrics for CATE estimation

Evaluation metrics	Descriptive
$corr\left(\hat{\tau}(X), \tau(X)\right)$	Pearson correlation between true CATE and estimated CATE
$agree\left(\hat{S}, S_{true}\right) = \frac{n\left(\hat{S} \cap S_{true}\right)}{n\left(\hat{S} \cup S_{true}\right)}$	Jaccard similarity coefficient
$ATE\left(\hat{S}\right) = E_X\{\tau(X) \mid \hat{\tau}(X) > 0\}$	True average treatment effect on estimated subgroup
$\widehat{ATE}\left(\hat{S}\right) = E_X\{\hat{\tau}(X) \mid \hat{\tau}(X) > 0\}$	Estimated average treatment effect on estimated subgroup
$bias\{ATE\left(\hat{S}\right)\} = \widehat{ATE}\left(\hat{S}\right) - ATE\left(\hat{S}\right)$	Difference between estimated average treatment effect on estimated subgroup and true average treatment effect on estimated subgroup
$SD\{\widehat{ATE}\left(\hat{S}\right)\}$	Standard deviation of estimated average treatment effect on estimated subgroup
$\eta = ATE\left(\hat{S}\right) \times \frac{n(\hat{S})}{n}$	Subgroup's utility index
$MSE_k = \frac{1}{S_{test}} \sum_{i \in S_{test}} \left(\tau_i - \hat{\tau}_k(X_i)\right)^2$	Mean squared error of test dataset

* $S_{true}(X) = \{x : \tau(X) > 0\}$ and $\hat{S}(X) = \{x : \hat{\tau}(X) > 0\}$

We evaluated agreement between the true and estimated CATE using Pearson correlation. The Jaccard similarity coefficient measured the overlap between the true subgroup and estimated subgroup. A true/estimated subgroup is defined as the set of individuals whose true/estimated CATE is greater than 0. True average treatment effect on estimated subgroup represents the average treatment effect of the subgroup with a positive estimated CATE. Bias is difference between estimated average treatment effect on estimated subgroup and true average treatment effect on estimated subgroup. Standard deviation of estimated average treatment effect on estimated subgroup indicates the extent of dispersion or variability. Subgroup's utility index is equivalent to the difference between the value of the estimated treatment assignment rule and that of a fixed regimen that assigns everyone to the control. In other words, a higher value indicates that the treatment assignment rule is effective in assigning appropriate treatments to patients with a greater treatment effect across the entire population.

The figures consist of one box plot (A) and three scatter plots (B-D):
(A) A box plot of mean squared error (MSE)
(B) A scatter plot of the correlation between true CATE and estimated CATE versus the subgroup utility index
(C) A scatter plot of the true average treatment effect (ATE) on the estimated subgroup versus the estimated ATE on the same subgroup
(D) A scatter plot of the correlation between true CATE and estimated CATE versus the Jaccard similarity coefficient between the true subgroup and the estimated subgroup.

These plots aim to evaluate performance of methods for accurately estimating CATE and identifying subgroups with significant treatment effects across simulations. In the following section, we evaluate the CATE estimation methods based on the simulation results.

5.3 Simulation results

The results of the evaluation metrics for comparing CATE estimation methods are presented in Tables 3–14. Across all simulations, the Super Learner consistently demonstrated superior performance, showing subgroup utility index values that were relatively close to the true values. When considering both the subgroup utility index and the bias in the average treatment effect of the estimated subgroup, the proposed method exhibited a higher subgroup utility index and relatively smaller bias compared to other methods.

Except for the A-learning and W-learning methods, the Pearson correlation and Jaccard similarity coefficient appeared to fall within a similar range and indicated a linear correlation in each Figure (D). However, the A-learning and W-learning methods showed poor CATE estimation results across multiple evaluation metrics. Augmentation methods that transform the form of the outcome occasionally demonstrated good performance depending on the scenario.

From the perspective of the treatment-to-control ratio, the bias in the average treatment effect of the estimated subgroup decreased as the treatment proportion became closer to the control proportion (Tables 3–5). This suggests that a balanced treatment-to-control ratio contributes to more stable and accurate estimations.

In the simulations, we used the S-learner and T-learner as substitutes for the true CATE. Although the T-learner showed slightly larger values under the same settings, overall simulation results showed no significant difference between the two learners, indicating that the plug-in estimator did not have notable impact on performance metrics.

In terms of MSE, the Super Learner method demonstrated significantly smaller values and ranges compared to other methods, while A-learning and W-learning exhibited very large values, indicating instability in their estimates. Additionally, in each Figure (B), the Pearson correlation and subgroup utility index appeared to have a linear relationship in

most cases. Through plot (C), it was possible to examine whether the estimated ATE in the estimated subgroup was underestimated or overestimated compared to the true ATE in the estimated subgroup. Most estimates tended to be overestimated. Estimates closer to the diagonal line were considered more accurate, with the Super Learner, S-learner, and X-learner, in that order, showing closer.

Considering multiple evaluation metrics comprehensively, the Super Learner demonstrated robust and superior performance in CATE estimation compared to other methods.

The additional simulation results for the null model indicate that in most scenario settings, the S-learner, A-learning, and W-learning methods showed relatively higher mean squared errors, as shown in plot (A). Since the null model represents a case where the true CATE is absent, various performance metrics cannot be comprehensively evaluated. However, as presented in plot (B), estimated average treatment effect on estimated subgroup values are ranked in descending order as causal forest, Super Learner, and S-learner in Appendix Figure A2-A13.

Table 3. Evaluation metrics for Scenario 1-1-1: Data generating process 1 with a 3:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.549	0.439	0.569	0.271	0.299	0.072	0.136
S-learner	0.652	0.498	0.345	0.315	0.030	0.075	0.161
R-learner	0.696	0.494	0.394	0.262	0.132	0.082	0.140
X-learner	0.586	0.431	0.332	0.257	0.075	0.073	0.136
Causal Forest	0.704	0.419	0.128	0.365	-0.236	0.073	0.137
A-learning	-0.028	0.295	3.791	0.008	3.782	4.933	0.007
A-learning aug*	0.676	0.450	0.469	0.227	0.241	0.083	0.137
W-learning	0.003	0.326	4.295	0.014	4.281	4.136	0.013
W-learning aug	0.761	0.505	0.620	0.283	0.337	0.069	0.164
Causal stacking	0.695	0.486	0.399	0.250	0.150	0.082	0.138
Super Learner**	0.773	0.525	0.404	0.348	0.055	0.094	0.185

* aug; augmentation

** Super Learner; proposed method

Table 4. Evaluation metrics for Scenario 1-2-1: Data generating process 1 with a 2:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.578	0.452	0.521	0.282	0.240	0.059	0.141
S-learner	0.630	0.488	0.307	0.302	0.005	0.062	0.155
R-learner	0.748	0.521	0.379	0.260	0.119	0.082	0.141
X-learner	0.644	0.462	0.311	0.290	0.021	0.062	0.151
Causal Forest	0.715	0.442	0.131	0.358	-0.226	0.067	0.142
A-learning	-0.003	0.283	4.619	0.008	4.610	4.839	0.006
A-learning aug*	0.667	0.449	0.434	0.224	0.210	0.073	0.135
W-learning	0.013	0.322	4.406	0.013	4.394	3.910	0.012
W-learning aug	0.723	0.483	0.560	0.255	0.304	0.064	0.152
Causal stacking	0.743	0.510	0.387	0.245	0.142	0.079	0.138
Super Learner**	0.758	0.511	0.379	0.334	0.045	0.075	0.180

* aug; augmentation

** Super Learner; proposed method

Table 5. Evaluation metrics for Scenario 1-3-1: Data generating process 1 with a 1:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.564	0.445	0.505	0.272	0.233	0.053	0.138
S-learner	0.577	0.455	0.280	0.270	0.010	0.065	0.143
R-learner	0.791	0.537	0.372	0.275	0.097	0.072	0.150
X-learner	0.698	0.480	0.329	0.315	0.014	0.059	0.164
Causal Forest	0.695	0.418	0.149	0.327	-0.178	0.075	0.146
A-learning	0.007	0.277	4.166	0.014	4.152	5.165	0.009
A-learning aug*	0.618	0.429	0.410	0.212	0.198	0.061	0.130
W-learning	-0.005	0.285	4.658	0.012	4.647	5.265	0.008
W-learning aug	0.595	0.430	0.483	0.215	0.268	0.056	0.131
Causal stacking	0.781	0.531	0.374	0.269	0.105	0.073	0.148
Super Learner**	0.729	0.496	0.341	0.341	0.001	0.065	0.181

* aug; augmentation

** Super Learner; proposed method

Table 6. Evaluation metrics for Scenario 1-1-2: Data generating process 1 with a 3:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.550	0.450	0.555	0.270	0.285	0.071	0.136
S-learner	0.646	0.506	0.333	0.312	0.021	0.069	0.159
R-learner	0.691	0.496	0.383	0.255	0.128	0.087	0.137
X-learner	0.572	0.435	0.325	0.258	0.067	0.078	0.133
Causal Forest	0.692	0.433	0.124	0.362	-0.238	0.069	0.136
A-learning	-0.014	0.286	5.480	0.007	5.473	6.687	0.005
A-learning aug*	0.676	0.457	0.458	0.217	0.241	0.077	0.132
W-learning	0.025	0.330	4.698	0.013	4.685	4.384	0.013
W-learning aug	0.761	0.506	0.607	0.271	0.336	0.069	0.159
Causal stacking	0.687	0.485	0.392	0.243	0.149	0.088	0.134
Super Learner**	0.778	0.528	0.390	0.346	0.044	0.088	0.183

* aug; augmentation

** Super Learner; proposed method

Table 7. Evaluation metrics for Scenario 1-2-2: Data generating process 1 with a 2:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.567	0.453	0.527	0.276	0.252	0.065	0.139
S-learner	0.628	0.500	0.311	0.303	0.007	0.077	0.156
R-learner	0.764	0.538	0.387	0.275	0.112	0.090	0.146
X-learner	0.647	0.471	0.315	0.289	0.026	0.071	0.149
Causal Forest	0.730	0.438	0.127	0.376	-0.249	0.071	0.140
A-learning	-0.036	0.291	4.430	0.006	4.423	5.017	0.005
A-learning aug*	0.673	0.459	0.436	0.220	0.216	0.075	0.135
W-learning	0.008	0.328	4.309	0.013	4.296	3.704	0.012
W-learning aug	0.723	0.494	0.568	0.255	0.313	0.077	0.152
Causal stacking	0.761	0.532	0.391	0.264	0.127	0.089	0.145
Super Learner**	0.763	0.527	0.381	0.354	0.028	0.080	0.185

* aug; augmentation

** Super Learner; proposed method

Table 8. Evaluation metrics for Scenario 1-3-2: Data generating process 1 with a 1:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.563	0.450	0.505	0.274	0.231	0.053	0.138
S-learner	0.574	0.459	0.292	0.271	0.021	0.066	0.143
R-learner	0.786	0.543	0.369	0.279	0.091	0.071	0.151
X-learner	0.699	0.489	0.327	0.321	0.006	0.062	0.164
Causal Forest	0.707	0.431	0.146	0.342	-0.196	0.072	0.148
A-learning	0.004	0.279	3.861	0.015	3.846	5.019	0.010
A-learning aug*	0.629	0.432	0.405	0.209	0.196	0.060	0.130
W-learning	0.006	0.293	4.118	0.017	4.101	4.907	0.012
W-learning aug	0.602	0.432	0.479	0.205	0.274	0.058	0.128
Causal stacking	0.781	0.540	0.371	0.276	0.095	0.071	0.151
Super Learner**	0.723	0.507	0.343	0.341	0.002	0.068	0.179

* aug; augmentation

** Super Learner; proposed method

Table 9. Evaluation metrics for Scenario 2-1-1: Data generating process 2 with a 3:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.155	0.292	1.507	0.082	1.425	0.144	0.042
S-learner	0.391	0.392	0.435	0.194	0.241	0.083	0.105
R-learner	0.371	0.365	0.442	0.107	0.335	0.107	0.066
X-learner	0.233	0.316	0.508	0.118	0.390	0.083	0.062
Causal Forest	0.161	0.238	0.229	0.097	0.131	0.189	0.028
A-learning	-0.011	0.287	4.893	0.010	4.884	5.567	0.007
A-learning aug*	0.645	0.432	0.467	0.192	0.275	0.088	0.122
W-learning	0.012	0.321	5.145	0.013	5.132	4.688	0.013
W-learning aug	0.745	0.493	0.618	0.254	0.364	0.072	0.154
Causal stacking	0.364	0.365	0.462	0.098	0.364	0.117	0.063
Super Learner**	0.660	0.473	0.447	0.276	0.171	0.120	0.154

* aug; augmentation

** Super Learner; proposed method

Table 10. Evaluation metrics for Scenario 2-2-1: Data generating process 2 with a 2:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.159	0.299	1.331	0.081	1.249	0.101	0.041
S-learner	0.395	0.403	0.382	0.195	0.187	0.074	0.103
R-learner	0.407	0.389	0.403	0.114	0.289	0.102	0.067
X-learner	0.272	0.333	0.500	0.130	0.369	0.073	0.066
Causal Forest	0.154	0.254	0.238	0.074	0.165	0.163	0.027
A-learning	0.007	0.290	4.906	0.012	4.894	5.553	0.008
A-learning aug*	0.635	0.449	0.441	0.208	0.232	0.074	0.126
W-learning	0.017	0.323	4.687	0.013	4.674	4.350	0.012
W-learning aug	0.698	0.487	0.563	0.246	0.317	0.066	0.146
Causal stacking	0.405	0.387	0.423	0.105	0.318	0.108	0.064
Super Learner**	0.651	0.471	0.424	0.273	0.150	0.103	0.151

* aug; augmentation

** Super Learner; proposed method

Table 11. Evaluation metrics for Scenario 2-3-1: Data generating process 2 with a 1:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.157	0.296	1.237	0.082	1.155	0.084	0.041
S-learner	0.374	0.380	0.352	0.189	0.163	0.070	0.099
R-learner	0.458	0.405	0.387	0.149	0.238	0.116	0.085
X-learner	0.330	0.348	0.496	0.173	0.323	0.076	0.083
Causal Forest	0.175	0.220	0.178	0.129	0.048	0.134	0.031
A-learning	0.000	0.270	3.895	0.014	3.880	4.646	0.009
A-learning aug*	0.622	0.431	0.405	0.208	0.197	0.066	0.128
W-learning	-0.004	0.289	3.883	0.016	3.867	4.070	0.011
W-learning aug	0.595	0.429	0.479	0.208	0.271	0.066	0.129
Causal stacking	0.455	0.402	0.392	0.148	0.244	0.116	0.085
Super Learner**	0.557	0.427	0.363	0.267	0.096	0.091	0.146

