American Journal of Epidemiology

Vol. 192, No. 7
https://doi.org/10.1093/aje/kwad043
Advance Access publication:
February 24, 2023

Practice of Epidemiology

Practical Guide to Honest Causal Forests for Identifying Heterogeneous Treatment Effects

Neal Jawadekar*, Katrina Kezios, Michelle C. Odden, Jeanette A. Stingone, Sebastian Calonico, Kara Rudolph, and Adina Zeki Al Hazzouri

* Correspondence to Neal Jawadekar, Department of Epidemiology, Mailman School of Public Health, Columbia University, 722 West 168th Street, Room 513, New York, NY 10032 (e-mail: nj2415@cumc.columbia.edu).

Initially submitted November 1, 2021; accepted for publication February 20, 2023.

"Heterogeneous treatment effects" is a term which refers to conditional average treatment effects (i.e., CATEs) that vary across population subgroups. Epidemiologists are often interested in estimating such effects because they can help detect populations that may particularly benefit from or be harmed by a treatment. However, standard regression approaches for estimating heterogeneous effects are limited by preexisting hypotheses, test a single effect modifier at a time, and are subject to the multiple-comparisons problem. In this article, we aim to offer a practical guide to honest causal forests, an ensemble tree-based learning method which can discover as well as estimate heterogeneous treatment effects using a data-driven approach. We discuss the fundamentals of tree-based methods, describe how honest causal forests can identify and estimate heterogeneous effects, and demonstrate an implementation of this method using simulated data. Our implementation highlights the steps required to simulate data sets, build honest causal forests, and assess model performance across a variety of simulation scenarios. Overall, this paper is intended for epidemiologists and other population health researchers who lack an extensive background in machine learning yet are interested in utilizing an emerging method for identifying and estimating heterogeneous treatment effects.

data science; effect modifiers; epidemiologic methods; honest causal forests; machine learning; precision medicine

Abbreviations: AIPW, augmented inverse probability weighting; ATE, average treatment effect; CART, classification and regression trees; CATE, conditional average treatment effect; EMSE, expected mean squared error; RCT, randomized controlled trial; VIF, variable importance factor.

Epidemiologists and population health researchers are often interested in estimating causal effects. An average causal effect, also known as an average treatment effect (ATE), is a population or samplewide measure that represents the effect of a specific treatment (or exposure) on an outcome of interest (see Table 1). An ATE (on the absolute scale) can be defined as the average of the difference in potential outcomes, comparing results that would be observed for an entire sample if everyone in the sample were treated with those that would be observed for the same sample if it were untreated (1).

While an ATE indicates the average expected change in the risk of disease that would result from treatment across an entire sample, this value does not necessarily correspond to how a subgroup within the same sample would respond to that treatment. To better capture this, one can instead estimate the conditional average treatment effect (CATE), which refers to the ATE specific to a subgroup defined by a vector $\boldsymbol{x}$ of covariates (see Table 1). While CATEs are scale-dependent—that is, whether the CATE is equal to or different from the ATE depends on whether it was measured on the relative or absolute scale—for the entirety of this paper, we focus on estimates measured on the absolute scale.

Knowledge of CATEs is beneficial when the effect of a treatment on disease varies substantially between subgroups of individuals within a population. For instance, the effectiveness of statin medications in reducing cholesterol levels differs based on the presence of certain single

Table 1. Definitions of Estimands of Interest in Epidemiologic Studies

Estimand	Mathematical Expression[a]	Definition
ATE	$\frac{1}{N}\sum_i \left[Y_i^{a=1} - Y_i^{a=0}\right]$	The average of the difference in potential outcomes in a sample where everyone is treated versus the same sample where everyone is untreated.
CATE	$\frac{1}{N}\sum_i \left[Y_i^{a=1} - Y_i^{a=0} \mid X = \mathbf{x}\right]$	The average of the difference in potential outcomes in a specific stratum (defined by a vector of covariates) where everyone in that stratum is treated versus everyone in that stratum being untreated.

Abbreviations: ATE, average treatment effect; CATE, conditional average treatment effect.
[a] N, number of individuals in the sample; Y, potential outcome; $a = 0$, not treated; $a = 1$, treated; i, each individual in the sample; $\mathbf{x}$, vector of covariates.

nucleotide polymorphisms (2–5). Similarly, the effectiveness of 2-dose severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) mRNA vaccination on immune response may depend on age (6) and coronavirus disease 2019 (COVID-19) recovery status (7). Accordingly, one should not always expect a given treatment to have the same effect on everyone.

To detect whether CATEs vary across subgroups (i.e., to identify the presence of treatment effect heterogeneity or effect-measure modification), researchers commonly include a treatment-modifier product term in a regression model or estimate treatment effects within modifier strata of one to several a priori hypothesized modifiers of interest. While these traditional approaches can help *detect* the presence of prespecified effect heterogeneity, their reliance on a priori hypotheses makes them ill-suited for *discovering* novel variables which most contribute to such heterogeneity (8). Furthermore, conducting hypothesis tests on many potential effect modifiers is subject to the multiple-comparisons problem. Data-driven hypothesis-generating approaches, on the other hand, may help to circumvent this problem, while also enabling the discovery of novel hypotheses on heterogeneous effects across nonprespecified subgroups.

One category of data-driven methods, *tree-based methods* (9, 10), refers to a group of algorithms which iteratively partition (i.e., split) a sample into subgroups based on the values of certain variables. To date, tree-based methods have typically been used for the purpose of prediction. However, in recent years, researchers have also described their promise for uncovering and estimating heterogeneous treatment effects in a causal setting (11–16). "Honest causal forests" are one such tree-based method proposed for this purpose (17). Several applications of the causal forest have generated new hypotheses on potential effect modifiers, although, in general, there remains a dearth of real-world applications of honest causal forests to health-related questions (18–20). Past applications include a post hoc analysis of a randomly allocated weight-loss intervention on cardiovascular disease–related mortality (18), a secondary analysis of the effect of the Systolic Blood Pressure Intervention Trial (SPRINT) on cardiovascular disease outcomes (19), and a retrospective study of the effects of intensive glycemic control on all-cause mortality within US-based diabetes treatment trials (20).

