Trustworthy assessment of heterogeneous treatment effect estimator via analysis of relative error

Anonymous Author
Anonymous Institution

Abstract

Accurate heterogeneous treatment effect (HTE) estimation is essential for personalized recommendations, making it important to evaluate and compare HTE estimators. Traditional assessment methods are inapplicable due to missing counterfactuals. Current HTE evaluation methods rely on additional estimation or matching on test data, often ignoring the uncertainty introduced and potentially leading to incorrect conclusions. We propose incorporating uncertainty quantification into HTE estimator comparisons. In addition, we suggest shifting the focus to the estimation and inference of the relative error between methods rather than their absolute errors. Methodology-wise, we develop a relative error estimator based on the efficient influence function and establish its asymptotic distribution for inference. Compared to absolute error-based methods, the relative error estimator (1) is less sensitive to the error of nuisance function estimators, satisfying a "global double robustness" property, and (2) its confidence intervals are often narrower, making it more powerful for determining the more accurate HTE estimator. Through extensive empirical study of the ACIC challenge benchmark datasets, we show that the relative error-based method more effectively identifies the better HTE estimator with statistical confidence, even with a moderately large test dataset or inaccurate nuisance estimators.

Preliminary work. Under review by AISTATS 2025. Do not distribute.

1 INTRODUCTION

The estimation of heterogeneous treatment effects (HTE) under the Neyman-Rubin potential outcome framework is becoming increasingly prominent, driven by the need for tailored approaches in areas such as personalized medicine, personalized education, and personalized advertising (Bennett and Lanning 2007; Lesko 2007; Low, Gallego, and Shah 2016; Murphy, Redding, and Twyman 2016; Splawa-Neyman, Dabrowska, and Speed 1990).

A variety of machine learning tools are being employed to estimate HTE, including LASSO, random forests, gradient boosting, and neural networks (see Künzel et al. 2019 for a review). Despite a rich body of research on HTE estimation, evaluating and comparing these estimators remain less investigated. Evaluating the performance of HTE estimators is essential for identifying a better candidate, especially considering the wide range of HTE estimation methods available.

One significant challenge in assessing an HTE estimator arises from the inherent missingness in potential outcome model. Provided with a test dataset, standard evaluation of a predictor compares the actual observations with the predicted values. However, in the potential outcome model, each observed response corresponds to one potential outcome (treatment or control), and the HTE—the difference between two potential outcomes—is not directly observed.

To deal with the missingness of HTE, existing methods of HTE assessment typically involves additional steps performed on the test dataset, such as matching (Rolling and Yang 2014) or nuisance function estimation (Alaa and Van Der Schaar 2019), to construct pseudo-observations of the HTE (Section 2.3). The additional steps could introduce substantial randomness, which may even dominate the actual error difference between two HTE estimators. Simply ignoring this source of randomness and outputting the estimator with the lower point error estimate could result in incorrect decisions with a significant probability.

In this paper, we advocate that the comparison of HTE estimators should account for the randomness introduced during the evaluation stage. Rather than providing a point estimate of the error, we suggest constructing a confidence interval for the evaluation error, which contains the true error value with a pre-specified probability. We then demonstrate that the confidence interval for the absolute error of an HTE estimator (1) can be sensitive to nuisance function estimation on the test data; (2) for two similar HTE estimators, their absolute error confidence intervals do not account for the similarity of the HTE estimators and could be too wide to determine the more accurate estimator. To address the issues of uncertainty quantification for absolute errors, we propose to directly construct confidence intervals for the *relative* error between two HTE estimators, rather than their individual absolute errors. Methodologically, we derive an efficient estimator for the relative error using influence functions and characterize its asymptotic distribution to facilitate the confidence interval construction. Theoretically, we prove that our confidence interval of relative errors is valid under weaker assumptions regarding the quality of the nuisance function estimators compared to that of absolute errors, and is guaranteed to be narrow when the HTE estimators for comparison are similar. Empirically, we show that the relative error confidence intervals achieves better coverage as well as are more powerful in identifying the better HTE estimator. Beyond the difference in conditional means, our proposal can also be generalized to compare estimators of a broader class of heterogeneous treatment effect estimands more suitable for quantifying treatment effects for non-continuous outcomes.

Contributions.

1. For comparing HTE estimators, we propose to account for the uncertainty in estimating the error of HTE estimators, which is largely ignored in the literature. Taking the uncertainty into consideration yields more trustworthy conclusion about the quality comparison of HTE estimators.

2. We suggest constructing confidence intervals of the relative error between two HTE estimators rather than their absolute errors. We propose a one-step correction estimator and the associated confidence interval for the relative error based on the efficient influence function, and prove the asymptotic validity and optimality of the proposed confidence interval. We prove the relative error estimator is less sensitive to nuisance function estimators, enjoys a global doubly robustness property, and is useful in selecting the better HTE estimator.

3. We provide off-the-shelf implementations of our relative error-based confidence intervals for HTE estimator comparison[1].

Organization. In Section 2, we present the problem formulation and provide a review of the relevant literature. In Section 3, we provide the efficient estimator, confidence interval of absolute errors, and discuss the issues therein. In Section 4, we focus on relative errors, presenting the efficient estimator and its associated confidence interval, and explain how the issues with absolute error confidence intervals are avoided. We also discuss the generalization of our proposal based on relative error to a broader class of causal estimands. In Section 5, we compare the confidence intervals for absolute and relative errors on a benchmark dataset from the 2016 ACIC challenge. In Section 6, we include future research directions. All proofs are provided in the supplementary materials.