* aug; augmentation

** Super Learner; proposed method

Table 12. Evaluation metrics for Scenario 2-1-2: Data generating process 2 with a 3:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.155	0.293	1.498	0.080	1.418	0.145	0.041
S-learner	0.406	0.394	0.408	0.198	0.210	0.070	0.105
R-learner	0.372	0.367	0.431	0.104	0.327	0.097	0.066
X-learner	0.256	0.322	0.514	0.120	0.394	0.097	0.062
Causal Forest	0.165	0.244	0.215	0.087	0.128	0.150	0.022
A-learning	-0.016	0.292	5.805	0.006	5.799	7.536	0.005
A-learning aug*	0.666	0.452	0.467	0.211	0.256	0.080	0.132
W-learning	0.002	0.327	5.228	0.013	5.215	5.514	0.012
W-learning aug	0.749	0.509	0.616	0.271	0.345	0.069	0.159
Causal stacking	0.359	0.364	0.497	0.094	0.404	0.147	0.061
Super Learner**	0.613	0.446	0.488	0.244	0.244	0.149	0.142

* aug; augmentation

** Super Learner; proposed method

Table 13. Evaluation metrics for Scenario 2-2-2: Data generating process 2 with a 2:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.166	0.295	1.357	0.087	1.270	0.100	0.044
S-learner	0.405	0.395	0.393	0.203	0.190	0.065	0.107
R-learner	0.419	0.387	0.420	0.126	0.294	0.093	0.075
X-learner	0.289	0.338	0.503	0.145	0.358	0.079	0.073
Causal Forest	0.178	0.249	0.227	0.099	0.128	0.178	0.029
A-learning	-0.004	0.287	4.894	0.010	4.884	5.701	0.008
A-learning aug*	0.641	0.452	0.453	0.227	0.226	0.067	0.137
W-learning	0.019	0.324	4.754	0.014	4.740	4.334	0.013
W-learning aug	0.702	0.484	0.570	0.253	0.317	0.064	0.150
Causal stacking	0.411	0.384	0.462	0.113	0.349	0.120	0.069
Super Learner**	0.633	0.458	0.434	0.256	0.178	0.100	0.148

* aug; augmentation

** Super Learner; proposed method

Table 14. Evaluation metrics for Scenario 2-3-2: Data generating process 2 with a 1:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

Method	$\mathbf{corr}(\hat{\tau}, \tau)$	$\mathbf{agree}(\hat{S}, S)$	$\widehat{\mathbf{ATE}}(\hat{S})$	$\mathbf{ATE}(\hat{S})$	$\mathbf{bias}\{\mathbf{ATE}(\hat{S})\}$	$\mathbf{SD}\{\widehat{\mathbf{ATE}}(\hat{S})\}$	$\boldsymbol{\eta}$
T-learner	0.152	0.296	1.245	0.079	1.166	0.096	0.040
S-learner	0.338	0.370	0.349	0.167	0.182	0.067	0.089
R-learner	0.442	0.407	0.382	0.124	0.258	0.094	0.075
X-learner	0.308	0.345	0.509	0.152	0.357	0.078	0.077
Causal Forest	0.189	0.229	0.191	0.124	0.068	0.141	0.032
A-learning	-0.011	0.281	3.072	0.007	3.065	3.613	0.005
A-learning aug*	0.599	0.421	0.406	0.184	0.221	0.062	0.118
W-learning	-0.001	0.295	3.451	0.012	3.439	3.570	0.009
W-learning aug	0.576	0.422	0.480	0.185	0.295	0.060	0.119
Causal stacking	0.438	0.405	0.387	0.122	0.265	0.095	0.075
Super Learner**	0.530	0.417	0.371	0.250	0.121	0.118	0.133

* aug; augmentation

** Super Learner; proposed method

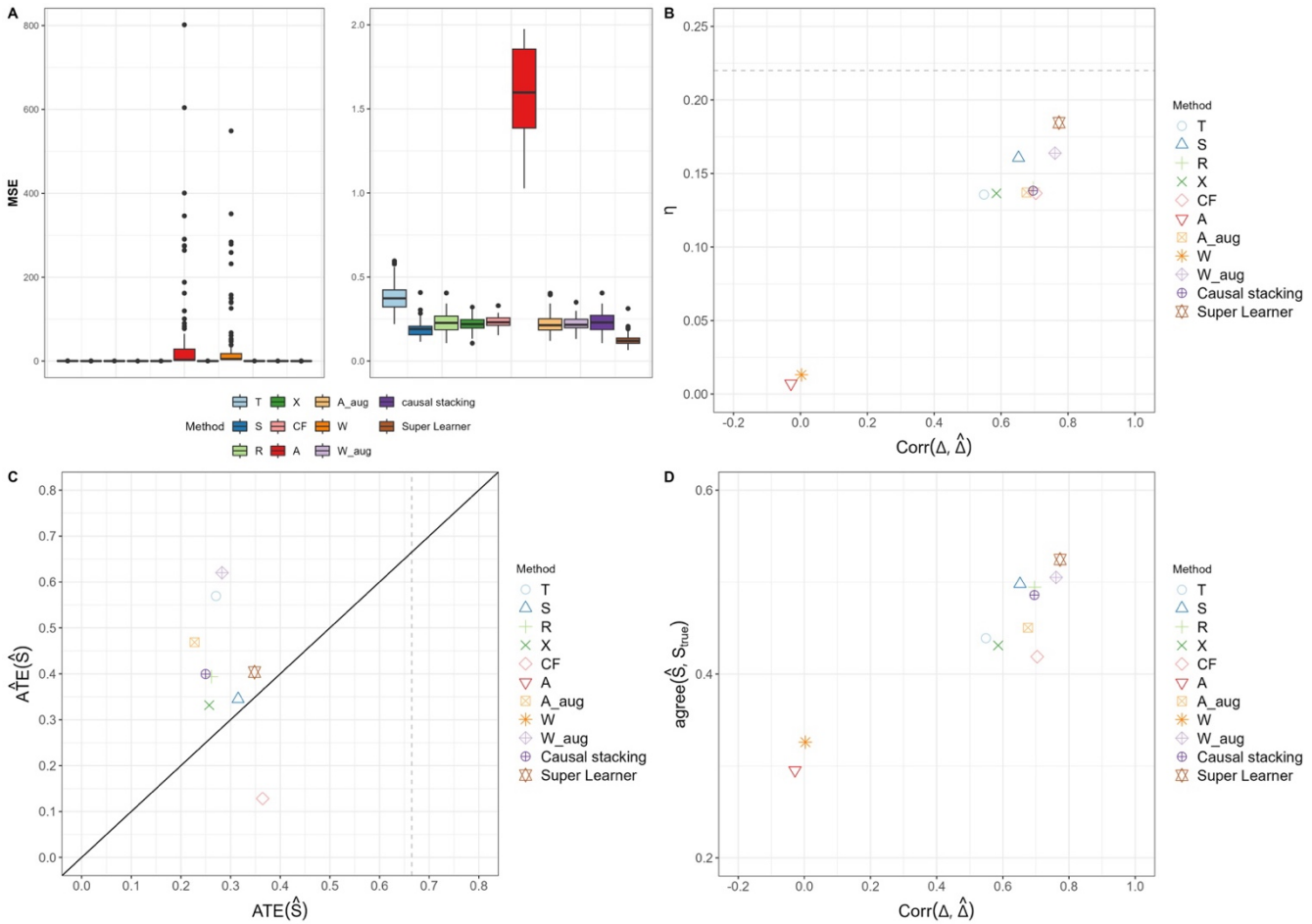


Figure 1. Visualization of model performance: Evaluation metrics for scenario 1-1-1: Data generating process 1 with a 3:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

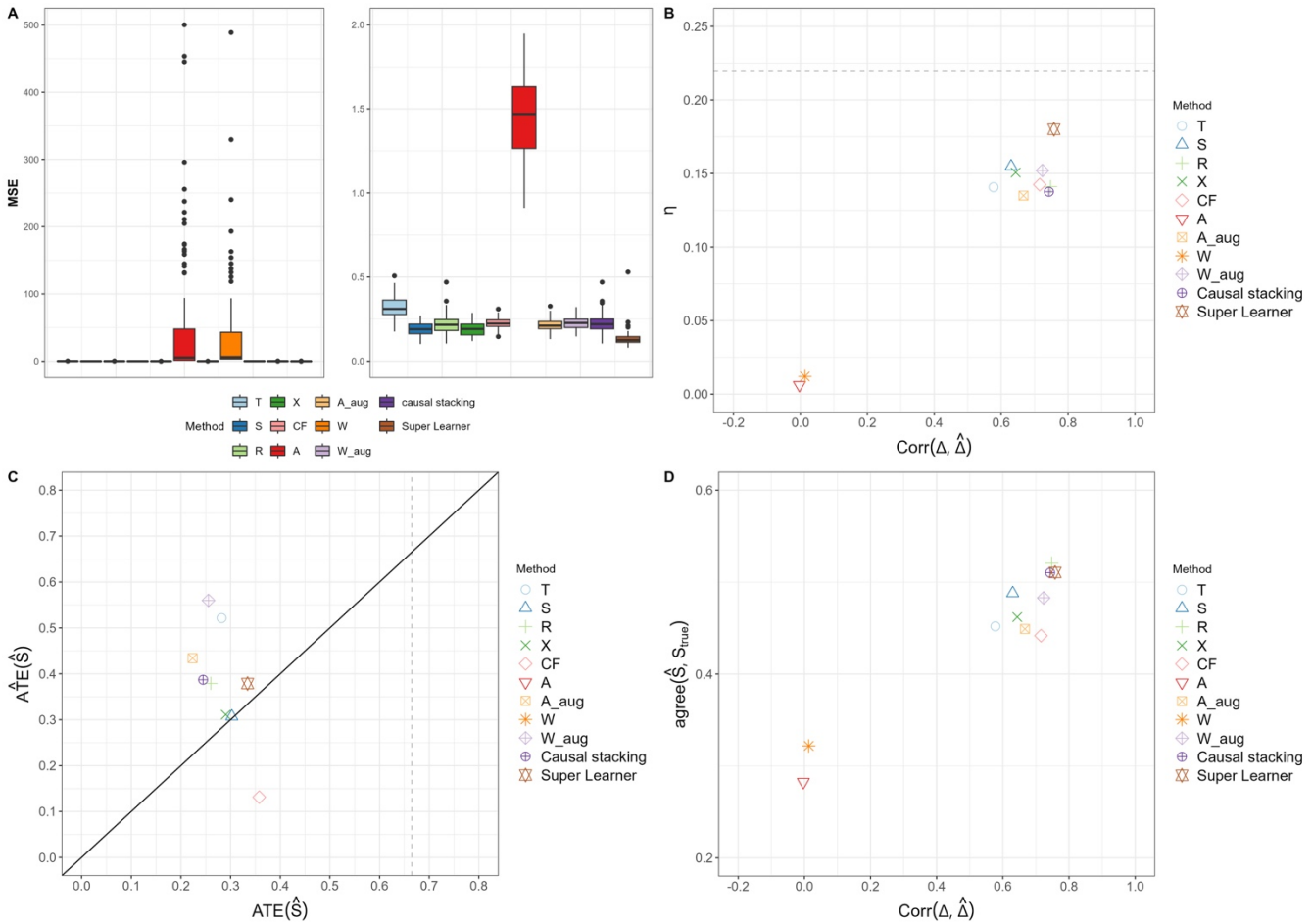


Figure 2. Visualization of model performance: Evaluation metrics for scenario 1-2-1: Data generating process 1 with a 2:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

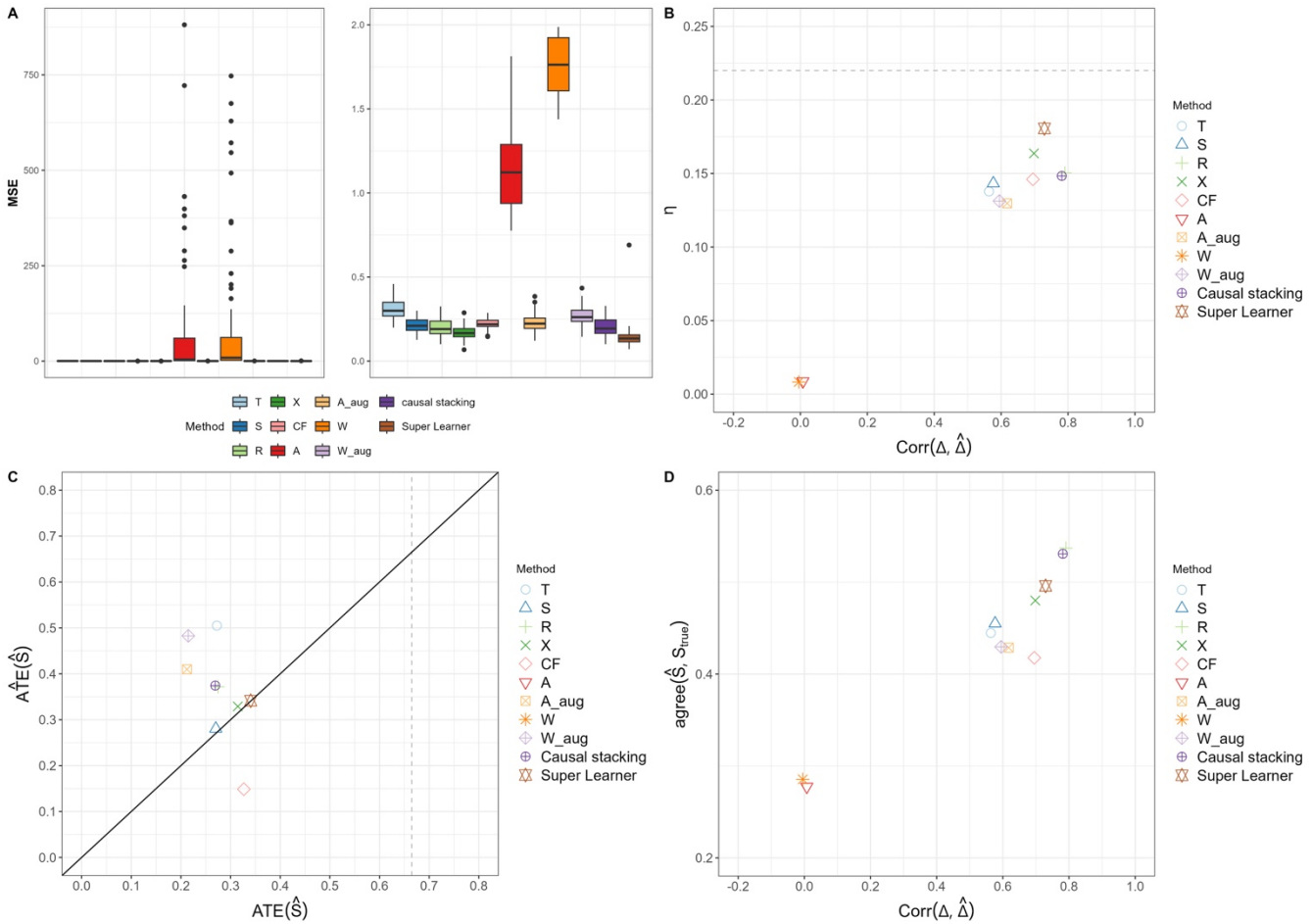


Figure 3. Visualization of model performance: Evaluation metrics for scenario 1-3-1: Data generating process 1 with a 1:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