Below, we discuss the honest causal forests method in detail, beginning with an introduction to the collection of methods from which it is derived: tree-based methods for prediction (including classification and regression trees (CART) and random forests). Understanding tree-based methods for prediction helps ground the discussion for their application in a causal inference setting. We then describe both honest causal trees and honest causal forests, before ending with an application of honest causal forests to simulated data (in both randomized and observational settings). Our practical guide serves as a conceptual and instructional tool for epidemiologists interested in this method.

METHODS

Tree-based learning methods for prediction: CART and random forest

Overview. Tree-based methods are data-driven algorithms often utilized for *predicting* an outcome Y. The most basic type of tree-based methods is CART, whereby classification trees predict *categorical* outcomes and regression trees predict *continuous* outcomes (21); here, we focus our discussion primarily on classification trees. A large advantage of tree-based methods is their ability to flexibly account for complex (including nonlinear) relationships between multiple variables in a manner that is more intuitive and easier to visualize than most other models.

Recursive partitioning and the structure of trees. An example of a CART prediction algorithm learned from data is depicted in Figure 1. In this illustration, an original sample of 700 individuals (located at the top-most node, known as the root node) was initially split on the variable "age." This first split resulted in 2 mutually exclusive subsamples beneath it (known as *child* nodes, in relation to the *parent* node). From top to bottom, the tree was split into additional pairs of nodes based on algorithm-selected levels of different variables, with the goal of predicting an outcome. At the bottom of the tree in the terminal nodes (also known as leaf nodes), a final prediction is made for each individual. When the outcome is dichotomous (yes/no), the prediction is either yes or no, and when using a majority voting classifier, this prediction will depend on the most common outcome among individuals within that leaf. When the outcome is

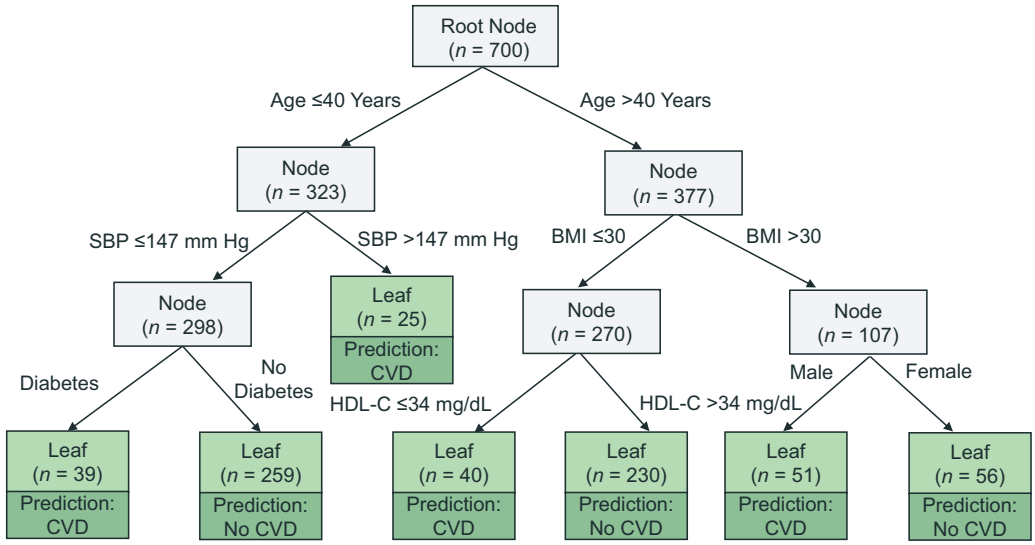

Figure 1. Hypothetical example of a classification and regression trees (CART) prediction algorithm. Individual nodes (pictured here as rectangles), which contain individuals within a given sample, are iteratively split into a pair of nodes beneath them (known as child nodes). These splits occur across values of covariates selected by the algorithm, with the goal of predicting an outcome (which in this example is cardiovascular disease (CVD)). Nodes without any arrows emanating from them are known as leaf nodes. BMI, body mass index; HDL-C, high-density lipoprotein cholesterol; SBP, systolic blood pressure.

continuous, the prediction is typically computed as the expectation of Y among individuals in that leaf. Accordingly, every individual belonging to the same leaf receives the same prediction.

The goal of CART is to build a tree through the minimization of a loss function tailored to maximize homogeneity within nodes. As such, when learning a tree, splits are performed with the objective of grouping together individuals who share similar outcome values. During this iterative process, the CART algorithm will select the variable split (among all potential splits) at each node which maximizes improvement in *node purity*, defined as how homogenous the observed outcomes are within individual nodes (22). A commonly utilized measure of node purity is the *Gini index*, as defined in equation 1 (10, 23). Here, a Gini index value of 0 corresponds to a perfectly homogenous (i.e., pure) node, whereas values closer to 0.5 indicate considerable heterogeneity within a node.

The Gini index (G) is defined as follows.

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right) \quad (1)$$

In equation 1, $\hat{p}_{mk}$ is the proportion of observations in a given node m belonging to class k.

There are 2 general approaches for creating splits within a CART algorithm, "greedy" and "nongreedy" algorithms, which prioritize purity at different nodes. In greedy algorithms—which are most commonly used due to the computational intensity of nongreedy ones—each split is chosen to optimize, for example, node purity at each of the 2 child nodes relative to its parent node. Meanwhile, nongreedy algorithms choose splits with the goal of *globally* optimizing node purity throughout the entire tree (24). This splitting process continues down the tree until reaching a stopping criterion. The stopping criterion is determined by a variety of hyperparameters that are prespecified by the user, including but not limited to the minimum number of people allowed in each leaf and the maximum number of leaves allowed in the tree (9). If none of the potential splits can meet the criteria of the stopping criterion, then no more splits are made, resulting in the final model.

To help ensure optimal model performance, an approach known as *pruning* can be incorporated into the CART model-building process. A detailed overview of pruning can be found in *An Introduction to Statistical Learning* by James et al. (22). In short, pruning is an approach which reduces the size of the tree by allowing splits to occur only if they result in an improvement in node purity that exceeds some large threshold. A commonly cited advantage of pruning is that it can help to reduce overfitting (9). In an overfitted tree, the algorithm is fitted well (i.e., low bias) to the training subsample but is unlikely to perform favorably in external populations.

