Recursive partitioning for heterogeneous causal effects

Susan Athey[a,1] and Guido Imbens[a]

[a]Stanford Graduate School of Business, Stanford University, Stanford, CA 94305

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved May 20, 2016 (received for review June 25, 2015)

In this paper we propose methods for estimating heterogeneity in causal effects in experimental and observational studies and for conducting hypothesis tests about the magnitude of differences in treatment effects across subsets of the population. We provide a data-driven approach to partition the data into subpopulations that differ in the magnitude of their treatment effects. The approach enables the construction of valid confidence intervals for treatment effects, even with many covariates relative to the sample size, and without "sparsity" assumptions. We propose an "honest" approach to estimation, whereby one sample is used to construct the partition and another to estimate treatment effects for each subpopulation. Our approach builds on regression tree methods, modified to optimize for goodness of fit in treatment effects and to account for honest estimation. Our model selection criterion anticipates that bias will be eliminated by honest estimation and also accounts for the effect of making additional splits on the variance of treatment effect estimates within each subpopulation. We address the challenge that the "ground truth" for a causal effect is not observed for any individual unit, so that standard approaches to cross-validation must be modified. Through a simulation study, we show that for our preferred method honest estimation results in nominal coverage for 90% confidence intervals, whereas coverage ranges between 74% and 84% for nonhonest approaches. Honest estimation requires estimating the model with a smaller sample size; the cost in terms of mean squared error of treatment effects for our preferred method ranges between 7–22%.

heterogeneous treatment effects | causal inference | cross-validation | supervised machine learning | potential outcomes

In this paper we study two closely related problems: first, estimating heterogeneity by covariates or features in causal effects in experimental or observational studies, and second, conducting inference about the magnitude of the differences in treatment effects across subsets of the population. Causal effects, in the Rubin causal model or potential outcome framework we use here (1–3), are comparisons between outcomes we observe and counterfactual outcomes we would have observed under a different regime or treatment. We introduce data-driven methods that select subpopulations to estimate treatment effect heterogeneity and to test hypotheses about the differences between the effects in different subpopulations. For experiments, our method allows researchers to identify heterogeneity in treatment effects that was not specified in a preanalysis plan, without concern about invalidating inference due to searching over many possible partitions.

Our approach is tailored for applications where there may be many attributes of a unit relative to the number of units observed, and where the functional form of the relationship between treatment effects and the attributes of units is not known. The supervised machine learning literature (e.g., ref. 4) has developed a variety of effective methods for a closely related problem, the problem of predicting outcomes as a function of covariates in similar environments. The most popular approaches [e.g., regression trees (5), random forests (6), LASSO (7), support vector machines (8), etc.] entail building a model of the relationship between attributes and outcomes, with a penalty parameter that penalizes model complexity. Cross-validation is often used to select the optimal level of complexity (the one that maximizes predictive power without "overfitting").

Within the prediction-based machine learning literature, regression trees differ from most other methods in that they produce a partition of the population according to covariates, whereby all units in a partition receive the same prediction. In this paper, we focus on the analogous goal of deriving a partition of the population according to treatment effect heterogeneity, building on standard regression trees (5, 6). Whether the ultimate goal in an application is to derive a partition or fully personalized treatment effect estimates depends on the setting; settings where partitions may be desirable include those where decision rules must be remembered, applied, or interpreted by human beings or computers with limited processing power or memory. Examples include treatment guidelines to be used by physicians or even online personalization applications where having a simple lookup table reduces latency for the user. We show that an attractive feature of focusing on partitions is that we can achieve nominal coverage of confidence intervals for estimated treatment effects even in settings with a modest number of observations and many covariates. Our approach has applicability even for settings such as clinical trials of drugs with only a few hundred patients, where the number of patient characteristics is potentially quite large. Our method may also be viewed as a complement to the use of "preanalysis plans" where the researcher must commit in advance to the subgroups that will be considered. It enables researchers to let the data discover relevant subgroups while preserving the validity of confidence intervals constructed on treatment effects within subgroups.

A first challenge for our goal of finding a partition and then testing hypotheses about treatment effects is that many existing machine learning methods cannot be used directly for constructing confidence intervals. This is because the methods are "adaptive": They use the training data for model selection, so that spurious correlations between covariates and outcomes affect the selected model, leading to biases that disappear only slowly as the sample size grows. In some contexts, additional assumptions such as "sparsity" (only a few covariates affect the outcomes) can be applied to guarantee consistency or asymptotic (large sample) normality of predictions (9). In this paper, we use an alternative approach that places no restrictions on model complexity, which we refer to as "honesty." We say that a model is "honest" if it does not use the same information for selecting the model structure (in our case, the partition of the covariate space) as for estimation given a model structure. We accomplish this by splitting the training sample into two parts, one for constructing the tree (including the cross-validation step) and a second for estimating treatment effects within leaves of the tree.

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Drawing Causal Inference from Big Data," held March 26–27, 2015, at the National Academies of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Big-data.

Author contributions: S.A. and G.I. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

Conflict of interest statement: The authors received funding from Microsoft Research.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: athey@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1510489113/-/DCSupplemental.

Honesty has the implication that the asymptotic properties of treatment effect estimates within the partitions are the same as if the partition had been exogenously given. Although there is a loss of precision due to sample splitting (which reduces sample size in each step of estimation), there is a benefit in terms of eliminating bias that offsets at least part of the cost.

A key contribution of this paper is to show that criteria for both constructing the partition and cross-validation change when we anticipate honest estimation. In the first stage of estimation, the criterion is the expectation of the mean squared error (MSE) when treatment effects are reestimated in the second stage. Crucially, we anticipate that second-stage estimates of treatment effects will be unbiased in each leaf, because they will be performed on an independent sample. In that case, splitting and cross-validation criteria are adjusted to ignore systematic bias in estimation and focus instead on the tradeoff between more tailored prediction (smaller leaf size) and the variance that will arise in the second (honest estimation) stage due to noisy estimation within small leaves.

A second and perhaps more fundamental challenge to applying machine learning methods such as regression trees (5) off-the-shelf to the problem of causal inference is that regularization approaches based on cross-validation typically rely on observing the "ground truth," that is, actual outcomes in a cross-validation sample. However, if our goal is to minimize the MSE of treatment effects, we encounter what Holland (2) calls the "fundamental problem of causal inference": The causal effect is not observed for any individual unit, and so we do not directly have a ground truth. We address this by proposing approaches for constructing unbiased estimates of the MSE of the causal effect of the treatment.

Using theoretical arguments and a simulation exercise, we compare our approach with previously proposed ones. Relative to approaches that focus on goodness of fit in model selection, our approach yields substantial improvements in the MSE of treatment effects (ranging from 43% to 210%). We also examine the costs and benefits of honest estimation relative to adaptive estimation. In the settings we consider, honest estimation leads to approximately nominal coverage of confidence intervals across estimation methods and settings, whereas for adaptive estimation approaches coverage can be as low as 69%. The cost of honest estimation in terms of MSE of treatment effects (where for adaptive estimation, we have a larger sample size available for training) ranges from 7% to 22% for our preferred model.