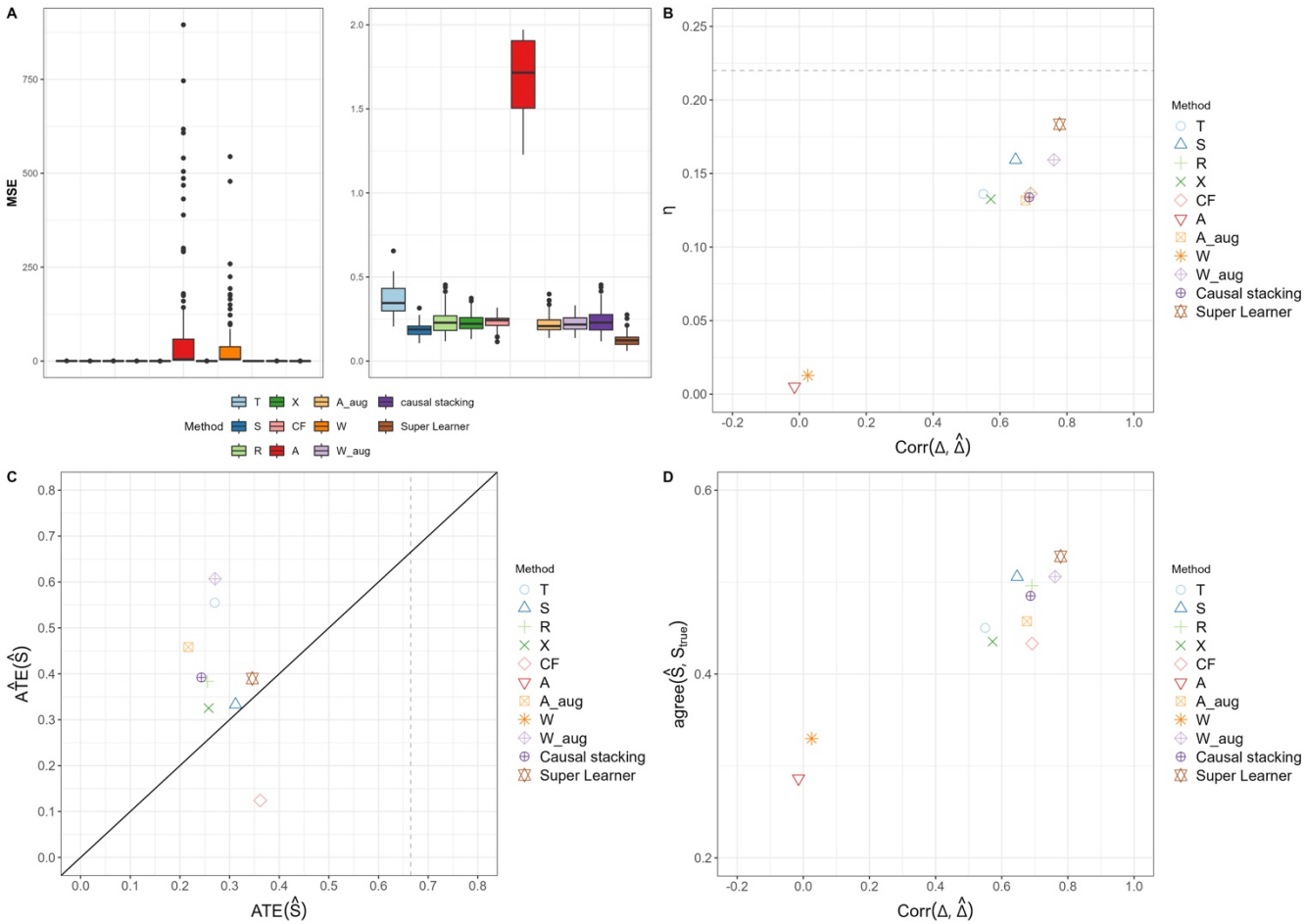


Figure 4. Visualization of model performance: Evaluation metrics for scenario 1-1-2: Data generating process 1 with a 3:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

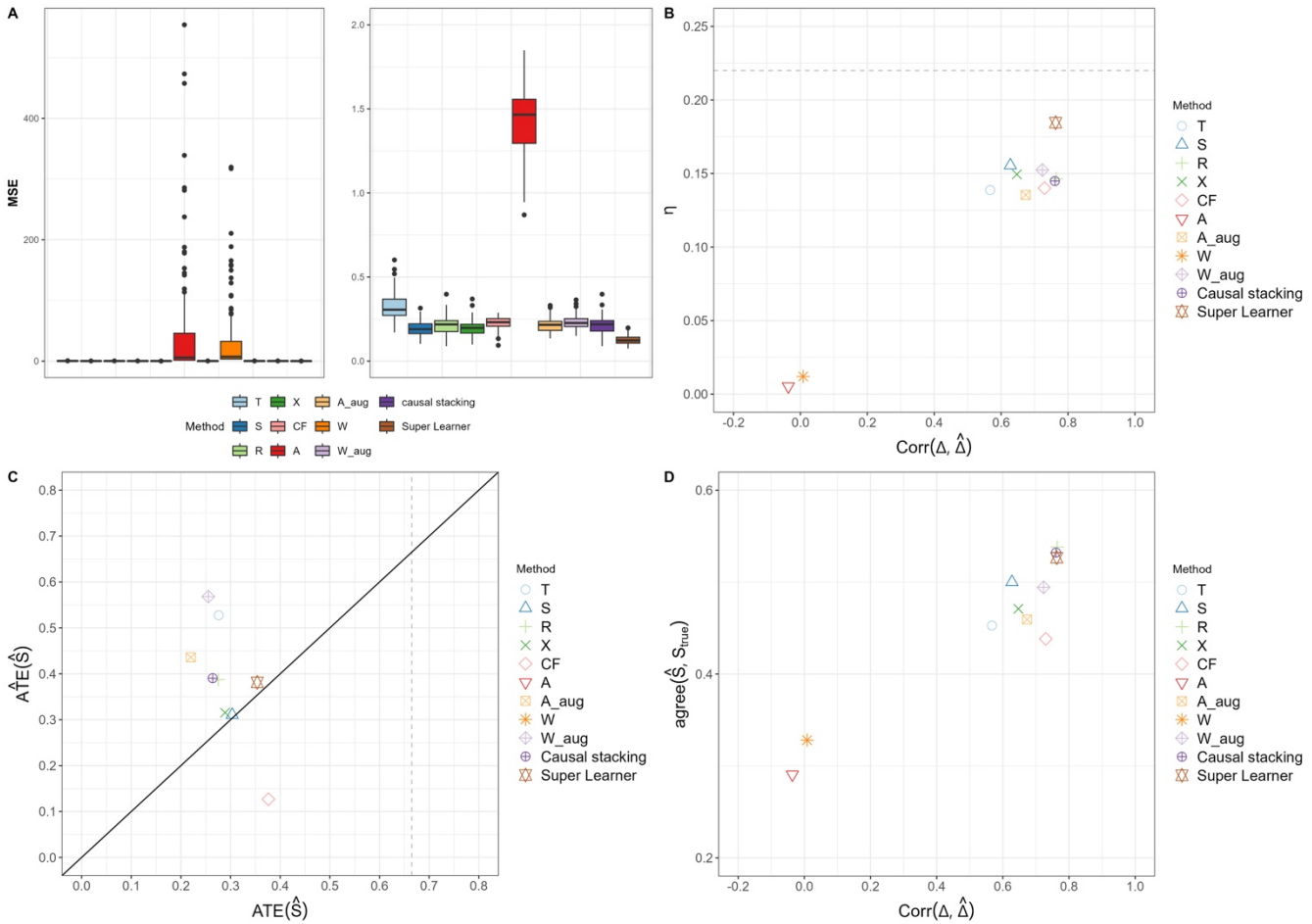


Figure 5. Visualization of model performance: Evaluation metrics for scenario 1-2-2: Data generating process 1 with a 2:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

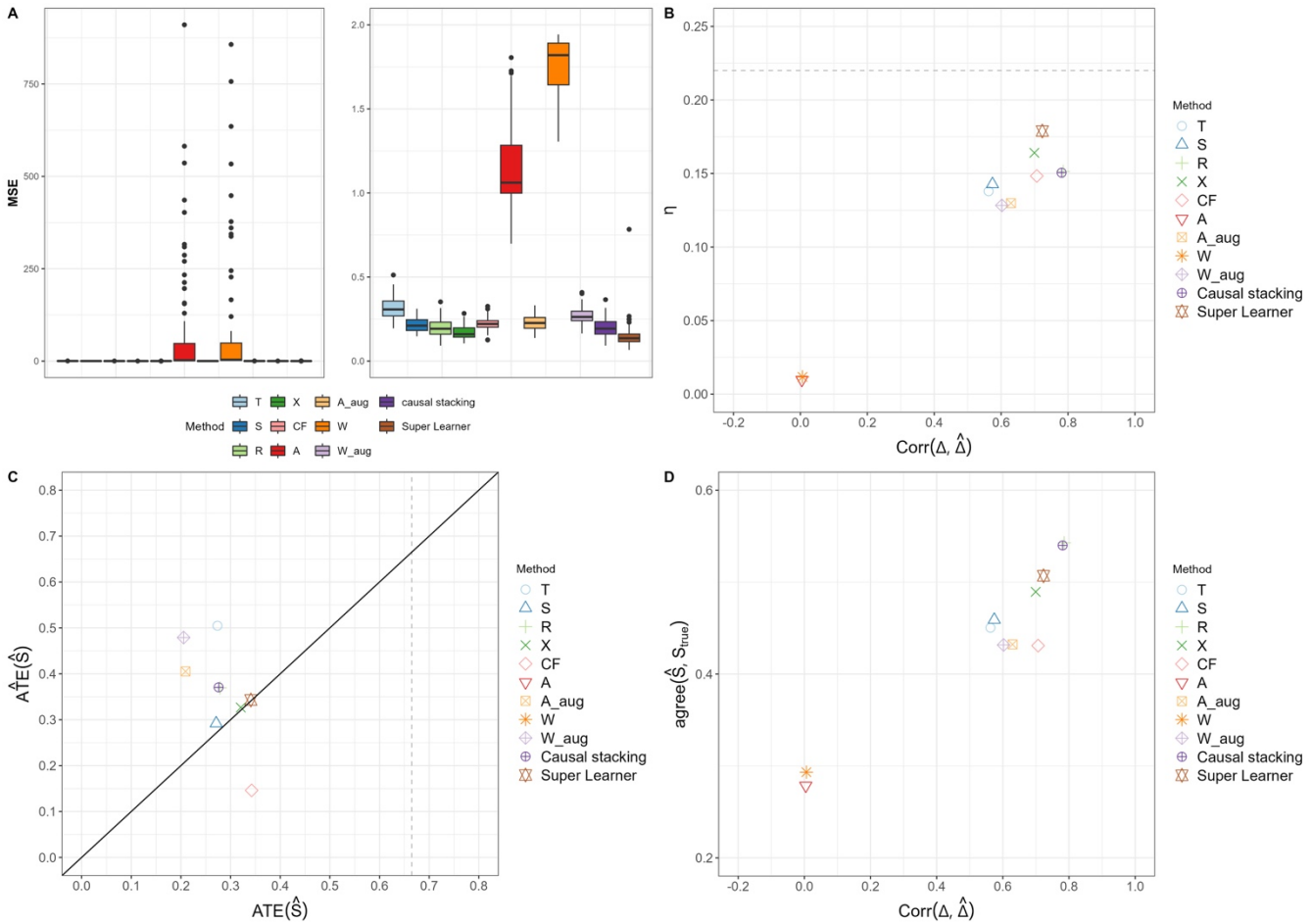


Figure 6. Visualization of model performance: Evaluation metrics for scenario 1-3-2: Data generating process 1 with a 1:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

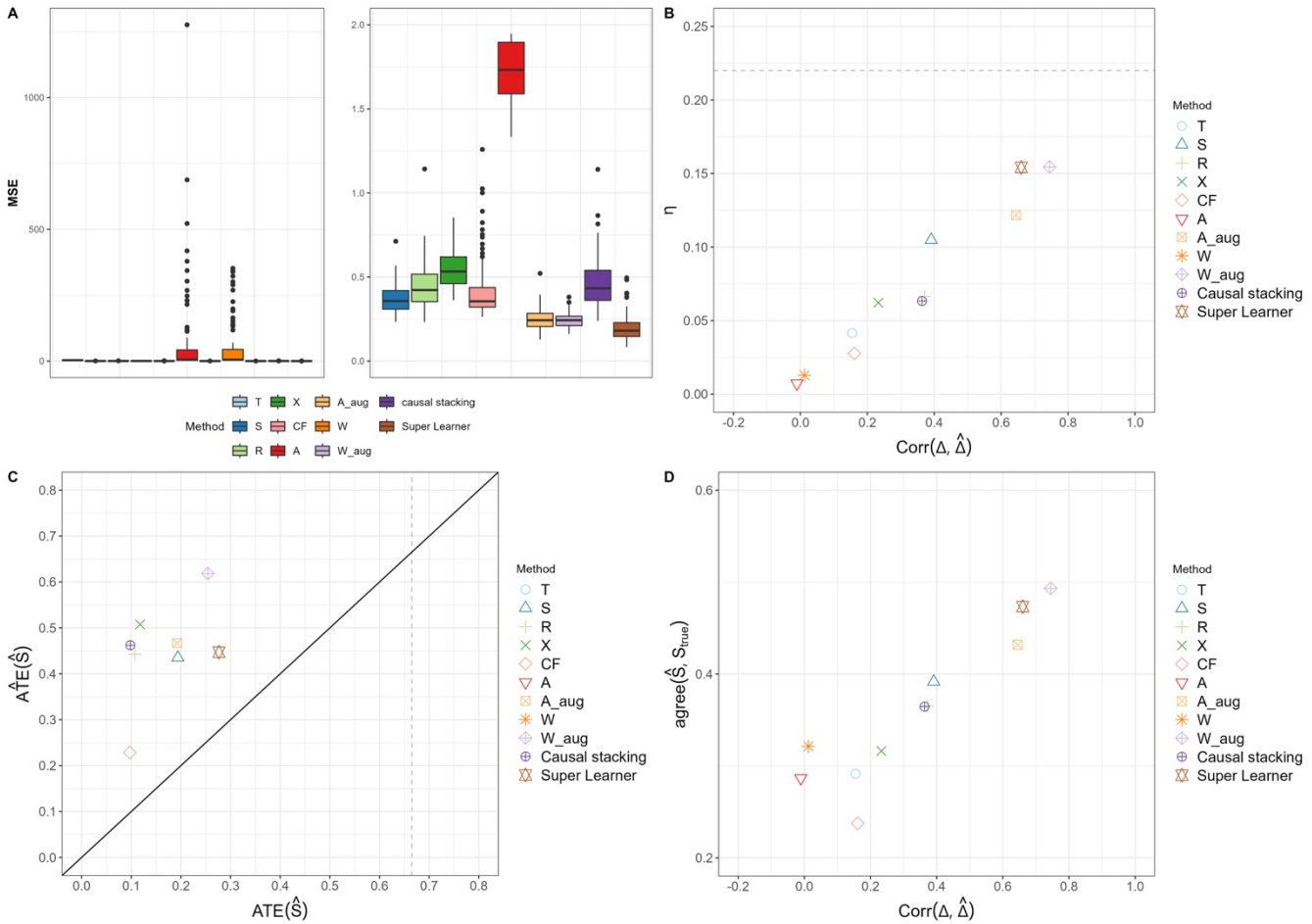


Figure 7. Visualization of model performance: Evaluation metrics for scenario 2-1-1: Data generating process 2 with a 3:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