2 Formulation and background

2.1 Potential outcome model

We follow the Neyman-Rubin potential outcome model (Rubin 1974) with a treatment condition and a control condition. For unit i, there is a d-dimensional covariate vector X_i, a binary treatment assignment $W_i \in \{0, 1\}$, and two potential outcomes: $Y_i(0)$ under the control condition and $Y_i(1)$ under the treatment condition. We denote the observed outcome by Y_i, which equals $Y_i(1)$ if the unit is under treatment, and $Y_i(0)$, otherwise. We use $Z_i = (X_i, W_i, Y_i)$ to denote all data of unit i. We make the conventional assumptions of SUTVA, overlap, and unconfoundedness.

To define our causal estimand, we introduce the super-population. We assume individuals are sampled i.i.d. from a super-population, denoted by $\mathbb{P}$. In particular, the covariates X_i are sampled from an unknown distribution $\mathbb{P}_X$. Given the covariates X_i, a binary group assignment $W_i \in \{0, 1\}$ is generated from a Bernoulli distribution with mean $e(X_i)$ (also known as the propensity score). We assume there exists $\eta > 0$ such that $\eta < e(x) < 1 - \eta$. The potential outcomes are modeled by

$$Y_i(1)|X_i = \mu_1(X_i) + \epsilon_i,$$
$$Y_i(0)|X_i = \mu_0(X_i) + \epsilon_i,$$

where $\mu_0(x)$, $\mu_1(x)$ represent the conditional mean function of the control and the treatment group respectively, and ϵ_i denotes the error term, assumed to

[1] All code has been deposited on GitHub, but access is restricted to maintain anonymity during the review process. Once the paper is accepted, we will release the address of the repository.

be zero-mean and independent of X_i. The estimand HTE is defined as

$$\tau(x) = \mathbb{E}\left[Y(1) - Y(0) \mid X = x\right] = \mu_1(x) - \mu_0(x),$$

which can also be expressed as the difference of the two group conditional mean functions.

2.2 Evaluation task and absolute/relative error

In this paper, we focus on the evaluation and comparison of the HTE estimators using a test dataset of size n drawn from the super-population $\mathbb{P}$. We use $\hat{\tau}_1(x)$, $\hat{\tau}_2(x)$ to denote HTE estimators derived independently of the test dataset. When there is only one HTE estimator, we drop the subscript and use $\hat{\tau}(x)$. We highlight that the HTE estimators are provided to us and the HTE estimation problem itself is not the focus of this paper.

To quantify the accuracy of an HTE estimator $\hat{\tau}(x)$, the absolute error is defined as

$$\phi(\hat{\tau}(x)) := \mathbb{E}\left[(\hat{\tau}(X) - \tau(X))^2\right], \tag{1}$$

where the expectation is evaluated at the distribution $\mathbb{P}_X$ of the covariates. A smaller absolute error suggests a more accurate HTE estimator, and a zero error implies $\hat{\tau}(X) = \tau(X)$ with probability one. The relative error of two estimators $\hat{\tau}_1(x)$ and $\hat{\tau}_2(x)$ is quantified as the difference between their absolute errors

$$\begin{aligned} &\delta(\hat{\tau}_1, \hat{\tau}_2) := \phi(\hat{\tau}_1(x)) - \phi(\hat{\tau}_2(x)) \\ =&\mathbb{E}\left[\hat{\tau}_1^2(X) - \hat{\tau}_2^2(X) - 2(\hat{\tau}_1(X) - \hat{\tau}_2(X))\tau(X)\right]. \end{aligned} \tag{2}$$

A negative $\delta(\hat{\tau}_1, \hat{\tau}_2)$ indicates that $\hat{\tau}_1(x)$ is more accurate; otherwise, $\hat{\tau}_2(x)$ is more accurate. In Section 4.3, we consider the treatment effects defined on the natural parameter scale suitable for binary, count, and survival responses, and extend the results regarding Eq. (1) to the corresponding errors.

For standard predictor evaluation, absolute prediction errors are more commonly displayed than relative prediction errors. However, for HTE estimators, we have the perhaps surprising observation that the relative error can often be approximated more accurately. Intuitively, this is because the relative error $\delta(\hat{\tau}_1, \hat{\tau}_2)$ is linear in the unobserved $\tau(x)$, while the absolute error $\phi(\hat{\tau}(x))$ also depends on the second moment of $\tau(x)$, and estimating the first moment of $\tau(x)$ is relatively easier that of the second moment. We provide rigorous characterizations and empirical comparison of the observation in the following sections.

2.3 Literature

In this paper, we consider comparing HTE estimators with statistical confidence through relative error evaluation, which is different from the mainstream literature focusing on evaluating absolute errors without uncertainty quantification. Nevertheless, we provide a brief overview of existing methods on assessing the absolute performance of an HTE estimator. One simple and common approach targets the observed response but not the treatment effect, and uses the standard prediction error of the response as the error measurement. However, an accurate predictor of the treatment effect may not necessarily be associated with precise response predictors (Curth and Van Der Schaar 2023). Furthermore, for HTE estimators that directly estimate the difference without estimating the response, such as causal trees (Athey and Imbens 2016), the response prediction error can not be computed. Another thread of assessment approaches involves creating "virtual twins" by matching treated and control units and use the response difference of a pair as a pseudo-observation of the treatment effect (Rolling and Yang 2014). However, matching is often computationally intensive (Rosenbaum 1989). Moreover, the complexity of matching algorithms makes it difficult to analyze statistically and perform downstream inference. A third thread of methods estimate the HTE on the test dataset and compare it with the provided HTE estimators, which we refer to as plug-in estimators. This can be problematic because the evaluation error is affected by the error from the HTE estimator obtained from the test set, a nuisance function in this causal assessment task. To reduce the impact of the error of the HTE estimator from the test data, bias correction methods based on influence functions have been developed (Alaa and Van Der Schaar 2019). The method is related to ours, but it does not address the uncertainty quantification of the estimated error, and their influence function is different from our efficient proposal (comparison in Section 5).