The Problem

Setup. We consider a setup where there are N units, indexed by $i = 1, \ldots, N$. We postulate the existence of a pair of potential outcomes for each unit, $(Y_i(0), Y_i(1))$, following the potential outcome or Rubin causal model (1–3), with the unit-level causal effect defined as the difference in potential outcomes, $\tau_i = Y_i(1) - Y_i(0)$. Let $W_i \in \{0,1\}$ be the binary indicator for the treatment, with $W_i = 0$ indicating that unit i received the control treatment and $W_i = 1$ indicating that unit i received the active treatment. The realized outcome for unit i is the potential outcome corresponding to the treatment received:

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Let X_i be a K-component vector of features, covariates, or pretreatment variables, known not to be affected by the treatment. Our data consist of the triple $(Y_i^{\text{obs}}, W_i, X_i)$, for $i = 1, \ldots, N$, which are regarded as an independent and identically distributed sample drawn from a large population. Expectations and probabilities will refer to the distribution induced by the random sampling, or by the (conditional) random assignment of the treatment. We assume that observations are exchangeable, and that there is no interference [the stable unit treatment value assumption (10)]. This assumption may be violated in settings where some units are connected through networks. Let $p = \text{pr}(W_i = 1)$ be the marginal treatment probability, and let $e(x) = \text{pr}(W_i = 1 | X_i = x)$ be the conditional treatment probability (the "propensity score" as defined by ref. 11). In a randomized experiment with constant treatment assignment probabilities $e(x) = p$ for all values of x.

Unconfoundedness. Throughout the paper, we maintain the assumption of randomization conditional on the covariates, or "unconfoundedness" (11), formalized as given below.

Assumption 1 (Unconfoundedness).

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \mid X_i,$$

using the symbol $\perp\!\!\!\perp$ to denote (conditional) independence of two random variables. This assumption is satisfied in a randomized experiment without conditioning on covariates but also may be justified in observational studies if the researcher is able to observe all of the variables that affect the unit's receipt of treatment and are associated with the potential outcomes.

To simplify exposition, in the main body of the paper we maintain the stronger assumption of complete randomization, whereby $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1), X_i)$. Later, we show that by using propensity score weighting (1) we can adapt all of the methods to that case.

Conditional Average Treatment Effects and Partitioning. Define the conditional average treatment effect

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x].$$

A large part of the causal inference literature (e.g., refs. 3 and 12–14) is focused on estimating the population (marginal) average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)]$. The main focus of the current paper is on obtaining accurate estimates of and inferences for the conditional average treatment effect $\tau(x)$. We are interested in estimators $\hat{\tau}(\cdot)$ [in general we use the $\hat{\cdot}$ symbol to denote estimators for a population quantity—in this case $\tau(x)$] that are based on partitioning the feature space and do not vary within the partitions.

Honest Inference for Population Averages

Our approach departs from conventional classification and regression trees (CART) in two fundamental ways. First, we focus on estimating conditional average treatment effects rather than predicting outcomes. Conventional regression tree methods are therefore not directly applicable because we do not observe unit-level causal effects for any unit. Second, we impose a separation between constructing the partition and estimating effects within leaves of the partition, using separate samples for the two tasks, in what we refer to as honest estimation. We contrast honest estimation with adaptive estimation used in conventional CART, where the same data are used to build the partition and estimate leaf effects. In this section we introduce the changes induced by honest estimation in the context of the conventional prediction setting; in the next section we consider causal effects. In the discussion in this section we observe for each unit i a pair of variables (Y_i, X_i), with the interest in the conditional expectation $\mu(x) \equiv \mathbb{E}[Y_i | X_i = x]$.

Setup. We begin by defining key concepts and functions. First, a tree or partitioning Π corresponds to a partitioning of the feature space $\mathbb{X}$, with $\#(\Pi)$ the number of elements in the partition. We write

$$\Pi = \{\ell_1, \ldots, \ell_{\#(\Pi)}\}, \text{ with } \cup_{j=1}^{\#(\Pi)} \ell_j = \mathbb{X}.$$

Let $\mathbb{P}$ denote the space of partitions. Let $\ell(x; \Pi)$ denote the leaf $\ell \in \Pi$ such that $x \in \ell$. Let $\mathbb{S}$ be the space of data samples from a

population. Let $\pi : \mathbb{S} \mapsto \mathbb{P}$ be an algorithm that on the basis of a sample $\mathcal{S} \in \mathbb{S}$ constructs a partition. As a very simple example, suppose the feature space is $\mathbb{X} = \{L, R\}$. In this case there are two possible partitions, $\Pi_N = \{L, R\}$ (no split) or $\Pi_S = \{\{L\}, \{R\}\}$ (full split), and so the space of trees is $\mathbb{P} = \{\Pi_N, \Pi_S\} = \{\{L, R\}, \{\{L\}, \{R\}\}\}$. Given a sample $\mathcal{S}$, the average outcomes in the two subsamples are $\bar{Y}_L$ and $\bar{Y}_R$. A simple example of an algorithm is one that splits if the difference in average outcomes exceeds a threshold c:

$$\pi(\mathcal{S}) = \begin{cases} \{\{L, R\}\} & \text{if } \bar{Y}_L - \bar{Y}_R \leq c, \\ \{\{L\}, \{R\}\} & \text{if } \bar{Y}_L - \bar{Y}_R > c. \end{cases}$$

The potential bias in leaf estimates from adaptive estimation can be seen in this simple example. Whereas $\bar{Y}_L - \bar{Y}_R$ is in general an unbiased estimator for the difference in the population conditional means $\mu(L) - \mu(R)$, if we condition on finding that $\bar{Y}_L - \bar{Y}_R \geq c$ in a particular sample, we expect that $\bar{Y}_L - \bar{Y}_R$ is larger than the population analog.

Given a partition Π, define the conditional mean function $\mu(x; \Pi)$ as

$$\mu(x; \Pi) \equiv \mathbb{E}[Y_i | X_i \in \ell(x; \Pi)] = \mathbb{E}[\mu(X_i) | X_i \in \ell(x; \Pi)],$$

which can be viewed as a step-function approximation to $\mu(x)$. Given a sample $\mathcal{S}$ the estimated counterpart is

$$\hat{\mu}(x; \mathcal{S}, \Pi) \equiv \frac{1}{\#(i \in \mathcal{S} : X_i \in \ell(x; \Pi))} \sum_{i \in \mathcal{S} : X_i \in \ell(x; \Pi)} Y_i,$$

which is unbiased for $\mu(x; \Pi)$. We index this estimator by the sample because we need to be precise about which sample is used for estimation of the regression function.

The Honest Target. A central concern in this paper is the criterion used to compare alternative estimators; following much of the literature, we focus on MSE criteria, but we will modify these criteria in a variety of ways.

For the prediction case, we adjust the MSE by $\mathbb{E}[Y_i^2]$; because this does not depend on an estimator, subtracting it does not affect how the criterion ranks estimators. Given a partition Π, define the MSE, where we average over a test sample $\mathcal{S}^{\text{te}}$ and the conditional mean is estimated on an estimation sample $\mathcal{S}^{\text{est}}$, as

$$\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi) \equiv \frac{1}{\#(\mathcal{S}^{\text{te}})} \sum_{i \in \mathcal{S}^{\text{te}}} \left\{ \left(Y_i - \hat{\mu}(X_i; \mathcal{S}^{\text{est}}, \Pi) \right)^2 - Y_i^2 \right\}.$$

The (adjusted) expected MSE is the expectation of $\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)$ over test and estimation samples:

$$\text{EMSE}_\mu(\Pi) \equiv \mathbb{E}_{\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}} \left[\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi) \right],$$

where the test and estimation samples are independent. In the algorithms we consider, we will consider a variety of estimators for the (adjusted) EMSE, all of which take the form of MSE estimators $\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)$, evaluated at the units in sample $\mathcal{S}^{\text{te}}$, with the estimates based on sample $\mathcal{S}^{\text{est}}$ and the tree Π. For brevity in this paper we will henceforth omit the term "adjusted" and abuse terminology slightly by referring to these objects as MSE functions.