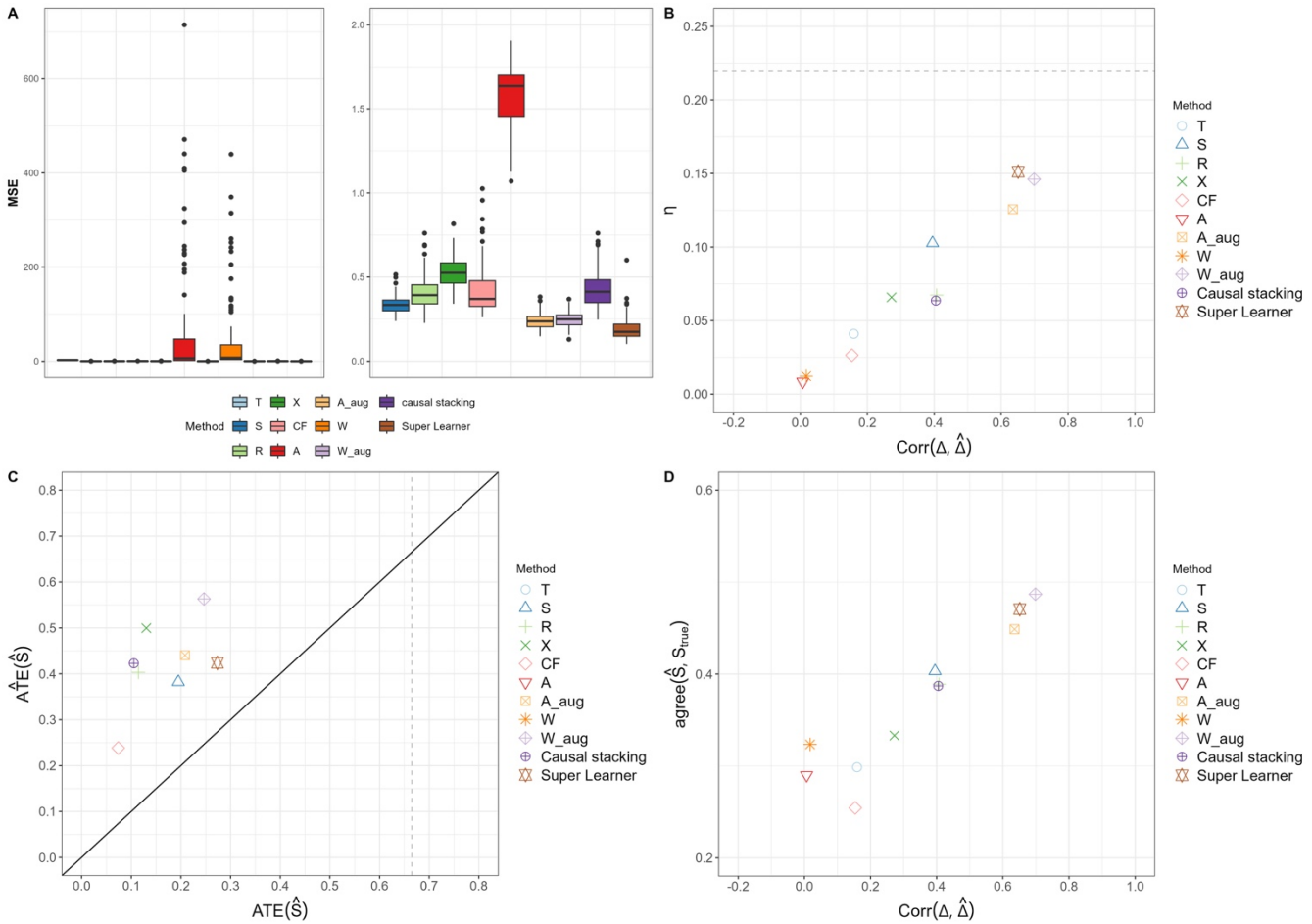


Figure 8. Visualization of model performance: Evaluation metrics for scenario 2-2-1: Data generating process 2 with a 2:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

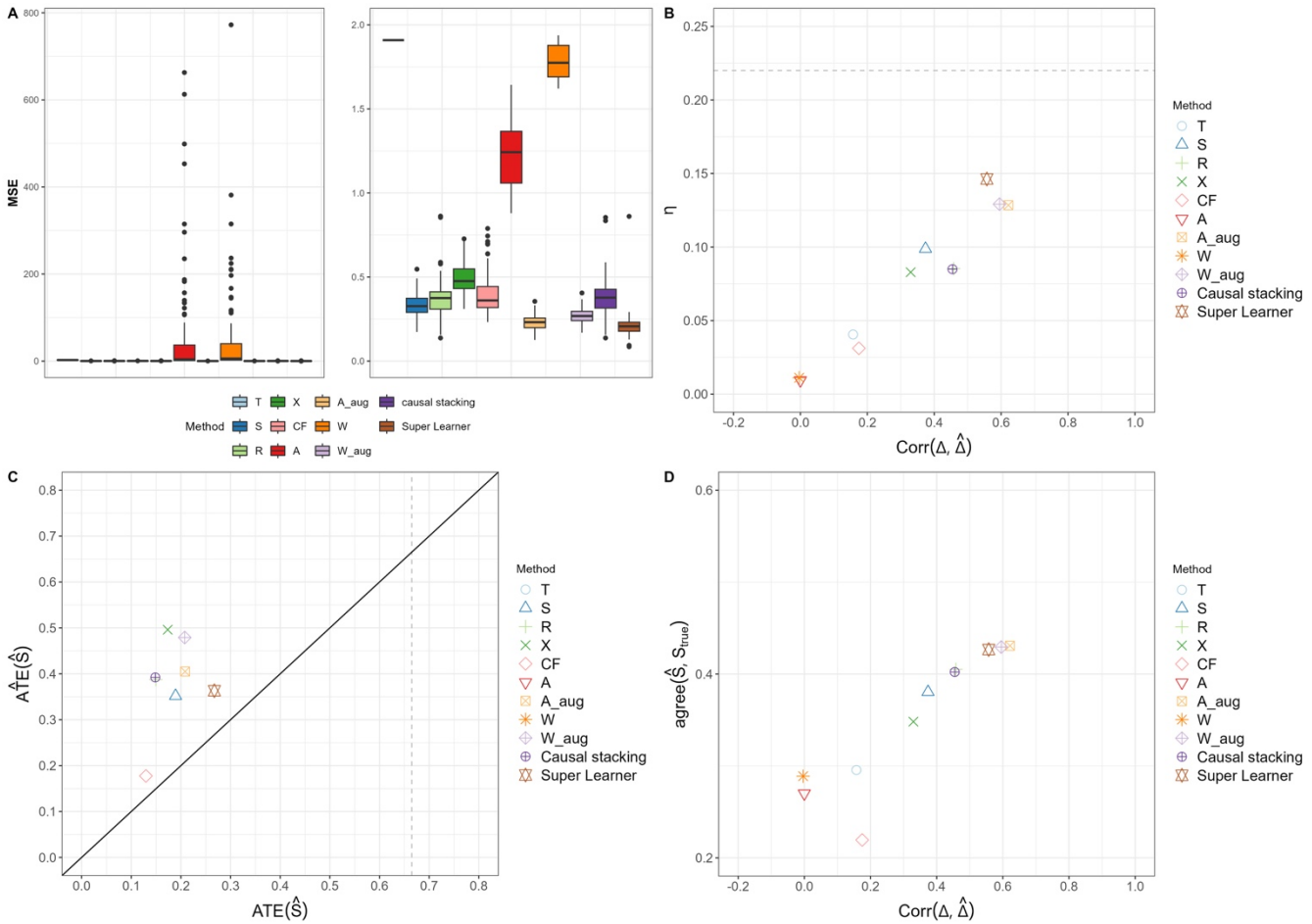


Figure 9. Visualization of model performance: Evaluation metrics for scenario 2-3-1: Data generating process 2 with a 1:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

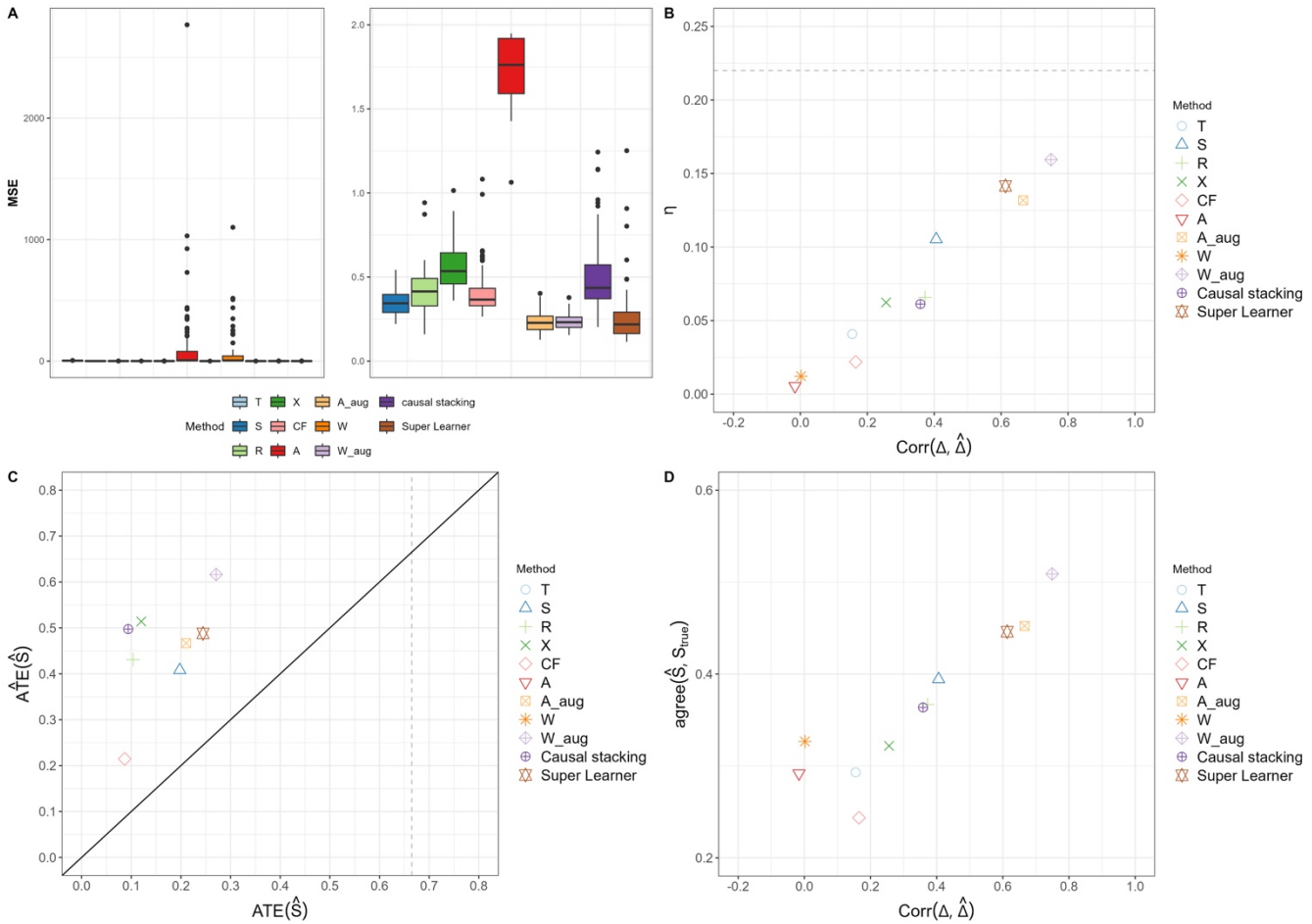


Figure 10. Visualization of model performance: Evaluation metrics for scenario 2-1-2: Data generating process 2 with a 3:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

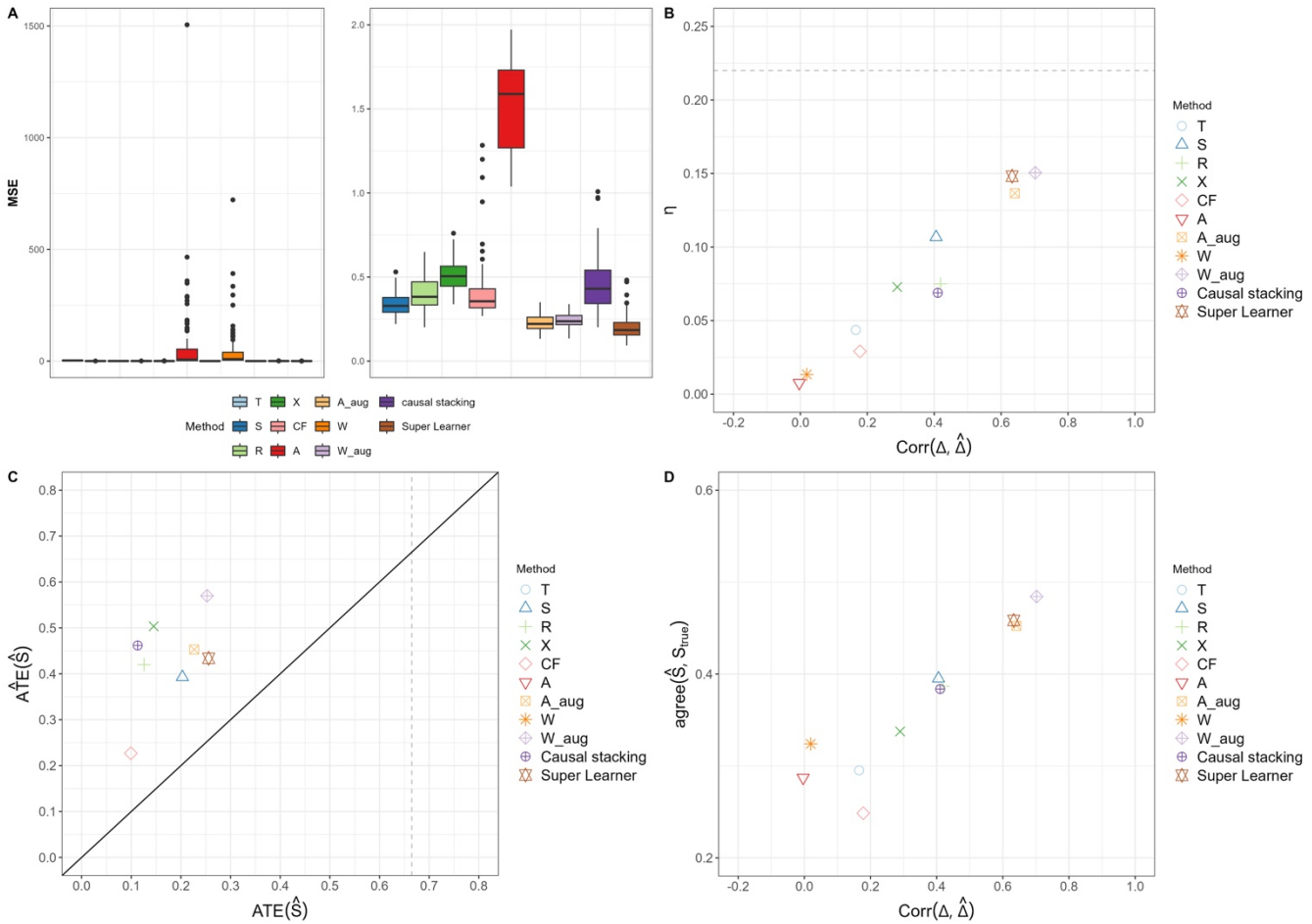


Figure 11. Visualization of model performance: Evaluation metrics for scenario 2-2-2: Data generating process 2 with a 2:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

