JUSTC

Article

http://justc.ustc.edu.cn

Received: March 21, 2022; Accepted: May 25, 2022

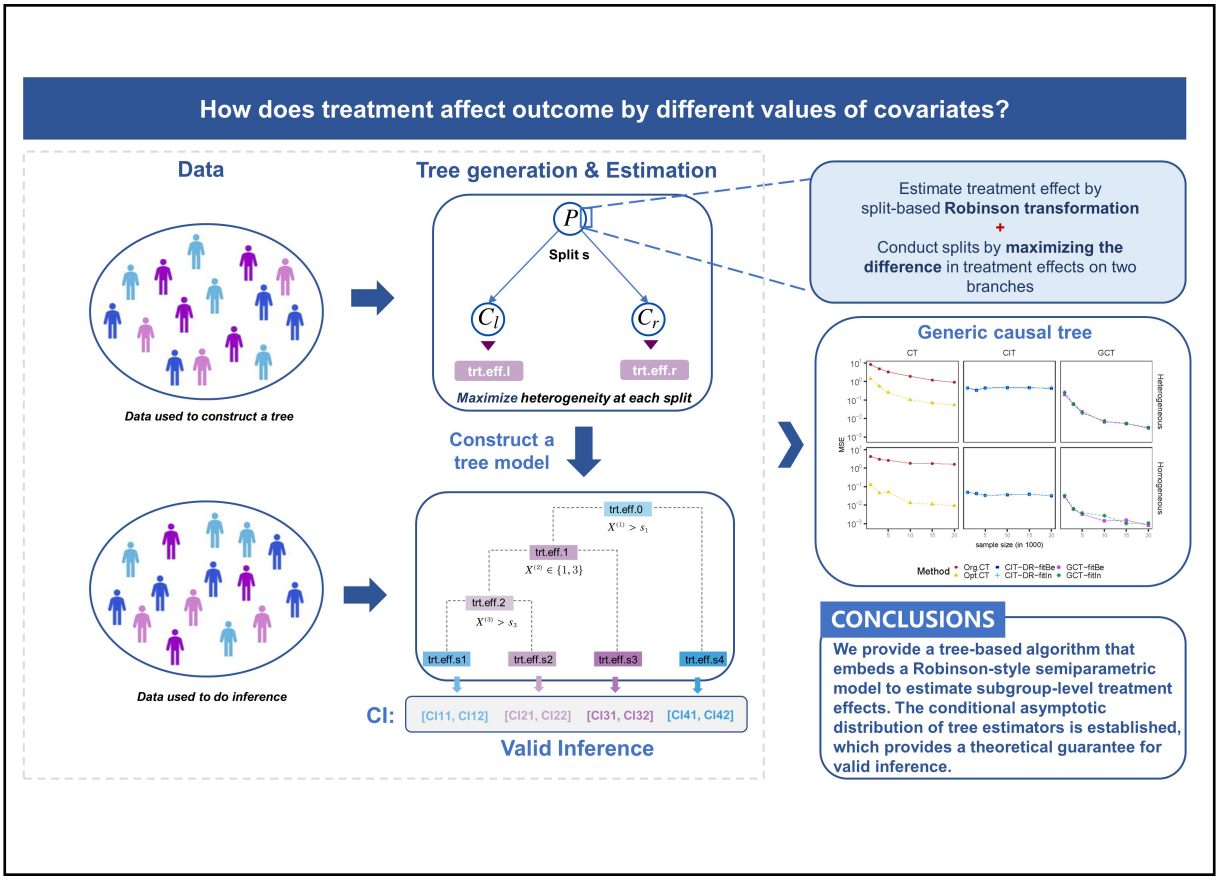
Inference of subgroup-level treatment effects via generic causal tree in observational studies

Caiwei Zhang[1], and Zemin Zheng[2] ✉

[1]*School of Data Science, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Management, University of Science and Technology of China, Hefei 230026, China*

✉Correspondence: Zemin Zheng, E-mail: zhengzm@ustc.edu.cn

Graphical abstract



A tree-based algorithm for subgroup identification with high interpretability allows valid inference for tree estimators.

Public summary

- We provide a tree-based algorithm for subgroup identification that embeds a Robinson-style semiparametric model to estimate subgroup-level treatment effects.
- The method is stepwise convex, computationally stable, efficient and scalable.
- Both theoretical and simulation results verify the feasibility of our method.

Citation: Zhang C W, Zheng Z M. Inference of subgroup-level treatment effects via generic causal tree in observational studies. *JUSTC*, **2023**, 53(11): 1102. DOI: 10.52396/JUSTC-2022-0054

JUSTC

Article

http://justc.ustc.edu.cn

Received: March 21, 2022; Accepted: May 25, 2022

Inference of subgroup-level treatment effects via generic causal tree in observational studies

Caiwei Zhang[1], and Zemin Zheng[2] ✉

[1]*School of Data Science, University of Science and Technology of China, Hefei 230026, China;*
[2]*School of Management, University of Science and Technology of China, Hefei 230026, China*

✉Correspondence: Zemin Zheng, E-mail: zhengzm@ustc.edu.cn

Cite This: *JUSTC*, **2023**, 53(11): 1102 (12pp) Read Online SI Supporting Information

Abstract: Exploring heterogeneity in causal effects has wide applications in the field of policy evaluation and decision-making. In recent years, researchers have begun employing machine learning methods to study causality, among which the most popular methods generally estimate heterogeneous treatment effects at the individual level. However, we argue that in large sample cases, identifying heterogeneity at the subgroup level is more intuitive and intelligble from a decision-making perspective. In this paper, we provide a tree-based method, called the generic causal tree (GCT), to identify the subgroup-level treatment effects in observational studies. The tree is designed to split by maximizing the disparity of treatment effects between subgroups, embedding a semiparametric framework for the improvement of treatment effect estimation. To accomplish valid statistical inference of the tree-based estimators of treatment effects, we adopt honest estimation to separate tree-building process and inference process. In the simulation, we show that the GCT algorithm has distinct advantages in subgroup identification and gives estimation with higher accuracy compared with the other two benchmark methods. Additionally, we verify the effectiveness of statistical inference by GCT.