Our ultimate goal is to construct and assess algorithms $\pi(\cdot)$ that maximize the honest criterion

$$Q^H(\pi) \equiv -\mathbb{E}_{\mathcal{S}^{\text{est}}, \mathcal{S}^{\text{est}}, \mathcal{S}^{\text{tr}}} \left[\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi(\mathcal{S}^{\text{tr}})) \right].$$

Note that throughout the paper we focus on maximizing criterion functions, which typically involve the negative of MSE expressions.

The Adaptive Target. In the conventional CART approach the target is slightly different:

$$Q^C(\pi) \equiv -\mathbb{E}_{\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{tr}}} \left[\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{tr}}, \pi(\mathcal{S}^{\text{tr}})) \right],$$

where the same training sample is used to construct and estimate the tree. Compared with our target $Q^H(\pi)$ the difference is that in our approach different samples $\mathcal{S}^{\text{tr}}$ and $\mathcal{S}^{\text{est}}$ are used for construction of the tree and estimation of the conditional means, respectively.

We refer to the conventional CART approach as adaptive and to our approach as honest. In practice there will be costs and benefits of the honest approach relative to the adaptive approach. The cost is sample size; given a dataset, putting some data in the estimation sample leaves fewer units for the training dataset, leading to higher expected MSE. The advantage of honest estimation is that it avoids a problem of adaptive estimation, which is that spurious extreme values of Y_i are likely to be placed into the same leaf as other extreme values by the algorithm $\pi(\cdot)$, and thus the sample means (in sample $\mathcal{S}^{\text{tr}}$) of the elements of $\pi(\mathcal{S}^{\text{tr}})$ are more extreme than they would be in an independent sample. This shows up in the poor coverage properties of confidence intervals for adaptive estimation methods relative to the honest methods.

The Implementation of CART. There are two distinct parts of the conventional CART algorithm, initial tree building and cross-validation to select a complexity parameter used for pruning. Each part of the algorithm relies on a criterion function based on MSE. In this paper we will take as given the overall structure of the CART algorithm (e.g., refs. 4 and 5), and our focus will be on modifying the criteria.

In the tree-building phase, CART recursively partitions the observations of the training sample. For each leaf, the algorithm evaluates all candidate splits of that leaf (which induce alternative partitions Π) using a "splitting" criterion that we refer to as the "in-sample" goodness-of-fit criterion $-\text{MSE}_\mu(\mathcal{S}^{\text{tr}}, \mathcal{S}^{\text{tr}}, \Pi)$.

It is well understood that the conventional criterion leads to overfitting, a problem that is solved by cross-validation to select a penalty on tree depth. The in-sample goodness-of-fit criterion will always improve with additional splits, even though additional refinements of a partition Π might in fact increase the expected MSE, especially when the leaf sizes become small. The reason is that the criterion ignores the fact that smaller leaves lead to higher-variance estimates of leaf means. To account for this factor, the conventional approach to avoiding overfitting is to add a penalty term to the criterion that is equal to a constant times the number of splits, so that essentially we only consider splits where the improvement in a goodness-of-fit criterion is above some threshold. The penalty term is chosen to maximize a goodness-of-fit criterion in cross-validation samples. In the conventional cross-validation the training sample is repeatedly split into two subsamples, the $\mathcal{S}^{\text{tr,tr}}$ sample that is used to build a new tree as well as estimate the conditional means and the $\mathcal{S}^{\text{tr,cv}}$ sample that is used to evaluate the estimates. We "prune" the tree using a penalty parameter that represents the cost of a leaf. We choose the optimal penalty parameter by evaluating the trees associated with each value of the penalty parameter. The goodness-of-fit criterion for cross-validation can be written as $-\text{MSE}_\mu(\mathcal{S}^{\text{tr,cv}}, \mathcal{S}^{\text{tr,tr}}, \Pi)$. Note that the cross-validation criterion directly addresses the issue we highlighted with the in-sample goodness-of-fit criterion, because $\mathcal{S}^{\text{tr,cv}}$ is independent of $\mathcal{S}^{\text{tr,tr}}$, and thus too-extreme estimates of leaf means will be penalized. The issue that smaller leaves lead to noisier estimates of leaf means is implicitly incorporated by the fact that a smaller leaf penalty will lead to deeper trees and thus smaller leaves, and the noisier estimates will lead to larger average MSE across the cross-validation samples.

Honest Splitting. In our honest estimation algorithm, we modify CART in two ways. First, we use an independent sample $\mathcal{S}^{\text{est}}$

instead of $\mathcal{S}^{\text{tr}}$ to estimate leaf means. Second (and closely related), we modify our splitting and cross-validation criteria to incorporate the fact that we will generate unbiased estimates using $\mathcal{S}^{\text{est}}$ for leaf estimation (eliminating one aspect of overfitting), where $\mathcal{S}^{\text{est}}$ is treated as a random variable in the tree building phase. We explicitly incorporate the fact that finer partitions generate greater variance in leaf estimates.

To begin developing our criteria, let us expand $\text{EMSE}_\mu(\Pi)$:

$$\begin{aligned}-\text{EMSE}_\mu(\Pi) &= -\mathbb{E}_{(Y_i,X_i),\mathcal{S}^{\text{est}}}\left[(Y_i-\mu(X_i;\Pi))^2 - Y_i^2\right] \\ &\quad -\mathbb{E}_{X_i,\mathcal{S}^{\text{est}}}\left[(\hat{\mu}(X_i;\mathcal{S}^{\text{est}},\Pi)-\mu(X_i;\Pi))^2\right] \\ &= \mathbb{E}_{X_i}\left[\mu^2(X_i;\Pi)\right] - \mathbb{E}_{\mathcal{S}^{\text{est}},X_i}\left[\mathbb{V}\left(\hat{\mu}^2(X_i;\mathcal{S}^{\text{est}},\Pi)\right)\right],\end{aligned}$$

where we exploit the equality $\mathbb{E}_{\mathcal{S}}[\hat{\mu}(x;\mathcal{S},\Pi)]=\mu(x;\Pi)$.

We wish to estimate $-\text{EMSE}_\mu(\Pi)$ on the basis of the training sample $\mathcal{S}^{\text{tr}}$ and knowledge of the sample size of the estimation sample N^{est}. To construct an estimator for the second term, observe that within each leaf of the tree there is an unbiased estimator for the variance of the estimated mean in that leaf. Specifically, to estimate the variance of $\hat{\mu}(x;\mathcal{S}^{\text{est}},\Pi)$ on the training sample we can use

$$\hat{\mathbb{V}}(\hat{\mu}(x;\mathcal{S}^{\text{est}},\Pi)) \equiv \frac{S^2_{\mathcal{S}^{\text{tr}}}(\ell(x;\Pi))}{N^{\text{est}}(\ell(x;\Pi))},$$

where $S^2_{\mathcal{S}^{\text{tr}}}(\ell)$ is the within-leaf variance, to estimate the variance. We then weight this by the leaf shares p_ℓ to estimate the expected variance. Assuming the leaf shares are approximately equal in the estimation and training samples, we can approximate this variance estimator by

$$\hat{\mathbb{E}}\left[\mathbb{V}\left(\hat{\mu}^2(X_i;\mathcal{S}^{\text{est}},\Pi)\right)\middle| i\in\mathcal{S}^{\text{te}}\right] \equiv \frac{1}{N^{\text{est}}}\cdot\sum_{\ell\in\Pi} S^2_{\mathcal{S}^{\text{tr}}}(\ell).$$

