Assessing effect heterogeneity of a randomized treatment using conditional inference trees

Ashwini Venkatasubramaniam¹, Brandon Koch², Lauren Erickson³, Simone French⁴, David Vock⁵, and Julian Wolfson⁵

Abstract
Treatment effect heterogeneity occurs when individual characteristics influence the effect of a treatment. We propose a novel approach that combines prognostic score matching and conditional inference trees to characterize effect heterogeneity of a randomized binary treatment. One key feature that distinguishes our method from alternative approaches is that it controls the Type I error rate, that is, the probability of identifying effect heterogeneity if none exists and retains the underlying subgroups. This feature makes our technique particularly appealing in the context of clinical trials, where there may be significant costs associated with erroneously declaring that effects differ across population subgroups. Treatment effect heterogeneity trees are able to identify heterogeneous subgroups, characterize the relevant subgroups and estimate the associated treatment effects. We demonstrate the efficacy of the proposed method using a comprehensive simulation study and illustrate our method using a nutrition trial dataset to evaluate effect heterogeneity within a patient population.

Keywords
Causal effects, conditional inference trees, matching, treatment effect heterogeneity

Introduction
Under mild assumptions, randomized experiments estimate the average causal effect (ACE) of an intervention, also referred to as the average treatment effect (ATE). However, individuals may vary in their response to intervention so that the ATE is a poor representation of some individuals’ expected benefit (or harm) from the intervention, a phenomenon often referred to as treatment effect heterogeneity.¹ For a given intervention, two heterogeneity-related questions arise: (1) Is the effect of the intervention heterogeneous? (2) If the intervention effect is heterogeneous, how does it vary across individuals? Many heterogeneity-focused statistical methods address both questions using a single model or procedure. Traditional approaches to characterizing treatment effect heterogeneity have primarily centered around regression modeling with interaction terms between the treatment and covariates. In such models, interaction terms can be used to assess whether treatment effect heterogeneity exists and also to characterize its magnitude. Alternatively, formal nonparametric tests have been developed to test a null hypothesis of zero average treatment effect for any subpopulation defined by covariates and whether the average treatment effect is identical for all subpopulations.²
Existing methods for the identification of treatment effect heterogeneity

A rapidly growing toolkit of flexible machine learning techniques has produced several data-driven methods for characterizing treatment effect heterogeneity. Flexible methods for heterogeneous treatment effect estimation include methods based on random forests, the LASSO, recursive partitioning, boosting, Bayesian frameworks, combination frameworks, a generic machine learning approach and a general class of two-stage algorithms that seek to address concerns about adapting machine learning for effect estimation. Decision tree frameworks retain the covariate relationship structures (unlike ensemble structures) that determine subgroups and recursively split the dataset using selected covariates at each node of the tree. These tree-determined covariate relationships and their thresholds have been utilized for various purposes including the development of bio-marker signatures to detect two subgroups corresponding to subjects with positive and negative treatment effects (e.g. Sequential BATTing). Other closely related methods include the AIM-rule, patient rule induction method (PRIM), a model-based recursive partitioning method (MOB) and a flexible thresholding approach that utilizes a two-stage change point detection method to set the splitting rule as a combination of multivariate predictors (rather than a single covariate leading to two subgroups).

While many of the aforementioned techniques have shown impressive abilities to identify heterogeneous subgroups in situations where heterogeneity exists, they are often overly aggressive in identifying treatment effect heterogeneity when, in truth, there is none. Put more simply, most existing procedures for detecting treatment effect heterogeneity do not control Type I error. This lack of control of Type I error is particularly problematic in the context of randomized trials, where false declarations of treatment effect heterogeneity for a therapeutic agent could lead to wasteful follow-up studies and inappropriate off-label use in identified sub-populations. Several recently-proposed approaches have taken initial steps to address this issue: two new splitting criteria were proposed within a CART-like framework to maintain a balance between minimizing error in estimating the treatment effect and maximizing heterogeneity; machine learning approaches were compiled to formally test for the presence of effect heterogeneity, evaluate the predictive benefits of the machine learning algorithm over a traditional statistical model and exert control over the type I error rate; an associated measure of uncertainty for each subgroup was recommended to account for the consequences of false discovery and a matching plus classification and regression tree (mCART) was introduced to reduce the potential for falsely detecting treatment effect heterogeneity. Other approaches have also offered a framework to formally test for the overall effect heterogeneity. However, no flexible, unified tree-based framework has yet emerged that combines formal inferential testing for treatment effect heterogeneity with a focus on characterization of subgroups experiencing differential treatment effects.

Our contribution

In this paper, we propose a novel approach to testing for and characterizing effect heterogeneity of a randomized binary treatment on a continuous outcome, while explicitly controlling the Type I error rate. Our Treatment Effect Heterogeneity Tree (TEHTree) method involves building a conditional inference tree using pairs of individuals matched on the prognostic score. After describing the TEHTree method and providing theoretical motivation for matching based on prognostic scores, we present the results of a substantial simulation study demonstrating TEHTree’s Type I error control (power) in the absence (presence) of effect heterogeneity and compare its performance to other established techniques. This simulation study also evaluates subgroup identification and associated characterization for a unified framework (as in Causal Tree and TEHTree) against a two-stage method, which first generates individual treatment effects from an ensemble framework and then runs the effects through a conditional inference tree to determine subgroups. We also offer a comparison of the real-world performance of the two methods using data from a recent randomized trial in nutrition.

