Flexible Machine Learning Estimation of Conditional Average Treatment Effects

A Blessing and a Curse

Richard A. J. Post, Marko Petkovic, Isabel L. van den Heuvel, and Edwin R. van den Heuvel

Abstract: Causal inference from observational data requires testable identification assumptions. If these assumptions apply, machine learning methods can be used to study complex forms of causal effect heterogeneity. Recently, several machine learning methods were developed to estimate the conditional average treatment effect (ATE). If the features at hand cannot explain all heterogeneity, the individual treatment effects can seriously deviate from the conditional ATE. In this work, we demonstrate how the distributions of the individual treatment effect and the conditional ATE can differ when a causal random forest is applied. We extend the causal random forest to estimate the difference in conditional variance between treated and controls. If the distribution of the individual treatment effect equals that of the conditional ATE, this estimated difference in variance should be small. If they differ, an additional causal assumption is necessary to quantify the heterogeneity not captured by the distribution of the conditional ATE. The conditional variance of the individual treatment effect can be identified when the individual effect is independent of the outcome under no treatment given the measured features. Then, in the cases where the individual treatment effect and conditional ATE distributions differ, the extended causal random forest can appropriately estimate the variance of the individual treatment effect distribution, whereas the causal random forest fails to do so.

Keywords: Causal inference; Heterogeneity of treatment effects; Individual treatment effect; Machine learning; Unmeasured effect modifiers

(Epidemiology 2024;35: 32–40)

Submitted November 19, 2022; accepted September 26, 2023
From the Department of Mathematics and Computer Science, Eindhoven University of Technology, the Netherlands; and Department of Preventive Medicine and Epidemiology, School of Medicine, Boston University, Boston, MA. The authors report no funding or conflicts of interest.
This manuscript was previously posted to arXiv: https://doi.org/10.48550/arxiv.2210.16547.
The R code used for the data simulation and analysis can be found online at https://github.com/RAJP93/CATE.
Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).
Correspondence: Richard A. J. Post, Department of Mathematics and Computer Science, Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven, the Netherlands. E-mail: r.a.j.post@tue.nl.
Copyright © 2023 Wolters Kluwer Health, Inc. All rights reserved.
ISSN: 1044-3983/24/351-3240
DOI: 10.1097/EDE.0000000000001684

The increasing availability of observational data has tremendously boosted the field of machine learning. Machine learning provides us with flexible, nonparametric methods to study the observed outcome $Y$, given features $X$, that may involve a treatment (or exposure) $A$, by statistical inference on the (conditional) distributions $Y | X, A = a$. Therefore, machine learning methods are excellent at predicting future observations that arise from the same (factual) distribution. However, it is essential to realize that these models cannot be automatically used to answer “what if” questions for the treatment $A$, as associations found are not necessarily causal. Statistical inference of associations is thus only one step in causal inference and, as such, in counterfactual prediction.

The critical step for causal inference is linking the distribution of outcomes in a universe where everyone was treated with $a$, that is potential outcomes $Y_a$, to the distribution of the observed data. Working with observational data, we have to make causal identification assumptions that cannot be verified with the data, so machine learning is insufficient. Instead, we have to rely on expert knowledge. If the identification assumptions apply, distributions of potential and observed outcomes can be linked, and observable analogs of the causal estimands can be derived. Accurate statistical inference on these analogs is also necessary for causal inference. When the identification assumptions apply, but the statistical inference is off, the causal inference will also be invalid. The flexibility offered by machine learning methods can improve statistical inference. More precisely, machine learning methods can be exploited to learn nuisance parameters of the data generating distribution, such as conditional means and propensity scores, that can be used to estimate the causal estimand, as is, for example, done in targeted maximum likelihood estimation.

The increasing availability of diverse data makes studying effect heterogeneity among individuals more feasible. The field of precision medicine aims to understand this heterogeneity to improve individual treatment decisions. The average treatment effect (ATE), $\mathbb{E} [Y_1 - Y_0]$, might seriously differ from the individual treatment effect, that is the actual change in outcome caused by the exposure for a particular individual $Y_i - Y_i$. However, it is well known that an individual treatment effect is not identifiable because of the fundamental problem of causal inference.
it is impossible to observe the different potential outcomes for one individual jointly. On the other hand, marginalized effects like the ATE and the conditional ATE become identifiable in the absence of unmeasured confounding. In randomized experiments, this unconfoundedness assumption holds by design. Treatment effect heterogeneity studies thus focus on the estimation of the individualized conditional ATEs, \( \mathbb{E}[Y^1 - Y^0 | \mathbf{X}] \), given measured features \( \mathbf{X} \), but aggregated over remaining unmeasured features, as a proxy for the individual treatment effects. The functional form of effect modification by different levels of the measured features might be very complex, so machine learning methods are promising tools for estimating conditional ATEs.  

In recent years several meta-learning strategies for conditional ATE estimation have been proposed. These strategies decompose the conditional ATE estimation into regression problems that can be solved with any suitable machine learning method (see Caron et al. for a detailed review). T-learners fit separate models for treated and controls and estimate conditional ATEs as the plug-in difference of the conditional mean estimates. The performance of T-learners will depend on the levels of sparsity and smoothness of conditional means for treated and controls, and the choice of the base learner, and is low when the treated and control samples differ in size. S-learners include treatment assignment as another covariate and the conditional ATE is estimated as the difference of the estimated conditional means for treated and controls. Estimation with S-learners might suffer from serious finite-sample bias because they do not involve the conditional ATE directly, a problem also known for ATE estimation. The R-learner directly identifies the conditional ATE by regressing transformed outcomes on transformed treatment assignment using estimates of nuisance parameters in the first step (as we will elaborate on in the methods section). The R-learner is also called “double machine learning” and may give unbiased estimates of the average causal effect for finite samples. At the same time, a one-step approach (S-learner) would still be biased. Similarly, the DR-learner deals with augmented inverse probability weighted transformation of observations after constructing estimates of the propensity score and conditional means in the first step. The cost of making weaker modeling assumptions using flexible machine learning methods is slower convergence rates for the estimators, known as the curse of dimensionality. Therefore, much of the ongoing research is focused on comparing the different methods for conditional ATE estimation to derive whether and when they are optimal. 