Our approach for drawing inference on relative error evaluation builds on the rapidly advancing body of work in semi-parametrics, particularly influence functions (Robins et al. 2008; Vaart 2000) (see Kennedy 2022 for a review). In many causal problems, the causal quantity of interest is typically a scalar or low-dimensional, but the model contains infinite dimensional nuisance functions, making it a semi-parametric problem. The influence function is a powerful tool for constructing the so-called one-step correction estimators that is more robust to the error of the estimators of nuisance components. Given a specific function class that the true distribution belongs to, the estimator that attains the minimal asymptotic variance is known as the semi-parametrically efficient estimator, and the corresponding influence function is referred to as the efficient influence function. Despite the appealing statistical properties of efficient influence func-

tions, the derivation of efficient influence functions is often case-specific, with only a few general rules and standard examples available (Kennedy 2022).

3 INFERENCE OF ABSOLUTE ERROR AND ISSUES

3.1 Absolute error inference via influence functions

Provided with an HTE estimator $\hat{\tau}(x)$, we adopt the following influence function for $\phi(\hat{\tau})$,

$$\begin{aligned}\psi(\phi(\hat{\tau}); Z) &:= ((\mu_1(X) - \mu_0(X)) - \hat{\tau}(X))^2 \\ &+ 2\left((\mu_1(X) - \mu_0(X)) - \hat{\tau}(X)\right) \\ &\cdot \left(\frac{W(Y - \mu_1(X))}{e(X)} - \frac{(1-W)(Y - \mu_0(X))}{1 - e(X)}\right) - \phi(\hat{\tau}).\end{aligned} \tag{3}$$

Our derivation of the efficient influence function aligns with the variable importance measurement for heterogeneous treatment effects (Hines, Diaz-Ordaz, and Vansteelandt 2022). The one-step correction estimator associated with (3) is

$$\begin{aligned}\hat{\phi}(\hat{\tau}) &= \frac{1}{n}\sum_{i=1}^{n} ((\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)) - \hat{\tau}(X_i))^2 \\ &+ 2\left((\tilde{\mu}_1(X_i) - \tilde{\mu}_0(X_i)) - \hat{\tau}(X_i)\right) \\ &\cdot \left(\frac{W_i(Y_i - \tilde{\mu}_1(X_i))}{\tilde{e}(X_i)} - \frac{(1-W_i)(Y_i - \tilde{\mu}_0(X_i))}{1 - \tilde{e}(X_i)}\right).\end{aligned} \tag{4}$$

Here $\tilde{\mu}_0(x)$, $\tilde{\mu}_1(x)$, and $\tilde{e}(x)$ are estimators of the nuisance functions $\mu_0(x)$, $\mu_1(x)$, and $e(x)$ obtained on the test dataset[2]. The variance of the estimated evaluation error, denoted by $V(\hat{\phi}(\hat{\tau}))$, can be approximated by the empirical variance of $\hat{\phi}_i(\hat{\tau})$, denoted by $\hat{V}(\hat{\phi}(\hat{\tau}))$. The entire algorithm is summarized in Algorithm 1.

Theorem 1. *Assume the following conditions.*

(a) Y is bounded, $\eta < e(X) < 1 - \eta$ for some $\eta > 0$.

(b) The nuisance function estimators $\tilde{\mu}_0(x)$, $\tilde{\mu}_1(x)$, $\tilde{e}(x)$ obtained from the test data[3] satisfy $\mathbb{E}[(\tilde{\mu}_1(X) - \mu_1(X))^2]^{1/2}$, $\mathbb{E}[(\tilde{\mu}_0(X) - \mu_0(X))^2]^{1/2}$, $\mathbb{E}[(\hat{e}(X) - e(X))^2]^{1/2} = o_p(n^{-1/4})$, $1 \le k \le K$.

(c) The true absolute error $\phi(\hat{\tau}) > 0$.

Then $\hat{\phi}(\hat{\tau})$, $\hat{V}(\hat{\phi}(\hat{\tau}))$ of Algorithm 1 satisfy

$$\frac{\hat{\phi}(\hat{\tau}) - \phi(\hat{\tau})}{\sqrt{n\hat{V}(\hat{\phi}(\hat{\tau}))}} \xrightarrow{d} \mathcal{N}(0,1).$$

In addition, the estimator $\hat{\phi}(\hat{\tau})$ is semi-parametrically efficient regarding the nonparametric model.

The proof is provided in the appendix. Our influence function and the estimator is different from that of Alaa and Van Der Schaar 2019, Theorem 2. Since our proposal $\hat{\phi}(\hat{\tau})$ is semi-parametrically efficient, the confidence interval is proved to be no wider than that in Alaa and Van Der Schaar 2019 (see Section 5 for numerical evidence). According to Theorem 1, the convergence of $\hat{\phi}(\hat{\tau})$ at the parametric rate of $n^{-1/2}$ only requires all nuisance function estimators to converge at a rate no slower than $n^{-1/4}$, that is the locally doubly robust property (Chernozhukov et al. 2018).