Keywords: causal inference; tree-based algorithm; subgroup identification; semiparametric estimation; heterogeneous treatment effects

CLC number: O213 **Document code:** A

2020 Mathematics Subject Classification: 62H30

1 Introduction

Correlation analysis is used as the primary way to discover relationships between variables; it measures increasing or decreasing trends quantified using a correlation coefficient[1]. However, just looking at correlation may lead to confusing conclusions. For instance, we know that blood circulating vitamin D levels correlate with a decreased risk of colorectal cancer. While saying a high dose of vitamin D intake could reduce the risk of colorectal cancer is implausible because the relationship between them can be an illusion under the influence of other factors[2]. Correlation has limitations in answering the question if A causes B without confounding, which is of more interest in today's research, and that is what the causation does.

Exploring causality has drawn unprecedented attention in many fields, such as public policy, econometrics, and medicine. From the feasibility of practical application, the scenarios of conducting causal inference are increasingly shifting from randomized experiments to observational studies. Since tailored randomized experiments are usually time-consuming and expensive, all covariates other than the "cause" should be ensured to be similar in treated and control groups. However, the observed data are easily accessible and economical. Furthermore, with soaring data size and increasing diversity of samples, treatment will inevitably lead to different treatment effects for individuals, the so-called heterogeneity. Heterogeneity in treatment effects essentially arises from the interactions between treatment and covariates, making various magnitudes of treatment effects appear in the population. Specifically, treatment may have a negative impact on the whole but a positive influence on specific subpopulations. Thus, learning heterogeneous treatment effects (HTE) is particularly instructive in modern data analysis.

Owing to the strong capabilities of machine learning (ML) in data mining, several meritorious ML algorithms have been developed to estimate HTE over the last few years. A brief introduction to related ML methodologies is shown in Section 2. Among the literature, the most popular ML methods generally give pointwise treatment effect estimators, e.g., estimating the treatment effect at the individual level, such as generalized random forest[3] or meta-learners[4]. We argue that individual treatment effects (ITEs) are too trivial to utilize in large-scale data analysis, where researchers are more concerned about subpopulations with distinct characteristics and treatment effects. Meanwhile, ITEs cannot provide an intuitive explanation of the heterogeneity mechanism for decision-makers. Nevertheless, there is no unified operation to derive subgroup-level treatment effects (STEs) by individual treatment effects, especially since the statistical properties of

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

derived STEs may be ambiguous. For this reason, gathering similar individuals and directly estimating a uniform amount to approximate their treatment effects will be more effective.

This paper aims to estimate STEs automatically with observational data and provide intelligible insight into the heterogeneity mechanism for practitioners. Trees have advantages in capturing interactions between variables, recursively partitioning the entire covariate space into distinct subspaces. The intuitive idea is to set a well-performed splitting criterion involving treatment effects in nodes to guide tree generation. We note that the method used to estimate the treatment effect will influence the performance of the tree, including the correctness of the partition and the accuracy of the STE estimators. Meanwhile, we hope to carry out statistical inference for the final subgroup-level treatment effects. To our knowledge, the work of Athey and Imbens[5] is the first tree-based method that allows for statistical inference of treatment effects on subgroups in randomized experiments by conducting regression after tree building. In their paper, trees grow by the causal tree criterion, which may pursue smaller variances in each split and tend to split even when there is no heterogeneity. On the other hand, they use the difference in the empirical mean of outcome between treated and control groups as treatment effect estimation, which may lead to unsatisfactory performance even by propensity score weighting for observational data.

This article develops a tree-based algorithm to automatically estimate heterogeneous treatment effects in subgroups with observational data, referred to as the generic causal tree (GCT) algorithm. We use a straightforward criterion that aims to maximize the heterogeneity in each split and apply the Robinson-style[6] treatment effect estimator to improve the performance of the treatment effect estimators in observational studies.

The rest of this paper is organized as follows. In Section 2, we review the related literature. In Section 3, we expound the GCT algorithm with a new subgroup treatment effect estimator. In Section 4, we introduce the honest version of our GCT algorithm and state the asymptotic properties of the STE estimators. We compare the performance of our GCT algorithms against the other two previous tree-based methods in simulation, as shown in Section 5. An application of GCT is provided in Section 6 to illustrate the feasibility of GCT in real data. Finally, our results are summarized in Section 7. The proofs of theorems, additional experiments, and supplementary material of data analysis are provided in Supported information.