The aim of this work is different, as we want to emphasize the difference between the conditional ATE and the individual treatment effect. The conditional ATE is much more personalized than the ATE and, thus, an important step towards precision medicine. However, it concerns us that the conditional ATE is sometimes seen as the individual treatment effect. Whether the conditional ATE can appropriately approximate the individual treatment effect depends on the remaining variability of causal effects given the considered modifiers, for example, a conditional ATE \( \geq 0 \) given \( \mathbf{X} = \mathbf{x}^a \) does not imply that all individual treatment effects \( \geq 0 \) for those individuals. In this work, we investigate whether we can use a causal random forest to estimate the variance of the marginal individual treatment effect distribution. More specifically, we investigate the performance of the causal random forest to estimate \( \text{var}(Y^1 - Y^0 | \mathbf{X} = \mathbf{x}) \) and \( \text{var}(Y^1 - Y^0) \) for data simulated from a causal system based on a real case study.

To open up the field of individual treatment effect distribution estimation, we derive identification assumptions, additionally to those necessary for marginal causal inference, to identify other characteristics of the conditional individual treatment effect distribution. To give an idea of how such assumptions on the joint distribution of potential outcomes can evolve the field of treatment effect heterogeneity, we extend the causal random forest algorithm to estimate the variance of the individual treatment effect given the measured features. First, we introduce our notation and present the identification assumptions necessary for conditional ATE estimation. Subsequently, we present the results of fitting the causal random forest to datasets simulated under different settings. Thereafter, we introduce the causal assumption for the identification of the conditional variance of the individual treatment effect. Moreover, we extend the causal random forest to estimate the latter and present its performance on the simulated datasets. Finally, we present some concluding remarks.

### NOTATION AND METHODS

Probability distributions of factual and counterfactual outcomes are defined in the potential outcome framework. Let \( Y_i \) and \( A_i \) represent the (factual) stochastic outcome and the random treatment assignment level of the individual \( i \). Let \( Y^a_i \) equal the potential outcome under an intervention on the treatment to level \( a \) (is counterfactual when \( A_i \neq a \)). We thus rely on a deterministic potential outcome framework, where each level of treatment corresponds to only one outcome for each individual (but its value typically differs between individuals).

We will consider only two treatment levels \{0, 1\} with 0 indicating no treatment. The individual causal effect of an arbitrary individual \( i \) equals \( Y^1_i - Y^0_i \). When we discuss the random variable describing the heterogeneity in the population, we do not subscript. Identification assumptions are necessary to relate the distribution of potential outcomes to the distribution of observed outcomes. First of all, it is necessary to have access to a set of measured features \( \mathbf{X} \) so that the treatment assignment is conditionally independent of the potential outcomes.

### Assumption 1. Conditional Exchangeability

\[ A \perp Y^0, Y^1, \mathbf{X} \]

This independence is called conditional exchangeability (or unconfoundedness) and implies the absence of unmeasured confounding that cannot be verified with observational data. Then there are no features, other than \( \mathbf{X} \), that \( Y^0 \) or \( Y^1 \) depend on and that differ in distribution between individuals with \( A = 1 \) and \( A = 0 \). Because we are interested in causal effect heterogeneity,
the set of features will also $X$ contain modifiers $X_m$ (i.e., $\exists x_1, x_2: \mathbb{E}[Y^1 - Y^0 | X_m = x_1] \neq \mathbb{E}[Y^1 - Y^0 | X_m = x_2]$) next to the confounders that are necessary to obtain independence. A feature can be only a modifier, only a confounder, or both, all on the additive scale. For a feature $L$ that is only a confounder but not a modifier $\forall l: \mathbb{E}[Y^1 - Y^0 | L = l, X_m = x] = \mathbb{E}[Y^1 - Y^0 | X_m = x]$, where $X_m$ represents the feature set $X$ without $L$.

Furthermore, we need to assume that an observed outcome equals the potential outcome for the assigned treatment, referred to as causal consistency.\(^47\)

Assumption 2. Causal Consistency

$Y_i = Y_i^0,$

Causal consistency is also referred to as the stable unit treatment value assumption.\(^48\) Causal consistency implies that potential outcomes are independent of the treatment levels of other individuals (no interference) and that there are no different versions of the exposure levels. Causal consistency can also not be verified with data.