Based on the $1-\alpha$ confidence interval[4] of the absolute errors for two HTE estimators $\hat{\tau}_1$, $\hat{\tau}_2$, if the absolute error interval of $\hat{\tau}_1$ lies entirely to the right of that of $\hat{\tau}_2$, we can conclude with at least $1 - 2\alpha$ confidence that the estimation error of $\hat{\tau}_1$ is greater than that of $\hat{\tau}_2$, and therefore $\hat{\tau}_2$ should be selected. If the two intervals overlap, we are unable to confidently decide which estimator is more accurate.

3.2 Issues of absolute error

(i) Sensitivity to errors of nuisance function estimators. In Figure 1, errors in $\tilde{\mu}_1(x)$, $\tilde{\mu}_0(x)$, $\tilde{e}(x)$ can introduce significant bias into the absolute error estimator, even resulting in negative estimates conflicting with the fact that error should always be non-negative.

(ii) Correlation across the estimated absolute errors of different HTE estimators. The estimated absolute error of different HTE estimators are correlated because they are based on the same validation data and share the nuisance function estimators. Theorem 1 does not directly address the correlation between the estimated absolute errors.

[2]We use cross-fitting (Chernozhukov et al. 2018) to ensure the independence of $\tilde{\mu}_0(x)$, $\tilde{\mu}_1(x)$, and $\tilde{e}(x)$ used by $\hat{\phi}_i(\hat{\tau})$ are independent of Y_i, W_i, X_i therein when computing $\hat{\phi}(\hat{\tau})$.

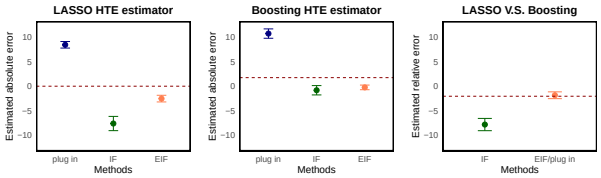
[3]When cross-fitting is used to compute the absolute error, we require (b) to hold for all $\tilde{\mu}_0^{-k}(x)$, $\tilde{\mu}_1^{-k}(x)$, $\tilde{e}^{-k}(x)$, $1 \le k \le K$.

[4]The $1-\alpha$ confidence interval of the absolute error takes the form

$$\begin{aligned}&[\underline{\hat{\phi}}(\hat{\tau}; 1-\alpha),\ \overline{\hat{\phi}}(\hat{\tau}; 1-\alpha)] \\ &:= \Big[\hat{\phi}(\hat{\tau}) - q_{1-\alpha/2}\hat{V}^{1/2}(\hat{\phi}(\hat{\tau})), \\ &\quad \hat{\phi}(\hat{\tau}) + q_{1-\alpha/2}\hat{V}^{1/2}(\hat{\phi}(\hat{\tau}))\Big],\end{aligned} \tag{5}$$

where $q_{1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of a standard normal random variable.

(iii) Degenerate null. Condition (c) of Theorem 1 indicates the asymptotic distribution may be invalid for the degenerate case $\mathbb{P}_X(\hat{\tau}(X) = \tau(X)) = 1$. One solution is analyzing the asymptotic distribution of the higher-order pathwise derivatives of $\phi(\hat{\tau})$ to derive the asymptotic distribution of (4) when $\tau(x) = \hat{\tau}(x)$, which remains generally an open problem (Hines, Diaz-Ordaz, and Vansteelandt 2022; Hudson 2023).



Absolute error Relative error

Figure 1: Comparison of estimated absolute and relative errors with inaccurate nuisance function estimators. We compare two HTE estimators: LASSO and Boosting. We implement three approaches to evaluate the absolute and the relative errors: (1) the plug-in method, (2) the estimator by Alaa and Van Der Schaar 2019 (IF), and (3) our proposal (EIF). For each evaluation, we plot the estimated error as well as the 90% confidence interval. The IF and EIF estimators for absolute error are negative, which conflicts with the non-negative nature of prediction errors. Only the confidence interval of the EIF relative error correctly captures the true value (third panel, orange).

4 INFERENCE OF RELATIVE ERROR

4.1 Relative error estimation via influence functions

Provided with two HTE estimators $\hat{\tau}_1(x)$, $\hat{\tau}_2(x)$, we adopt the following influence function for the relative error $\delta(\hat{\tau}_1, \hat{\tau}_2)$,

$$\begin{aligned}&\psi(\delta(\hat{\tau}_1,\hat{\tau}_2);Z) := \hat{\tau}_1^2(X) - \hat{\tau}_2^2(X)\\ &- 2\left(\hat{\tau}_1(X) - \hat{\tau}_2(X)\right)\cdot\left(\frac{W(Y-\mu_1(X))}{e(X)} + \mu_1(X)\right.\\ &\left.-\frac{(1-W)(Y-\mu_0(X))}{1-e(X)} - \mu_0(X)\right) - \delta(\hat{\tau}_1,\hat{\tau}_2).\end{aligned} \tag{6}$$

The implied one-step correction estimator is

$$\begin{aligned}&\hat{\delta}(\hat{\tau}_1,\hat{\tau}_2) = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_1^2(X_i) - \hat{\tau}_2^2(X_i)\\ &- 2\left(\hat{\tau}_1(X_i) - \hat{\tau}_2(X_i)\right)\cdot\left(\frac{W_i(Y_i-\tilde{\mu}_1(X_i))}{\tilde{e}(X_i)} + \tilde{\mu}_1(X_i)\right.\\ &\left.-\frac{(1-W_i)(Y_i-\tilde{\mu}_0(X_i))}{1-\tilde{e}(X_i)} - \tilde{\mu}_0(X_i)\right).\end{aligned} \tag{7}$$

The algorithm can be summarized as Algorithm 1 combined with (7).