To estimate the average of the squared outcome $\mu^2(x;\Pi)$ (the first term of the target criterion), we can use the square of the estimated means in the training sample $\hat{\mu}^2(x;\mathcal{S}^{\text{tr}},\Pi)$, minus an estimate of its variance,

$$\hat{\mathbb{E}}\left[\mu^2(x;\Pi)\right] = \hat{\mu}^2(x;\mathcal{S}^{\text{tr}},\Pi) - \frac{S^2_{\mathcal{S}^{\text{tr}}}(\ell(x;\Pi))}{N^{\text{tr}}(\ell(x;\Pi))}.$$

Combining these estimators leads to the following unbiased estimator for $\text{EMSE}_\mu(\Pi)$:

$$\begin{aligned}-\widehat{\text{EMSE}}_\mu(\mathcal{S}^{\text{tr}},N^{\text{est}},\Pi) \equiv \frac{1}{N^{\text{tr}}}\sum_{i\in\mathcal{S}^{\text{tr}}}\hat{\mu}^2(X_i;\mathcal{S}^{\text{tr}},\Pi) \\ -\left(\frac{1}{N^{\text{tr}}}+\frac{1}{N^{\text{est}}}\right)\cdot\sum_{\ell\in\Pi}S^2_{\mathcal{S}^{\text{tr}}}(\ell(x;\Pi)).\end{aligned}$$

Comparing this to the criterion used in the conventional CART algorithm, which can be written as

$$-\text{MSE}_\mu(\mathcal{S}^{\text{tr}},\mathcal{S}^{\text{tr}},\Pi) = \frac{1}{N^{\text{tr}}}\sum_{i\in\mathcal{S}^{\text{tr}}}\hat{\mu}^2(X_i;\mathcal{S}^{\text{tr}},\Pi),$$

the difference comes from the terms involving the variance. For a given x, $S^2_{\mathcal{S}^{\text{tr}}}(\ell(x;\Pi))$ is proportional to the MSE within the associated leaf; thus, the difference between the adaptive and honest criteria is how the within-leaf MSE is weighted, where the honest criterion penalizes small leaf size.

Honest Cross-Validation. Even though $\widehat{\text{EMSE}}_\mu(\mathcal{S}^{\text{tr}},N^{\text{est}},\Pi)$ is approximately unbiased as an estimator of our ideal criterion $\text{EMSE}_\mu(\Pi)$ for a fixed Π, it is not unbiased when we use it repeatedly to evaluate splits using recursive partitioning on the training data $\mathcal{S}^{\text{tr}}$. The reason is that initial splits tend to group together observations with similar, extreme outcomes. So, after the training data have been divided once, the sample variance of observations in the training data within a given leaf is on average lower than the sample variance would be in a new, independent sample. Thus, $\widehat{\text{EMSE}}_\mu(\mathcal{S}^{\text{tr}},N^{\text{est}},\Pi)$ is likely to overstate goodness of fit as we grow a deeper and deeper tree, implying that cross-validation can still play an important role with our honest estimation approach, although perhaps less so than in the conventional CART.

Because the conventional CART cross-validation criterion does not account for honest estimation we consider the analog of our unbiased estimate of the criterion, which accounts for honest estimation by evaluating a partition Π using only outcomes for units from the cross-validation sample $\mathcal{S}^{\text{tr,cv}}$:

$$-\widehat{\text{EMSE}}_\mu(\mathcal{S}^{\text{tr,cv}},N^{\text{est}},\Pi).$$

This estimator for the honest criterion is unbiased for fixed Π, although it may have higher variance than $\text{MSE}_\mu(\mathcal{S}^{\text{tr,cv}},\mathcal{S}^{\text{tr,tr}},\Pi)$ due to the small sample size of the cross-validation sample. Note that when we apply the formula for $\widehat{\text{EMSE}}_\mu$ in this case, we replace N^{tr} with $N^{\text{tr,cv}}$.

Honest Inference for Treatment Effects

In this section we change the focus to estimating conditional average treatment effects instead of estimating conditional population means. We refer to the estimators developed in this section as "causal tree" (CT) estimators.

The setting with treatment effects creates some specific problems because we do not observe the value of the treatment effect whose conditional mean we wish to estimate. This complicates the calculation of the criteria we introduced in the previous section. However, a key point of this paper is that we can estimate these criteria and use those estimates for splitting and cross-validation.

We now observe in each sample the triple $(Y_i^{\text{obs}},X_i,W_i)$. For a sample $\mathcal{S}$ let $\mathcal{S}_{\text{treat}}$ and $\mathcal{S}_{\text{control}}$ denote the subsamples of treated and control units, respectively, with cardinality N_{treat} and N_{control}, respectively, and let $p=N_{\text{treat}}/N$ be the share of treated units. The concept of a tree remains the same as in the previous section. Given a tree Π, define for all x and both treatment levels w the population average outcome

$$\mu(w,x;\Pi) \equiv \mathbb{E}[Y_i(w)|X_i\in\ell(x;\Pi)],$$

and the average causal effect

$$\tau(x;\Pi) \equiv \mathbb{E}[Y_i(1)-Y_i(0)|X_i\in\ell(x;\Pi)] = \mu(1,x;\Pi)-\mu(0,x;\Pi).$$

The estimated counterparts are

$$\hat{\mu}(w,x;\mathcal{S},\Pi) \equiv \frac{1}{\#(\{i\in\mathcal{S}_w : X_i\in\ell(x;\Pi)\})}\sum_{i\in\mathcal{S}_w:X_i\in\ell(x;\Pi)}Y_i^{\text{obs}},$$

$$\hat{\tau}(x;\mathcal{S},\Pi) \equiv \hat{\mu}(1,x;\mathcal{S},\Pi)-\hat{\mu}(0,x;\mathcal{S},\Pi).$$

Define the MSE for treatment effects as

$$\text{MSE}_\tau(\mathcal{S}^{\text{te}},\mathcal{S}^{\text{est}},\Pi) \equiv \frac{1}{\#(\mathcal{S}^{\text{te}})}\sum_{i\in\mathcal{S}^{\text{te}}}\left\{(\tau_i-\hat{\tau}(X_i;\mathcal{S}^{\text{est}},\Pi))^2-\tau_i^2\right\},$$

and define $\text{EMSE}_\tau(\Pi)$ to be its expectation over the estimation and test samples,

$$\mathrm{EMSE}_\tau(\Pi) \equiv \mathbb{E}_{\mathcal{S}^{\mathrm{te}},\mathcal{S}^{\mathrm{est}}}[\mathrm{MSE}_\tau(\mathcal{S}^{\mathrm{te}},\mathcal{S}^{\mathrm{est}},\Pi)].$$

A key challenge is that, in contrast to $\mathrm{MSE}_\mu(\mathcal{S}^{\mathrm{te}},\mathcal{S}^{\mathrm{est}},\Pi)$, the workhorse MSE function $\mathrm{MSE}_\tau(\mathcal{S}^{\mathrm{te}},\mathcal{S}^{\mathrm{est}},\Pi)$ is infeasible, because we do not observe τ_i. However, we show below that we can estimate it.