2 Related works

Much of the literature uses ML methods to estimate treatment effects at the individual level. The concept of constructing a forest for causal inference has been prevalent since the work of Wager and Athey[7], in which they aggregated regression or classification trees to construct a forest to allow for pointwise treatment effect estimation. Afterward, Athey et al.[3] proposed the forest by an adaptive-weighted method based on random forest, called generalized random forest (GRF). The target parameter can be obtained by solving a weighted-specific moment condition, where heterogeneity is embedded in the sample weights. Drawing inspiration from the idea of nearest neighbor, a debiased 1-nearest neighbor algorithm was proposed by Fan et al.[8]; they used a weighted sum of two L-statistics estimators with different sample sizes to eliminate the main order of bias. Künzel et al.[4] proposed a nonparametric framework called X-learner to avoid overfitting caused by the unbalanced sizes of treatment groups. In addition, several works[9–12] trained neural networks to fit counterfactual functions in a prelearned representation space, where the distributions of covariates are adjusted to be similar in both treatment groups.

However, in the preceding discussion, we have clarified the limits of pointwise estimation in discovering the heterogeneity mechanism. Nevertheless, several tree-based works allowed for it with reasonable interpretation. Su et al.[13] first advocated building trees for subgroup analysis in randomized trials. Yang et al.[14] generalized the interaction tree in Su et al.[13] to adapt observational data by replacing the treatment effect estimator with three adjusted estimators: inverse probability weighting estimator, G-formula estimator, and doubly robust estimator. Foster et al.[15] applied random forest as a tool to predict treatment effects and took them as the outcome into classification and regression tree[16] to detect the most affected group by treatment. Athey and Imbens[5] first proposed a sound tree-based method that enables valid statistical inference for average treatment effects in subspaces. The confidence intervals can be obtained through linear regression after the tree model is constructed.

The estimation method in this paper is inspired by the literature inferring STEs through linear models. Due to the challenge of inferring ITEs directly, Chernozhukov et al.[17] proposed a generic ML framework to infer the key features of ITEs instead, which includes the subgroup average treatment effects. They postprocessed rough estimations of ITEs to obtain STEs through a subgroup-specified linear model, where the subgroups are designated according to some rules, such as the quantiles of ITE estimators. Nevertheless, this will lead to an unreasonable grouping when the true subgroups are uneven, or the ITEs differ significantly. Park and Kang[18] proposed a sample splitting linear model to infer STEs, where subgroups are prespecified by the investigators' scientific grouping hypothesis or clustering algorithm, K-means. The former grouping method relies on prior knowledge, which may not be reliable in practical problems involving many features—misspecifying effect modifiers or setting wrong splitting points will cause varied results. For the latter data-driven approach, K-means cannot automatically highlight the vital role of effect modifiers when measuring the similarity of observations in the whole covariate space.

In general, defining a scientific grouping strategy while ensuring favorable statistical properties of STE estimators is our main purpose. The ideal situation is to perform subgroup identification based on a data-driven approach without additional prior expertise. Through the tree-based method, this goal could be reached more readily only by making trees adapt to statistical inference for STEs. Suppose the subgroups are determined, constructing confidence intervals reduces to conditional inference. Several papers give related references for

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

the application; see Section 4 for details. Our approach makes new contributions to subgroup identification through ML methods, automatically forming subgroups, and giving the treatment effect estimation and valid inferences. Tree-based methods have been questioned in term of their estimation accuracy since they use an average treatment effect to approximate the individual treatment effects of the units belonging to a subgroup. This may lead to the loss of accuracy compared with some modern methods for ITE estimation; however, it instead provides insights into heterogeneous mechanisms more intuitively. Thus, one of our main focuses is to improve the tree's estimation performance. In this paper, we seek to utilize the idea of partialling out by Robinson-style method[6] to give a better estimation, provably enhancing the performance of trees.

3 Treatment effect estimation on subgroups

3.1 Identification of treatment effect on subgroups

Suppose the observational data contain N i.i.d. samples indexed by $i = 1, 2, \cdots, N$. For each sample, we observe $(\boldsymbol{X}_i, A_i, Y_i) \in \mathbb{R}^p \times \{0,1\} \times \mathbb{R}$, where $\boldsymbol{X}_i = (X_i^{(1)}, \cdots, X_i^{(p)})$ denotes the p-dimensional covariates, A_i is a binary indicator that $A_i = 1$ means receiving treatment, while $A_i = 0$ means not receiving treatment, and Y_i is the outcome. Let Y^1 and Y^0 be the potential outcomes[19] under the treatment assignment $A = 1$ or 0, respectively.

Several assumptions are required to guarantee the identifiability of subgroup treatment effects in the observational study: a) Consistency assumption: $Y = Y^a = Y^1 A + Y^0(1-A)$, if $A = a$. The potential outcome Y^a is equal to the observed outcome Y if the actual treatment takes value a. b) Unconfoundedness assumption: $Y^1, Y^0 \perp A \mid \boldsymbol{X}$, namely conditional on covariates $\boldsymbol{X}$, the potential outcomes are independent of treatment assignment. c) Positivity assumption: $0 < P(A = a \mid \boldsymbol{X} = \boldsymbol{x}) < 1$ for all values for $\boldsymbol{X}$ and treatment A. This assumption stipulates that each sample has a nonzero probability of receiving both treatments.

Under these assumptions, the expected value of potential outcome $E(Y^a|\boldsymbol{X} = \boldsymbol{x})$ could be transformed as the conditional mean of observed outcome $E(Y|A = a, \boldsymbol{X} = \boldsymbol{x})$, which allows us to identify treatment effects using observed data (see Refs. [20, 21] for more details about identifiability conditions).