Method

Setup and notation

Let $Y = \{Y_1, \ldots, Y_n\}$ be a continuous response vector for $k$ subjects randomized to treatment $Z = 0$ and $n-k$ subjects randomized to treatment $Z = 1$. The treatment assignments for all subjects are denoted by $Z = \{Z_1, \ldots, Z_n\}$. An accompanying $n \times p$-dimensional matrix $X = \{X_1, \ldots, X_p\}$ contains the $p$ covariates for each of the $n$ subjects with $X_i = \{X_{i1}, \ldots, X_{ip}\}$.

In the counterfactual framework, each individual has a pair of counterfactual outcomes $(Y_0, Y_1)$, where $Y_{1i}$ is the outcome of subject $i$ if assigned to treatment $Z = 1$ and $Y_{0i}$ is the outcome if assigned to $Z = 0$. Hence, every individual has a counterfactual, causal treatment effect $Y_{1i} - Y_{0i}$, and the average treatment effect (ATE) is defined as the mean of these within-individual differences, $ATE = E[Y(1) - Y(0)]$ (we have switched to the parenthetical counterfactual notation $Y(Z)$ to denote the observation $Y_{Zi}$ for arbitrary $i$). One of the benefits of randomization is that, under often plausible assumptions,
the difference in means of randomized groups $E(Y|Z = 1) - E(Y|Z = 0)$ estimates the ATE. The stable unit treatment value assumption (SUTVA) bundles together two assumptions: (1) the treatment assigned to an individual affects only the outcome for that individual, and (2) there is only one “version” of treatment. In most randomized studies, SUTVA is plausible, a notable exception being studies of infectious diseases in closed populations. The other key assumption is ignorability, that is, that treatment assignment $Z$ is independent of the counterfactual pair $(Y(0), Y(1))$. While this “no unmeasured confounding” assumption is non-trivial in observational studies, it is satisfied by design in a randomized trial.

The counterfactual framework allows every individual $i$ to experience a different effect of treatment, but because study participants are typically assigned to either $Z = 1$ or $Z = 0$, these individual-level effects $Y_{i1} - Y_{i0}$ are unobserved. Instead, we can characterize treatment effect heterogeneity in the counterfactual framework by estimating the conditional average treatment effect

$$CATE(x) = E(Y(1) - Y(0)|X = x).$$

Because $Z$ is randomized, ignorability holds within any subset defined by $X = x$, and hence $CATE(x)$ can be estimated from $E(Y|Z = 1, X = x) - E(Y|Z = 0, X = x) = \mu_1(x) - \mu_0(x)$. Therefore, the key challenge to characterizing treatment effect heterogeneity in randomized studies is to identify distinct subgroups defined by $X$ with different CATEs. In the absence of treatment effect heterogeneity, the null hypothesis $H_0: CATE(x) = ATE \forall x$ holds. Most methods that seek to characterize how CATEs employ flexible semi- and non-parametric techniques in an attempt to identify regions of heterogeneity, but do not control the Type I error probability. In contrast, our approach embeds a classical parametric regression framework within a flexible tree model, allowing for both explicit control of the Type I error rate and characterization of treatment effect heterogeneity when it exists. The following two sections introduce the matching and conditional inference tree techniques that form the basis of our method.

### Matching

If we observed $Y_{i1}, Y_{i0},$ and $X_i$ for all $i$ then standard approaches to characterizing variability in a continuous outcome with respect to covariates could be used to estimate CATEs; for example, we could fit a regression tree using the differences $Y_{i1} - Y_{i0}$ as outcomes and $X_i$ as predictors. However, in most trials an individual’s outcome is observed under only one treatment, and hence $Y_{i1} - Y_{i0}$ is unobserved. So, we propose to impute it by matching each individual $i$ assigned to $Z_i = 1$ with a “similar” individual $j$ having $Z_j = 0$ and using $(Y_i - Y_j)$ to approximate $Y_{i1} - Y_{i0}$. If $j$ is an “exact” match for $i$ in the sense that $X_i = X_j$, then we can use $E(Y_i - Y_j|X_i = x)$ to estimate $CATE(x)$. When the number of covariates $p$ is even moderately large and/or elements of $X$ are continuous, it will typically be impossible to find exact matches for most individuals. One way of overcoming this problem is by deriving a single measure that characterizes the “distance” between individuals. If two individuals $i$ and $j$ with $Z_i = 1$ and $Z_j = 0$ have distance $d_{ij} = d$ between them

$$E(Y_i - Y_j|X_i = x, X_j = x', d_{ij} = d, Z_i = 1, Z_j = 0) = CATE(x) + \Delta_0^i(x, x', d)$$

(1)

(see Appendix A, Supplemental Material (SM) for the short proof). Hence, pairs matched according to $d_{ij}$ can be used to estimate CATEs provided $\Delta^0$ is small. Note that, in general, $\Delta^0$ may be non-zero even if $d = 0$; it is the price paid for reducing the multidimensional vectors $x$ and $x'$ to the scalar distance $d$.