Modifying Conventional CART for Treatment Effects. Consider first modifying conventional (adaptive) CART to estimate heterogeneous treatment effects. Note that in the prediction case, using the fact that $\hat{\mu}$ is constant within each leaf, we can write

$$\mathrm{MSE}_\mu(\mathcal{S}^{\mathrm{te}},\mathcal{S}^{\mathrm{tr}},\Pi) = -\frac{2}{N^{\mathrm{tr}}}\sum_{i\in\mathcal{S}^{\mathrm{te}}}\hat{\mu}\left(X_i;\mathcal{S}^{\mathrm{te}},\Pi\right)\cdot\hat{\mu}(X_i;\mathcal{S}^{\mathrm{tr}},\Pi) + \frac{1}{N^{\mathrm{tr}}}\sum_{i\in\mathcal{S}}\hat{\mu}^2(X_i;\mathcal{S}^{\mathrm{tr}},\Pi).$$

In the treatment effect case we can use the fact that

$$\mathbb{E}_{\mathcal{S}^{\mathrm{te}}}[\tau_i | i\in\mathcal{S}^{\mathrm{te}} : i\in\ell(x,\Pi)] = \mathbb{E}_{\mathcal{S}^{\mathrm{te}}}[\hat{\tau}(x;\mathcal{S}^{\mathrm{te}},\Pi)]$$

to construct an unbiased estimator of $\mathrm{MSE}_\tau(\mathcal{S}^{\mathrm{te}},\mathcal{S}^{\mathrm{tr}},\Pi)$:

$$\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{te}},\mathcal{S}^{\mathrm{tr}},\Pi) \equiv -\frac{2}{N^{\mathrm{tr}}}\sum_{i\in\mathcal{S}^{\mathrm{te}}}\hat{\tau}\left(X_i;\mathcal{S}^{\mathrm{te}},\Pi\right)\cdot\hat{\tau}(X_i;\mathcal{S}^{\mathrm{tr}},\Pi) + \frac{1}{N^{\mathrm{tr}}}\sum_{i\in\mathcal{S}^{\mathrm{te}}}\hat{\tau}^2(X_i;\mathcal{S}^{\mathrm{tr}},\Pi).$$

This leads us to propose, by analogy to CART's in-sample MSE criterion $-\mathrm{MSE}_\mu(\mathcal{S}^{\mathrm{tr}},\mathcal{S}^{\mathrm{tr}},\Pi)$,

$$-\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{tr}},\mathcal{S}^{\mathrm{tr}},\Pi) = \frac{1}{N^{\mathrm{tr}}}\sum_{i\in\mathcal{S}^{\mathrm{tr}}}\hat{\tau}^2(X_i;\mathcal{S}^{\mathrm{tr}},\Pi),$$

as an estimator for the infeasible in-sample goodness-of-fit criterion. For cross-validation we used in the prediction case $-\mathrm{MSE}_\mu(\mathcal{S}^{\mathrm{tr,cv}},\mathcal{S}^{\mathrm{tr,tr}},\Pi)$. Again, the treatment effect analog is infeasible, but we can use an unbiased estimate of it, which leads to $-\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{tr,cv}},\mathcal{S}^{\mathrm{tr,tr}},\Pi)$.

Modifying the Honest Approach. The honest approach described in the previous section for prediction problems also needs to be modified for the treatment effect setting. Using the same expansion as before, now applied to the treatment effect setting, we find

$$-\mathrm{EMSE}_\tau(\Pi) = \mathbb{E}_{X_i}\left[\tau^2(X_i;\Pi)\right] - \mathbb{E}_{\mathcal{S}^{\mathrm{est}},X_i}\left[\mathbb{V}\left(\hat{\tau}^2(X_i;\mathcal{S}^{\mathrm{est}},\Pi)\right)\right].$$

For splitting we can estimate both components of this expectation using only the training sample, yielding an estimator for the infeasible criterion that depends only on $\mathcal{S}^{\mathrm{tr}}$ and N^{est}:

$$-\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}^{\mathrm{tr}},N^{\mathrm{est}},\Pi) \equiv \frac{1}{N^{\mathrm{tr}}}\sum_{i\in\mathcal{S}^{\mathrm{tr}}}\hat{\tau}^2(X_i;\mathcal{S}^{\mathrm{tr}},\Pi) - \left(\frac{1}{N^{\mathrm{tr}}}+\frac{1}{N^{\mathrm{est}}}\right)\cdot\sum_{\ell\in\Pi}\left(\frac{S^2_{\mathcal{S}^{\mathrm{tr}}_{\mathrm{treat}}}(\ell)}{p}+\frac{S^2_{\mathcal{S}^{\mathrm{tr}}_{\mathrm{control}}}(\ell)}{1-p}\right).$$

For cross-validation we use the same expression, now with the cross-validation sample: $-\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}^{\mathrm{tr,cv}},N^{\mathrm{est}},\Pi)$.

These expressions are directly analogous to the criteria we proposed for the honest version of CART in the prediction case. The criteria reward a partition for finding strong heterogeneity in treatment effects and penalize a partition that creates variance in leaf estimates. One difference is that in the prediction case the two terms both tend to select features that predict heterogeneity in outcomes, whereas for the treatment effect case the two terms reward different types of features. It is possible to reduce the variance of a treatment effect estimator by introducing a split, even if both child leaves have the same average treatment effect, if a covariate affects the mean outcome but not treatment effects. In such a case, the split results in more homogeneous leaves, and thus lower-variance estimates of the means of the treatment group and control group outcomes. Thus, the distinction between adaptive and honest splitting criterion will be more pronounced for treatment effect estimation. As in the prediction case, the cross-validation criterion estimates treatment effects within leaves using the $\mathcal{S}^{\mathrm{tr,cv}}$ sample rather than $\mathcal{S}^{\mathrm{tr,tr}}$.

Four Partitioning Estimators for Causal Effects

In this section we briefly summarize our CT estimator and then describe three alternative types of estimators. We compare CT to the alternatives theoretically and through simulations. For each of the four types there is an adaptive version and an honest version, where the latter takes into account that estimation will be done on a sample separate from the sample used for constructing the partition, leading to a total of eight estimators. Note that further variations are possible; one could use adaptive splitting and cross-validation methods to construct a tree but still perform honest estimation on a separate sample. We do not consider such variations.

CTs. The discussion above developed our preferred estimator, CTs. To summarize, for the adaptive version of CTs, denoted CT-A, we use for splitting the objective $-\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{tr}},\mathcal{S}^{\mathrm{tr}},\Pi)$. For cross-validation we use the same objective function, but evaluated at the samples $\mathcal{S}^{\mathrm{tr,cv}}$ and $\mathcal{S}^{\mathrm{tr,tr}}$, namely $-\widehat{\mathrm{MSE}}_\tau(\mathcal{S}^{\mathrm{tr,cv}},\mathcal{S}^{\mathrm{tr,tr}},\Pi)$. For the honest version, CT-H, the splitting objective function is $-\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}^{\mathrm{tr}},N^{\mathrm{est}},\Pi)$. For cross-validation we use the same objective function, but now evaluated at the cross-validation sample, $-\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}^{\mathrm{tr,cv}},N^{\mathrm{est}},\Pi)$.

Transformed Outcome Trees. Our first alternative method is based on the insight that by using a transformed version of the outcome $Y_i^* = Y_i\cdot(W_i-p)/(p\cdot(1-p))$ it is possible to use off-the-shelf regression tree methods to focus splitting and cross-validation on treatment effects rather than outcomes. Similar approaches are used in refs. 15–18. Because $\mathbb{E}[Y_i^*|X_i=x]=\tau(x)$, off-the-shelf CART methods can be used directly, where estimates of the sample average of Y_i^* within each leaf can be interpreted as estimates of treatment effects. This ease of application is the key attraction of this method. The main drawback (relative to CT-A) is that in general it is not efficient because it does not use the information in the treatment indicator beyond the construction of the transformed outcome. For example, the sample average in $\mathcal{S}$ of Y_i^* within a given leaf $\ell(x;\Pi)$ will only be equal to $\hat{\tau}(x;\Pi,\mathcal{S})$ if the fraction of treated observations within the leaf is exactly equal to p. Because this method is primarily considered as a benchmark, in simulations we focus only on an adaptive version that can use existing learning methods entirely off-the-shelf. The adaptive version of the transformed outcome tree (TOT) estimator we consider, TOT-A, uses the conventional CART algorithm with the transformed outcome replacing the original outcome. The honest version, TOT-H, uses the same splitting and cross-validation criteria, so that it builds the same trees; it differs only in that a separate estimation sample is used to construct the leaf estimates. The treatment effect estimator within a leaf is the same as the adaptive method, that is, the sample mean of Y_i^* within the leaf.