At the same time, we can derive a useful equation by taking the expectation of consistency assumption,

$$\begin{aligned} E(Y|\boldsymbol{X}) =& E(Y^1 A|\boldsymbol{X}) + E(Y^0(1-A)|\boldsymbol{X}) = \\ & \text{(Consistency assumption)} \\ & E(Y^1|\boldsymbol{X})E(A|\boldsymbol{X}) + E(Y^0|\boldsymbol{X}) - E(Y^0|\boldsymbol{X})E(A|\boldsymbol{X}) = \\ & \text{(Unconfoundedness assumption)} \\ & E(Y^0|\boldsymbol{X}) + \theta(\boldsymbol{X})e(\boldsymbol{X}), \end{aligned} \tag{1}$$

where we denote $\theta(\boldsymbol{X}) = E(Y^1 - Y^0|\boldsymbol{X})$ as the treatment effect of $\boldsymbol{X}$, and $e(\boldsymbol{X}) = E(A|\boldsymbol{X}) = P(A = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ is the propensity score denoting the probability of $\boldsymbol{X}$ receiving treatment.

3.2 Estimator of subgroup-level treatment effect

We first construct suitable STE estimators that adapt to tree splitting. The observed outcome consists of two main effects implied on two paths: one is the baseline effect $E(Y^0|\boldsymbol{X})$ from $\boldsymbol{X}$ to Y, and the other is the treatment effect $A \cdot \theta(\boldsymbol{X})$ from A to Y, so we have

$$Y = E(Y^0|\boldsymbol{X}) + A \cdot \theta(\boldsymbol{X}) + \epsilon, \quad E(\epsilon \mid A, \boldsymbol{X}) = 0.$$

Furthermore, the baseline effect $E(Y^0|\boldsymbol{X})$ can be replaced by $E(Y|\boldsymbol{X}) - \theta(\boldsymbol{X})e(\boldsymbol{X})$ according to Eq. (1).

During tree growth, the observations falling into the same node could be viewed as a subgroup. Given a split s, a left node C_l and a right node C_r will be formed, then, the model for the observed outcome on the formed subgroups can be built as

$$Y = E(Y|\boldsymbol{X}) + (A - e(\boldsymbol{X})) \cdot I_s(\boldsymbol{X}) \cdot \boldsymbol{\theta}_s + \epsilon, \quad E(\epsilon \mid A, \boldsymbol{X}) = 0, \tag{2}$$

where we use the average treatment effect of the child node to approximate the treatment effects of units falling into it, e.g., $\theta(\boldsymbol{X}) = I_s(\boldsymbol{X}) \cdot \boldsymbol{\theta}_s$, $\boldsymbol{\theta}_s = (\theta_l, \theta_r)^{\mathsf{T}}$ is the vector of treatment effects on left and right nodes, and $I_s(\boldsymbol{X}) = (I(\boldsymbol{X} \in C_l), I(\boldsymbol{X} \in C_r))$ denotes the child node that $\boldsymbol{X}$ is falling in. Model (2) is constructed following the Robinson-style method[6]. There are several pieces of literature[18, 22, 23] estimating treatment effects based on the same idea from Robinson transformation, which eliminates the effect of covariates from treatment and outcome. From the perspective of the causal graph, this model blocks the backdoor path from treatment A to outcome Y by controlling the covariates $\boldsymbol{X}$.

For the nuisance functions in model (2), the conditional mean of the outcome $E(Y|\boldsymbol{X})$, and the propensity score $e(\boldsymbol{X})$, we use their leave-one-fold estimators as plug-ins into model (2). We randomly divide the indices set $I = \{1, 2, \cdots, N\}$ of the whole sample into K folds with equal sample sizes, $I = \bigcup_{k=1}^{K} I_k$, $|I_k| = N/K$. The nuisance functions are fitted on the sample indexed by the complement of I_k, denoted as I_k^C, and $\widehat{E}_k(Y_i|\boldsymbol{X}_i)$ and $\widehat{e}_k(\boldsymbol{X}_i)$ are estimated on I_k, $i \in I_k$. Iterating through $k = 1, 2, \cdots, K$, we can obtain the nuisance estimators for the entire sample. Given the estimated nuisances, the treatment effect $\boldsymbol{\theta}_s$ of subgroups formed by split s can be estimated as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_s = & \left\{ \sum_{k=1}^{K} \sum_{i \in I_k} (A_i - \widehat{e}_k(\boldsymbol{X}_i))^2 I_s(\boldsymbol{X}_i)^{\mathsf{T}} I_s(\boldsymbol{X}_i) \right\}^{-1} \cdot \\ & \left\{ \sum_{k=1}^{K} \sum_{i \in I_k} (Y_i - \widehat{E}_k(Y_i|\boldsymbol{X}_i)) \cdot (A_i - \widehat{e}_k(\boldsymbol{X}_i)) I_s(\boldsymbol{X}_i)^{\mathsf{T}} \right\}. \end{aligned} \tag{3}$$

To derive the properties for $\hat{\boldsymbol{\theta}}_s$, we put some limitations on the estimated nuisance functions. Here, we denote the $L^q(P)$-norm of random variable $\boldsymbol{Z}$ as $\|\boldsymbol{Z}\|_{P,q} = \{\int \|z\|_2^q \mathrm{d}P(z)\}^{1/q}$, where $P(z)$ is the probability distribution of $\boldsymbol{Z}$.

Assumption 3.1. (Bounded moment) The $L^q(P)$-norm of outcome $\|Y\|_{P,q}$ is bounded for all $q \geq 2$; $E(\epsilon^2 \mid \boldsymbol{X})$ is bounded.