Finally, the probability of receiving treatment should be bounded away from 0 and 1 for all levels of $X$, referred to as positivity\(^10\) and is also known as overlap.\(^48\)

Assumption 3. Positivity

$\forall x: 0 < \mathbb{P}(A = 1 | X = x) < 1$

As in Section 6 of Athey et al.,\(^41\) by causal consistency, we use the parameterization

$$Y_i = Y_i^0 + b_i A_i,$$

(1)

where $b_i$ is the individual treatment effect of individual $i$, so that $Y_i^1 = Y_i^0 + b_i.$ The conditional mean of $b_i$ given features $X_i$ equals the conditional ATE $\tau(X_i)$, where $\tau(x) = \mathbb{E}[Y^1 - Y^0 | X = x]$. The individual treatment effect can thus be divided into $\tau(X_i)$ and the individual deviation from the conditional ATE that is referred to as $U_i$. We rewrite Equation (1) as

$$Y_i = \theta_0(X_i) + N_i + (\tau(X_i) + U_i) A_i,$$

(2)

where $\theta_0(x) = \mathbb{E}[Y_i^0 | X_i = x]$, $N_i$ represents the deviation of $Y_i^0$ from $\theta_0(X_i)$, $\tau(x) = \mathbb{E}[b_i | X_i = x]$, $\mathbb{E}[N_i | X_i = x] = 0$ and $\mathbb{E}[U_i | X_i = x] = 0$. In this parameterization, the individual $Y_i^0$ and effect $b_i$ have been rewritten as the sum of their conditional expectations and zero mean deviations from these expectations. Note that other characteristics (different from the mean) of the $N_i | X_i = x$ and $U_i | X_i = x$ distributions can depend on the value of $x$. Furthermore, $U_i$ and $N_i$ can be dependent.

Case Study and Data Simulation

To illustrate how the random conditional expectation $\mathbb{E}[Y^1 - Y^0 | X]$ and $Y^1 - Y^0$ may differ in distribution, we simulate data based on the Framingham Heart Study.\(^50\) We focus on the heterogeneity in the effect of nonalcoholic fatty liver disease on a clinical precursor to heart failure, the left ventricular filling pressure.\(^50\) The association found in the original work was adjusted for several features. However, for this illustration, we assume that only gender (male = 0 and female = 1) and systolic blood pressure (SBP, mmHg) are confounders. We will simulate the following cause-effect relations:

$$A_i = \mathbb{I} \left\{ \frac{\exp(\alpha_0 + \alpha_{SBP} X_{SBP, i} + \alpha_{gen} X_{gen, i})}{1 + \exp(\alpha_0 + \alpha_{SBP} X_{SBP, i} + \alpha_{gen} X_{gen, i})} > N_{Ai} \right\}$$

$$Y_i^0 = \beta_0 + \beta_{SBP} X_{SBP, i} + N_{Yi};$$

$$Y_i^1 = Y_i^0 + (\tau_0 + \tau_{SBP} X_{SBP, i} + \tau_{gen} X_{gen, i} + \tau_{SBP, gen} X_{SBP, gen} + \tau_{SBP, gen} X_{SBP, gen} U_i) + N_{Yi},$$

where $X_{gen, i} \sim \text{Ber}(p), X_{SBP, j} \sim \mathcal{N}(0, 1), U_{i1} \sim \mathcal{N}(0, \sigma_1^2), U_{i2} \sim \mathcal{N}(0, \sigma_2^2), N_{Yi} \sim \text{Uni}[0, 1]$, and $U_i \perp N_{Yi}$. Moreover, there is no unmeasured confounding, that is $N_{Ai} \perp N_{Yi}, U_{i1}$ so that $A_i \perp (Y_i^1, Y_i^0) | X_{gen, i}, X_{SBP, i}$. By causal consistency, the observed outcome $Y_i = Y_i^0$ equals $Y_i^1$ when $A_i = 1$ and $Y_i^0$ when $A_i = 0$. The parameter values are obtained by fitting a linear mixed model for the relation of fatty liver disease and the left ventricular filling pressure adjusted for standarized SBP and gender,

$Y = \beta_0 + \beta_{SBP} X_{SBP, i} + \beta_{SBP, gen} X_{SBP, gen} + \beta_{SBP, gen} X_{SBP, gen} U_i$ + $\tau_0 + \tau_{SBP} X_{SBP, i} + \tau_{SBP, gen} X_{SBP, gen} + \tau_{SBP, gen} X_{SBP, gen} U_i$ $A_i$,

(4)

to the subset of the Framingham Heart Study participants n = 2356 as used by Chiuz et al.\(^50\) The distribution of $Y^1 - Y^0$ using the parameters obtained with PROC LOGISTIC and PROC MIXED in SAS is shown in Figure 1, where $\mathbb{E}[Y^1 - Y^0] = 0.5, \sqrt{\text{var}(Y^1 - Y^0)} = 1.41$ and $\mathbb{P}(Y^1 - Y^0 > 0) = 0.64$. Furthermore, the distribution of the conditional expectation $\mathbb{E}[Y^1 - Y^0 | X_{SBP, gen}]$ is shown with a standard deviation equal to 0.16 and $\mathbb{P}(\mathbb{E}[Y^1 - Y^0 | X_{SBP, gen}] > 0) = 1.00$. The conditional expectation distribution seriously differs from that of the individual treatment effect due to the unmeasured (remaining) effect heterogeneity ($U_i$). For completeness, the distributions of $Y^1$ and $Y^0$ are presented in eFigure 1 in eAppendix B; http://links.lww.com/EDE/C83. Moreover, we simulate $X_0$, which is a measured variable associated with the level of the individual modifier $U_i$,

$$(U_i, X_0)^T \sim \mathcal{N}\left(0, \left(\begin{array}{cc} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_2 \sigma_1 & \sigma_2^2 \end{array} \right)\right).$$

For $\rho > 0$,

$X_0$ is another measured modifier. Varying $\rho$ can thus be used to investigate cases where more of the latent individual effect modification can be explained, whereas preserving the distribution of $Y^1 - Y^0$. All programming codes used for this work can be found online at https://github.com/RAJP93/CATE.