Fit-Based Trees. We consider two additional alternative methods for constructing trees, based on suggestions in the literature. In

the first of these alternatives the choice of which feature to split on, and at what value of the feature to split, is based on comparisons of the goodness of fit (F) of the outcome rather than the treatment effect. In standard CART of course goodness of fit of outcomes is also the split criterion, but here we estimate a model for treatment effects within each leaf. Specifically, we have a linear model with an intercept and an indicator for the treatment as the regressors, rather than only an intercept as in standard CART. This approach is used in Zeileis et al. (19), who consider building general models at the leaves of the trees. Treatment effect estimation is a special case of their framework. Zeileis et al. (19) propose using statistical tests based on improvements in goodness of fit to determine when to stop growing the tree, rather than relying on cross-validation, but for ease of comparison with CART, in this paper we will stay closer to traditional CART in terms of growing deep trees and pruning them. We modify the MSE function:

$$\text{MSE}_{\mu,W}(\mathcal{S}^{\text{te}},\mathcal{S}^{\text{est}},\Pi) \equiv \sum_{i\in\mathcal{S}^{\text{te}}}\left(\left(Y_i^{\text{obs}}-\hat{\mu}_w(W_i,X_i;\mathcal{S}^{\text{est}},\Pi)\right)^2-Y_i^2\right).$$

For the adaptive version F-A we follow conventional CART, using the criterion $-\text{MSE}_{\mu,W}$ in place of $-\text{MSE}_{\mu}$ (that is, using $-\text{MSE}_{\mu,W}(\mathcal{S}^{\text{tr}},\mathcal{S}^{\text{tr}},\Pi)$ for splitting and $-\text{MSE}_{\mu,W}(\mathcal{S}^{\text{tr,cv}},\mathcal{S}^{\text{tr,tr}},\Pi)$ for cross-validation). For the honest version we use the analog of $-\widehat{\text{EMSE}}_{\mu}(\mathcal{S}^{\text{tr}},N^{\text{est}},\Pi)$, with $\hat{\mu}_w$ in place of $\hat{\mu}$, for training, and the same function evaluated at $(\mathcal{S}^{\text{tr,cv}},N^{\text{est}},\Pi)$ for cross-validation. To highlight the disadvantages of the F approach, consider a case where two splits improve the fit to an equal degree. In one case, the split leads to variation in average treatment effects, and in the other case it does not. The first split would be better from the perspective of estimating heterogeneous treatment effects, but the fit criterion would view the two splits as equally attractive.

Squared T-Statistic Trees. For the last estimator we look for splits with the largest value for the square of the t-statistic (TS) for testing the null hypothesis that the average treatment effect is the same in the two potential leaves. This estimator was proposed by Su et al. (20). If the two leaves are denoted L (Left) and R (Right), the square of the t-statistic is

$$T^2 \equiv N\cdot\frac{\left(\bar{Y}_L-\bar{Y}_R\right)^2}{S^2/N_L+S^2/N_R},$$

where S^2 is the conditional sample variance given the split. At each leaf, successive splits are determined by selecting the split that maximizes T^2. The concern with this criterion is that it places no value on splits that improve the fit, even though our characterization of EMSE_{τ} shows that improving fit has value through reduction of the variance of leaf estimates. Both the adaptive and honest versions of the TS approach use T^2 as the splitting criterion. For cross-validation and pruning, it is less obvious how to proceed. Zeileis et al. (19) suggest that when using a statistical test for splitting, if it is desirable to grow deep trees and then cross-validate to determine depth, then one can use a standard goodness-of-fit measure for pruning and cross-validation. However, this could undermine the key advantage of TS, to focus on heterogeneous treatment effects. For this reason, we instead propose to use the CT-A and CT-H criteria for cross-validation for TS-A and TS-H, respectively.

Comparison of the CTs, the F Criterion, and the TS Criterion. It is useful to compare our proposed criterion to the F and TS criteria in a simple setting to gain insight into the relative merits of the three approaches. We do so here focusing on a decision whether to proceed with a single possible split, based on a binary covariate $X_i\in\{L,R\}$. Let Π_N and Π_S denote the trees without and with the split, and let $\bar{Y}_w$, $\bar{Y}_{Lw}$ and $\bar{Y}_{Rw}$ denote the average outcomes for units with treatment status $W_i=w$. Let N_w, N_{Lw}, and N_{Rw} be the sample sizes for the corresponding subsamples. Let S_0^2 be the sample variance of the outcomes given a split, and let $\tilde{S}^2$ be the sample variance without a split. Define the squared t-statistics for testing that the average outcomes for control (treated) units in both leaves are identical:

$$T_0^2\equiv\frac{\left(\bar{Y}_{L0}-\bar{Y}_{R0}\right)^2}{S^2/N_{L0}+S^2/N_{R0}},\ T_1^2\equiv\frac{\left(\bar{Y}_{L1}-\bar{Y}_{R1}\right)^2}{S^2/N_{L1}+S^2/N_{R1}}.$$

Then, we can write the improvement in goodness of fit from splitting the single leaf into two leaves as

$$F=\tilde{S}^2\cdot\frac{2\cdot\left(T_0^2+T_1^2\right)}{1+2\cdot\left(T_0^2+T_1^2\right)/N}.$$

Ignoring degrees-of-freedom corrections, the change in our proposed criterion for the honest version of the CT in this simple setting can be written as a combination of the F and TS criteria:

$$\widehat{\text{EMSE}}_{\tau}(\mathcal{S},\Pi_N)-\widehat{\text{EMSE}}_{\tau}(\mathcal{S},\Pi_S)=\frac{\left(T^2-4\right)\left(\tilde{S}^2-F/N\right)+2\tilde{S}^2}{p\cdot(1-p)}.$$

The CT-H criterion focuses primarily on T^2. Unlike TS, however, it incorporates the benefits of improving fit.

Inference

Given the estimated conditional average treatment effect we also would like to do inference. Once constructed, the tree is a function of covariates, and if we use a distinct sample to conduct inference, then the problem reduces to that of estimating treatment effects in each member of a partition of the covariate space. For this problem, standard approaches are therefore valid for the estimates obtained via honest estimation and, in particular, no assumptions about model complexity are required. As our simulations below illustrate, for the adaptive methods standard approaches to confidence intervals are not generally valid for the reasons discussed above.

A Simulation Study

To assess the relative performance of the proposed algorithms we carried out a small simulation study with three distinct designs. In Table 1 we report a number of summary statistics from the simulations. We report averages; results for medians are similar. We report results for $N^{\text{tr}}=N^{\text{est}}$ with either 500 or 1,000 observations. When comparing adaptive to honest approaches, we report the ratio of the MSE_{τ} for adaptive estimation with $N^{\text{tr}}=1{,}000$ to MSE_{τ} for honest estimation with $N^{\text{tr}}=N^{\text{est}}=500$, to highlight the tradeoff between sample size and bias reduction that arises with honest estimation. We evaluate MSE_{τ} using a test sample with $N^{\text{te}}=8{,}000$ observations to minimize the sampling variance.