A number of distance measures for matching have been proposed, some of which we review briefly here. Broadly speaking, these measures can be broken down into three categories according to how they define similarity: based on the distance between covariate vectors (e.g. Mahalanobis distance), based on the probability of being treated (propensity score), and based on the predicted value of the outcome (prognostic score). The propensity score is unhelpful, when treatment is randomized, since by design the covariates are independent of treatment assignment and as a result propensity score matching does not make $\Delta^0$ small. Another way of summarizing the similarity between individuals is via the prognostic score. Individuals with similar prognostic scores have similar predicted values of the outcome under treatment $Z = 0$ (typically a control condition). Matching on prognostic scores is appealing in our context where the goal is approximate individual causal treatment effects $Y_{i1} - Y_{i0}$. If two individuals $i$ and $j$ with $Z_i = 1$ and $Z_j = 0$ have the same prognostic score $\phi(X_i) = \phi(X_j) = \phi$,

$$E(Y_i - Y_j | X_i = x, X_j = x', \phi, Z_i = 1, Z_j = 0) = CATE(x)$$

(2)

(see proof in Appendix A, Supplemental Material (SM)). This result immediately implies that if $d_{ij} = |\phi(X_i) - \phi(X_j)| = 0$, then $\Delta^0 = 0$ in equation (1). In other words, matching on the prognostic score yields pairs that can be used to estimate conditional average treatment effects. Note that this result holds if $X$ is replaced by any measurable function $m(X)$, so that if $m$ captures the way in which $X$ modifies the effect of treatment, then pairs matched on $\phi(X)$ retain information about effect modification.
Conditional inference trees

With matched pairs in hand that can be used to estimate $CATE(x)$, the next step is to characterize how the CATE varies with $x$. Our approach uses conditional inference trees, a variant of decision trees which we briefly introduce here. The most popular and commonly used technique for building decision trees, the classification and regression tree (CART) technique (originally in 1984). Because of its interpretability and flexibility, CART has also been incorporated into several methods for assessing treatment effect heterogeneity. For example, the Causal Tree optimizes for heterogeneity in treatment effects and uses a modified mean-squared criterion expression within a CART framework for both splitting and cross-validation. One of CART's drawbacks is that, because it considers many possible thresholds on all possible variables when searching for an optimal split, it has a tendency to overfit the data on hand and produce overly complex models. The overfitting tendency can be controlled somewhat by “pruning” trees based on a complexity parameter. However, as we show in our simulation study, even pruned CARTs do not control the Type I error for effect heterogeneity. Several related methods have demonstrated excellent performance in detecting the presence or absence of underlying treatment effect heterogeneity and estimating individual-level CATEs. For example, the Causal Forest algorithm, which was developed as an ensemble over Causal Trees, offers an associated formal test for the presence of effect heterogeneity and can estimate $CATE(x)$ for a wide variety of functional relationships between $x$ and the treatment effect. However, this and similar ensemble methods do not generally yield an interpretable set of subgroups characterizing effect heterogeneity; we highlight these differences in characterization in our simulation study in section “Simulation study.”

One alternative to a CART framework is the conditional inference tree (CTree). The main difference between CTrees and CARTs is in the splitting process: in CTrees, the processes for choosing a variable to split on or to stop splitting (the “variable selection” step) and choosing an optimal splitting threshold for the selected variable (the “splitting” step) occur sequentially, while in CART they happen simultaneously. In the variable selection step of CTree, the decision of whether not to continue splitting is based on a test of the global null hypothesis sequentially, while in CART they happen simultaneously. In the variable selection step of CTree, the decision of whether not to continue splitting is based on a test of the global null hypothesis $H_0: E(Y|X) = E(Y)$, which is tested by considering all marginal null hypotheses $H_{0m}^m: E(Y|X_m) = E(Y)$ for $m = 1, \ldots, p$. In a simple case, each $H_{0m}^m$ can be assessed by calculating the $p$-value for the slope term from a univariate regression model of $Y$ on $X_m$. More generally, this step can accommodate a wide variety of models and test statistics; even if a statistic’s sampling distribution is unknown, permutation tests can be used to calculate $p$-values for each partial null hypothesis.

Since the global null $H_0 = \bigcap_{m=1}^p H_{0m}^m$ is rejected if the minimum $p$-value for all of the partial null hypotheses is less than a pre-specified level of significance, control of Type I error can be achieved by setting this level using an appropriate multiplicity adjustment to account for the testing of the $p$ partial null hypotheses (see section “Testing partial null hypotheses”). If the minimum $p$-value exceeds the threshold, the tree does not split the given subset further. Otherwise, the partial null hypothesis $H_{0m}^m$ that results in the smallest $p$-value will indicate the covariate $X_m$ that is most strongly associated with the outcome $Y$ and the algorithm proceeds to the next step to determine how to optimally threshold $X_m$. In our method, we calculate $p$-values associated with the (fixed) slope term from univariate linear mixed models. Once covariate $X_m$ has been selected for splitting, the second step of the CTree algorithm is to find the threshold $c$ that maximizes the discrepancy $|E(Y|X_m \leq c) - E(Y|X_m > c)|$. This two-phase splitting procedure is repeated on the resulting partitions until no more subsets are eligible for splitting.

Treatment effect heterogeneity trees (TEHTrees)

We propose a two-stage approach to assessing treatment effect heterogeneity in randomized studies. In the first stage, the prognostic scores are calculated and every treated subject is matched to a control subject (with replacement) based on the prognostic score. In the second stage, within-pair differences in the outcome along with the covariate values of the treated member of each pair are used as inputs to a conditional inference tree. The nodes of the fitted conditional inference tree identify subgroups across which the causal effect of treatment varies. The full algorithm is as follows; in the sections that follow, we provide details about its key steps.