Based on the $1-\alpha$ confidence interval of the relative error, if the interval lies entirely to the right of zero, we can conclude with at least $1-\alpha$ confidence that the estimation error of $\hat{\tau}_1$ is greater than that of $\hat{\tau}_2$, and therefore $\hat{\tau}_2$ should be selected. If the interval lies entirely to the left of zero, we can conclude with the same confidence that the estimation error of $\hat{\tau}_2$ is greater than that of $\hat{\tau}_1$, and $\hat{\tau}_1$ should be chosen. If the interval contains zero, we are unable to confidently decide which estimator is superior.

Algorithm 1 Absolute (relative) error

1: **Input**: An HTE estimator $\hat{\tau}(x)$, test data $Z_i = (X_i, W_i, Y_i)$, $1 \le i \le n$, methods of estimating nuisance functions $\mu_0(x)$, $\mu_1(x)$, $e(x)$, number of folds K for cross-fitting, confidence level $1-\alpha$.
2: Randomly split the test dataset into K folds of approximately equal size. Denote the k-th fold by D_k.
3: **for** $k = 1, \ldots, K$ **do**
4: Apply the nuisance function estimators to test folds $\cup_{j\neq k} D_j$ and obtain $\tilde{\mu}_0^{-k}(x)$, $\tilde{\mu}_1^{-k}(x)$, and $\tilde{e}^{-k}(x)$.
5: **end for**
6: Compute the one-step correction estimator based on Equation (4) (Equation (7)).
7: Compute the estimator of the variance $\hat{V}(\hat{\phi}(\hat{\tau}))$ $(\hat{V}(\hat{\delta}(\hat{\tau}_1,\hat{\tau}_2)))$.
8: **Output**: Estimated error $\hat{\phi}(\hat{\tau})$, the estimator of its variance $\hat{V}(\hat{\phi}(\hat{\tau}))$, $1-\alpha$ confidence interval $\left[\hat{\phi}(\hat{\tau}) - q_{1-\alpha/2}\hat{V}^{1/2}(\hat{\phi}(\hat{\tau})),\ \hat{\phi}(\hat{\tau}) + q_{1-\alpha/2}\hat{V}^{1/2}(\hat{\phi}(\hat{\tau}))\right]$.

Theorem 2. *Assume the following conditions.*

(a) Y is bounded, $\eta < e(X) < 1-\eta$ for some $\eta > 0$.

(b) The nuisance function estimators obtained from the test data satisfy $\|\tilde{\mu}_1(X) - \mu_1(X)\|_2$, $\|\tilde{\mu}_0(X) - \mu_0(X)\|_2$, $\|\hat{e}(X) - e(X)\|_2 = o_p(1)$, and $\mathbb{E}[|(\tilde{\mu}_1(X) - \mu_1(X))(\hat{e}(X) - e(X))|]$, $\mathbb{E}[|(\tilde{\mu}_0(X) - \mu_0(X))(\hat{e}(X) - e(X))|] = o_p(n^{-1/2})$.

(c) The true relative error $\mathbb{E}[(\hat{\tau}_1(X) - \hat{\tau}_2(X))^2] \neq 0$.

Then $\hat{\delta}(\hat{\tau}_1,\hat{\tau}_2)$, $\hat{V}(\hat{\delta}(\hat{\tau}_1,\hat{\tau}_2))$ of Algorithm 1 satisfy

$$\frac{\hat{\delta}(\hat{\tau}_1,\hat{\tau}_2) - \delta(\hat{\tau}_1,\hat{\tau}_2)}{\sqrt{n\hat{V}(\hat{\delta}(\hat{\tau}_1,\hat{\tau}_2))}} \xrightarrow{d} \mathcal{N}(0,1).$$

The estimator $\hat{\delta}(\hat{\tau}_1,\hat{\tau}_2)$ is efficient regarding the nonparametric model.

Further assume

(d) $\hat{e}(x) = e(x)$ *or* $\tilde{\mu}_1(x) = \mu_1(x)$, $\tilde{\mu}_0(x) = \mu_0(x)$.

Then the estimator $\hat{\delta}(\hat{\tau}_1, \hat{\tau}_2)$ *is unbiased, i.e.,* $\mathbb{E}\left[\hat{\delta}(\hat{\tau}_1, \hat{\tau}_2)\right] = \delta(\hat{\tau}_1, \hat{\tau}_2)$.

4.2 Advantages of relative error

(i) Condition (b) of Theorem 2 only requires the nuisance function estimators to be consistent and the product of the error of $\hat{e}(x)$ and $\tilde{\mu}_1(x)$, $\tilde{\mu}_0(x)$ to converge at $n^{-1/2}$. This condition of relative error is strictly weaker than the condition (b) of the absolute error in theorem 1. Particularly, when the treatment assignment is known, such as in randomized trials, then the validity of Theorem 2 only requires the consistency of $\tilde{\mu}_1(x)$, $\tilde{\mu}_0(x)$.

In addition, condition (d) implies that the relative error estimator satisfies a global doubly robust property, meaning that if $\hat{e}(x) = e(x)$, then $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ can be arbitrary, even inconsistent, and the estimated relative error will still be unbiased. Similarly, the relative error remains unbiased if $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ are correct, regardless of the estimator $\hat{e}(x)$ used. This property does not hold for the absolute error estimator.