In all designs the marginal treatment probability is $P=0.5$, K denotes the number of features, we have a model $\eta(x)$ for the mean effect and $\kappa(x)$ for the treatment effect, and the potential outcomes are written, for $w=0,1$,

$$Y_i(w)=\eta(X_i)+\frac{1}{2}\cdot(2w-1)\cdot\kappa(X_i)+\epsilon_i,$$

where $\epsilon_i\sim\mathcal{N}(0,.01)$, and the X_i are independent of ϵ_i and one another, and $X_i\sim\mathcal{N}(0,1)$. The designs follow:

$$1: K=2; \eta(x)=\frac{1}{2}x_1+x_2; \kappa(x)=\frac{1}{2}x_1.$$

$$2: K=10; \eta(x)=\frac{1}{2}\sum_{k=1}^{2}x_k+\sum_{k=3}^{6}x_k; \kappa(x)=\sum_{k=1}^{2}1\{x_k>0\}\cdot x_k$$

$$3: K=20; \eta(x)=\frac{1}{2}\sum_{k=1}^{4}x_k+\sum_{k=5}^{8}x_k; \kappa(x)=\sum_{k=1}^{4}1\{x_k>0\}\cdot x_k.$$

In each design, there are some covariates that affect treatment effects (κ) and mean outcomes (η), some covariates that enter η but not κ; and some covariates that do not affect outcomes at all ("noise" covariates). Design 1 does not have noise covariates. In designs 2 and 3, the first few covariates enter κ, but only when their signs are positive, whereas they affect η throughout their range. Different criterion will thus lead to different optimal splits, even within a covariate; F will focus more on splits when the covariates are negative.

The first section of Table 1 compares the number of leaves in different designs and different values of $N^{\text{tr}}=N^{\text{est}}$. Recalling that TOT-A and TOT-H have the same splitting method, we see that it tends to build shallow trees. The failure to control for the realized value of W_i leads to additional noise in estimates, which tends to lead to aggressive pruning. For the TS and CT estimators, the adaptive versions lead to shallower trees than the honest versions, because the honest versions anticipate correcting for bias in leaf estimates and thus prune less; the main cost of small leaf size is high variance in leaf estimates. F-A and F-H are very similar; the splitting criteria are similar, and further, the F estimators are less prone to overfitting treatment effects, because they split based upon overall model fit. We also observe that the F estimators build the deepest trees; they reward splitting on covariates that affect mean outcomes as well as treatment effects.

The second section of Table 1 examines the performance of the alternative honest estimators, as evaluated by the infeasible criterion MSE_τ. We report the ratio of the average of MSE_τ for a given estimator to MSE_τ for our preferred estimator, CT-H. The TOT-H estimator performance is within 10% of CT in designs 2 and 3 but suffers in design 1. In design 1, the variance of Y_i conditional on (W_i, X_i) is very low at 0.01, and so the failure of TOT to account for the realization of W_i results in a noticeable loss of performance. The F-H estimator suffers in all three designs; all designs give the F-H criterion attractive opportunities to split based on covariates that do not enter κ. F-H would perform better in alternative designs where $\eta(x)=\kappa(x)$; F-H also does well at avoiding splits on noise covariates. The TS-H estimator performs well in design 1, where x_1 affects η and κ the same way, so that the CT-H criterion is aligned with TS-H. Designs 2 and 3 are more complex, and the ideal splits from the perspective of balancing overall MSE of treatment effects (including variance reduction) are different from those favored by TS-H. Thus, TS performs worse, and the difference is exacerbated with larger sample size in design 3, where there are more opportunities for the estimators to build deeper trees and thus to make different choices. We also calculate comparisons based on a feasible criterion, the average squared difference between the transformed outcome Y_i^* and the estimated treatment effect $\hat{\tau}_i$. For details see *SI Appendix*. The results are consistent with those from the infeasible criterion, but the feasible criterion compresses the performance differences.

The third section of Table 1 explores the costs and benefits to honest estimation. The table reports the ratio of $\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}\cup\mathcal{S}^{\text{tr}}, \pi^{\text{Estimator-A}}(\mathcal{S}^{\text{est}}\cup\mathcal{S}^{\text{tr}}))$ to $\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator-H}}(\mathcal{S}^{\text{tr}}))$ for each estimator. The adaptive version uses the union of the training and estimation samples for tree building, cross-validation, and leaf estimation, yielding double the sample size (1,000 observations) at each step. The honest version uses 500 of the observations in training and cross-validation, with the complement used for estimating treatment effects within leaves. The results show that in most cases there is a cost to honest estimation in terms of MSE_τ, varying by design and estimator. The cost is large for the fit estimator in design 1; with a smaller sample size it largely ignores treatment effect heterogeneity in splitting. For CT, the cost ranges from 6.8 to 21.5%.

The final two sections of Table 1 show the coverage rate for 90% confidence intervals. We achieve nominal coverage rates for honest methods in all designs, where, in contrast, the adaptive methods have coverage rates substantially below nominal rates. The fit estimator has the highest adaptive coverage rates; it does not focus on treatment effects and thus is less prone to overstating that heterogeneity through adaptive estimation. Thus, our simulations bear out the tradeoff that honest estimation sacrifices some goodness of fit (of treatment effects) in exchange for valid confidence intervals.

Table 1. Simulation study

$N^{\text{tr}}=N^{\text{est}}$	Design 1		Design 2		Design 3	
Estimator	500	1,000	500	1,000	500	1,000
			No. of leaves			
TOT	2.9	3.2	2.9	3.5	3.6	5.4
F-A	6.1	13.1	6.3	13.0	6.2	13.0
TS-A	4.0	5.4	3.4	5.1	3.4	6.6
CT-A	4.0	5.5	3.2	3.7	3.5	5.4
F-H	6.0	12.9	6.3	13.0	6.3	13.1
TS-H	4.3	7.8	5.6	11.4	5.9	12.4
CT-H	4.2	7.6	5.6	11.4	6.1	12.5
			Infeasible MSE divided by infeasible MSE for CT-H*			
TOT-H	1.554	1.938	1.089	1.069	1.081	1.042
F-H	1.790	1.427	1.983	2.709	1.502	2.085
TS-H	0.971	0.963	1.183	1.145	1.178	1.338
			Ratio of infeasible MSE: Adaptive to honest†			
TOT-A/TOT-H		1.021		0.754		0.717
F-A/F-H		0.491		0.985		0.993
T-A/T-H		0.935		0.841		0.918
CT-A/CT-H		0.929		0.851		0.785
			Coverage of 90% confidence intervals – adaptive			
TOT-A	0.82	0.85	0.78	0.81	0.69	0.74
F-A	0.89	0.89	0.83	0.84	0.82	0.82
TS-A	0.84	0.84	0.78	0.82	0.75	0.75
CT-A	0.83	0.84	0.78	0.82	0.76	0.79
			Coverage of 90% confidence intervals – honest			
TOT-H	0.90	0.90	0.90	0.89	0.89	0.90
F-H	0.90	0.90	0.90	0.90	0.90	0.90
TS-H	0.90	0.90	0.91	0.91	0.89	0.90
CT-H	0.89	0.90	0.90	0.90	0.89	0.90

*$\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator}}(\mathcal{S}^{\text{tr}}))/\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{CT-H}}(\mathcal{S}^{\text{tr}}))$.
†$\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}\cup\mathcal{S}^{\text{tr}}, \pi^{\text{Estimator-A}}(\mathcal{S}^{\text{est}}\cup\mathcal{S}^{\text{tr}}))/\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator-H}}(\mathcal{S}^{\text{tr}}))$.