**TEHTree algorithm**

1. Separate the dataset into a training and holdout set for the purpose of constructing a tree (steps 2–5 below) and for estimating treatment effects (step 6).
2. Fit a model to calculate prognostic scores $\phi$ using individuals in the training data with treatment status $Z = 0$, and obtain estimated prognostic scores $\hat{\phi}$ for each individual in the sample. Model details for prognostic score estimation are provided in section “Estimating prognostic scores.”
3. Form a set of matched pairs $(i_1, j_1), (i_2, j_2), \ldots, (i_k, j_k)$ from the training data by matching each treated ($Z = 1$) subject with one control ($Z = 0$) subject, with replacement, based on $\hat{\phi}$. Ties are broken randomly.
4. For each pair \((i, j)\), calculate the within-pair difference in the outcome, \( \delta_{ij} = Y_i - Y_j \). Each pair can now be viewed as a single “pseudo-individual” represented by the scalar continuous outcome \( \delta_{ij} \) and the covariate vector \( \mathbf{X}_i \equiv \mathbf{X}_j \).

5. Use the pseudo-individual data \((\delta_{ij}, \mathbf{X}_i)\) created in the previous step and the desired Type I error rate as inputs to create a Treatment Effect Heterogeneity Tree (TEHTree), as described in section “Testing partial null hypotheses.”

6. Estimate the treatment effect within each terminal node of the fitted TEHTree as described in section “Treatment effect estimation.”

The preceding algorithm assumes that a sufficient amount of data is available to create a holdout test set of sufficient size to accurately estimate treatment effects within subgroups defined by each terminal node. If the sample size is limited, steps 2–6 can be carried out on the entire dataset with a single-sample estimation approach for step 6 as described in section “Treatment effect estimation.”

**Estimating prognostic scores**

To provide robustness against misspecification of the prognostic score model, we apply the Super Learner\(^{38}\) to estimate the prognostic score \( \phi(\mathbf{X}) = E(Y \mid Z = 0, \mathbf{X}) \) using data from the untreated \((Z = 0)\) group. The base learners in our application consist of the sample mean, a linear model (with and without interaction terms), a generalized additive model, a random forest, stepwise regression (with and without interaction terms), and “polymars” (multivariate adaptive polynomial spline regression) as base learners.

**Testing partial null hypotheses**

Because the outcome values \( \delta_{ij} \) are derived from pairs formed by matching with replacement, inputs to the conditional inference tree are correlated and hence a standard univariate linear regression-based approach to evaluating the partial null hypotheses \( \{H_{l0} : E(Y \mid \mathbf{X}_l) = E(Y)\} \) will produce invalid \( p \)-values. Instead, we test partial null hypotheses by fitting univariate linear mixed models of the form

\[
E(\delta_{ij} \mid \mathbf{X}_{im}, b_j) = \beta_0 + \beta_1 \mathbf{X}_{im} + b_j
\]

where \( b_j \sim N(0, \tau^2) \) is a random intercept corresponding to the control subject in each pair. Similar models are used to determine the optimal splitting for selected covariate \( \mathbf{X}_s \), replacing \( \mathbf{X}_{im} \) in equation (3) by \( 1[\mathbf{X}_s \geq c] \). To establish proof of concept for our method, we used the Bonferroni method to adjust the marginal hypothesis test \( p \)-values for multiple comparisons, which sets the significance threshold at \( \frac{\alpha}{p} \) for desired Type I error rate \( \alpha \). Other less conservative adjustment methods could also be applied; see the section “Discussion” for more details.

**Treatment effect estimation**

We propose a double-sample (test/holdout set) approach to estimating heterogeneous treatment effects using TEHTree that parallels the one used in the Causal Tree method. After having generated the TEHTree based on the training set, we determine the TEHTree terminal node that each individual in the holdout set belongs to. Each terminal node \( T \) consists of the union of two subsets \( T_1 = \{ i : Z_i = 1, i \in T \} \) and \( T_0 = \{ i : Z_i = 0, i \in T \} \). We compute the treatment effect estimate as \( \Delta(T) = \frac{1}{|T_1|} \sum_{l \in T_1} Y_{il} \) \( - \frac{1}{|T_0|} \sum_{l \in T_0} Y_{il} \), where \( Y_{il} \) refers to outcomes in the test/holdout set. Note that, for a 1:1 randomized treatment, large discrepancies between \( |T_1| \) and \( |T_0| \) are unlikely, and hence the precision of \( \Delta(T) \) will be approximately proportional to \( \frac{1}{\sqrt{|T|}} \). If limited sample size precludes carving out an independent holdout set from the original data, a straightforward single-sample estimation approach can be used. Let \( T \) denote the set of matched pairs from the data belonging to each terminal node. Then, the single-sample treatment effect estimate is simply \( \Delta(T) = \frac{1}{|T|} \sum_{i \in T} \delta_i \). Estimation of the precision of the single-sample estimate of \( \Delta(T) \) is complicated (relative to the double-sample approach) by the need to consider the correlation between matched pairs. In section “Illustration,” we apply this single-sample approach to the data illustration in the presence of a limited sample size.

**Implementation**

We implemented TEHTree in R\(^{39}\) using a modified conditional inference tree framework and relevant functions in the partykit package.\(^{40}\) Matching was conducted using the Matching package,\(^{41}\) and all linear mixed models were fit using the nlme package.\(^{42}\) The Super Learner was used to estimate the prognostic score. Code for implementing TEHTree can be found at https://github.com/AshwiniKV/TEHTree. The Causal Tree and Causal Forest were implemented using the causalTree and grf packages.\(^{43}\)

**Simulation study**

We conducted simulations to evaluate the TEHTree method and compare its performance to other approaches, including several variants of the Causal Tree technique. We evaluated the Type I error, power, and other statistical
properties of the techniques for the tree types over different data generating scenarios that are documented in the Supplemental Material (SM). Factors that were varied over the scenarios include sample size \(N = 100, 200, 500, 1000,\) and 2000), number of covariates, type of covariates (binary and continuous), coefficients, and pairwise correlation among covariates. The treatment variable \(Z\) was generated such that \(N/2\) subjects received treatment \((Z = 1)\) and \(N/2\) subjects received control \((Z = 0)\). All results described in this simulation study are based on 1000 simulations per scenario. Continuous covariates were generated from multivariate normal distributions with mean zero, unit variance, and varying pairwise correlations. Binary covariates were generated as independent Bernoulli(0.5). Continuous outcomes were generated as independent Normal with unit variance and means depending on the relevant scenarios.