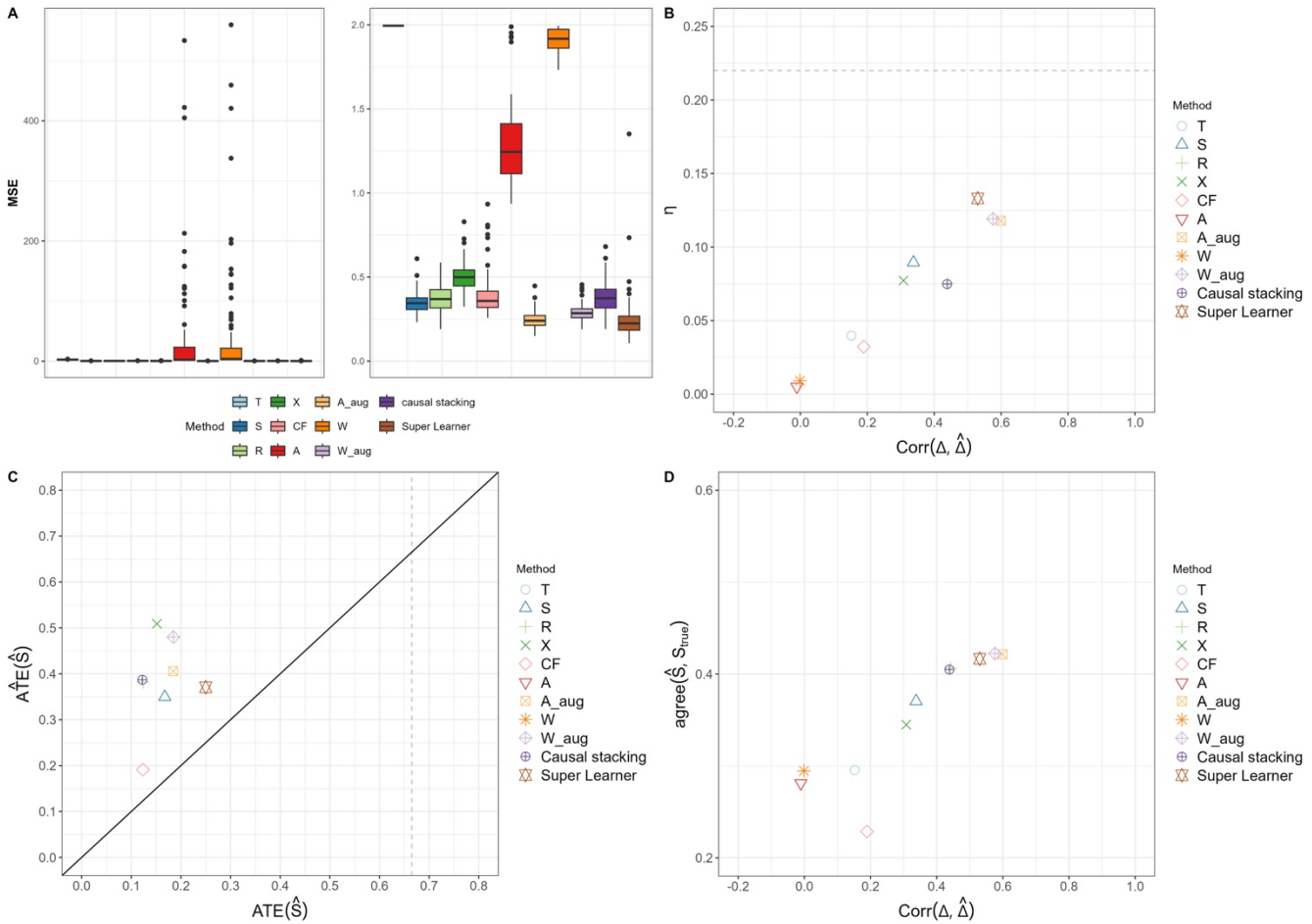


Figure 12. Visualization of model performance: Evaluation metrics for scenario 2-3-2: Data generating process 2 with a 1:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

Chapter 6

Conclusion and Discussion

In causal inference, developing a robust and consistent method for estimating conditional average treatment effects (CATE) is critically important. Existing methods often rely on varying assumptions, which can lead to inconsistent results. If the CATE estimation results differ across methods, clinicians may face difficulties in deciding whether to administer treatment to patients with specific covariates.

To address this issue, this study introduces a Super Learner-based CATE estimation method that demonstrates robust and reliable performance, even in challenging and complex estimation settings. The proposed method uses cross validation to efficiently utilize data, overcoming the limitations of individual methods and combining the strengths of each method to produce more stable and accurate estimates.

In the simulation study, the proposed method outperformed other methods across multiple evaluation metrics, including mean squared error (MSE), bias, and subgroup utility index. It demonstrated enhanced accuracy and robustness, highlighting its potential as an effective approach for estimating CATE, particularly in scenarios where traditional methods struggle due to data complexity or heterogeneity. These findings suggest that the Super Learner-based approach can effectively address key challenges in CATE estimation, enhancing the consistency and interpretability of the results. Moreover, the null model simulation further demonstrated that in the absence of heterogeneous treatment effects, most methods produce HTE estimates close to zero.

Despite these strengths, the study also underscores certain limitations. The reliance on a plug-in estimator as a substitute for the true CATE leads to inherent uncertainty due to the unobservability of the true values. This limitation emphasizes the need for further investigation of alternative substitutes for the true CATE. Comparing models using multiple substitutes could help reduce potential biases and enhance the robustness of the estimation procedure.

When applying various meta-learners or methods such as causal forests, this study employed XGBoost; however, alternative machine learning methods could also be applied. In such cases, the results may vary depending on the dataset used in simulations or real-world applications. Further research could explore which machine learning methods yield better performance in outcome modeling.

Further studies should extend the proposed method to real-world applications, such as randomized clinical trials (RCTs) and observational studies, with more pronounced heterogeneity in data and treatment effects. Additionally, refining methodologies to support treatment recommendations based on heterogeneous treatment effects represents a significant area for further investigation. These efforts will be essential for advancing both methodological theory and practical applications of CATE estimation, particularly in fields like precision medicine and policymaking.

Continued advancements in this area, especially in mitigating inherent uncertainties and enhancing the robustness of estimation methods, will be crucial for improving the reliability and applicability of causal inference methodologies. This study contributes to these ongoing efforts by establishing a rigorous framework for more effective and reliable CATE estimation in diverse and complex settings.

Bibliography

Aronow, P. M., & Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, *1*(1), 135-154.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests.

Bouvier, F., Peyrot, E., Balendran, A., Ségalas, C., Roberts, I., Petit, F., & Porcher, R. (2024). Do machine learning methods lead to similar individualized treatment rules? A comparison study on real data. *Stat Med*, *43*(11), 2043-2061.

Chen, S., Tian, L., Cai, T., & Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, *73*(4), 1199-1209.

Foster, J. C., Taylor, J. M., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*(24), 2867-2880.

Han, K. W., & Wu, H. (2022). Ensemble method for estimating individualized treatment effects. *arXiv preprint arXiv:2202.12445*.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008-3049, 3042.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), 4156-4165.

Lipkovich, I., Svensson, D., Ratitch, B., & Dmitrienko, A. (2023). Overview of modern approaches for identifying and evaluating heterogeneous treatment effects from clinical data. *Clin Trials*, *20*(4), 380-393.

Lipkovich, I., Svensson, D., Ratitch, B., & Dmitrienko, A. (2024). Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. *Statistics in Medicine*, *43*(22), 4388-4436.

Loh, W. Y., Cao, L., & Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(5), e1326.

Mahajan, D., Mitliagkas, I., Neal, B., & Syrgkanis, V. (2022). Empirical Analysis of Model Selection for Heterogeneous Causal Effect Estimation. *arXiv preprint arXiv:2211.01939*.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299-319.

Polley, E. C., & Van der Laan, M. J. (2010). Super learner in prediction.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322-331.

Saito, Y., & Yasui, S. (2020). Counterfactual cross-validation: Stable model selection procedure for causal inference models. International Conference on Machine Learning (pp.8398-8407), PMLR.

Van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical applications in genetics and molecular biology*, *6*(1).

Zhao, Q., & Panigrahi, S. (2019). Selective inference for effect modification: An empirical investigation. *Observational Studies*, *5*(2), 131-140.

Appendix

Table A1. CATE estimation method's tuning parameters

CATE estimation methods	Tuning parameters
S-learner	cvboost3
T-learner	k_folds = 5
R-learner	tree depth = {2,3,4}, eta = {0.0005, 0.01, 0.015, 0.025, 0.05, 0.08, 0.1, 0.2},
X-learner	ntree_max = 1000, early_stopping_rounds = 10,
	subsample = 0.9, colsample_bytree = 0.9
Causal Forest	num.trees = 10000
A-learning	max_depth = 5, eta = 0.01, nthread=1, booster = "gbtree", subsample = 0.90, colsample_bytree = 0.90, nrounds = 1000, nfold = 5, early_stopping_rounds = 50
A-learning aug	max_depth = 5, eta = 0.01, nthread=1, booster = "gbtree", subsample = 0.90, colsample_bytree = 0.90, nrounds = 1000, nfold = 5, early_stopping_rounds = 50, nfolds.crossfit = 5, augment.func = aug.func
W-learning	max_depth = 5, eta = 0.01, nthread=1, booster = "gbtree", subsample = 0.90, colsample_bytree = 0.90, nrounds = 1000, nfold = 5, early_stopping_rounds = 50
W-learning aug	max_depth = 5, eta = 0.01, nthread=1, booster = "gbtree", subsample = 0.90, colsample_bytree = 0.90, nrounds = 1000, nfold = 5, early_stopping_rounds = 50, nfolds.crossfit = 5, augment.func = aug.func
Causal stacking	train:validation = 2:1
Super Learner	5-folds

Table A2. Simulation scenarios for the data generating process $D1$ null: 4 prognostic covariates; $D2$ null: 9 prognostic covariates

Scenario	Data generating process	Treatment-to-control ratio	Surrogate CATE
1-1-1	$D1$ null	3:1	S-learner
1-2-1	$D1$ null	2:1	S-learner
1-3-1	$D1$ null	1:1	S-learner
1-1-2	$D1$ null	3:1	T-learner
1-2-2	$D1$ null	2:1	T-learner
1-3-2	$D1$ null	1:1	T-learner
2-1-1	$D2$ null	3:1	S-learner
2-2-1	$D2$ null	2:1	S-learner
2-3-1	$D2$ null	1:1	S-learner
2-1-2	$D2$ null	3:1	T-learner
2-2-2	$D2$ null	2:1	T-learner
2-3-2	$D2$ null	1:1	T-learner

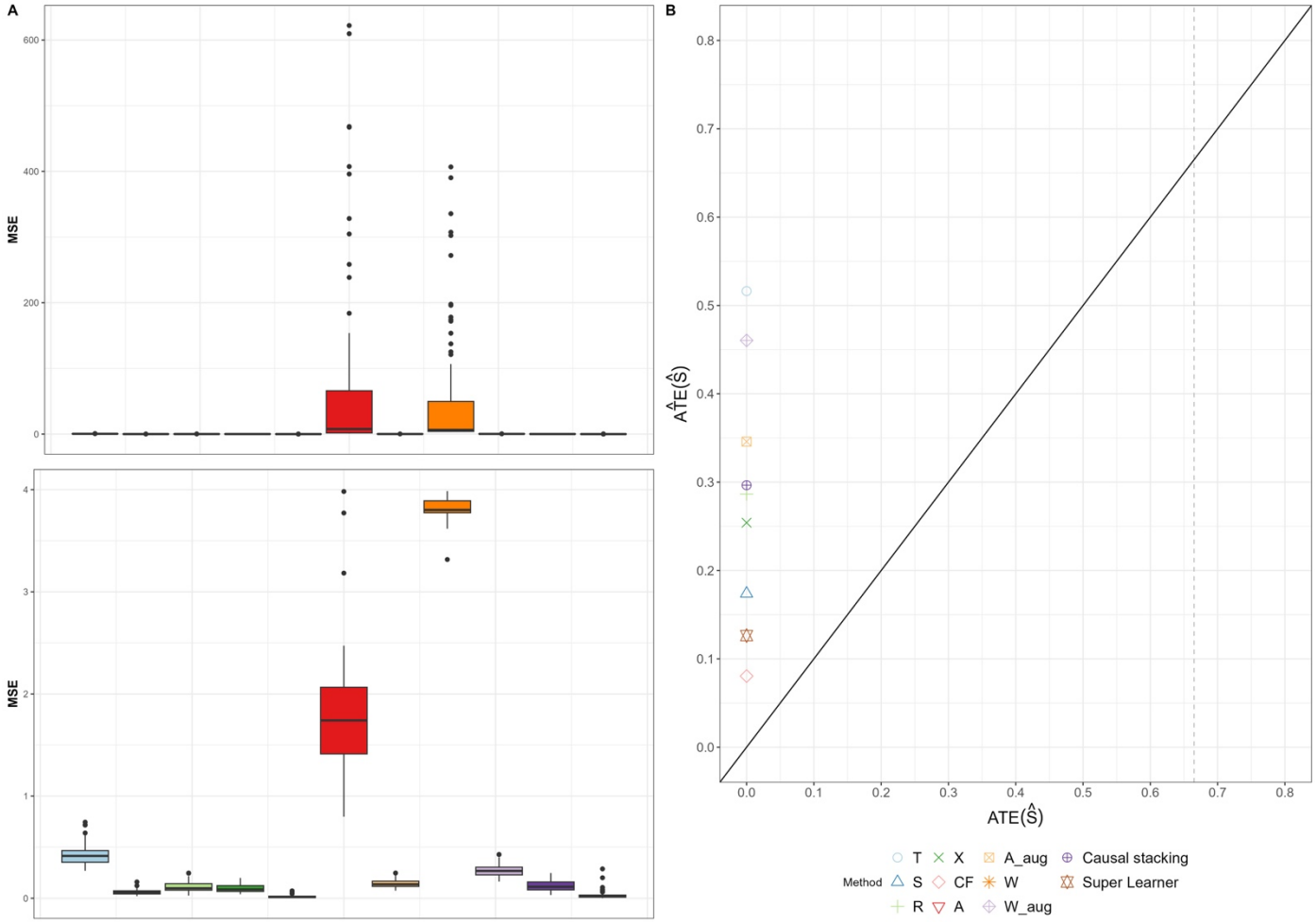


Figure A3. Visualization of model performance: Evaluation metrics for scenario 1-1-1 null model: Data generating process 1 with a 3:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