Random forest: an ensemble approach to CART. While pruning can help to reduce overfitting in CART to some extent, a commonly cited disadvantage of single tree-based algorithms is that they can still be unstable with small perturbations in data, leading to drastic changes in accuracy and error. To counteract such instabilities, ensemble approaches (i.e., those that utilize a combination of algorithms) can be used instead. One such commonly used ensemble approach is known as a *random forest* (25). A random forest is an ensemble of many classification or regression trees.

In a random forest algorithm, bootstrapping is performed, whereby repeated random subsamples of the data set are selected (with replacement) and used to create each of the different trees in the forest. In addition, the variables considered in each splitting step within each tree are randomly selected—hence the term "random" forest (10). Both of these elements are designed to help reduce the variance of predictions without significantly compromising on bias. The predictions that are made across each of the trees are then aggregated (typically through either averaging, majority voting, or some other prespecified approach) in order to make a final prediction for every individual. Random forests have generally been shown to yield superior model performance (i.e., predictive accuracy) in comparison with CART, as they harmonize results from multiple diverse trees that utilize a variety of different variables, variable splits, and observations (25, 26). Furthermore, the set of observations not included in a tree's bootstrapped sample is known as the "out-of-bag sample." The model performance of the random forest can be assessed by making predictions on this out-of-bag sample—that is, by using only the trees that did not include each observation in its bootstrapped sample (22). Cross-validation approaches (e.g., the validation set method and *k*-fold cross-validation) can alternatively be used to assess model performance (22).

After building a random forest, one might also be interested in understanding the variables that were most influential on the predictive model. To investigate this, one can assess "variable importance," which indicates the extent to which a variable contributed to predictive accuracy, node purity, or other relevant measures of model performance. There are several different ways in which variable importance can be measured, with the 2 most common ones being *permutation importance* (27) and *node impurity importance* (28). Regardless of the specific metric used, the most commonly used software implementations of random forest provide a ranked list of the "variable importance factor" (VIF), indicating the variables that most contributed to the model's performance.

Honest causal trees and honest causal forests

Over the past decade, scholars have developed modifications to the CART and random forest algorithms to make them relevant for identifying heterogeneity of treatment effects within samples. One tree-based method in particular, an "honest causal forest," can be used to help identify potential heterogeneous subgroups as well as estimate pointwise consistent CATEs with asymptotically normal confidence intervals (17).

Honest causal forests are conceptually similar to random forests, except that rather than maximizing the *homogeneity* of the *outcome within* nodes, they are designed to maximize the *heterogeneity* of *treatment effects across* nodes (11). An honest causal forest is made up of multiple "honest causal trees," and here "honesty" (also known as cross-fitting, which is utilized in a variety of machine learning methods (29–31)) refers to the requirement that the data used for making splits be distinct from the data that are used to estimate treatment effects. Most importantly, both honest causal trees and honest causal forests rely on the assumption of there being exchangeability within leaf nodes. These 2 algorithms will be further discussed below.

Honest causal trees. An honest causal tree is a single tree algorithm which structurally resembles a classification or regression tree, but it is unique in that it is designed to partition a sample in order to *maximize* heterogeneity of CATEs across the child nodes that result from each partition (11). Unlike a classification or regression tree, which focuses on the prediction of Y, an honest causal tree's main outcome of interest is the CATE at each leaf, which is defined as $E\left[Y^{a=1}|X=\boldsymbol{x}\right] - E\left[Y^{a=0}|X=\boldsymbol{x}\right]$ when the treatment is binary. The vector $\boldsymbol{x}$ that defines a given leaf or stratum is equivalent to the covariate splits that created it, where the goal is to create splits that maximize the heterogeneity of CATEs *between* nodes. However, valid inference of these conditional effects relies on several key assumptions—most notably, the assumption of exchangeability within leaf nodes (specific methods for addressing nonexchangeability are discussed below). If we assume that this exchangeability assumption can be met in the honest causal tree example depicted in Figure 2, along with the other causal assumptions, then the first estimated CATE of −0.19 means that for individuals who are under age 50 years, are negative for the ε4 allele of the apolipoprotein E gene (*APOE-4*), and have uncontrolled levels of low-density lipoprotein cholesterol in midlife, the effect of treatment ($A = 1$) results in an absolute risk reduction of 19 percentage points.

To help estimate valid CATEs, Wager and Athey (17) also encourage the use of an "honest" (i.e., cross-fitting) approach, whereby a random half of the data set (splitting subsample J) is used to build the tree and the other half of the data set (estimating subsample I) is kept for estimating the CATEs at each leaf. When applied to honest causal forests (described below), this process of randomly splitting the data set into 2 subsamples occurs iteratively across each of the causal trees within an honest causal forest. Such an arrangement helps to mitigate overfitting, since each tree utilizes a different subsample.

To be clear, identification of CATEs from observed data requires typical causal assumptions, including *consistency* (where the observed Y is equal to the potential outcome Y^a under the same treatment), *positivity* (where everyone must have a nonzero probability of receiving all possible values of the treatment, conditional on covariates x) (32), and *conditional exchangeability* (equation 2), whereby the potential outcomes $Y^{a=1}$ and $Y^{a=0}$ are independent of the treatment assignment A, conditional on the values of covariates x that define each leaf (11, 33).

Conditional exchangeability is expressed as

$$A \perp\!\!\!\perp (Y^{a=0}, Y^{a=1}) \mid X = \boldsymbol{x}. \tag{2}$$

In equation 2, A is the treatment assignment, $Y^{a=0}$ and $Y^{a=1}$ are potential outcomes, and $\boldsymbol{x}$ is a vector of covariates that defines a subgroup.