Assumption 3.2. (Consistency of nuisance estimators) The nuisance estimators $\widehat{E}_k(Y|\boldsymbol{X})$ and $\widehat{e}_k(\boldsymbol{X})$ are estimated on the

DOI: 10.52396/JUSTC-2022-0054
JUSTC, 2023, 53(11): 1102

subset I_k of sample $I=\{1,2,\cdots,N\}$, where $I=\bigcup_{k=1}^{K} I_k$ and $|I_k|=N/K$. For each $k=1,2,\cdots,K$, the nuisance estimators obey the following conditions:

① $\left\|\widehat{E}_k(Y|\boldsymbol{X})-E(Y|\boldsymbol{X})\right\|_{P,2}\to 0$, $\left\|\widehat{e}_k(\boldsymbol{X})-e(\boldsymbol{X})\right\|_{P,2}\to 0$;

②

$$\sqrt{N}\,\|\hat{e}_k(\boldsymbol{X})-e(\boldsymbol{X})\|_{P,2}\times$$
$$\left(\|\hat{e}_k(\boldsymbol{X})-e(\boldsymbol{X})\|_{P,2}+\|\hat{E}_k(Y|\boldsymbol{X})-E(Y|\boldsymbol{X})\|_{P,2}\right)\to 0.$$

Assumption 3.1 gives the bounded moment conditions of model (2); Assumption 3.2 restricts the convergence rate of nuisance estimators, and this rate could be available for many modern ML methods[22]. A simple way is to fit nuisance functions on the whole training data before tree building. As we obtain the nuisance estimation for each sample, storing and passing them into the tree-building process for subsequent use, can reduce the frequency of model fitting and ensure the full efficiency of the entire training data.

Theorem 3.1. Given a split s, suppose that Assumptions 3.1 and 3.2 hold. The treatment effect estimator $\hat{\boldsymbol{\theta}}_s=(\hat{\theta}_l,\hat{\theta}_r)^{\mathrm{T}}$ defined by Eq. (3) is asymptotically normal,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_s-\boldsymbol{\theta}_s)\xrightarrow{\mathrm{d}}N(0,\Sigma_s), \tag{4}$$

where $\Sigma_s=\mathrm{diag}(\sigma_l^2,\sigma_r^2)$ is a covariance matrix, and the asymptotic variance of the treatment effect of C_l and C_r can be estimated as

$$\widehat{\mathrm{Var}}(\hat{\theta}_j)=\frac{\sum_{k=1}^{K}\sum_{i\in I_k}(A_i-\widehat{e}_k(\boldsymbol{X}_i))^2\widehat{\epsilon}_i^2\,I(\boldsymbol{X}_i\in C_j)}{\left\{\sum_{k=1}^{K}\sum_{i\in I_k}(A_i-\widehat{e}_k(\boldsymbol{X}_i))^2\,I(\boldsymbol{X}_i\in C_j)\right\}^2},\quad j=l,r, \tag{5}$$

where $\widehat{\epsilon}_i=Y_i-\widehat{E}_k(Y_i|\boldsymbol{X}_i)-(A_i-\widehat{e}_k(\boldsymbol{X}_i))\hat{\theta}_j$.

3.3 Generic causal tree algorithm

In this section, we introduce the procedures of tree generation based on the above-defined STE estimators. We preserve the core frames of interaction trees[13, 14], including partitioning, pruning, and selection, while using a different splitting criterion and replacing the estimation idea. Since the semiparametric causal effect model in Section 3.2 is suitable for general scenarios with confounding interference and allows multiple machine learning algorithms to fit nuisance functions, and the framework of such tree architecture is generic, we call it a GCT.

3.3.1 Tree splitting

The central problem is to seek a split that immediately maximizes the heterogeneity on child nodes. For the disjoint subgroups in the left and right branches, we wonder if their treatment effects are significantly different. Analogous to the two-sample hypothesis test, we construct a splitting statistic:

$$T^2(s)=\left(\frac{\hat{\theta}_l-\hat{\theta}_r}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta}_l)+\widehat{\mathrm{Var}}(\hat{\theta}_r)}}\right)^2, \tag{6}$$

where $\hat{\theta}_l$ and $\hat{\theta}_r$ are the estimated treatment effects of the left and right child nodes formed by a given split s, respectively, and $\widehat{\mathrm{Var}}(\hat{\theta}_l)$, $\widehat{\mathrm{Var}}(\hat{\theta}_r)$ are their estimated variances. The estimators in the splitting statistic (6) can be estimated by Eqs. (3) and (5). Here, the split statistic imposes two constraints on the tree splits: i) maximize the heterogeneity of treatment effects in both subgroups; ii) minimize the uncertainty within each subgroup. We calculate split statistics at all candidate splits, then the split with maximum statistic T^2 will be conducted. It should be noted that, there exist some splits that violate the above model setting and assumptions during the split-searching procedure, we simply regard the estimators in Eq. (6) as rough approximations at such times.

We repetitively execute the splitting procedure on new-generated nodes until meeting a prespecified terminal condition, for example, reaching the minimum number limitation of observations in each node. Here, both the number of treated units and control units should be constrained to avoid poor estimation performance of the treatment effect in the nodes.

3.3.2 Pruning and tree selection

We adopt the pruning idea of CART[16] and generate a sequence of subtrees by gradually cutting the weakest branches of the initial tree. The goodness of a given tree Π is accumulative of splitting statistics at all its splits, reflecting the heterogeneity it carries:

$$G(\Pi)=\sum_{s}T^2(s). \tag{7}$$

Following the work of Su et al.[13], the heterogeneity-complexity for a given tree Π can be defined as

$$G^{(\lambda)}(\Pi)=G(\Pi)-\lambda\left|\widetilde{\Pi}\right|, \tag{8}$$

where $\widetilde{\Pi}$ represents a subtree containing all internal nodes of Π, $|\cdot|$ is the number of tree nodes, and λ is a penalty parameter.