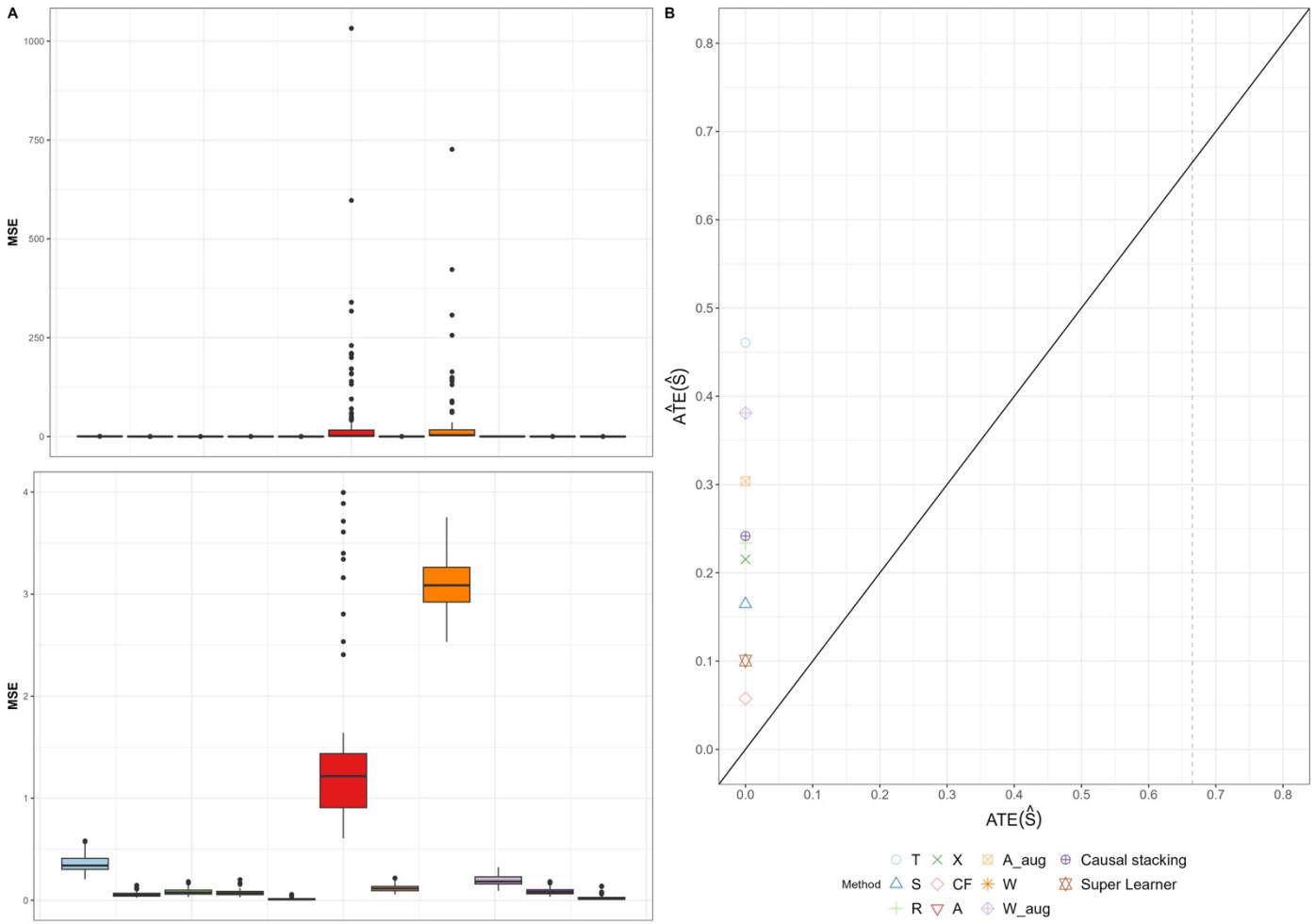


Figure A4. Visualization of model performance: Evaluation metrics for scenario 1-2-1 null model: Data generating process 1 with a 2:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

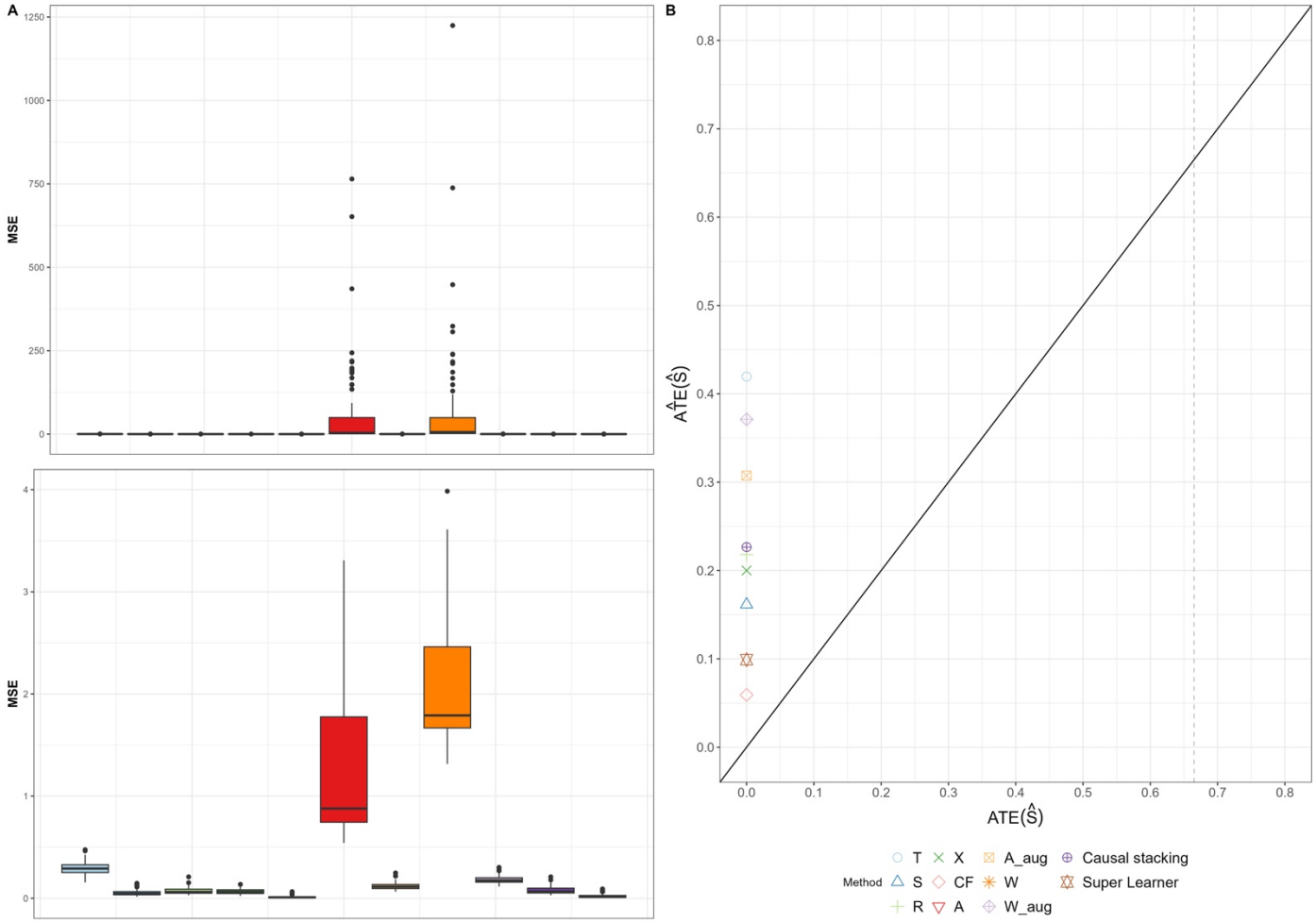


Figure A5. Visualization of model performance: Evaluation metrics for scenario 1-3-1 null model: Data generating process 1 with a 1:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

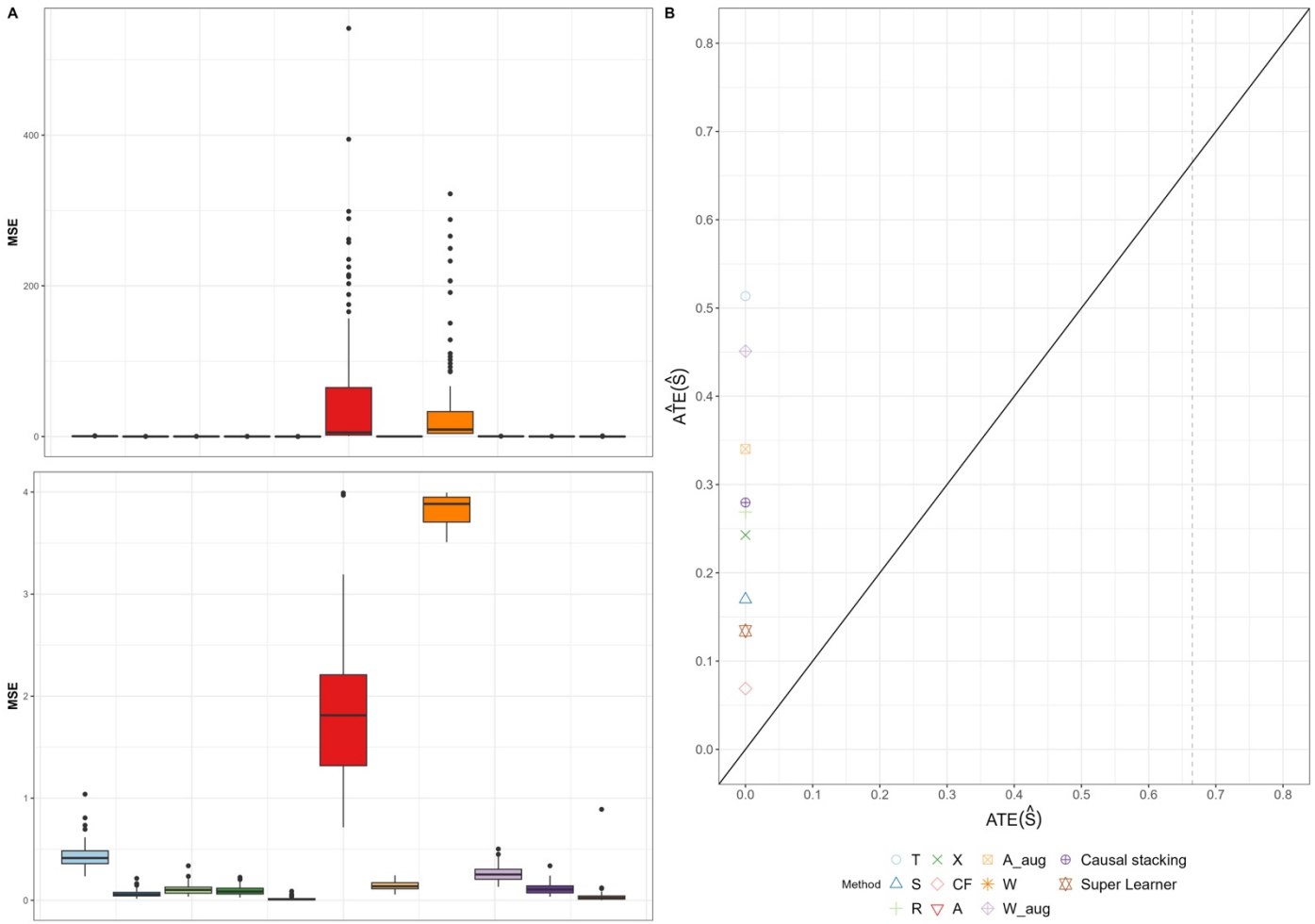


Figure A6. Visualization of model performance: Evaluation metrics for scenario 1-1-2 null model: Data generating process 2 with a 3:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

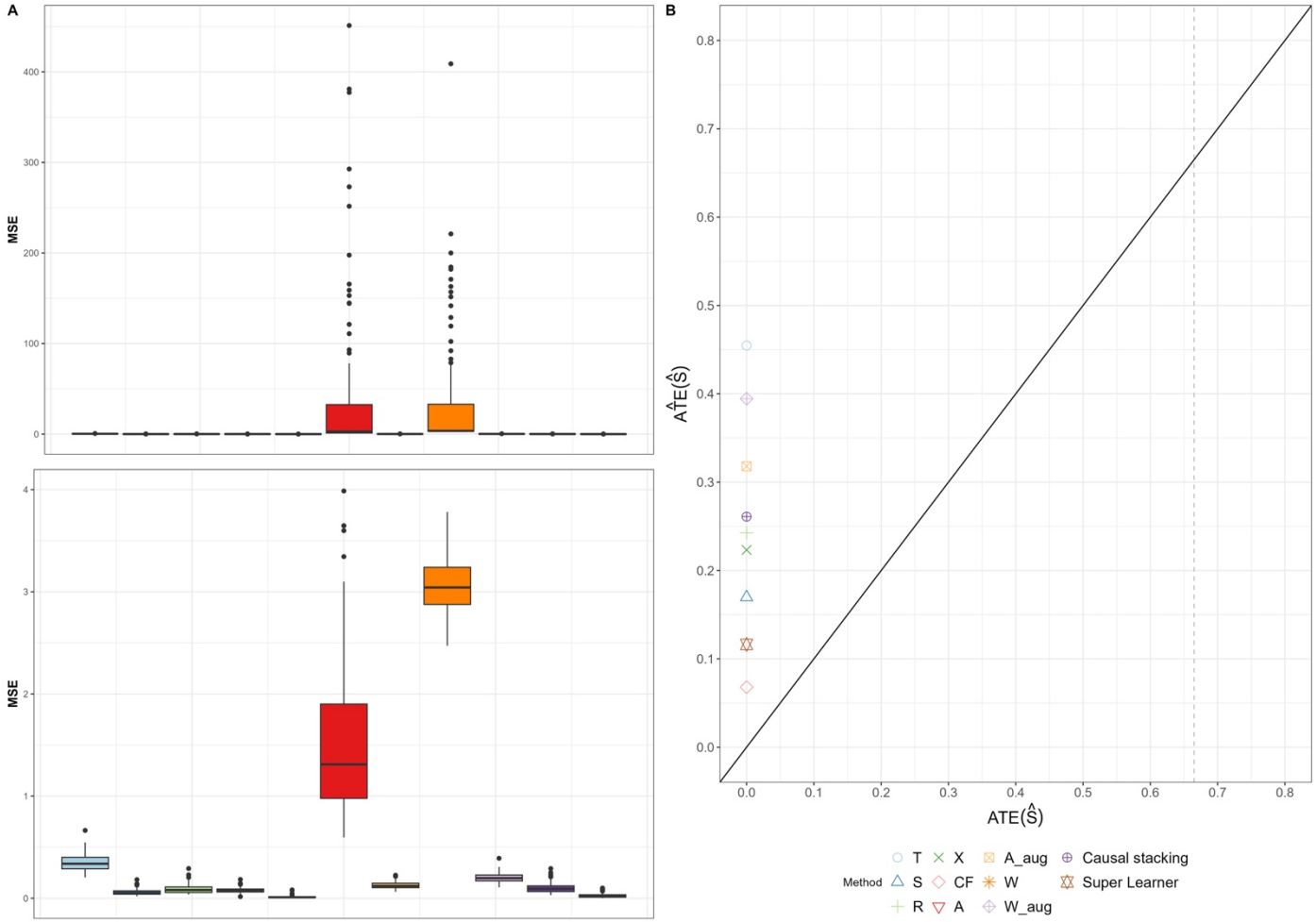


Figure A7. Visualization of model performance: Evaluation metrics for scenario 1-2-2 null model: Data generating process 2 with a 2:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