(ii) Relative error estimators directly estimates and performs inference for the relative difference between two models' performance, rather than dealing with each model's error separately and then comparing them. Therefore, unlike the absolute error estimation approach, there is no need to handle the dependency between two absolute error confidence intervals.

(iii) Degenerate null. Condition (c) of Theorem 1 $\hat{\tau}(x) \neq \tau(x)$ is hard to validate or falsify, since $\tau(x)$ is not directly observed and needs to be estimated. In contrast, condition (c) of Theorem 2 $\hat{\tau}_1(x) \neq \hat{\tau}_2(x)$ can be verified straightfowardly, since $\hat{\tau}_1(x)$ and $\hat{\tau}_2(x)$ are provided functions and there is no extra estimation required.

(iv) The relative error is more effective in identifying the better HTE estimators compared to the absolute error when $\hat{\tau}_1$, $\hat{\tau}_2$ are similar (see Figure 2 for a numerical example). When $\hat{\tau}_1(X)$ and $\hat{\tau}_2(X)$ are similar, the confidence intervals for the absolute errors of $\hat{\tau}_1(X)$ and $\hat{\tau}_2(X)$ can still be wide, and the absolute error confidence intervals are too wide to be informative of which estimator is better. In contrast, the width of the confidence interval for $\hat{\delta}(\hat{\tau}_1, \hat{\tau}_2)$ is proportional to $\sqrt{\mathbb{E}[(\hat{\tau}_1(X) - \hat{\tau}_2(X))^2]}$, which we prove in the appendix, and the confidence interval of the relative error can still be sufficiently narrow to reliably identify the more accurate estimator.

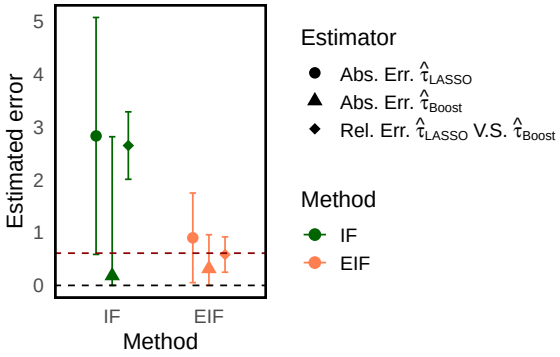


Figure 2: Relative error evaluation estimators are more effective in determining the better one of two similar HTE estimators. We consider two similar HTE estimators using LASSO for nuisance function estimation which only differ in the regularization hyperparameter. For evaluation methods, see the caption of Figure 1. The 90% absolute error confidence intervals of both IF and EIF are too wide to distinguish the two HTE estimators, while the confidence interval of the relative error estimator is significantly narrower and find the better estimator $\hat{\tau}_2$ (the true relative error is indicated in dark red). IF's relative error confidence interval does not contain the true relative error.

4.3 Beyond difference in conditional means

For responses generated from an exponential family or the Cox model, such as when $Y \mid X$ follows a Poisson distribution, the difference in the natural parameter functions between the treatment and the control has been proposed as the treatment effect estimand (Gao and Hastie 2022). The difference in natural parameter functions is better suited for modeling covariate dependence, as transforming to the natural parameter scale eliminates support constraints. Let $\eta(\cdot)$ be the link function transforming the conditional mean to the natural parameter scale, the estimand is

$$\tau(x) := \eta(E[Y(1) \mid X = x]) - \eta(E[Y(0) \mid X = x]).$$

In the appendix, we similarly define the absolute error (1), the relative error (2), and derive the efficient one-step correction estimators. The estimators recover (4) and (7) when the link function $\eta(\cdot)$ is identity. In the appendix, we also present analogous results to Theorem 1 and Theorem 2. The relative error of DINA maintains the advantages in Section 4.2.

5 HYPOTHETICAL REAL DATA ANALYSIS

The ACIC 2016 competition dataset (Dorie et al. 2019), derived from the real-world Collaborative Peri-

natal Project (Niswander, Gordon, and Gordon 1972), is used as a benchmark dataset for HTE estimation and related tasks. The dataset includes 55 real variables of various types with natural associations. Treatment assignments and potential outcomes are generated using a variety of models. Therefore, the true $\mu_0(x)$, $\mu_1(x)$ and thus true $\tau(x)$ are known. We select four representative scenarios varying two key factors: the model of the propensity score and the model of the group conditional functions. Specifically, the scenarios include: (a) linear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$; (b) nonlinear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$; (c) linear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$; (d) nonlinear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$. Here, linear functions refer to a linear combination of covariates x, while nonlinear functions incorporate quadratic or higher-order terms of the covariates. For both linear and non-linear functions, the nuisance functions only rely on less than 20% of the covariates. The total sample size is 4802. We use a randomly sampled 2802 data points to obtain the HTE estimators for comparison, and the rest 2000 data points to evaluate and compare the estimators' performance.

For the HTE estimators for comparison, We consider two T-learners (Künzel et al. 2019). The estimators first estimate $\mu_0(x)$, $\mu_1(x)$ separately and then take the difference $\hat{\mu}_1(x) - \hat{\mu}_0(x)$. The key difference between two HTE estimators lies in the machine learning algorithms used to obtain $\hat{\mu}_0(x)$, $\hat{\mu}_1(x)$: (i) LASSO. This method uses LASSO[5] to estimate $\mu_0(x)$ and $\mu_1(x)$. Preferred in scenarios (a) and (c), (ii) Boosting. This method uses gradient boosting[6] to estimate $\mu_0(x)$ and $\mu_1(x)$. Preferred in scenarios (b) and (d).