Causal Random Forest

Because the actual causal effects are not observed, defining an appropriate loss function is not straightforward, so using machine learning methods to study causal effect heterogeneity is challenging.\(^21\) Causal trees\(^21\) and causal forests\(^21\) have been
introduced to draw inferences in case of complex causal effect heterogeneity. These papers mainly focus on data from randomized experiments and suggest adding traditional propensity score methods\(^1\) or using a different algorithm less sensitive to the complexity of the treatment effect function\(^1\) for observational studies. The causal random forest procedure as implemented in the causal_forest() function from the R-package grf is an example of a generalized random forest\(^1\) developed for observational data in the absence of unmeasured confounding. This machine learning method is a random forest-based variant of the R-learner\(^2\) using the relation between normalized outcome and treatment assignment as originally used by Robinson\(^5\):

\[
Y_i - m(X_i) = (A_i - e(X_i)) \hat{\tau}(X_i) + \tilde{N}_Y,
\]

where \(m(x) = \mathbb{E}[Y|X = x], e(x) = \mathbb{E}[A|X = x]\) and \(\forall x, a: \mathbb{E}[N_Y | A_i = a, X_i = x] = 0\). Asociational quantities are presented with a tilde, and causal quantities are without. Using this convention, \(\hat{\tau}(x)\) represents the conditional association measure of \(A\) and \(Y\) given \(X = x\). This association will equal the conditional ATE for \(X = x\) \(\tau(x)\), under certain identification assumptions, as we will show next. Using parameterization (2) one can derive, for \(X_i = x\),

\[
Y_i - m(x) = (\theta_0(x) + N_{Y_i} + (\tau(x) + U_{i1}) A_i) - (\theta_0(x) + \mathbb{E}[(\tau(x) + U_{i1}) A_i|X_i = x])
\]

\[= (A_i - e(x)) \tau(x) + (A_i U_{i1} - \mathbb{E}[A_i U_{i1}|X_i = x] + N_{Y_i}),\]

where the distribution of \(A_i, U_{i1}\) and \(N_{Y_i}\) can depend on the value of \(x\). If \(\mathbb{E}[U_{i1} + N_{Y_i}|A = 1, X = x] \neq \mathbb{E}[N_{Y_i}|A = 0, X = x]\), then \(\hat{\tau}(x) \neq \tau(x)\). This can occur when there is remaining (unmeasured) confounding after adjusting for \(X\). However, in the absence of unmeasured confounding, that is \(A \perp N_{Y_i}, U_i\) for \(X_i = x\),

\[
Y_i - m(x) = (A_i - e(x)) \tau(x) + (A_i U_{i1} + N_{Y_i}),
\]

where \(\forall x: \mathbb{E}[U_{i1} + N_{Y_i}|A_i = 1, X_i = x] = \mathbb{E}[N_{Y_i}|A_i = 0, X_i = x] = 0\). Then, \(\hat{\tau}(x)\) equals \(\tau(x)\), the conditional ATE for \(X = x\).

Under Assumption 1, the R-learner thus allows us to estimate (or predict) the conditional ATEs, \(\tau(x)\), from observational data. We present a summary of the generalized random forest implementation of the causal random forest in eAppendix A; http://links.lww.com/EDE/C83 including the corresponding estimate of \(\tau(x)\) in Equation (10).

In this work, we fit a causal random forest to the simulated data as described in the previous section to estimate the ATE conditional on \((X_{gen_i}, X_{SBP_i}, X_0)\) for each individual. We vary the sample size, \(n \in \{200, 2000, 20000\}\), and the correlation between the unmeasured modifier \(U_i\) and the measured \(X_0, \rho \in \{0, 0.25, 0.5, 0.75, 1\}\), whereas fixing \(\delta = 2\). We use the default settings of the causal_forest() function, except for the \(n = 200\) settings where we set min.node.size = 1. For each simulation, we compute the empirical standard deviation (SD) and positive effect probability, \(\mathbb{P}(Y^1 - Y^0 > 0)\), of the estimated conditional ATE distribution to estimate the SD and positive effect probability of the individual treatment effect distribution, respectively. Furthermore, based on 1000 bootstrap samples, we estimate 95% confidence intervals (CIs) for all three characteristics. Based on 1000 simulations, we estimate the bias, mean squared error (MSE), and coverage for the different settings. Finally, we estimate the individual treatment effect distribution per simulation with a Gaussian kernel density estimator over the estimated conditional ATEs using the density() function in R with the default settings.

**RESULTS**

The bias, MSE, and coverage for the ATE, SD, and positive effect probability of the individual treatment effect distribution based on the conditional ATE distribution, estimated with the causal random forest, are presented in Table 1 for the different settings.

In the absence of features \((X_0)\) that are associated with the unmeasured modifier, that is when \(\rho = 0\), the variability in the individual treatment effect is seriously underestimated when using the conditional ATE distribution as shown in the first row of eFigure 2; http://links.lww.com/EDE/C83. Therefore, the SD and positive effect probability of the distribution of the conditional expectation \(\mathbb{E}[Y^1 - Y^0|X]\) are biased estimators of the characteristics of the individual treatment effect distribution. The bias is the lowest for a small sample size due to a finite-sample effect for both the SD and positive effect probability. For \(n = 200\), the coverage of the positive effect probability is not much off. For larger sample sizes, the conditional ATE distribution can be

**FIGURE 1.** The individual treatment effect (solid line) distribution for the cause-effect relations for \(\alpha_0 = -1.7, \alpha_{gen} = -0.1, \alpha_{SBP} = 0.4\) (so that \(P(A = 1 | X_{gen} = 1) = 0.15\) and \(P(A = 1 | X_{gen} = 0) = 0.16\)), \(\beta_0 = 5.9, \beta_{gen} = 0.8, \beta_{SBP} = 0.5, \tau_0 = 0.45, \tau_{gen} = 0.1, \tau_{SBP} = 0.15, \sigma_0^2 = 1.6^2, \) and \(\sigma_1^2 = 1.4^2\). The ATE is presented with a vertical solid line. The distribution of a random conditional ATE, thus for random \(X_{SBP}\) and \(X_{gen}\), is also presented (dotted line).
estimated more precisely. Then, the bias increases, and the coverage decreases.