In a typical randomized trial setting, the exchangeability assumption can be reasonably satisfied assuming that there

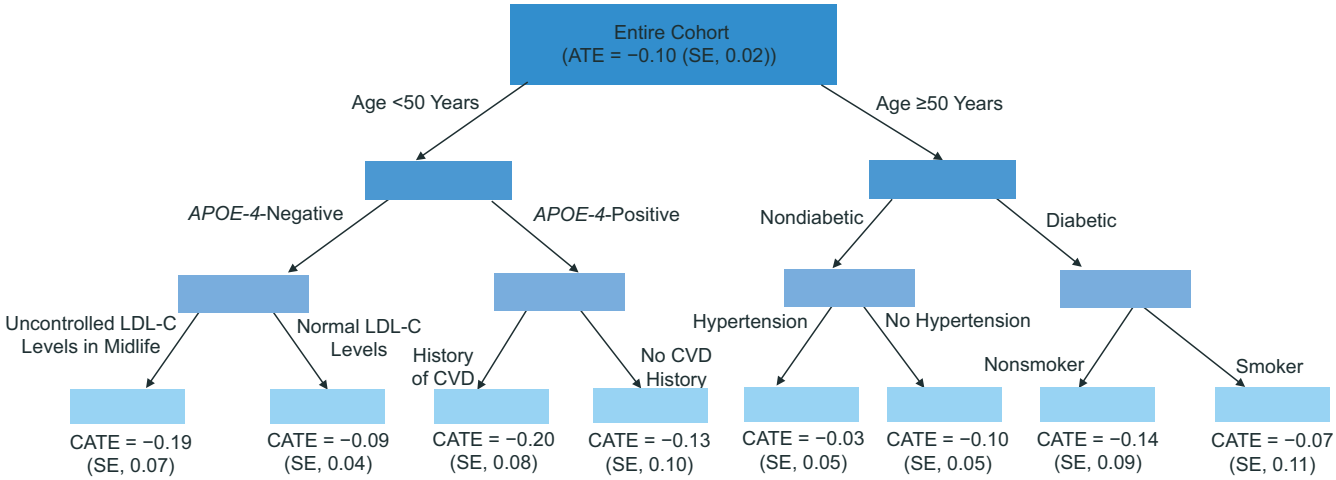

Figure 2. Hypothetical example of an honest causal tree. In an honest causal tree, individual nodes (pictured here as rectangles) are iteratively split into additional nodes, with the goal of maximizing heterogeneity of treatment effects between each of the nodes. Conditional average treatment effects (CATEs) for specific individual(s) can be estimated at the bottom of the tree (in the leaf nodes). *APOE-4*, ε4 allele of the apolipoprotein E gene; ATE, average treatment effect; CVD, cardiovascular disease; LDL-C, low-density lipoprotein cholesterol; SE, standard error.

are no other major sources of bias (e.g., attrition) and given that everyone has an equal probability of receiving the treatment. However, satisfying the assumption of exchangeability is more challenging in observational settings, where treatment assignment is *not* independent of potential outcomes to begin with. In an honest causal tree, the covariate values (x) that define a given node are chosen strictly on the basis of their ability to help maximize heterogeneity of treatment effects, regardless of their influence on achieving conditional exchangeability. Therefore, we cannot reasonably expect that the variables which influence heterogeneity of effects are the same as the variables that also produce conditional exchangeability within strata. However, there exist multiple approaches which can help reduce confounding and hence suggest that conditional exchangeability is closer to being met when estimating treatment effects in an honest causal tree (or honest causal forest).

In one such approach, known as R-learner (derived from Robinson's transformation (34)), rather than estimating the treatment effect of the observed treatment (A) on the observed outcome (Y), one instead models the effect of the *residual* treatment on the *residual* outcome, whereby the residual represents the difference between an individual's observed and predicted values of that variable (as a function of a set of covariates) (29, 34–36). If we assume that the predictive models for A and Y captured variables that comprise a sufficient set to establish conditional exchangeability and that these models were correctly specified (i.e., they eliminate backdoor paths between the exposure and the outcome), then an honest causal tree modeled on the residuals should be expected to isolate the sources of heterogeneity (in the absence of confounding). To further reduce the possibility of violating the conditional exchangeability assumption, one should carefully consider the variables investigated as potential effect modifiers, since stratification on colliders or mediators of the A-Y relationship could potentially threaten this assumption. A supplementary approach that has been applied to address confounding is augmented inverse probability weighting (AIPW); this method combines an inverse-probability-of-treatment weight with a weighted average of the outcome model, with both models being conditional on a set of covariates. In short, AIPW enables doubly robust estimation of treatment effects in honest causal forests. More information about AIPW can be found elsewhere (37–39).

To accomplish its goal of identifying heterogeneous subgroups, the honest causal tree algorithm creates splits by utilizing an expected mean squared error (EMSE) criterion equation, as described by Athey and Imbens (11) (equation 3).

The estimator of the EMSE for an honest causal tree is

$$-\widehat{\mathbf{EMSE}}_{\boldsymbol{\tau}}\left(S^{\mathrm{tr}}, N^{\mathrm{est}}, \Pi\right) = \frac{1}{N^{\mathrm{tr}}} \sum_{i \in S^{\mathrm{tr}}} \hat{\tau}^2\left(X_i; S^{\mathrm{tr}}, \Pi\right) - \left(\frac{1}{N^{\mathrm{tr}}} + \frac{1}{N^{\mathrm{est}}}\right) \times \sum_{l \in \Pi} \left(V_{S^{\mathrm{tr}}}(l)\right). \quad (3)$$

In equation 3, $-\widehat{\mathbf{EMSE}}_{\boldsymbol{\tau}}$ is the estimator of the EMSE of the treatment effect ($\boldsymbol{\tau}$), where $\boldsymbol{\tau}$ is defined as $E[Y_i(1) - Y_i(0) \mid X_i] \in l\,(x; \Pi)$, Y_i is the potential outcome, $\boldsymbol{X_i}$ is the vector of covariates, l is a given leaf node, Π is a specific partition, S^{tr} is the training sample, N^{tr} is the size of the training sample, i is a specific observation within the training sample, N^{est} is the size of the estimating sample, and $V_{S^{\mathrm{tr}}}(l)$ is the within-leaf variance for treated individuals.

Table 2. Difference Between a Regression Tree and an Honest Causal Tree[a]

Regression Tree	Honest Causal Tree
Outcome of interest is Y.	Outcome of interest is the CATE, $E\left[Y^{a=1} \mid X = \mathbf{x}\right] - E\left[Y^{a=0} \mid X = \mathbf{x}\right]$, within a given stratum defined by a vector of covariates.
Most common use case is for prediction.	Most common use case is for detecting effect modifiers and estimating heterogeneous causal effects.
Algorithm is designed to select covariates which minimize prediction error and maximize node purity.	Algorithm is designed to select covariates which maximize heterogeneity in treatment effects between the leaves, while minimizing variance within each treatment effect.
No exchangeability assumption is needed to predict Y.	Conditional exchangeability must be satisfied to validly estimate CATEs.

Abbreviation: CATE, conditional average treatment effect.
[a] Y, potential outcome; $a = 0$, not treated; $a = 1$, treated; $\mathbf{x}$, vector of covariates.