Let Π_h denote a subtree that is rooted in an internal node h and comprises all descendants of h. Subtree Π_h is going to be cut only when the heterogeneity and complexity of Π_h improves after pruning, leaving a single node h. Then we have

$$G(\Pi_h)-\lambda\left|\widetilde{\Pi}_h\right|=\underset{\text{(Before pruning)}}{G^{(\lambda)}(\Pi_h)}\leqslant\underset{\text{(After pruning)}}{G^{(\lambda)}(h)}=0\Rightarrow\lambda\geqslant\frac{G(\Pi_h)}{|\widetilde{\Pi}_h|}.$$

During pruning, we iterate through each internal node h of a given tree Π and calculate the corresponding

$$g(h)=G(\Pi_h)/\left|\widetilde{\Pi}_h\right|. \tag{9}$$

The branch with minimal $g(h)$, denoted as $g(h^*)$, will be cut off. That means once the degree of penalty λ grows exactly up to $g(h^*)$, the weakest split Π_{h^*} will be truncated while other splits will remain.

Continuing the pruning procedure until the initial tree Π_0 is pruned to be a root node Π_M, we can obtain a sequence of the subtree $\{\Pi_M,\cdots,\Pi_1,\Pi_0\}$. Given a penalty parameter λ①, we

① Ordinarily, we set $\lambda=\chi^2_{1-\alpha}(1)$, which is the $1-\alpha$ quantile of the asymptotic distribution of Eq. (6) when there is no significant difference of treatment effects in two child nodes.

DOI: 10.52396/JUSTC-2022-0054
JUSTC, 2023, 53(11): 1102

select the optimal tree by evaluating the heterogeneity and complexity (8) of each candidate subtree with validation data:

$$\widehat{\Pi} = \arg\max_{m\in\{0,1,\ldots,M\}} \{G^{(\iota)}(\Pi_m)\}. \tag{10}$$

The procedures for building the GCT are shown in Algorithms 1 and 2.

4 Honest estimation for subgroup treatment effect

One of our central concerns is finding a solution to construct confidence intervals for the tree-based treatment effect estimators. It is still challenging to attempt direct inference since the derived subgroups may change with different tree models. Refs. [24, 25] suggested using conditional inference by conditioning on the selected model to solve this issue since the uncertainty of targets of inference is induced by the randomness of the selected model. Additionally, Refs. [5, 17, 18] employed honest estimation, an approach based on data splitting that usually appears in the context of conditional inference, allowing for valid statistical inference with a given model that is built independently of the data on which inference is based Ref. [26]. This, combined with the theory of semiparametric framework[22], gives clear conditional distributions of our tree estimators and sheds light on valid inference.

The key point of honest estimation is employing independent data for different tasks. Recall that in the regular version of GCT, the same dataset is used to decide splits and estimate treatment effects, as described in Section 3.3. Here, in the honest version of GCT, we divide the whole sample into two equal-sized parts. Half of the sample will be used to generate the tree model; this procedure is the same as Algorithm 1; The other half of the sample, called estimation data, will be put into the given tree to estimate treatment effects and not participate in tree building.

Let Ω denote the covariate space. Given a specific partition $\widehat{\Pi}$ decided by data $\mathcal{S}^{\text{tr}}\cup\mathcal{S}^{\text{va}}$, the covariate space Ω is divided into G disjoint subspaces $\Omega=\bigcup_{g=1}^{G}\Omega_g$. The vector of treatment effect $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots,\theta_G)^{\mathsf{T}}$ on the formed subgroups will be estimated with the estimation data $\mathcal{S}^{\text{es}}$ as if subgroups are prespecified. For the observed outcome, we construct a model as follows:

$$\begin{aligned} &Y_i = E(Y_i\mid \boldsymbol{X}_i) + (A_i - e(\boldsymbol{X}_i))\cdot I_{\widehat{\Pi}}(\boldsymbol{X}_i)\cdot\boldsymbol{\theta}+\epsilon_i,\\ &E(\epsilon_i\mid A_i,\boldsymbol{X}_i)=0,\ \boldsymbol{X}_i\in\mathcal{S}^{\text{es}}, \end{aligned} \tag{11}$$

where $I_{\widehat{\Pi}}(\boldsymbol{X}_i)=(\hat{I}(\boldsymbol{X}_i\in\Omega_1),\ldots,\hat{I}(\boldsymbol{X}_i\in\Omega_G))$ is a subgroup indicator for indicating which subgroup $\boldsymbol{X}_i$ belongs to under map $\widehat{\Pi}$.

The nuisance functions can be replaced by their leave-one-fold estimators on estimation data $\mathcal{S}^{\text{es}}$, then, the treatment effect $\boldsymbol{\theta}$ can be estimated as

$$\begin{aligned} \hat{\boldsymbol{\theta}} = &\left\{\sum_{k=1}^{K}\sum_{i\in I_k}(A_i-\widehat{e}_k(\boldsymbol{X}_i))^2 I_{\widehat{\Pi}}(\boldsymbol{X}_i)^{\mathsf{T}} I_{\widehat{\Pi}}(\boldsymbol{X}_i)\right\}^{-1}\cdot\\ &\left\{\sum_{k=1}^{K}\sum_{i\in I_k}(Y_i-\widehat{E}_k(Y_i|\boldsymbol{X}_i))\cdot(A_i-\widehat{e}_k(\boldsymbol{X}_i))I_{\widehat{\Pi}}(\boldsymbol{X}_i)^{\mathsf{T}}\right\}, \end{aligned} \tag{12}$$

where I_k is the partition of estimation data.