We observe the same trend for $\rho = 0.25$. However, for $\rho \geq 0.50$, the bias is more extensive for small sample sizes. In the latter cases, the finite-sample effect of the conditional ATE distribution estimator no longer compensates for the difference between the individual treatment effect and conditional ATE distribution. Nevertheless, the uncertainty in the estimate for small sample sizes still results in higher coverage of the positive effect probability. For $\rho = 0.75$, the conditional ATE distribution becomes a reasonable proxy for the individual treatment effect distribution, as seen from the fourth row in eFigure 2; http://links.lww.com/EDE/C83. Finally, for $\rho = 1$, $U_i$ equals $X_0$, and there is thus no unmeasured effect modification. In this case, the flexible machine learning estimation of the conditional ATE distribution can be used to estimate the individual treatment effect distribution and understand the variability in the treatment effect. Indeed, the bias of the SD and the positive effect probability become small, and the coverage approaches the nominal probability. The bias of the SD is still not negligible, so the coverage of the SD deviates from the nominal probability.

From Conditional Means to Conditional Distributions

We presented examples in which the distribution of $\mathbb{E}[Y^1 - Y^0|X] = x$ differs from that of $Y^1 - Y^0$. To overcome this issue, we should consider the remaining effect heterogeneity. In this section, we show that the variance of $Y^1 - Y^0|X = x$ equals $\sigma^2_1(x) = \mathbb{E}[(U_1)^2|X = x]$. As derived in eAppendix C; http://links.lww.com/EDE/C83, under Assumption 1, the Robinson decomposition of the squared observations enables us to estimate

$$\Delta(x) = \tau(x)^2 + \sigma^2_1(x) + 2\tau(x)\theta_0(x) + 2\mathbb{E}[N_1 Y_1|X = x]$$

(7)

from observational data. If also Assumptions 2 and 3 apply, $\tau(x)$ and $\theta_0(x)$ are identifiable. By subtracting the estimates of their observable analogs $\hat{\tau}(x)$ and $\hat{\theta}_0(x) = m(x) - e(x)\hat{\tau}(x)$,

$$\hat{\sigma}^2_1(x) = \sigma^2_1(x) + 2\mathbb{E}[N_1 U_1|X = x]$$

(8)

can be estimated. Because $\mathbb{E}[U_1|X = x]$ and $\mathbb{E}[N_1 Y_1|X = x]$ equal 0, $\hat{\sigma}^2_1(x)$ represents the sum of the conditional variance of the individual treatment effect and twice the conditional covariance of $Y^0$ and the individual treatment effect. However, as a result of the fundamental problem of causal inference, $\mathbb{E}[N_1 U_1|X = x]$, the conditional expectation of the product of the deviation of $Y^0$ from $\theta(x)$ and the deviation of $Y^1 - Y^0$ from $\tau(x)$, is not identifiable. So, we cannot estimate $\hat{\sigma}^2_1(x)$ without an additional (cross-world) assumption.

If we can assume that $U_1 \perp N_1 | X = x$, then $\mathbb{E}[N_1 U_1|X = x] = 0$ and $\hat{\sigma}^2_1(x) = \sigma^2_1(x)$, so that the variance of the causal effect given $X$ becomes identifiable. The assumption implies conditional independence of the outcome under no treatment and the effect, i.e. the deviation of $Y^0 | X = x$ from $\mathbb{E}[Y^0|X = x]$ is independent of the deviation of $Y^1 - Y^0 | X = x$ from the conditional ATE. Therefore, we refer to this assumption as conditional independent effect deviation.

Assumption 4. Conditional Independent Effect Deviation

$Y^1 - Y^0 \perp Y^0 | X = x$
Conditional independent effect deviation implies that all features that affect both $Y^0$ and $Y^1 - Y^0$ should be contained in $X$. As an example, one could think of the effectiveness of medical drugs that depends on the amount of enzyme present for an individual, whereas the presence of the enzyme itself does not inform on the outcome of interest in the absence of the drug. The antiplatelet medicine clopidogrel reduces the risk of stroke and myocardial infarction in individuals with acute coronary syndrome, but its effect depends on its conversion to an active metabolite which is accomplished by the cytochrome P450 2C19 (CYP2C19) enzyme. For individuals with a CYP2C19 gene mutation, the drug is known to have a reduced antiplatelet effect; the CYP219 gene thus results in effect heterogeneity. However, there is no reason to believe that the gene affects platelet aggregation in the absence of the drug. In cases where $Y^0$ is still expected to inform on the value of $Y^1 - Y^0$ given the levels of $X$, the identification assumption does not apply. If the CYP219 gene had also affected the platelet aggregation in the absence of the drug, Assumption 4 would have been violated. More generally, Assumption 4 is violated when there exists a feature that despite conditioning on $X$ affects both $Y^1 - Y^0$ and $Y^0$. This applies when such a feature cannot be (or is not) measured.