This equation shows that for a given tree's partition, Π, using training sample S^{tr} and an estimation sample of size N^{est}, $-\widehat{\mathbf{EMSE}}_{\boldsymbol{\tau}}$ is an estimator which represents the modified mean squared error that is expected as the result of a specific partition. This estimator is comprised of 2 main terms. First, the term $\frac{1}{N^{\text{tr}}} \sum_{i \in S^{\text{tr}}} \hat{\tau}^2 \left(\boldsymbol{X_i}; S^{\text{tr}}, \Pi\right)$ represents the amount of heterogeneity in treatment effects that exists *across* the different leaves. Furthermore, the *penalty term*, $-\left(\frac{1}{N^{\text{tr}}} + \frac{1}{N^{\text{est}}}\right) \times \sum_{l \in \Pi} \left(V_{S^{\text{tr}}}(l)\right)$, assesses how much variance exists *within* individual leaf (l) estimates. An honest causal tree ultimately chooses the covariate split which maximizes the heterogeneity *between* treatment effects of leaves, while also minimizing the variance of conditional estimates *within* leaves. Because this criterion includes a term that penalizes variance, this will result in a relatively greater number of observations within each leaf, by default, than would otherwise arise.

A summary of some key differences between a regression tree and an honest causal tree can be found in Table 2.

Honest causal forests. An honest causal forest is an ensemble (i.e., a method that combines many models (22)) of multiple honest causal trees. Like a random forest, honest causal forests utilize random samples of observations in each tree, as well as a random selection of covariates at each split. Furthermore, estimation of CATEs in honest causal forests differs slightly from that of honest causal trees. Rather than estimating the CATEs for a subgroup at an individual leaf of the tree (which occurs in an honest causal tree), in an honest causal forest, a CATE for a unique set of covariate values ($\boldsymbol{x}$) is estimated by averaging the treatment effect of that subgroup across all of the trees (17). This estimation occurs by taking one observation or a set of observations and then running each of them through the trees in the honest causal forest which did not utilize them to build the tree (i.e., out-of-bag observations are used to obtain causal estimates). Equation 4 shows how honest causal forests calculate CATEs by averaging estimates, $\hat{\tau}$, across those B honest causal trees. Standard errors for these estimates, along with 95% confidence intervals, can also be computed (17).

CATE estimation in the honest causal forest method is performed as follows.

$$\hat{\boldsymbol{\tau}}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{\tau}}_{\boldsymbol{b}^{\text{est}}}(x) \tag{4}$$

In equation 4, $\hat{\boldsymbol{\tau}}$ is the estimated CATE for individual(s) defined by their covariates ($\boldsymbol{x}$). This estimation occurs by averaging tree-specific CATEs of those individuals ($\hat{\boldsymbol{\tau}}_{\boldsymbol{b}^{\text{est}}}$) across B honest causal trees (with each honest causal tree built on a Bth bootstrapped sample). Here, only trees which did *not* contain such individuals in their training sample are utilized.

While an honest causal tree is easy to visualize (because it is only 1 tree), honest causal forests have been alternatively proposed because of their characteristics, which are more conducive to producing pointwise consistent estimates with asymptotically normal properties (17). An honest causal forest's use of honest splitting (also known as cross-fitting) is intended to address a key requirement for consistent estimation, which is for the outcome and treatment model estimators to not be too adaptive (i.e., the "Donsker condition") (17, 29, 30). Although some literature suggests that honest causal forests may still not yield sufficient coverage of confidence intervals (40), particularly in settings with high dimensions, other simulations have demonstrated that under certain conditions (e.g., $\leq$15 covariates) they can perform better—in terms of bias and confidence interval coverage—than other data-driven methods, such as the k-nearest neighbors method (17). Some advantages and disadvantages of honest causal forests are summarized in Table 3.

Finally, in contrast to random forests, where the VIF typically assesses the extent to which specific variables contributed to the predictive accuracy of the outcome, variable importance in honest causal forests can be calculated by assessing the degree to which each covariate contributed to *heterogeneity* between CATEs for the research question of interest (i.e., the effect of A on Y). Here, variable importance is typically quantified by weight-summing the number of times each variable was used to split the sample throughout

Table 3. Advantages and Disadvantages of Honest Causal Forests

Advantages	Disadvantages
Algorithm identifies unspecified heterogeneous subgroups in an intuitive manner.	Like random forests, not as easy to visualize as a single tree.
Treatment effects are pointwise consistent, with normally distributed and asymptotic confidence intervals.	Correct identification of effect modifiers may depend upon sufficient sample size.
R-learner and AIPW combined with honest causal forest may help satisfy the conditional exchangeability assumption.	Conditional exchangeability may be especially challenging to satisfy in an observational setting.

Abbreviation: AIPW, augmented inverse-propensity weighting.

the forest (weighted by the depth of the tree where each split occurred, with splits near the top of a tree corresponding to larger weights) (41). While variable importance may serve as a useful tool for identifying potential effect modifiers, Athey et al. (42) have also described alternative approaches for summarizing heterogeneity of CATEs learned by honest causal forests. For example, they suggest comparing the average values of covariates within quartiles of the CATEs that were estimated across the honest causal forest.

Implementation in R

In this implementation, we walk through an example application of honest causal forests to simulated data. We provide readers with a list of steps required to simulate data sets, build honest causal forests on those data sets, and assess model performance across a variety of simulation scenarios (see the Web Appendix, available at https://doi.org/10.1093/aje/kwad043). Each simulated data set is comprised of a dichotomous treatment assignment (A), 20 covariates ($\boldsymbol{X}$) including 10 dichotomous variables and 10 continuous variables, and a dichotomous outcome (Y), with no individuals lost to follow-up. The covariates ($\boldsymbol{X}$) include a categorical effect modifier (variable B) and a continuous effect modifier (variable N).