Observational Studies with Unconfoundedness

The discussion so far has focused on the setting where the assignment to treatment is randomized. The proposed methods can be adapted to observational studies under the assumption of unconfoundedness. In that case we need to modify the estimates within leaves to remove the bias from simple comparisons of treated and control units. There is a large literature on methods for doing so (e.g., ref. 3). For example, as in ref. 21 we can do so by propensity score weighting. Efficiency will improve if we renormalize the weights within each leaf and within the treatment and control group when estimating treatment effects.

Crump et al. (22) propose approaches to trimming observations with extreme values for the propensity score to improve robustnesses. Note that there are some additional conditions required to establish asymptotic normality of treatment effect estimates when propensity score weighting is used (see, e.g., ref. 21); these results apply without modification to the estimation phase of honest partitioning algorithms.

The Literature

A small but growing literature seeks to apply supervised machine learning techniques to the problem of estimating heterogeneous treatment effects. Beyond those previously discussed, Tian et al. (23) transform the features rather than the outcomes and then apply LASSO to the model with the original outcome and the transformed features. Foster et al. (24) estimate $\mu(w,x) = \mathbb{E}[Y_i(w)|X_i = x]$ for $w = 0,1$ using random forests, then calculate $\hat{\tau}_i = \hat{\mu}(1,X_i) - \hat{\mu}(0,X_i)$. They use machine learning algorithms to estimate $\hat{\tau}_i$ as a function of the units' attributes, X_i. Imai and Ratkovic (25) use LASSO to estimate the effects of both treatments and attributes, but with different penalty terms for the two types of features to allow for the possibility that the treatment effects are present but the magnitudes of the interactions are small. Their approach is similar to ours in that they distinguish between the estimation of treatment effects and the estimation of the impact of other attributes of units. Green and Kern (26) use Bayesian additive regression trees to model treatment effect heterogeneity. Taddy et al. (27) consider a model with the outcome linear in the covariates and the interaction with the treatment variable. Using Bayesian nonparametric methods, they project estimates of heterogeneous treatment effects onto the feature space using LASSO-type regularization methods to get low-dimensional summaries of heterogeneity. Dudik et al. (16) and Beygelzimer and Langford (15) propose a related approach for finding the optimal treatment policy that combines inverse propensity score methods with "direct methods" [e.g., directly estimating $\mu(w,x)$] that predict the outcome as a function of the treatment and the unit attributes. The methods can be used to evaluate the average difference in outcomes from any two policies that map attributes to treatments, as well as to select the optimal policy function. They do not focus on hypothesis testing for heterogeneous treatment effects, and they use conventional approaches for cross-validation. Also related is targeted learning (28), which modifies the loss function to increase the weight on the parts of the likelihood that concern parameters of interest, and work on experimental design optimized to find subpopulations with positive treatment effects (29). Finally, Wager and Walther (30) adjust confidence intervals to account for adaptive estimation, and List et al. (31) adjust for exhaustively searching the space of simple partitions.

Conclusion

In this paper we introduce methods for constructing trees for causal effects that allow us to do valid inference for the causal effects in randomized experiments and in observational studies satisfying unconfoundedness. These methods provide valid confidence intervals without restrictions on the number of covariates or the complexity of the data-generating process. Our methods partition the feature space into subspaces. The output of our method is a set of treatment effects and confidence intervals for each subspace.

A potentially important application of the techniques is to "data mining" in randomized experiments. Our method can be used to explore any previously conducted randomized controlled trial, for example, medical studies or field experiments in development economics. Our methods can discover subpopulations with lower-than-average or higher-than-average treatment effects while producing confidence intervals for these estimates with nominal coverage, despite having searched over many possible subpopulations.

ACKNOWLEDGMENTS. We are grateful for comments provided at seminars at the National Academy of Sciences Sackler Colloquium, the Southern Economics Association, the Stanford Conference on Causality in the Social Sciences, the MIT Conference in Digital Experimentation, Harvard University, University of Washington, Microsoft Research, Facebook, KDD, the AAAI Embedded Machine Learning Conference, the University of Pennsylvania, the California Econometrics Conference, the Collective Intelligence Conference, the University of Arizona, the Paris DataLead conference, Cornell University, Carnegie Mellon University, University of Bonn, University of California, Berkeley, the DARPA conference on Machine Learning and Causal Inference, and the NYC Data Science Seminar Series. Part of this research was conducted while the authors were visiting Microsoft Research.

1. Rubin D (1974) Estimating causal effects of treatments in randomized and non-randomized studies. *Educ Psychol* 66(5):688–701.
2. Holland P (1986) Statistics and causal inference (with discussion). *J Am Stat Assoc* 81(396):945–970.
3. Imbens G, Rubin D (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge Univ Press, Cambridge, UK), p 159.
4. Hastie T, Tibshirani R, Friedman J (2011) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York), 2nd Ed.
5. Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and Regression Trees* (Wadsworth, Belmont, CA).
6. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32.
7. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc B* 58(1):267–288.
8. Vapnik V (1998) *Statistical Learning Theory* (Wiley, New York).
9. Wager S, Athey S (2015) Estimation and inference of heterogeneous treatment effects using random forests. Available at arxiv.org/abs/1510.04342.
10. Rubin D (1978) Bayesian inference for causal effects: The role of randomization. *Ann Stat* 6(1):34–58.
11. Rosenbaum P, Rubin D (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.
12. Abadie A, Imbens G (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–267.
13. Pearl J (2000) *Causality: Models, Reasoning and Inference* (Cambridge Univ Press, Cambridge, UK).
14. Rosenbaum P (2002) *Observational Studies* (Springer, New York).
15. Beygelzimer A, Langford J (2009) The offset tree for learning with partial labels. arxiv: 0812.4044.
16. Dudik M, Langford J, Li L (2011) Doubly robust policy evaluation and learning. *Proceedings of the 28th International Conference on Machine Learning* (International Machine Learning Society).
17. Sigovitch J (2007) Identifying informative biological markers in high-dimensional genomic data and clinical trials. PhD thesis (Harvard Univ, Cambridge, MA).
18. Weisberg HI, Pontes VP (2015) Post hoc subgroups in clinical trials: Anathema or analytics? *Clin Trials* 12(4):357–364.
19. Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. *J Comput Graph Stat* 17(2):492–514.
20. Su X, Tsai C, Wang H, Nickerson D, Li B (2009) Subgroup analysis via recursive partitioning. *J Mach Learn Res* 10:141–158.
21. Hirano K, Imbens G, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–1189.
22. Crump R, Hotz J, Imbens G, Mitnik O (2008) Nonparametric tests for treatment effect heterogeneity. *Rev Econ Stat* 90(3):389–405.
23. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R (2014) A simple method for estimating interactions between a treatment and a large number of covariates. *J Am Stat Assoc* 109(508):1517–1532.
24. Foster JC, Taylor JM, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. *Stat Med* 30(24):2867–2880.
25. Imai K, Ratkovic M (2013) Estimating treatment effect heterogeneity in randomized program evaluation. *Ann Appl Stat* 7(1):443–470.
26. Green D, Kern H (2012) Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opin Q* 76(3):491–511.
27. Taddy M, Gardner M, Chen L, Draper D (2015) Heterogeneous treatment effects in digital experimentation. arXiv:1412.8563.
28. Van Der Laan M, Rose S (2011) *Targeted Learning: Causal Inference for Observational and Experimental Data* (Springer, New York).
29. Rosenblum M, Van der Laan MJ (2011) Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika* 98(4):845–860.
30. Wager S, Walther G (2015) Uniform convergence of random forests via adaptive concentration. arxiv:1503.06388.
31. List J, Shaikh A, Xu Y (2016) Multiple hypothesis testing in experimental economics. NBER Working Paper No. 21875 (National Bureau of Economic Research, Cambridge, MA).