Similar to Assumption 1, Assumption 4 cannot be verified with data as it concerns unmeasured features that affect both $Y^0$ and $Y^1 - Y^0$ and should be based on expert knowledge. However, in contrast to Assumption 1, no reason guarantees that Assumption 4 applies in a randomized experiment.

Also, in the case where sufficient features are measured so that the individual treatment effect follows the conditional ATE, Assumption 4 applies as $\forall i: U_{i+} = 0$ and thus independent of $Y^0$. If Assumption 4 (in addition to 1, 2, and 3) applies, then $\text{var}(Y^1 - Y^0|X = x)$ can be estimated with an extended causal random forest as presented in Algorithm 1.

Algorithm 1 provides us with an estimate for both the conditional ATE and the conditional variance of the individual treatment effect given the measured features. The ATE estimate remains the same as for the original causal random forest. The SD of the effect equals

$$\sqrt{\mathbb{E}\left[\text{var}(Y^1 - Y^0|X) + \mathbb{E}[Y^1 - Y^0|X]^2\right] - \mathbb{E}[Y^1 - Y^0]^2}$$

and is therefore estimated as

$$\sqrt{\text{max}\left\{0, n^{-1}\left(\sum_{i=1}^n \tilde{\sigma}_i^2(\hat{\tau}(x_i)) + \hat{\tau}(x_i)^2\right) - \text{ATE}^2\right\}}.$$ 

Only when the conditional individual treatment effect distribution can be well approximated with a Gaussian distribution the distribution of $Y^1 - Y^0|X = x$ is identified by the conditional ATE and the conditional variance. Then, by the law of total probability,

$$\mathbb{P}(Y^1 - Y^0 \leq y) = \int \mathbb{P}(Y^1 - Y^0 \leq y|X = x) dF_X(x)$$

(9)

For illustration, we will assume the Gaussianity of the conditional individual treatment effect distribution to use the extended causal random forest to estimate the individual treatment effect distribution from the simulated datasets. The positive effect probability is now estimated as $n^{-1}\sum_{i=1}^n \mathbb{P}(Z_i > 0)$, where

$$Z_i \sim \mathcal{N}\left(\hat{\tau}(x_i), \max\left\{0, \tilde{\sigma}_i^2(x_i)\right\}\right).$$

The Gaussianity assumption plays a different role than the identification Assumptions 1 to 4. The focus of this work is on the conditional variance (and the conditional ATE) that is only identifiable under Assumptions 1 to 4. We resort to the Gaussianity assumption to also estimate the conditional effect distribution. In eAppendix D; http://links.lww.com/EDE/C83, we show that under violation of the Gaussianity

---

Algorithm 1 Extended causal random forest

1: input features $x$
2: Run the traditional causal random forest
   i: Save predicted conditional ATE, $\hat{\tau}(x)$
   ii: $\forall i$: save the predicted $\hat{k}^{-i}(x_i)$
   iii: Save the predicted $Y^0$, $\hat{\theta}_0(x)$
   iv: Save the similarity weights $\alpha_i(x)$
   ▶ When $\exists j$: $x_i = x_j$ then $\hat{\theta}_0(x_i)$ equals $\hat{\mu}_0(x_i)$ as defined after Equation (11) in eAppendix A; http://links.lww.com/EDE/C83 by only considering those trees that did not use individual $i$ for training.
3: Fit a separate regression forest to estimate $h(x_i) = \mathbb{E}(Y_i|x_i = x_i)$
   i: $\forall i$: obtain the out of bag predictions of $(Y_i)^2$, $\hat{k}^{-i}(x_i)$
4: Estimate $\mathbb{E}[(Y^1)^2 - (Y^0)^2|X = x]$ as $\hat{\Delta}(x) = \frac{\sum_{i \in \mathcal{L}} \alpha_i(x)(Y_i)^2 - \hat{k}^{-i}(x_i)(4\hat{\Delta}(x) + \hat{\tau}(x_i)))}{\sum_{i \in \mathcal{L}} \alpha_i(x)(4\hat{\Delta}(x) + \hat{\tau}(x_i)))}$
5: Estimate $\mathbb{E}[|U_i|^2|X = x] + 2\mathbb{E}[Y_i|U_i|X = x]$ as $\hat{\sigma}_i^2(x) = \hat{\Delta}(x) - \hat{\tau}(x_i)^2 - 2\hat{\tau}(x_i)\hat{\theta}_0(x_i)$.
6: return $\left(\hat{\tau}(x), \hat{\sigma}_i^2(x)\right)$
assumption, the SD of the effect can still be appropriately estimated with Algorithm 1, but the positive effect probability and individual treatment effect distribution estimates will be off.

**Results Resumed**

The bias, MSE, and coverage for the ATE ($\rho = 0.5$), SD ($\rho = 1.41$), and positive effect probability ($\rho = 0.64$) of the individual treatment effect distribution, respectively, using the extended causal random forest, whereas assuming Gaussian distributed $Y^1 - Y^0 | X = x$ are presented in Table 2 for the different settings of the simulation study described before.

In the case of remaining heterogeneity ($\rho < 1$), the bias of the SD estimator using the extended causal random forest is much lower than using the causal random forest. For larger sample sizes ($n = 2000$ and $n = 20,000$), the small bias is of the opposite sign to the one using the causal random forest. The bias of the positive effect probability is also seriously decreased. For all settings, the MSE of the extended estimator is smaller for both the SD and positive effect probability. Also, the coverage of the SD and positive effect probability did considerably improve. However, for $n = 20,000$ and $\rho = 0.50$ or $\rho = 0.75$, the coverage probability of the SD did deviate from the nominal level due to the small bias and the narrow CIs.