**Type I error**

For a case when there is no treatment effect heterogeneity, we will say that a tree-based method for detecting heterogeneity has committed a Type I error when the tree generates more than one terminal node, incorrectly implying that treatment heterogeneity exists. We generated data under two sets of scenarios where there was no treatment effect heterogeneity. In the first set of scenarios, outcomes were generated from a linear model with main effects for treatment and covariates (Model M1 in SM Table 1). In these scenarios, a simple linear model including treatment and covariates correctly specifies the prognostic score; in other words, the SuperLearner ensemble used to estimate the prognostic score contains the correct model. Figure 1(a) displays the Type I error (in logarithmic scale) of TEHTrees and Causal Trees for data simulated under these scenarios. In all cases, the Type I error of TEHTrees is less than the desired 0.05 level, while the Type I error of Causal Trees is greater than 0.05 in

![Figure 1a](image-a.png)

**Figure 1.** Results of simulations conducted over different scenarios to evaluate the Type I error rate for TEHTrees and Causal Trees. The plots use data generated from Models M1 and M2, as specified in SM Table 1. The different scenarios associated with each model and their parameters are also further summarized in SM Table 2. The error rates are displayed at sample sizes \(N = 100, 200, 500, 1000,\) and 2000.

![Figure 1b](image-b.png)
every scenario, usually substantially so. As the sample size increases, the Type I error of Causal Trees increases and is approximately 1 at \( N = 2000 \) for all three scenarios; the Type I error of TEHTrees stays roughly constant. In the second set of scenarios, outcomes were generated from a linear model with the main effects of treatment and covariates, along with additional effects for thresholded versions of continuous covariates (Model M2 in SM Table 1). In these scenarios, the SuperLearner ensemble does not contain the true model. Figure 1(b) displays the Type I error (in logarithmic scale) for these scenarios. The Type I error rate of TEHTrees tends to increase with sample size and is no longer below the desired 0.05 in every scenario. However, the Type I error rate using TEHTrees is still much lower than the Type I error rate using causal trees.

We note that this is a particularly challenging scenario for an approach based on prognostic score matching; even modest misspecification of the prognostic score could markedly increase the proportion of matched pairs where one individual has \( X > 0 \) and the other has \( X \leq 0 \), leading to the (erroneous) conclusion that treatment effects are heterogeneous in \( X \). When TEHTree is used with a correctly-specified prognostic score model (light gray points and lines), the Type I error rate is once again controlled.

In this implementation, TEHTree utilizes eight different algorithms (including generalized linear model, random forest and generalized additive model) to predict the outcome. The SuperLearner seeks to minimize the cross-validated risk and determine the optimal combination of algorithms for the ensemble framework. To account for misspecification of the prognostic score, it might be useful to modify the list of prediction algorithms and include other methods (e.g. Bayesian additive regression tree (BART) and Extreme gradient boosting). This is particularly the case for complex relationships that may not be adequately represented by the suggested algorithms.

**Power**

We characterized the performance of TEHTrees and Causal Trees under a number of different data generating scenarios where treatment effect heterogeneity is driven by binary (dichotomized) covariates only, and a mixture of binary and continuous covariates. We defined the power for detecting treatment effect heterogeneity as the probability that a tree produced a split on a variable having a non-zero interaction with treatment (i.e. one that is responsible for producing treatment effect heterogeneity). Note that while this definition of power does not credit trees with splitting on variables unrelated to heterogeneity, in cases where the treatment effect interacts with several different covariates, a tree is credited with rejecting the null hypothesis if it splits on any of these covariates.

Figure 2(a) to (d) summarizes the results of scenarios where heterogeneity is determined by a single covariate \( (X_1) \) (Models M3, M4, M5, and M7 in SM Table 1), and hence the power is the probability of splitting on \( X_1 \). TEHTrees and Causal Trees have similar power when heterogeneity is determined by \( I(X_1 > 0) \) (Figure 2(a)) and by continuous \( X_1 \) (Figure 2(b)); the power of TEHTrees is modestly lower for small sample sizes when both the indicator and continuous value contribute to heterogeneity (Figure 2(c)). TEHTrees and Causal Trees have similar power at higher sample sizes when heterogeneity is induced by \( \sin(\eta X_1) \) (Figure 2(d), at \( \eta = 1.5 \)).

SM Figure 2(a) is notable as it shows that TEHTree has very low power when effect heterogeneity is driven by \( I(-0.5 < X_1 < 0.5) \). This result is due to the fact that, in our implementation, the null hypothesis \( H_0: E(\delta|X) = E(\delta_j) \) is tested via the coefficient of the main effect of \( X \) in a (mixed) linear model. When heterogeneity is due to the above indicator, this coefficient will often be estimated as being close to zero and hence the null hypothesis is unlikely to be rejected. An implementation which used a more robust test for variation of the mean with \( X \) (e.g. by using a joint test for higher-order polynomial terms) would perform better in this case at the cost of greater computational complexity. As with the utilization of the appropriate functional form and its positive impact on performance, TEHTree is likely to have lower power if important predictor variables are simply not measured.