Assumption 4.1. (Nonempty partition) Let $\widehat{\Pi}$ denote the

Algorithm 1. Generic causal tree for subgroup identification

Input: Sample $\mathcal{S}$, the minimum sample size minsize of treated and control units in nodes, ML methods FitMethod used to fit nuisance, and the number of folds K.
Output: Optimal tree Π^*
1: Training sample $\mathcal{S}^{\text{tr}}$ and validation sample $\mathcal{S}^{\text{va}}$ ⟵ SampleSplit($\mathcal{S}$)
2: Nuisance estimator $\hat{\boldsymbol{\eta}}^{\text{tr}}=(\widehat{E}_k(Y|\boldsymbol{X}),\widehat{e}_k(\boldsymbol{X})),\boldsymbol{X}\in\mathcal{S}^{\text{tr}}$ ⟵ LeaveOneFold($\mathcal{S}^{\text{tr}},K$, FitMethod)
3: Initial tree InitialTree ⟵ INITIALTREEBUILD($\mathcal{S}^{\text{tr}},\hat{\boldsymbol{\eta}}^{\text{tr}}$, minsize) ▷ See Algorithm 2
4: Initialize a list of subtrees subtreeList
5: Initialize TreeToPrune $\dot{\Pi}$ ⟵ InitialTree
6: **while** $|\dot{\Pi}|>1$ **do**
7: Initialize the threshold vector g_h ⟵ Zeros($|\dot{\Pi}|$)
8: **for** node $h\in\dot{\Pi}$ **do**
9: **if** h is leaf **then**
10: g_h ⟵ AddToVector($g_h,+\infty$) ▷ See Eq. (9)
11: **else**
12: g_h ⟵ AddToVector($g_h,G(\Pi_h)/|\widetilde{\Pi}_h|$) ▷ See Eq. (9)
13: Target node $h^*=\arg\min_h g_h$
14: TreeToPrune $\dot{\Pi}$ ⟵ Prune($\dot{\Pi}$, Π_{h^*})
15: subtreeList ⟵ AddToList(subtreeList, $\dot{\Pi}$)
16: Initialize a complexity vector $\mathcal{G}$ ⟵ Zeros(|subtreeList|)
17: **for** $t=1$ to |subtreeList| **do**
18: TreeToValid Π_t ⟵ subtreeList[t]
19: Vector $\mathcal{G}$ ⟵ CalculateComplexity($\Pi_t,\mathcal{S}^{\text{va}}$) ▷ See Eq. (8)
20: The index of optimal tree $t^*=\arg\max_t\mathcal{G}$
21: Optimal tree Π^* = subtreeList[t^*]
22: **return** optimal tree Π^*
* Note: The function LeaveOneFold($\cdot$) gives leave-one-fold estimators of nuisance; CalculateComplexity($\cdot$) gives heterogeneity and complexity by Eq. (8).

Algorithm 2. Initial tree building

Procedure INITIALTREEBUILD ($\mathcal{S}^{\text{tr}}$, minsize, $\hat{\boldsymbol{\eta}}^{\text{tr}}$)
Root node P_0 ⟵ CreateRoot($\mathcal{S}^{\text{tr}}$)
Queue Q ⟵ CreateQueue(P_0)
Parent node P ⟵ P_0
while NotNull(P) **do**
Extract the data on P, $\mathcal{S}^P$ ⟵ $\mathcal{S}^{\text{tr}}$, $\hat{\boldsymbol{\eta}}^P$ ⟵ $\hat{\boldsymbol{\eta}}^{\text{tr}}$
if N^P_{trt}, N^P_{ctl} >= minsize **then**
Initialize the split statistics $t_s\leftarrow 0$
for candidate split s **do**
$(\hat{\theta}_l,\hat{\theta}_r,\hat{\sigma}^2_{\theta_l},\hat{\sigma}^2_{\theta_r})$ ⟵ DoEstimation($\hat{\boldsymbol{\eta}}^P$, $\mathcal{S}^P$) ▷ See Eqs. (3) and (5)
t_s ⟵ max{t_s, SplitStatistic($\hat{\theta}_l,\hat{\theta}_r,\hat{\sigma}^2_{\theta_l},\hat{\sigma}^2_{\theta_r}$)} ▷ See Eq. (6)
Best split s^* ⟵ $\arg\max_s t_s$
Child nodes C_l,C_r ⟵ TreeSplit(P, s^*)
Queue Q ⟵ AddNodes(Q,C_l,C_r)
Queue Q ⟵ Remove(P), remove node P from queue Q
Parent node P ⟵ the earliest node in Q
return InitialTree
* Note: N^P_{trt} and N^P_{ctl} are the sample sizes of the treated units and control units in node P. The function DoEstimation($\cdot$) estimate the treatment effect and variance estimator on a node by Eqs. (3) and (5). Splitting statistic is calculated at each candidate split by SplitStatistic($\cdot$) through Eq. (6).