Within our simulations, we generate 2 distinct settings: a randomized controlled trial (RCT) setting and an observational setting. In the RCT setting, in expectation, treatment was independent of all covariates, and in the observational setting we simulated confounding of the treatment-outcome relationship. Within each setting, we run 6 different scenarios of sample size and number of trees, across 2 different types of data sets based on realistic settings: one in which there is high correlation between covariates and small differences between CATEs, and another in which there is low correlation between covariates and large differences between CATEs. These scenarios are selected to cover the default number of trees in the *causal_forest* function of the *grf* package in R (2,000 trees per honest causal forest) and are also grounded in a realistic range of sample sizes (41, 43–45) and covariate correlations (46–48) from prior literature. Furthermore, in our observational setting, we implement each of these scenarios across 2 doubly robust estimators: one in which the propensity models for the treatment and outcome variables are appropriately adjusted for covariates and the other in which those same models are unadjusted. Specifically, for proof-of-concept purposes, the unadjusted models are intentionally misspecified by excluding covariates from them; by ignoring these covariates, we expect there to be bias in the CATEs under study. Overall, we examine a total of 24 observational simulation scenarios and 12 RCT simulation scenarios. For each of these 36 scenarios, we run 1,000 simulations. Web Table 1 displays complete details regarding the parameterization of each simulation scenario.

We build honest causal forests on each simulated data set using the *grf* package (41) and the *causal_forest* function in R (R Foundation for Statistical Computing, Vienna, Austria); in addition, we use R-learner (36) and AIPW (39) to enable doubly robust estimation of treatment effects. The model performance of the honest causal forests built on each of our simulated data sets was assessed by summarizing the extent to which the algorithm 1) correctly identified the covariates (variables B and N) contributing to heterogeneity and 2) accurately estimated the 4 prespecified CATEs, in terms of bias of the point estimates as well as the coverage provided by the confidence intervals. The identification of covariates was assessed using the *grf* package's default variable importance metric—a type of node impurity importance—which weight-sums each variable's number of splits in the causal forest by the depth at which each split occurred. As a supplementary approach to identify potential effect modifiers, we also assess the average values of covariates within quartiles of the CATEs that were estimated across the honest causal forest (42). Meanwhile, the CATEs and corresponding 95% confidence intervals were estimated using the doubly robust AIPW estimator built within the package. Our simulations were run using R 4.1.1; the R software code for this implementation, along with accompanying documentation, can be found on GitHub (49) and in the Web Appendix.

RESULTS

For each of the described scenarios in Web Table 1 (i.e., 12 RCT scenarios and 24 observational scenarios), we simulated 1,000 randomized controlled data sets and built 1,000 corresponding honest causal forests. In Web Table 2, for each of the simulation scenarios, we display ranked lists of the "important" variables (i.e., the covariates most contributing to heterogeneity of treatment effect estimates)

that were calculated on the basis of the VIF averaged across all 1,000 simulations. For the RCT simulations, among scenarios with *high* correlation between covariates and *small* differences in CATEs, 0 out of 2 (0%) simulations using a sample size of 1,000, 2 out of 2 (100%) simulations using a sample size of 10,000, and 2 out of 2 (100%) simulations using a sample size of 40,000 were able to correctly identify the 2 prespecified effect modifiers (variables *B* and *N*) as the 2 variables most highly ranked by the VIF. For comparison, among RCT simulation scenarios with *low* correlation between covariates and *large* differences in CATEs, 6 out of 6 (100%) simulations correctly identified the continuous effect modifier *N*. However, within these same RCT simulations (low correlation and large differences in CATEs), only scenarios using a sample size of 40,000 correctly identified categorical variable *B* as one of the 2 most highly ranked variables.

In the observational setting, 0 out of 8 (0%) simulations using a sample size of 1,000, 0 out of 8 (0%) simulations using a sample size of 10,000, and 8 out of 8 (100%) simulations using a sample size of 40,000 were able to correctly identify both prespecified effect modifiers (variables *B* and *N*). However, the continuous effect modifier (*N*) was identifiable in all of the observational simulations with at least 10,000 individuals in each sample. Furthermore, within these observational scenario simulations, there were limited differences in the model's ability to identify the effect modifiers between the versions where the doubly robust estimator was derived from a covariates-adjusted model versus when it was based on an unadjusted model. However, we caution that these results may be attributable to the simple data-generating mechanism which we implemented for demonstration purposes.

In Web Table 3, we report the CATEs and corresponding 95% confidence intervals (defined across levels of our 2 prespecified effect modifiers) that were estimated, on the average, within each simulation scenario. As shown, across the various RCT simulations, the absolute value of the percent difference between the average observed CATE and the corresponding true CATE never exceeded 18%. Furthermore, among RCT simulations utilizing a sample size of 40,000, the absolute value of the percent difference between the average observed CATE and the corresponding true CATE never exceeded 4%. These RCT-specific results demonstrate the extent to which honest causal forests could accurately estimate CATEs across levels of the prespecified effect modifiers, assuming those effect modifiers (and corresponding a priori hypotheses) could first be identified.

Meanwhile, in the observational simulations where the doubly robust estimator was adjusted, the absolute value of the percent difference between the average observed CATE and the corresponding true CATE never exceeded 19%. However, as anticipated, among the observational simulations where the doubly robust estimator was *unadjusted* (i.e., confounding was not accounted for), this absolute value of the percent difference was as great as 84%. Furthermore, confidence interval coverage performed well in observational simulation scenarios using an adjusted doubly robust estimator; coverage never fell below 86.1% in such scenarios, whereas coverage fell as low as 48.7% in observational scenarios which did not use this adjusted doubly robust estimator.

DISCUSSION

In this practical guide, we discussed the foundations of tree-based algorithms, described the methodological basis of honest causal forests, and walked through an implementation of this method in simulated data. Our conceptual overview and simulations suggest that honest causal forests could be useful for generating new hypotheses on heterogeneous treatment effects in a variety of research settings. In our simulation study, while CATE estimation was biased and coverage probabilities were poor in observational settings using misspecified (i.e., unadjusted) propensity models, such performance improved when appropriately controlling for confounding (through R-learner and AIPW) (36, 39), and they were additionally improved by the presence of lower levels of heterogeneity between the CATEs of interest. Furthermore, given a sufficient sample size, the honest causal forests could successfully identify the true effect modifiers across a variety of randomized and observational simulations. While the honest causal forests performed unexpectedly well at identifying relevant effect modifiers in observational settings that were unadjusted, we acknowledge that these favorable results may be a consequence of the simple data-generating scheme we implemented for demonstration purposes (i.e., to show using a toy example how confounder adjustment can be incorporated into the approach to reduce bias in the estimated CATEs).