We implement five assessment methods: three based on absolute error estimation and two based on relative error estimation. For absolute error estimators, we consider the one-step correction estimator based on our efficient influence function (EIF), as defined in (4), the one-step correction estimator by Alaa and Van Der Schaar 2019 (IF), and the plug-in estimator (plug in): $\sum_{i=1}^{n}(\hat{\tau}(X_i) - \tilde{\tau}(X_i))^2/n$, where $\tilde{\tau}(x)$ is the AIPW estimator of the HTE $\tau(x)$ obtained from the test dataset. For relative error estimators, we consider the one-step correction estimator based on our efficient influence function (EIF) in (7)[7], and the one-step correction estimator of the relative error based on the influence function in Alaa and Van Der Schaar 2019. We use 2-fold cross-fitting. All five methods share the same nuisance function estimators for fair comparison, which we specify below.

To compute the coverage and average width of the confidence intervals (90% confidence level), and the probability of selecting the correct model (selection accuracy), for scenarios (a) through (d), we generate 100 different realizations of $\mu_0(x)$, $\mu_1(x)$, and $e(x)$, adhering to the linear/nonlinear constraints aforementioned. For each realization of $\mu_0(x)$, $\mu_1(x)$, and $e(x)$, we further simulate Y and Z independently 100 times to compute the coverage. Similarly for other metrics above.

5.1 Confidence interval coverage

In Figure 3, we examine the simplest scenario (a), where both the propensity score function $e(x)$ and the outcome functions $\mu_0(x)$ and $\mu_1(x)$ are linear. We compare the five evaluation methods equipped with three nuisance functions: (i) the true nuisance functions, (ii) a well-specified linear model, and (iii) an erroneous gradient boosting learner.

For relative errors, the EIF confidence interval is consistently valid, robust to the nuisance function estimator provided. In contrast, the coverage of the relative error IF confidence interval fails to achieve the correct coverage. For absolute errors, the coverage depends heavily on the accuracy of the nuisance learners. When the true nuisance estimators are used, the coverage of the EIF method for both the LASSO HTE estimator and the Boosting HTE estimator is close to the target level. However, as we shift to the nuisance estimator obtained using linear regression (which is well-specified but not as accurate), the coverage decreases, and it almost never covers the true value when the inaccurate gradient boosting with underfitting is used. The absolute error estimator IF consistently undercovers, even with true nuisance functions. Additionally, as shown in the appendix, the width of the confidence intervals based on IF, for both relative and absolute error estimators, is significantly larger than those based on EIF, indicating less precision.

5.2 Probability of selecting the winner

We present the probability of selecting the correct model in Figure 4. The analysis covers settings (a) through (d), where in (a) and (c), the LASSO HTE estimator outperforms the Boosting HTE estimator by a moderate margin, whereas in (b) and (d), the Boosting HTE estimator significantly outperforms the LASSO HTE estimator. For absolute error confidence intervals, we select the better estimator if the two confidence intervals at the $\alpha/2$ level do not overlap. Otherwise, no selection is made. In contrast, for relative

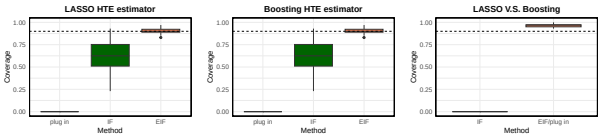
[5]We use the R package GLMNET for implementation. In particular, we use CV.GLMNET to choose the regularization parameter.

[6]We use XGBOOST for implementation.

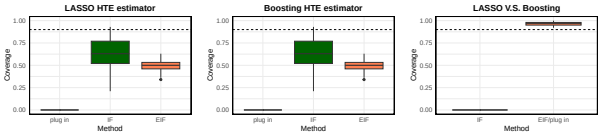
[7]We note that the relative error estimator implied by the plug-in estimator of the absolute error is equivalent to EIF relative error estimator. Therefore, we merge the two methods in display.

error confidence intervals, we select the better estimator if the confidence interval for the relative error does not contain zero.

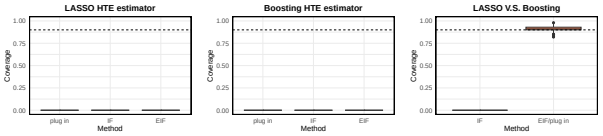
For the relative error-based methods, we observe that across the four scenarios, the EIF relative error method achieves the highest selection accuracy, followed by the IF relative error method. For the absolute error-based methods, the selection accuracy of the EIF absolute error method is lower in settings (a) and (c). This is because the improvement of the LASSO HTE estimator over the Boosting counterpart is relatively modest, and the gap between them is overshadowed by the wide confidence intervals, resulting in the methods being unable to confidently make a selection. In settings (b) and (d), the gap between the two HTE estimators is more pronounced, leading to a higher probability of correct selection. for the plug-in and IF absolute error methods, the confidence intervals are too wide, preventing confident conclusions made across scenarios.



(i) True



(ii) Linear regression

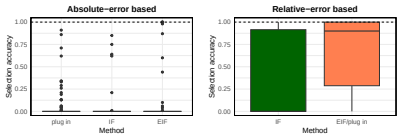


(iii) Gradient boosting with underfitting

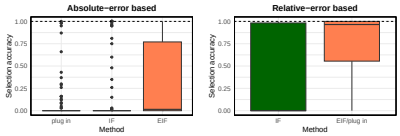
Figure 3: Coverage of the estimated absolute (LASSO, Boosting)/relative (LASSO V.S. Boosting) error's 90% confidence intervals across three methods (plug in, IF, EIF) over three nuisance functions ((i) true, (ii) estimated by linear regression, (iii) estimated by gradient boosting with underfitting).