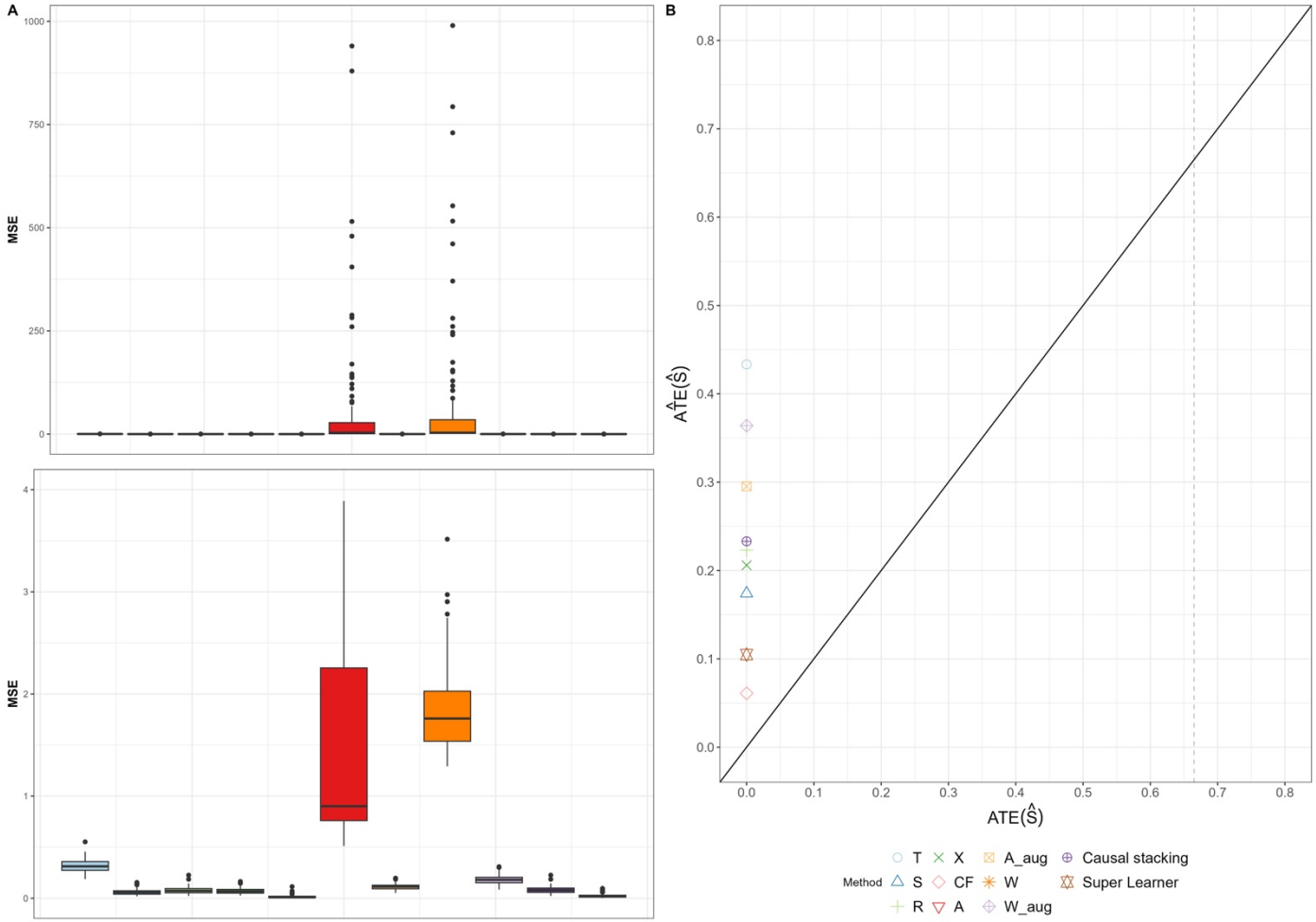


Figure A8. Visualization of model performance: Evaluation metrics for scenario 1-3-2 null model: Data generating process 2 with a 1:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

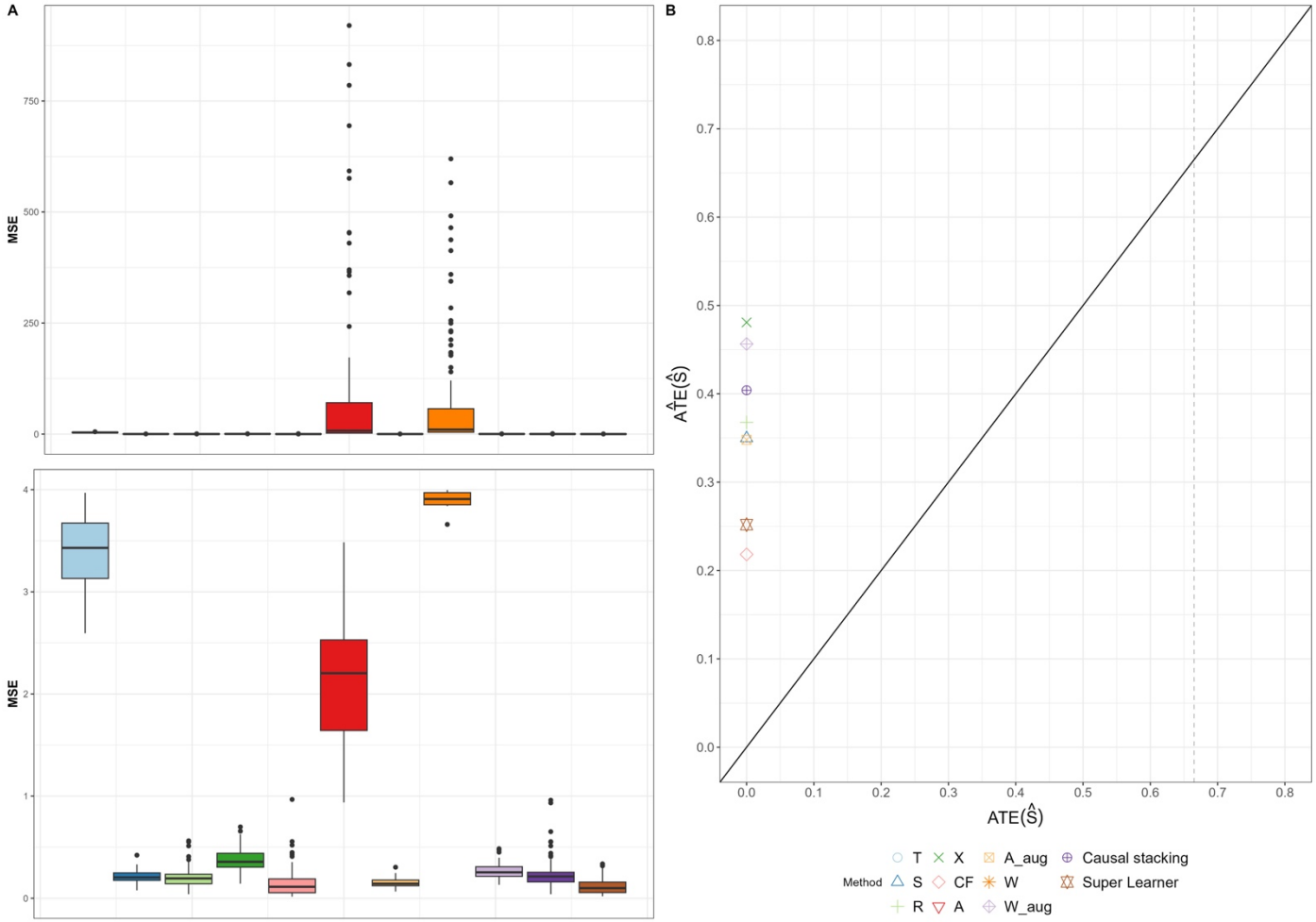


Figure A9. Visualization of model performance: Evaluation metrics for scenario 2-1-1 null model: Data generating process 1 with a 3:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

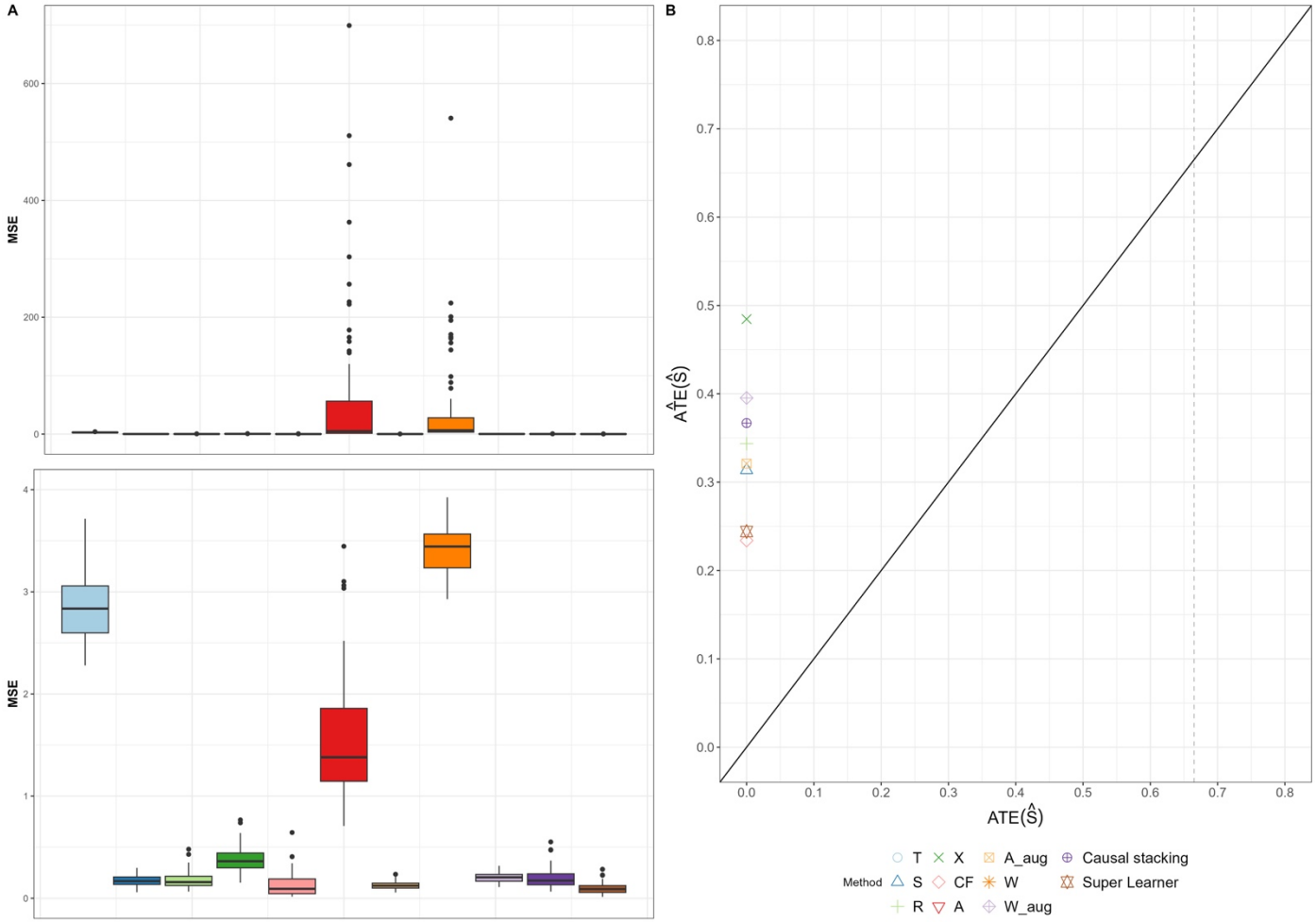


Figure A10. Visualization of model performance: Evaluation metrics for scenario 2-2-1 null model: Data generating process 1 with a 2:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

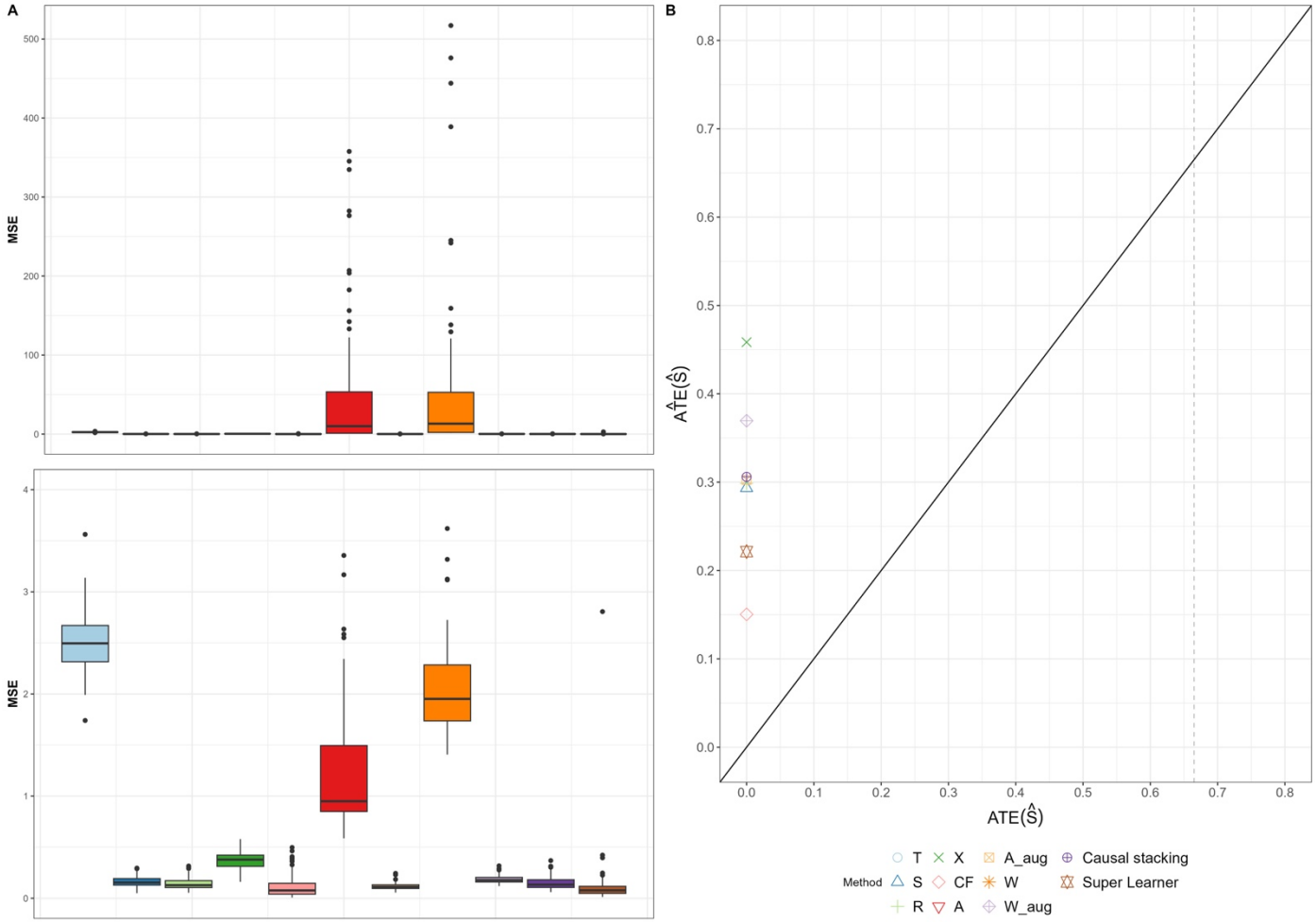


Figure A11. Visualization of model performance: Evaluation metrics for scenario 2-3-1 null model: Data generating process 1 with a 1:1 treatment-to-control ratio using S-learner to estimate surrogate CATE over 100 iterations