Our paper expands on the existing literature in a variety of ways. First, we offer a practical guide to honest causal forests, tailored for epidemiologists and other health researchers with limited prior knowledge of machine learning. Second, we explored the practical operating properties of this method in a wide range of scenarios (with varying levels of heterogeneity, correlations between covariates, sample sizes, doubly robust estimators, and research settings). Further, in contrast to Athey et al. (11, 50), we assessed the extent to which the method could correctly identify effect modifiers, which may help inform the specific types of scenarios in which honest causal forests may perform best (in terms of CATE estimation and effect-modifier discovery). For example, we demonstrate that sample size has more of an impact on performance than the number of trees; in all of the scenarios in which there were 40,000 individuals, honest causal forests could correctly identify the 2 prespecified effect modifiers of interest. We also showed that in certain settings (e.g., a sample size of 10,000), we were better able to identify the continuous effect modifier (*N*) than the categorical effect modifier (*B*). Although one limitation of this study is that we only explored scenarios in which there were 20 covariates, Athey et al. have already demonstrated the performance of estimating CATEs across varying numbers of dimensions (50). Future work could also benefit from explorations into how the number of dimensions additionally affects one's ability to correctly identify effect modifiers. Lastly, we acknowledge that while we performed a variety of simulations in a number of scenarios, we cannot practically guarantee that performance from our study will reflect how

honest causal forests will perform in the real world, especially in observational settings, given that, for illustrative purposes, we used simplistic data-generating mechanisms.

The unique contribution of honest causal forests is that they offer a rigorous approach to the identification of potential treatment effect heterogeneity when the goal is hypothesis generation. This tree-based algorithm avoids the problem of multiple hypothesis-testing altogether by using an iterative partitioning rule to select the covariates that maximize heterogeneity of effects. This feature makes it a particularly appropriate method in research settings where there is limited understanding of heterogeneity of effects, or in scenarios where there exist numerous covariates for which it would be overly burdensome to conduct traditional subgroup analyses across each variable. In our simulation study, although we used a simple data-generating scheme, we showed that the effect modifiers of interest could be correctly identified (using the VIF metric) across multiple randomized and observational scenarios. Even still, we acknowledge that in practice, VIF should not be directly interpreted as the ground truth for causal effect modification, particularly because variables correlated with true effect modifiers could still plausibly be picked (e.g., we found that covariates which were correlated with the "true" effect modifiers were often found near the top of the VIF lists). As such, we caution that our approach for identifying potential effect modifiers should be considered a hypothesis-generation tool and not a causal discovery tool. Additionally, to augment our approach for identifying potential effect modifiers, which may be vulnerable to spurious effect modification, in our tutorial we demonstrate a supplemental method for identifying potential effect modifiers recommended by Athey et al. (42), whereby the average values of covariates within quartiles of the estimated CATEs are visualized (Web Figure 1).

Amidst the rise of "big data" science and high-dimensional data sets, honest causal forests hold great potential for helping researchers understand how the effect of a treatment may vary across a population. Yet, honest causal forests have not been widely adopted by epidemiologists thus far, perhaps because it is a newly developed and complex algorithm, and few examples of its implementation exist in the field, particularly using observational data. To address this gap, this paper serves as a practical resource guide for researchers seeking to better understand how honest causal forests work, prior to applying them towards the identification of potential effect modifiers. While underutilized in public health and the medical sciences thus far, honest causal forests offer promise for the future. We hope our walk-through of this method helps to inspire greater awareness and adoption of honest causal forests, which could ultimately lead to important discoveries in health care.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York, United States (Neal Jawadekar, Katrina Kezios, Jeanette A. Stingone, Kara Rudolph, Adina Zeki Al Hazzouri); Department of Epidemiology and Population Health, School of Medicine, Stanford University, Stanford, California, United States (Michelle C. Odden); and Department of Health Policy and Management, Mailman School of Public Health, Columbia University, New York, New York, United States (Sebastian Calonico).

K.R. and A.Z.A.H. share senior authorship.

The simulated data utilized for this study can be reproduced using the R software code located on GitHub (https://github.com/njawadekar/Simulation-HCF), as well as in our Web Appendix.

We thank Dr. Erik Sverdup for responding to several questions regarding the *grf* package.

This work was presented at a Methods for Longitudinal Studies in Dementia (MELODEM) working group meeting (virtual), December 16, 2021.

The views expressed in this article are those of the authors.

Conflict of interest: none declared.

REFERENCES

1. Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health.* 2004;58(4):265–271.
2. Chasman DI, Posada D, Subrahmanyan L, et al. Pharmacogenetic study of statin therapy and cholesterol reduction. *JAMA.* 2004;291(23):2821–2827.
3. Donnelly LA, Doney AS, Dannfald J, et al. A paucimorphic variant in the HMG-CoA reductase gene is associated with lipid-lowering response to statin treatment in diabetes: a GoDARTS study. *Pharmacogenet Genomics.* 2008;18(12):1021–1026.
4. Elens L, Becker ML, Haufroid V, et al. Novel *CYP3A4* intron 6 single nucleotide polymorphism is associated with simvastatin-mediated cholesterol reduction in the Rotterdam Study. *Pharmacogenet Genomics.* 2011;21(12):861–866.
5. Fiegenbaum M, da Silveira FR, Van der Sand CR, et al. The role of common variants of *ABCB1*, *CYP3A4*, and *CYP3A5* genes in lipid-lowering efficacy and safety of simvastatin treatment. *Clin Pharmacol Ther.* 2005;78(5):551–558.
6. Collier DA, Ferreira IATM, Kotagiri P, et al. Age-related immune response heterogeneity to SARS-CoV-2 vaccine BNT162b2. *Nature.* 2021;596(7872):417–422.
7. Lozano-Ojalvo D, Camara C, Lopez-Granados E, et al. Differential effects of the second SARS-CoV-2 mRNA vaccine dose on T cell immunity in naive and COVID-19 recovered individuals. *Cell Rep.* 2021;36(8):109570.
8. VanderWeele TJ, Luedtke AR, van der Laan MJ, et al. Selecting optimal subgroups for treatment using many covariates. *Epidemiology.* 2019;30(3):334–341.
9. Venkatasubramaniam A, Wolfson J, Mitchell N, et al. Decision trees in epidemiological research. *Emerg Themes Epidemiol.* 2017;14(1):11.
10. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods.* 2009;14(4):323–348.

11. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proc Natl Acad Sci.* 2016;113(27):7353–7360.
12. Odden MC, Rawlings AM, Khodadadi A, et al. Heterogeneous exposure associations in observational cohort studies: the example of blood pressure in older adults. *Am J Epidemiol.* 2020;189(1):55–67.
13. Du J, Linero AR. Interaction detection with Bayesian decision tree ensembles. Presented at the *22nd International Conference on Artificial Intelligence and Statistics*, Naha, Okinawa, Japan, April 16–18, 2019.
14. Su X, Peña AT, Liu L, et al. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Stat Med.* 2018;37(17):2547–2560.
15. Yang J, Dahabreh IJ, Steingrimsson JA. Causal interaction trees: tree-based subgroup identification for observational data [preprint]. *arXiv.* 2020. (https://doi.org/10.48550/arXiv.2003.03042). Accessed March 1, 2020.
16. Hu L, Ji J, Li F. Estimating heterogeneous survival treatment effect in observational data using machine learning. *Stat Med.* 2021;40(21):4691–4713.
17. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc.* 2018;113(523):1228–1242.
18. Baum A, Scarpa J, Bruzelius E, et al. Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD Trial. *Lancet Diabetes Endocrinol.* 2017;5(10):808–815.
19. Scarpa J, Bruzelius E, Doupe P, et al. Assessment of risk of harm associated with intensive blood pressure management among patients with hypertension who smoke: a secondary analysis of the Systolic Blood Pressure Intervention Trial. *JAMA Netw Open.* 2019;2(3):e190005–e190005.
20. Raghavan S, Josey K, Bahn G, et al. Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control. *Ann Epidemiol.* 2021;65:101–108.
21. Breiman L, Freidman JH, Olshen RA, et al. *CART: Classification and Regression Trees*. 1st ed. New York, NY: Taylor & Francis Group; 1984.
22. James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning: With Applications in R*. 1st ed. New York, NY: Springer Publishing Company; 2013.
23. Reis I, Baron D, Shahaf S. Probabilistic random forest: a machine learning algorithm for noisy datasets. *Astron J.* 2018;157(1):16.
24. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry.* 2015;27(2):130–135.
25. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32.
26. Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn.* 2003;51(2):181–207.
27. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 1st ed. Munich, Germany: Leanpub; 2019.
28. Louppe G, Wehenkel L, Sutera A, et al. Understanding variable importances in forests of randomized trees. Presented at the *27th Conference on Neural Information Processing Systems (NIPS)*, Stateline, Nevada, December 5–10, 2013.
29. Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased machine learning for treatment and structural parameters. *Econom J.* 2018;21(1):C1–C68.
30. Balzer LB, Westling T. Demystifying statistical inference when using machine learning in causal research [published online ahead of print July 15, 2021]. *Am J Epidemiol.* . https://doi.org/10.1093/aje/kwab200.
31. van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York, NY: Springer Publishing Company; 2011.
32. Hernán MA. Beyond exchangeability: the other conditions for causal inference in medical research. *Stat Methods Med Res.* 2012;21(1):3–5.
33. Knaus MC, Lechner M, Strittmatter A. Machine learning estimation of heterogeneous causal effects: empirical Monte Carlo evidence. *Econom J.* 2021;24(1):134–161.
34. Robinson PM. Root-N-consistent semiparametric regression. *Econometrica.* 1988;56(4):931–954.
35. Oprescu M, Syrgkanis V, Wu Z. Orthagonal random forest for causal inference. Presented at the *36th International Conference on Machine Learning (ICML)*, Long Beach, California, June 9–15, 2019.
36. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika.* 2021;108(2):299–319.
37. Kurz CF. Augmented inverse probability weighting and the double robustness property. *Med Decis Making.* 2021;42(2):156–167.
38. Athey S, Wager S. Estimating treatment effects with causal forests: an application. *Obs Stud.* 2019;5(2):37–51.
39. Robins J, Rotnitzky AG, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc.* 1994;89(427):846–866.
40. Lei L, Candès EJ. Conformal inference of counterfactuals and individual treatment effects. *J R Stat Soc Series B Stat Methodol.* 2021;83(5):83–938.
41. Tibshirani J. Package 'grf'. (Version 2.2.0). https://cran.r-project.org/web/packages/grf/index.html. Published August 6, 2022. Accessed September 6, 2022.
42. Athey S, Wager S, Hadad V, et al. Estimation of heterogeneous treatment effects. (Prepared for the class "Machine Learning and Causal Inference"). https://gsbdbi.github.io/ml_tutorial/hte_tutorial/hte_tutorial.html. Published May 7, 2020. Accessed March 1, 2022.
43. Chou R, Dana T, Blazina I, et al. Statins for prevention of cardiovascular disease in adults: evidence report and systematic review for the US Preventive Services Task Force. *JAMA.* 2016;316(19):2008–2024.
44. Thompson AM, Hu T, Eshelbrenner CL, et al. Antihypertensive treatment and secondary prevention of cardiovascular disease events among persons without hypertension: a meta-analysis. *JAMA.* 2011;305(9):913–922.
45. McDonald I, Murray SM, Reynolds CJ, et al. Comparative systematic review and meta-analysis of reactogenicity, immunogenicity and efficacy of vaccines against SARS-CoV-2. *NPJ Vaccines.* 2021;6(1):74.
46. Walters SJ, Jacques RM, Dos Anjos Henriques-Cadby IB, et al. Sample size estimation for randomised controlled trials with repeated assessment of patient-reported outcomes: what correlation between baseline and follow-up outcomes should we assume? *Trials.* 2019;20(1):566–566.
47. Polack FP, Thomas SJ, Kitchin N, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N Engl J Med.* 2020;383(27):2603–2615.
48. Singh BM, Lamichhane HK, Srivatsa SS, et al. Role of statins in the primary prevention of atherosclerotic cardiovascular disease and mortality in the population with mean cholesterol

in the near-optimal to borderline high range: a systematic review and meta-analysis. *Adv Prev Med.* 2020;2020: 6617905.
49. Jawadekar N, Kezios K, Odden M, et al. Simulation-HCF: R code for running honest causal forests on simulated data. (Version 1.0.0). https://github.com/njawadekar/Simulation-HCF. Published November 24, 2022. Accessed November 24, 2022.
50. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Stat.* 2019;47(2):1179–1203.