The extended causal random forest still performs well for the $\rho = 1$ case, where all variability in causal effect could be explained with the measured features. Only when $n = 20,000$ the bias of the SD estimator using the extended causal random forest is slightly higher than for the original estimator due to an overspecified model. The difference is so small that the MSE is of the same magnitude. In this case, the coverage again deviates from the nominal level and is now slightly lower than the coverage using the traditional causal random forest.

The pointwise mean (and 95% CI), from 1000 simulations, of the estimated probability density function of the individual treatment effect, is presented in eFigure 3; http://links.lww.com/EDE/C83 for the different settings.

**DISCUSSION**

Machine learning methods are of great value in understanding effect heterogeneity using conditional ATEs. However, we have illustrated that the individualized conditional ATE can still seriously differ from the individual treatment effect. Studying the remaining effect heterogeneity is challenging as the fundamental problem of causal inference prevents us from learning the joint distribution of potential outcomes. Nevertheless, the conditional second moments of the treated and the controls should be similar under remaining effect homogeneity. We have extended the causal random forest algorithm of Athey et al.41 to estimate the difference in conditional variance between treated and controls. If variances are different, under Assumptions 1 to 3, the individual treatment effect distribution cannot be explained by the conditional ATEs alone. The increased variance among the treated is due to the conditional variance of $Y^1 - Y^0$ and the covariance of $Y^1 - Y^0$ and $Y^0$. To identify the (conditional) variance of the individual treatment effect, the dependence structure of $Y^1 - Y^0$ and $Y^0$ should be assumed. When conditional independent effect deviation applies, the conditional variance of the individual treatment effect can be estimated next to the expected effect for each individual. As a result, in contrast to the causal random forest, the extended causal random forest can be used

| TABLE 2. Bias, Mean Squared Error (MSE), and Coverage of the Estimated Average Treatment Effect (ATE), Standard Deviation (SD), and Positive Effect Probability (PEP) of the Individual Treatment Effect Distribution Using the Extended Causal Random Forest Based on 1000 Simulated Samples Per Setting |
|---|---|---|---|---|---|---|---|---|
| $\rho$ | $n$ | Bias | | MSE | | Coverage | |
| | | ATE | SD | PEP | ATE | SD | PEP | |
| 0 | 200 | 0.05 | -0.06 | 0.04 | 0.20 | 0.25 | 0.02 | 0.95 | 0.90 | 0.91 |
| 0 | 2000 | -0.00 | 0.04 | 0.01 | 0.02 | 0.02 | 0.00 | 0.94 | 0.93 | 0.90 |
| 0.25 | 20,000 | 0.00 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 | 0.95 | 0.89 | 0.82 |
| 0.25 | 2000 | 0.01 | -0.08 | 0.03 | 0.20 | 0.29 | 0.02 | 0.94 | 0.91 | 0.92 |
| 0.50 | 20,000 | 0.00 | 0.04 | 0.01 | 0.02 | 0.02 | 0.00 | 0.95 | 0.93 | 0.90 |
| 0.50 | 200 | 0.02 | -0.07 | 0.03 | 0.19 | 0.27 | 0.02 | 0.95 | 0.90 | 0.92 |
| 0.50 | 2000 | 0.01 | 0.05 | 0.01 | 0.02 | 0.02 | 0.00 | 0.95 | 0.94 | 0.91 |
| 0.50 | 20,000 | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.94 | 0.80 | 0.92 |
| 0.75 | 200 | 0.04 | -0.07 | 0.03 | 0.16 | 0.26 | 0.02 | 0.95 | 0.92 | 0.92 |
| 0.75 | 2000 | 0.01 | 0.07 | -0.00 | 0.01 | 0.03 | 0.00 | 0.95 | 0.91 | 0.94 |
| 0.75 | 20,000 | 0.00 | 0.07 | -0.00 | 0.00 | 0.01 | 0.00 | 0.94 | 0.71 | 0.93 |
| 1.00 | 200 | 0.04 | -0.03 | 0.02 | 0.14 | 0.26 | 0.02 | 0.95 | 0.91 | 0.92 |
| 1.00 | 2000 | 0.01 | 0.09 | -0.01 | 0.01 | 0.03 | 0.00 | 0.96 | 0.90 | 0.93 |
| 1.00 | 20,000 | 0.00 | 0.09 | -0.01 | 0.00 | 0.01 | 0.00 | 0.95 | 0.56 | 0.89 |
to estimate the individual treatment effect distribution’s variance unbiasedly. It should be clear that for settings where Assumption 4 is violated, the estimate of the individual treatment effect variance based on the extended causal random forest will be biased as illustrated in eAppendix D.1; http://links.lww.com/EDE/C83. When assuming that the conditional individual treatment effect distribution can be approximated with Gaussian distributions, the individual treatment effect distribution can also be estimated. Otherwise, the distribution is not captured by the conditional ATE and conditional variance alone. Then, estimates of other distributional properties like the positive effect probability will be off, as demonstrated for a scenario with non-Gaussian conditional treatment effects in eAppendix D.2; http://links.lww.com/EDE/C83 with confounders that are no effect modifiers and in eAppendix D.5; http://links.lww.com/EDE/C83 without confounders.