Figure 2(e) also displays the power of TEHTrees and Causal Trees when heterogeneity in the treatment effect is due to two variables, \( X_1 \) and \( X_2 \). For this Model M9, we considered two definitions of power: splitting on either \( X_1 \) or \( X_2 \), and splitting on both \( X_1 \) and \( X_2 \). In general, both types of power are higher for Causal Trees than TEHTrees. Results for Models M8 and M10 are available in Supplemental Material (SM Figure 2(b)-(c)). While the preceding results show that the Causal Tree approach has higher power than our proposed TEHTree method under many data generating scenarios, this comparison (like most power comparisons) is somewhat misleading since TEHTree controls the Type I error rate while Causal Tree does not. SM Table 6 shows how the power of both trees vary in Models M4, M6, and M8 as the parameters that determine heterogeneity range from 0 (no heterogeneity) to larger values (substantial heterogeneity).

**Treatment effect estimation**

Figures 2(f) to (h) present the mean squared error (MSE) of the treatment effect estimates across data generating scenarios and for different sample sizes. The scenarios are defined using Models M5, M7, and M9. Across most scenarios,
Causal Trees and TEHTrees show similar MSEs and Causal Forests generally show lower MSEs at larger sample sizes. Characteristics of TEHTrees and Causal Trees using data generated by Models M3 and M7 are summarized in SM Figure 1 and include both the proportion of split points within a given distance of the true split point and the number of terminal nodes.

Subgroup identification

As demonstrated above, single-method techniques which do not control Type I error tend to "oversplit" and identify many subgroups when heterogeneity is truly defined by only a few. Figure 3(a) describes these differences in comparison to the Causal Tree and TEHTree algorithms using data generated from Model M1 (simulation results for other models can be

![Graphs showing power and mean squared error (MSE) for TEHTrees and Causal Trees.](image)

**Figure 2.** (a) to (e) Values of power for TEHTrees and Causal Trees. The plots use data generated from Models M3, M4, M5, M7, and M9. (f) to (h) The mean squared error (MSE) of the estimated average treatment effect. The plots use data generated from Models M5, M7, and M9. In both cases, data are generated according to different scenarios with sample sizes $N = 100, 200, 500, 1000,$ and $2000$. See Supplemental Material for details of simulation settings and additional results.

Causal Trees and TEHTTrees show similar MSEs and Causal Forests generally show lower MSEs at larger sample sizes. Characteristics of TEHTTrees and Causal Trees using data generated by Models M3 and M7 are summarized in SM Figure 1 and include both the proportion of split points within a given distance of the true split point and the number of terminal nodes.

**Subgroup identification**

As demonstrated above, single-method techniques which do not control Type I error tend to "oversplit" and identify many subgroups when heterogeneity is truly defined by only a few. Figure 3(a) describes these differences in comparison to the Causal Tree and TEHTree algorithms using data generated from Model M1 (simulation results for other models can be
Across sample sizes, TEHTree retains the Type I error rate control while Causal Forest generated effects leads to trees that continue to split in the absence of effect heterogeneity.

Figure 4 illustrates this phenomenon by showing the sample partitions defined by the terminal nodes of three techniques: TEHTree, Causal Tree, and a decision tree with treatment effects from Causal Forest used as inputs. Each column in the panel corresponds to a tree type and each row is a separate realization of a simulated dataset. In these data, the only driver of effect heterogeneity is \( I(X_1 > 0) \), and the methods are given data on two covariates, \( X_1 \) and \( X_2 \). TEHTree consistently identifies the correct subgroups by splitting in the neighborhood of \( X_1 = 0 \); in four out of five simulated datasets, this is the only split TEHTree produces. The Causal Tree also regularly identifies a split near \( X_1 = 0 \), but also frequently identifies additional, non-heterogeneity-inducing splits on both \( X_1 \) and \( X_2 \). Estimated effects by Causal Forest are much more heterogeneous, decision trees based on these effects produce multiple splits on both \( X_1 \) and \( X_2 \).

Illustration

We illustrate the application of TEHTree, Causal Tree, and Causal Forest to data from the Box Lunch Study (BLS), a randomized controlled trial to evaluate the effect of receiving daily boxed lunches of three different portion sizes (400, 800, and 1600 kcal) on daily energy intake and body weight of working adults. The BLS study enrolled 233 subjects including a group that did not receive any boxed lunches; we analyzed a complete-case version of the data where 156 subjects were assigned to receive one of the three fixed portion size lunches. More specifically, we consider the problem of characterizing heterogeneity across the effect of “treatment” (defined as the 800 and 1600 kcal boxed lunches, \( n = 107 \)) versus “control” (the 400 kcal boxed lunch, \( n = 49 \)) on daily caloric intake 6 months after randomization. The average

Figure 3. Type I error and power of different tree types when data are generated over sample sizes \( N = 100, 200, 500, 1000, \) and 2000. The data to demonstrate the error rates and power are generated using Models M1 and M3, respectively. For a given simulation, these measures are defined over the probability of the tree splitting at \( X_1 \) at the root node, the tree splitting at \( X_1 \) at any node and the probability of a tree splitting at all.
The treatment effect (ATE) is a difference of 193 kcal/day. For this analysis, we consider how the ATE varies with four baseline covariates: BMI, age, a measure of hunger, and EDEQ-14.0, a measure of loss of control over eating in the past 28 days.