6 DISCUSSIONS

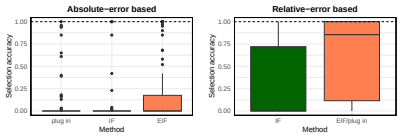
We advocate that the comparison of HTE estimators should account for the randomness incurred in the test stage to determine the more accurate estimator with statistical confidence. We propose to achieve this goal by constructing confidence intervals for the relative error between two HTE estimators rather than their absolute errors. Explicitly, we derive a one-step correction estimator based on the efficient influence function, and provide the asymptotically narrowest confidence interval of the relative error. The relative error confidence interval is less sensitive to nuisance function



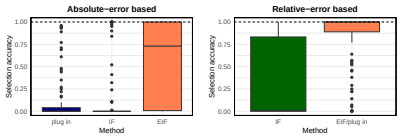
(a) Linear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$



(b) Nonlinear $\mu_0(x)$, $\mu_1(x)$, linear $e(x)$



(c) Linear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$



(d) Nonlinear $\mu_0(x)$, $\mu_1(x)$, nonlinear $e(x)$

Figure 4: The probability of correctly selecting the better HTE estimator by comparing absolute/relative error's confidence intervals across three methods (plug in, IF, EIF) over four scenarios ((a) to (d)).

estimation errors and is more powerful in identifying the better HTE estimator when the candidate estimators are similar.

We discuss potential directions for future research.

- An intriguing direction for future exploration is how the relative error evaluation method can be integrated into the training process to produce a more accurate HTE estimator. Specifically, one potential approach is to use cross-validation in the HTE estimation, where the validation step is carried out using our proposed evaluation method.
- Note that some HTE estimator could be overall accurate but considerably less precise in underrepresented groups due to insufficient training data. This presents a fairness concern, especially since HTE estimators could influence downstream policy decisions. To address this, it would be beneficial to extend the error averaged over the entire population to that averaged over certain subgroups.
- In this paper, we bypass the degeneracy issue of the asymptotic distribution of the one-step correction estimator by shifting attention from the absolute estimand (absolute error) to its relative counterpart (relative error). It is of interest whether this approach could be generalized to address the degeneracy issue in other scenarios.

References

Alaa, Ahmed and Mihaela Van Der Schaar (2019). "Validating causal inference models via influence functions". In: *International Conference on Machine Learning*. PMLR, pp. 191–201.

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.

Bennett, James and Stan Lanning (2007). "The netflix prize". In: *Proceedings of the KDD Cup Workshop*. Vol. 2007. New York, NY, USA., pp. 3–6.

Chernozhukov, Victor et al. (2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21 (1).

Curth, Alicia and Mihaela Van Der Schaar (2023). "In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation". In: *International Conference on Machine Learning*. PMLR, pp. 6623–6642.

Dorie, Vincent et al. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition". In: *Statistical Science* 34 (1).

Gao, Zijun and Trevor Hastie (2022). "Estimating Heterogeneous Treatment Effects for General Responses". In: *submitted to Journal of the Royal Statistical Society: Series B*.

Hines, Oliver, Karla Diaz-Ordaz, and Stijn Vansteelandt (2022). "Variable importance measures for heterogeneous causal effects". In: *arXiv preprint arXiv:2204.06030*.

Hudson, Aaron (2023). "Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space". In: *arXiv preprint arXiv:2306.07492*.

Kennedy, Edward H (2022). "Semiparametric doubly robust targeted double machine learning: a review". In: *arXiv preprint arXiv:2203.06469*.

Künzel, Sören R. et al. (2019). "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the National Academy of Sciences* 116.10, pp. 4156–4165. ISSN: 0027-8424. DOI: 10.1073/pnas.1804597116. eprint: https://www.pnas.org/content/116/10/4156.full.pdf. URL: https://www.pnas.org/content/116/10/4156.

Lesko, LJ (2007). "Personalized medicine: elusive dream or imminent reality?" In: *Clinical Pharmacology and Therapeutics* 81.6, pp. 807–816.

Low, Yen Sia, Blanca Gallego, and Nigam Haresh Shah (2016). "Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records". In: *Journal of comparative effectiveness research* 5.2, pp. 179–192.

Murphy, Marilyn, Sam Redding, and J.S. Twyman (2016). *Handbook on personalized learning for states, districts, and schools*. Center for Innovations in Learning, Philadelphia, PA.

Niswander, Kenneth R, Myron J Gordon, and Myron Gordon (1972). *The women and their pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. Vol. 73. 379. National Institute of Health.

Robins, James et al. (2008). "Higher order influence functions and minimax estimation of nonlinear functionals". In: 2, pp. 335–422.

Rolling, Craig A and Yuhong Yang (2014). "Model selection for estimating treatment effects". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76.4, pp. 749–769.

Rosenbaum, Paul R (1989). "Optimal matching for observational studies". In: *Journal of the American Statistical Association* 84.408, pp. 1024–1032.

Rubin, Donald B (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Splawa-Neyman, Jerzy, D.M. Dabrowska, and T.P. Speed (1990). "On the application of probability theory to agricultural experiments. Essay on Principles, Section 9". In: *Statistical Science*, pp. 465–472.

Vaart, Aad W Van der (2000). *Asymptotic statistics*. Vol. 3. Cambridge university press.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]