We have presented the extended causal random forest just as an example, and other machine learning algorithms can be extended in a similar way to estimate the (conditional) variance of causal effects. To do so, the machine learning method should appropriately estimate the conditional ATEs and not be prone to overfitting. Furthermore, it is important to realize that the objective function is typically chosen to maximize heterogeneity in conditional ATEs, so that confounders that are not effect modifiers (on the additive scale) are not adjusted for. One should always discuss whether the distribution of potential outcomes is correctly linked to the observed distribution. For the causal random forest, this is due to the orthogonalization step as explained in eAppendix D.4; http://links.lww.com/EDE/C83 with confounders that are no effect modifiers and in eAppendix D.5; http://links.lww.com/EDE/C83 without confounders.

Although the extended causal random forest algorithm can be used in practice, our main aim was to emphasize that the conditional ATE and individual treatment effect distributions can differ and to present identification assumptions for the (conditional) variance of the causal effect. Because assumptions on the conditional dependence of and the individual treatment effect cannot be tested with factual data, it will be challenging for field experts to review such assumptions. It will be impossible for some applications to do so, but in others, for example, the clopidogrel example mentioned in this paper, it might be possible. Reasoning about more examples will be an important topic of future interdisciplinary research.

With this paper, we hope to open up the field of ITE distribution estimation under assumptions like conditional independent effect deviation. The latter is necessary to understand to what degree individualized conditional ATEs are informative at the actual individual level.

REFERENCES

1. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Ann Rev Public Health*. 2018;39:95–112.
2. Dickerman BA, Hernán MA. Counterfactual prediction is not only for causal inference. *Eur J Epidemiol*. 2020;35:615–617.
3. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance*. 2019;32:42–49.
4. Prosperi M, Guo Y, Sperrin M, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell*. 2020;2:369–375.
5. van Geloven N, Swanson SA, Ramspeck CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol*. 2020;35:619–630.
6. Mooney SJ, Keil AP, Westreich DJ. Thirty questions about using machine learning in causal research (You Won’t believe the answer to number 10!). *Am J Epidemiol*. 2021;190:1476–1482.
7. Cui P, Athey S. Stable learning establishes some common ground between causal inference and machine learning. *Nat Mach Intell*. 2022;4:110–115.
8. Dickerman BA, Dahabreh IJ, Cantos KV, et al. Predicting counterfactual risks under hypothetical treatment strategies: an application to HIV. *Eur J Epidemiol*. 2022;37:367–376.
9. Balzer LB, Petersen ML. Invited commentary: machine learning in causal inference—how do I love thee? let me count the ways. *Am J Epidemiol*. 2021;190:1483–1487.
10. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton, FL: Chapman & Hall/CRC; 2020.
11. Robins JM, Wasserman L. On the impossibility of inferring causation from association without background knowledge. In: *Computation, Causation, and Discovery*. AAAI Press; 1999.
12. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *Int J Epidemiol*. 2019;49:2058–2064.
13. van der Laan MJ, Rose S. Targeted Learning: Causal Inference for Observational and Experimental Data. New York: Springer; 2011.
14. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. *Am J Epidemiol*. 2017;185:65–73.
15. Kosorok MR, Laber EB. Precision medicine. *Annu Rev Stat Appl*. 2019;6:263–286.
16. Kravitz R, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q*. 2004;82:661–687.
17. Holland PW. Statistics and causal inference. *J Am Stat Assoc*. 1986;81:688–960.
18. Robertson SE, Leith A, Schmid CH, Dahabreh IJ. Assessing heterogeneity of treatment effects in observational studies. *Am J Epidemiol*. 2020;190:1088–1100.
19. Bica I, Alaa AM, Lambert C, van der Schaar M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin Pharmacol Ther*. 2021;109:87–100.
20. Caron A, Baio G, Manolopoulos I. Estimating individual treatment effects using non-parametric regression models: a review. *J R Stat Soc Ser A Stat Soc*. 2022;185:1115–1149.
21. Athey S, Imbens GW. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci*. 2016;113:7353–7360.
22. Powers S, Qian J, Jung K, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. *Stat Med*. 2018;37:1767–1787.
23. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci*. 2019;116:4156–4165.
24. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20:217–240.
25. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med*. 2011;30:2867–2880.
26. Green DP, Kern HL. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opin Q*. 2012;76:491–511.
27. Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat*. 2013;7:443–470.

© 2023 Wolters Kluwer Health, Inc. All rights reserved.

www.epidem.com | 39

Copyright © 2023 Wolters Kluwer Health, Inc. Unauthorized reproduction of this article is prohibited.
28. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J*. 2018;21:C1–C68.

29. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*. 2020;108:299–319.

30. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89:846–866.

31. Kennedy EH. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*. 2020.

32. Fan Q, Hsu YC, Lieli RP, Zhang Y. Estimation of conditional average treatment effects with high-dimensional data. *J Bus Econ Stat*. 2022;40:313–327.

33. Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *Am J Epidemiol*. 2021.

34. Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat Med*. 2018;37:3309–3324.

35. Knaus MC, Lechner M, Strittmatter A. Machine learning estimation of heterogeneous causal effects: empirical Monte Carlo evidence. *Econom J*. 2020;24:134–161.

36. Curth A, van der Schaar M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In: Banerjee A, Fukumizu K (eds) Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR, Proceedings of Machine Learning Research; 2021;130:1810–1818.

37. Kennedy EH, Balakrishnan S, Robins JM, Wasserman L. Minimax rates for heterogeneous causal effect estimation. *arXiv preprint arXiv:2203.00837*. 2023.

38. Lu M, Sadqi S, Feaster DJ, Ishwaran H. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat*. 2018;27:209–219.

39. Talisa VB, Chang CCH. Learning and confirming a class of treatment responders in clinical trials. *Stat Med*. 2021;40:4872–4889.