We applied Causal Forest to estimate individual treatment effects and empirically evaluate the presence of effect heterogeneity. The Causal Forest was fit to a training dataset of 105 subjects (including the four baseline patient characteristics) and the relevant causal treatment effects were estimated on 51 test set subjects (displayed in SM Figure 3). The results suggest that the subjects experience a relatively homogeneous response to treatment, with respect to these four covariates. The method does not explicitly identify covariates (and their split criterions) which induce heterogeneity.

**Figure 4.** Series of plots showing partitions determined by different tree types (for a continuous outcome, continuous covariates, and 1000 corresponding simulated observations using Model M3). The first split at each root node and the criterion is labeled within each plot and a tree determined partition is represented by a bold dashed line. In the first row, the first plot displays partitions for a TEHTree (root node splits at \( X_1 \leq 0.06 \)), the second plot for a Causal Tree (root node splits at \( X_2 \geq 0.55 \)) and the third plot displays partitions extracted using a conditional inference tree over treatment effects estimated from a Causal Forest (root node splits at \( X_1 \leq 0.03 \)). Heterogeneity is driven only by \( I(X_1 > 0) \) and a true partition over the simulated data points corresponds to values where \( X_1 > 0 \) and \( X_1 \leq 0 \). This panel displays partitions determined by three different tree types for five simulated datasets (as corresponding to the five rows and three columns).
Next, we compared the results of applying TEHTree and Causal Tree to the data. Figure 5(a) and (b) shows the TEHTree and Causal Tree that result from analyzing the BLS data. TEHTree was implemented using the modifications specified in section “Treatment effect estimation,” and the double-sample Causal Tree method used the default parameters with the exception of setting a minimum size for splitting to 10 treated and 10 controls. While both trees identify covariates that induce heterogeneity, the covariates identified and their split points are quite different. Figure 5(c) and (d) represents the results of applying both tree types to a permuted version of the dataset in which the rows of the covariate matrix were permuted to remove associations between covariates and the outcomes but retain within-covariate correlation. In these data, there should be no association between covariates and treatment effects, however, Causal Tree identifies heterogeneous subgroups (here defined by age and BMI). TEHTree does not generate any splits, correctly reflecting the true lack of heterogeneity in the permuted data.

Discussion

Characterizing treatment effect heterogeneity is becoming a common target of secondary analyzes of data from randomized controlled trials. As an alternative to methods that require that the nature of potential subgroups be pre-specified (e.g. via covariate interactions with treatment), several methods have been recently proposed to detect treatment effect heterogeneity.
in a more data-driven manner.\textsuperscript{13,8,14,19} However, while most of these methods incorporate procedures for preventing over-fitting, they do not offer any guarantees about Type I error, that is, the probability of identifying heterogeneous subgroups in the absence of treatment effect heterogeneity. Particularly in the context of randomized trials, explicit control of Type I error may increase the willingness of researchers to apply treatment effect heterogeneity techniques. In this paper, we propose TEHTrees, a novel method that uses a conditional inference tree framework to characterize effect heterogeneity of a binary treatment while controlling the Type I error rate.

As shown in our simulation study, existing methods often yield much higher than nominal Type I error rates, with this error rate generally increasing with sample size. In contrast, TEHTree maintains the specified Type I error rate across most scenarios. In some scenarios, the power of TEHTrees to detect true heterogeneity is competitive with Causal Tree; in other scenarios, TEHTrees has lower power, but these discrepancies mostly arise in scenarios where Causal Tree has very high Type I error rates, that is, its power curve lies above that of TEHTrees for both null and alternative hypotheses. Causal Trees displayed lower MSE than TEHTrees in multiple scenarios, which was most likely due to greater variability in split points for TEHTree structures. We conjecture that this variability can be attributed to the bias introduced by the matching estimator. Decreasing bias in the matching estimator, or using an alternative approach to estimating the outcomes that are used as inputs in the conditional inference tree of TEHTrees, may improve estimation of treatment effects with TEHTrees when there are continuous covariates.

TEHTrees offer a flexible approach to detecting effect heterogeneity, and its various building blocks allow for numerous modifications including the choice of a different matching algorithm, an alternative prognostic score model, the utilization of other criteria to select the splitting variable or its split point, or the implementation of a different estimation technique. For example, the Bonferroni correction method used to find the splitting variable in TEHTrees is likely too conservative to detect small treatment effects when there are a large number of covariates in the study. Alternative multiple comparison adjustment methods should be explored in such cases, or example, controlling the false discovery rate may be a desirable alternative. In addition, the TEHTree approach could be modified to use other regression models for determining splitting variables; indeed, any technique that tests the relevant marginal null hypotheses while accounting for the matching-induced correlation could be easily incorporated in the TEHTree framework. Our method also inherits the desirable properties of conditional inference trees including the fact that unlike CART-based Causal Trees, TEHTrees do not favor the inclusion of continuous variables with many potential split points over categorical variables with fewer.\textsuperscript{45}

Other modifications and extensions could also improve the robustness of TEHTrees. Like many matching-based approaches, our method does not account for the variability introduced by the matching estimator, so an extra step may be required to control for the inflation of Type I error that might occur in situations when good matches are difficult to obtain due to lack of overlap in covariate distributions between treatment groups. However, in randomized studies covariates are, on average, balanced between treatment groups and hence lack of overlap between the supports of the covariate distributions is unlikely to be a problem. TEHTree’s performance depends on the accuracy of the prognostic score model. As shown in the simulation, using an ensemble approach to estimating the prognostic score provides a degree of (but not total) robustness against prognostic score misspecification. One type of misspecification that we did not consider in the simulation was omission of predictors of the outcome; as with most matching-based methods, omission of such variables will decrease the quality of the resulting matches and hence lead to poorer performance. Though we assume treatment is randomized throughout this paper, TEHTrees could be extended for use with observational data involving non-randomized treatments or exposures. However, additional assumptions and modifications to the method would be required to achieve covariate balance and ensure unbiased estimation of treatment effects.