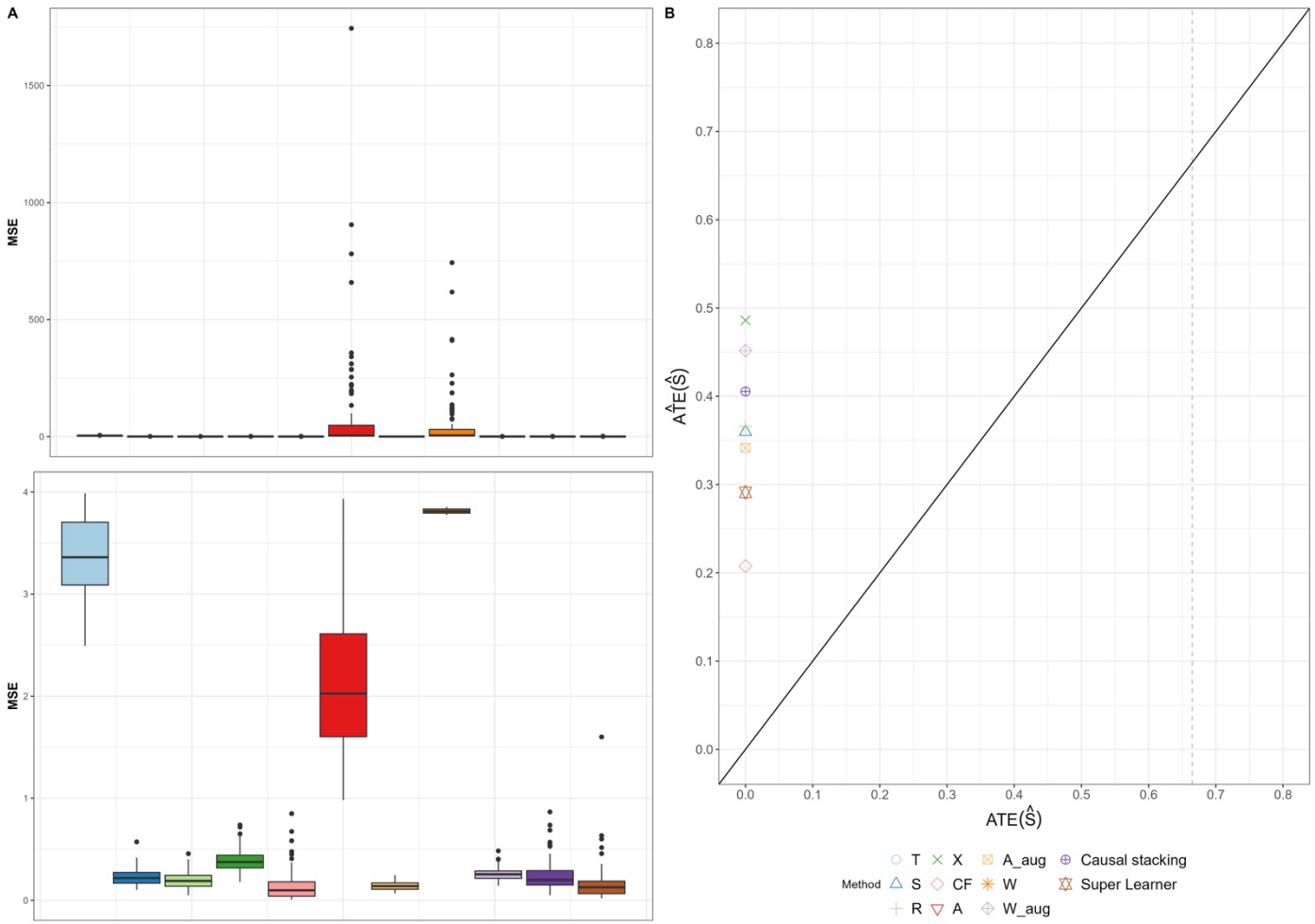


Figure A12. Visualization of model performance: Evaluation metrics for scenario 2-1-2 null model: Data generating process 2 with a 3:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

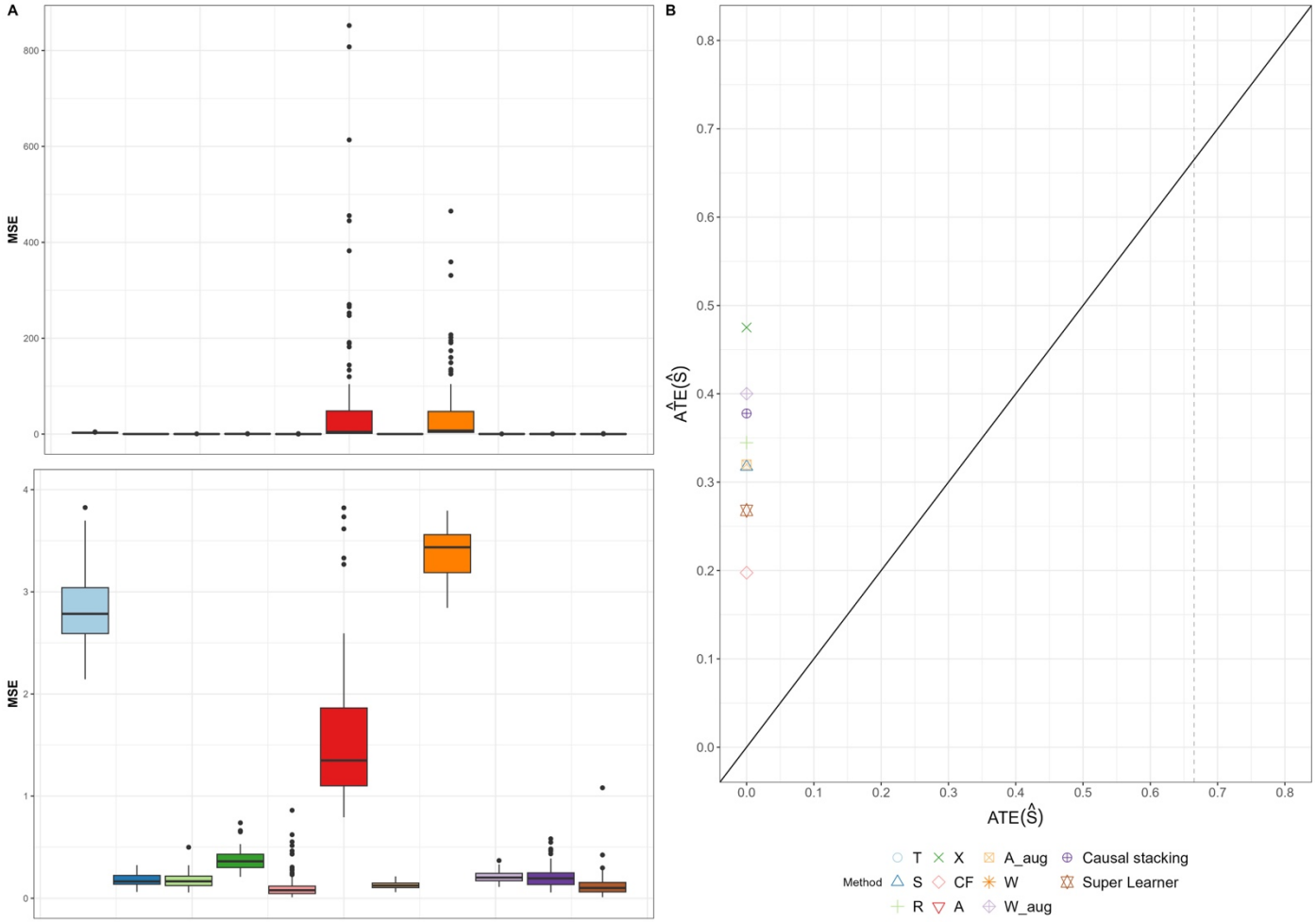


Figure A13. Visualization of model performance: Evaluation metrics for scenario 2-2-2 null model: Data generating process 2 with a 2:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

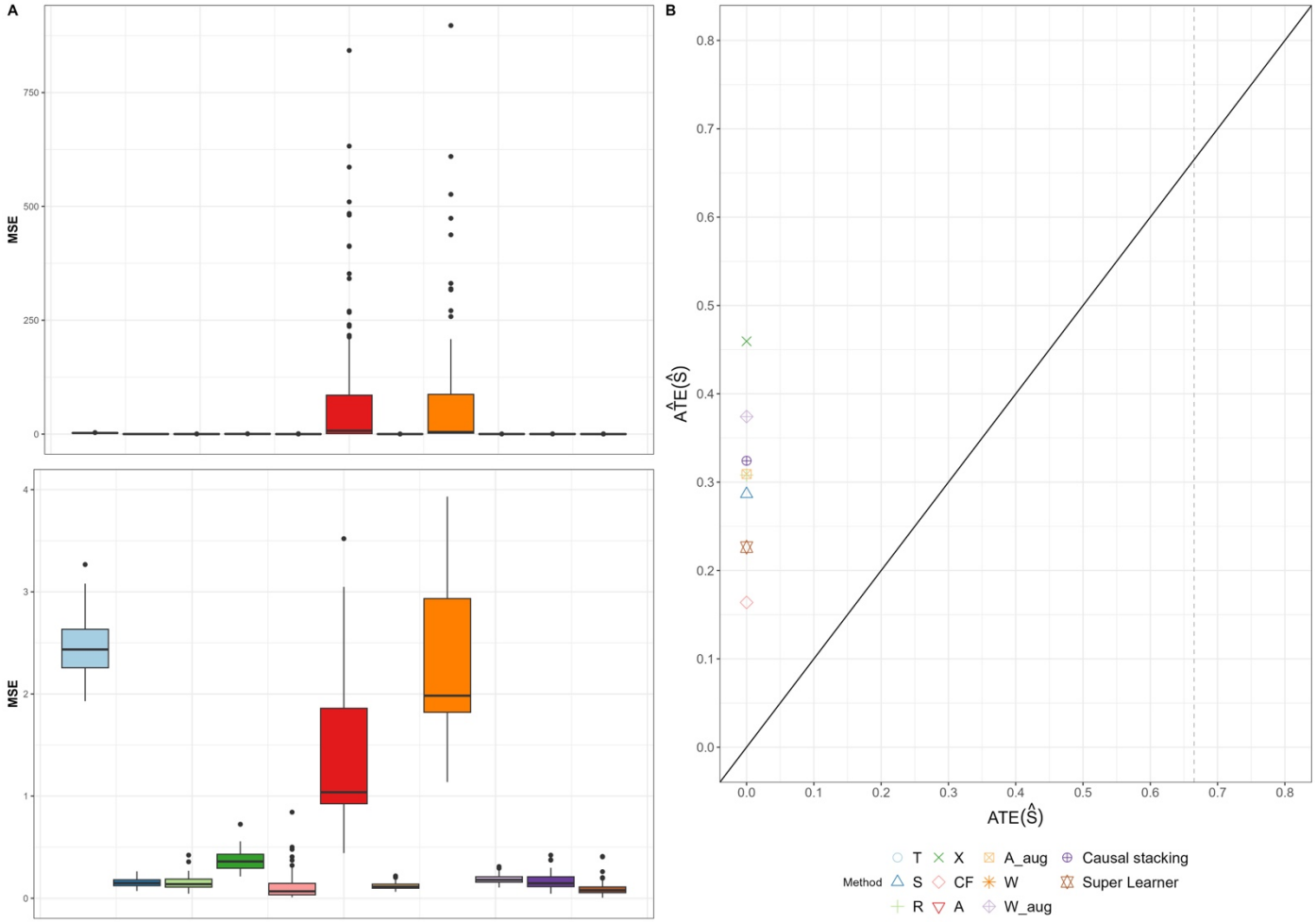


Figure A14. Visualization of model performance: Evaluation metrics for scenario 2-3-2 null model: Data generating process 2 with a 1:1 treatment-to-control ratio using T-learner to estimate surrogate CATE over 100 iterations

Abstract in Korean

RCT에서 Super Learner를 활용한 CATE 추정에 대한 앙상블 접근법

이전 연구들은 인과 추론에서 개별 치료 효과보다는 평균 치료 효과에 초점을 맞춰왔다. 그러나 정밀 의학에 대한 관심이 높아짐에 따라, 최근에는 조건부 평균 처치효과(CATE) 추정과 개별 치료 규칙(ITR)에 대한 연구가 크게 증가하고 있다. CATE 추정은 동일한 특성 속성을 가진 집단에 대한 평균 치료 효과를 추정하는 방법이다.

CATE 추정을 위해 다양한 모수적(parametric) 및 비모수적(non-parametric) 방법이 제안되었으나, 최근 연구들은 특정 기준에서 모든 방법보다 항상 우수한 방법은 없다는 것을 보여준다. 이러한 한계를 해결하기 위해, 한 연구에서는 인과 스태킹(causal stacking)과 같은 앙상블 방법을 적용하여 CATE 추정의 일관성을 향상시키고자 하였다. 이러한 맥락에서, 우리는 다양한 성능평가 지표에서 우수한 결과를 얻기 위해 Super Learner를 활용한 CATE 추정방법을 제안했다. Super Learner는 데이터를 교차 검증(cross-validation)을 사용하여 분할하는 장점이 있어 과적합(overfitting)을 방지하고, 스태킹이나 다른 개별 추정방법들에 비해 최적의 결과를 도출할 수 있다.

시뮬레이션 결과, 본 논문에서 제안한 방법은 다른 방법들에 비해 MSE가 더 낮았으며 다양한 성능 지표에서 우수한 성능을 보였다. 이러한 결과는 치료 결정을 위해 단일 CATE 추정 방법에 의존하기보다는 여러 방법의 결과를 결합하기 위해 Super Learner를 활용하는 것이 더 견고하고 신뢰할 수 있는 환자 치료 최적화 프레임워크를 제공함을 시사한다. 결과적으로 Super Learner 접근법은 개별화된 치료 규칙을 개발하는 데 있어 실질적이고 효과적인 도구가 될 수 있고 환자 치료 최적화를 위한 상당한 가능성을 제공할 것이라 예상한다.

핵심되는 말 : 인과추론, 이질적 처치효과, 조건부 평균 처치효과, 개별 치료 규칙, 플러그인 추정량