**Acknowledgements**

The authors are grateful to Dr Simone French for permitting use of the Box Lunch Study data.

**Declaration of conflicting interests**

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**ORCID iDs**

Ashwini Venkatasubramaniam https://orcid.org/0000-0002-6043-6506
Brandon Koch https://orcid.org/0000-0002-9585-9891
Supplemental Material
Supplemental material for this article is available online.

References
1. Longford NT. Selection bias and treatment heterogeneity in clinical trials. Stat Med 1999; 18: 1467–1474.
2. Crump RK, Hotz VJ, Imbens GW, et al. Nonparametric tests for treatment effect heterogeneity. Rev Econ Stat 2008; 90: 389–405.
3. Lipkovich I, Dmitrienko A and B D’Agostino R Sr. Tutorial in biostatistics: datadriven subgroup identification and analysis in clinical trials. Stat Med 2017; 36: 136–196.
4. Loh WY, Cao L and Zhou P. Subgroup identification for precision medicine: A comparative review of 13 methods. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2019; 9: e1326.
5. Goldstein BA and Rigdon J. Using machine learning to identify heterogeneous effects in randomized clinical trials Moving beyond the forest plot and into the forest. JAMA Network Open 2019; 2: e190047.
6. Dorie V, Hill J, Shalit U, et al. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Stat Sci 2019; 34: 43–68.
7. Athey S, Tibshirani J and Wager S. Generalized random forests. The Annals of Statistics 2019; 47: 1148–1178.
8. Wager S and Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 2018; 113: 1228–1242.
9. Su X, Pea AT, Liu L, et al. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. Stat Med 2018; 37: 2547–2560.
10. Imai K and Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat 2013; 7: 443–470.
11. Su X, Tsai CL, Wang H, et al. Subgroup analysis via recursive partitioning. J Mach Learn Res 2009; 10: 141–158.
12. Foster JC, Taylor JM and Ruberg SJ. Subgroup identification from randomized clinical trial data. Stat Med 2011; 30: 2867–2880.
13. Athey S and Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci USA 2016; 113: 7353–7360.
14. Powers S, Qian J, Jung K, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. Stat Med 2018; 37: 1767–1787.
15. Green DP and Kern HL. Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. Public Opin Q 2012; 76: 491–511.
16. Hill JL. Bayesian nonparametric modeling for causal inference. J Comput Graph Stat 2011; 20: 217–240.
17. Hahn PR, Carvalho CM, Puelz D, et al. Regularization and confounding in linear regression for treatment effect estimation. Bayesian Anal 2018; 13: 163–182.
18. Hahn PR, Murray JS and Carvalho C. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. arXiv preprint arXiv:1706.09523. 2017 Jun 29.
19. Knezel SR, Sekhon JS, Bickel PJ, et al. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci USA 2019; 116: 4156–4165.
20. Luedtke AR and van der Laan MJ. Super-learning of an optimal dynamic treatment rule. Int J Biostat 2016; 12: 305–332.
21. Chernozhukov V, Demirer M, Duch E, et al. Generic machine learning inference on heterogeneous treatment effects in randomized experiments. National Bureau of Economic Research, 2018.
22. Nie X and Wager S. Quasi-oracle estimation of heterogeneous treatment effects. arXiv preprint arXiv:1712.04912. 2017 Dec 13.
23. Zhang W, Li J, Li Y, et al. A modelbased multithreshold method for subgroup identification. Stat Med 2019; 38: 2605–2631.
24. Hansen BB. The prognostic analogue of the propensity score. Biometrika 2008; 95: 481–488.
25. Rubin DB. Randomization analysis of experimental data: The fisher randomization test comment. J Am Stat Assoc 1980; 75: 591–593.
26. Alberto A and Imbens Guido W. Simple and Bias-Corrected Matching Estimators: For Average Treatment Effects. NBER Technical working paper 283, 2002.
27. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41–55.
35. Stuart EA, Lee BK and Leacy FP. Prognostic scorebased balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. J Clin Epidemiol 2013; 66: S84–90.
36. Breiman L, Friedman J, Stone CJ, et al. Classification and regression trees. Boca Raton: CRC press, 1984.
37. Hothorn T, Hornik K and Zeileis A. Unbiased recursive partitioning: A conditional inference framework. J Comput Graph Stat 2006; 15: 651–674.
38. Van der Laan MJ, Polley EC and Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007; 6.
39. Team RC. R: A language and environment for statistical computing.
40. Hothorn T and Zeileis A. partykit: A modular toolkit for recursive partitioning in R. J Mach Learn Res 2015; 16: 3905–3909.
41. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. J Stat Softw, Forthcoming. 2008.
42. Pinheiro J, Bates D, DebRoy S, et al. Team RC. nlme: Linear and nonlinear mixed effects models R package version 2013; 3: 111.
43. Tibshirani J, Athey S, Wager S, et al. SystemRequirements GN. Package grf.
44. French SA, Mitchell NR, Wolfson J, et al. Portion size effects on weight gain in a free living setting. Obesity 2014; 22: 1400–1405.
45. Song HI, Song ET and Song MS. A study on the bias reduction in split variable selection in CART. Commun Stat Appl Methods 2004; 11: 553–562.