DOI: 10.52396/JUSTC-2022-0054
JUSTC, 2023, 53(11): 1102

final tree model. The covariate space is divided into G disjoint subspaces $\Omega = \bigcup_{g=1}^{G} \Omega_g$. We assume that there exists a constant $\delta > 0$, such that the given tree $\widehat{\Pi}$ satisfies the

$$\delta < P\left(\widehat{\Pi}(\boldsymbol{X}_i) = g\right) < 1 - \delta. \tag{13}$$

That is, given a specified partition $\widehat{\Pi}$, $\int_{\Omega_g} \mathrm{d}P(\boldsymbol{x}) > \delta$.

Note that the properties of tree estimators are generally derived based on the simplification of fixed partitions, and the following theorem is conditional on a built tree model.

Theorem 4.1. Given a specified partition $\widehat{\Pi}$, suppose that Assumptions 3.1, 3.2, and 4.1 hold, the estimators of subgroup-level treatment effects are asymptotically normal,

$$\sqrt{N_e}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \mid \widehat{\Pi} \xrightarrow{\mathrm{d}} N(0, \Sigma_{\widehat{\Pi}}), \tag{14}$$

where $\Sigma_{\widehat{\Pi}} = \mathrm{diag}(\sigma^2_{1,\widehat{\Pi}}, \ldots, \sigma^2_{G,\widehat{\Pi}})$ is the diagonal covariance matrix. For subgroup Ω_g, the variance estimator $\sigma^2_{g,\widehat{\Pi}}$ can be consistently estimated by

$$\widehat{\sigma}^2_{g,\widehat{\Pi}} = \frac{\dfrac{1}{N_e}\displaystyle\sum_{k=1}^{K}\sum_{i\in I_k}(A_i - \widehat{e}_k(\boldsymbol{X}_i))^2 \widehat{\epsilon}^2_{k,i} \hat{I}(\boldsymbol{X}_i \in \Omega_g)}{\left\{\dfrac{1}{N_e}\displaystyle\sum_{k=1}^{K}\sum_{i\in I_k}(A_i - \widehat{e}_k(\boldsymbol{X}_i))^2 \hat{I}(\boldsymbol{X}_i \in \Omega_g)\right\}^2}, \tag{15}$$

where $\widehat{\epsilon}_{k,i} = Y_i - \widehat{E}_k(Y_i|\boldsymbol{X}_i) - (A_i - \widehat{e}_k(\boldsymbol{X}_i))I_{\widehat{\Pi}}(\boldsymbol{X}_i)\hat{\boldsymbol{\theta}}$.

5 Simulation

In this part, we conduct two simulation studies to test the performance of our GCT algorithm. The first study compares GCT with two benchmark methods: causal tree[5] and causal interaction tree[14], concentrating on their subgroup identification and estimation abilities. In the second study, we test the effectiveness of the GCT algorithm on statistical inference. We retain the experimental setup and performance metrics similar to those of Ref. [14].

5.1 Simulation setup

For the data generating process in the following experiments, we draw N i.i.d. samples of $\boldsymbol{O}_i = (\boldsymbol{X}_i, A_i, Y_i) \in \mathbb{R}^5 \times \{0,1\} \times \mathbb{R}$. The 5-dimensional covariate vector is generated from $\boldsymbol{X}_i \sim N(0, \Sigma_x)$, where Σ_x is a covariance matrix with 1 for diagonal elements and 0.3 for nondiagonal items. We set all variables as confounding to affect both treatment A and outcome Y. The propensity score $e(\boldsymbol{X})$ is designed as

$$e(\boldsymbol{X}_i) = \frac{1}{1 + \exp\left(-0.5X_i^{(1)} + 0.5X_i^{(2)} - 0.5X_i^{(3)} + 0.5X_i^{(4)} - 0.5X_i^{(5)}\right)}.$$

The treatment A_i for each individual is generated from a Bernouilli distribution Bern(p_i) with probability $p_i = e(\boldsymbol{X}_i)$. Consider the following two cases:

(i) Homogeneous setting with the same treatment effect for all units. The outcome is generated as follows:

$$Y_i = 2 + 2A_i + \left(X_i^{(1)}\right)^2 + \exp\left(X_i^{(2)}\right) + 2X_i^{(3)} + 3I\left(X_i^{(4)} > 0\right) + \left(X_i^{(5)}\right)^3 + \epsilon,$$

where $\epsilon \sim N(0,1)$. The treatment effect is equal to 2 on the whole population.

(ii) Heterogeneous setting with treatment effects varying from the values of the variable $X^{(4)}$. The outcome is generated as follows:

$$Y_i = 2 + 2A_i + \left(X_i^{(1)}\right)^2 + \exp\left(X_i^{(2)}\right) + 2X_i^{(3)} + 3A_i I\left(X_i^{(4)} > 0\right) + \left(X_i^{(5)}\right)^3 + \epsilon,$$

where $\epsilon \sim N(0,1)$. The treatment effect of $\boldsymbol{X}_i$ is equal to 5 if $X_i^{(4)} > 0$ and equal to 2 if $X_i^{(4)} < 0$.

5.2 Performance metrics

We evaluate these tree-based methods on test data with sample size $N_{\mathrm{te}} = 1000$. The following metrics will be reported:

The probability of correct trees (corr.tree): Trees that split at correct split variables with correct split times are considered as correct trees.

The number of leaf nodes (num.leaf): The correct number of leaf nodes of a tree should be 1 in the homogeneous case and 2 in the heterogeneous case.

The number of noise splits (num.noise): It is the number of times that tree splits at noise split variables. In homogeneous cases, all covariates are noise split variables; in heterogeneous cases, $\{X^{(1)}, X^{(2)}, X^{(3)}, X^{(5)}\}$ are noise split variables.

Accuracy of the first split (fir.splt.acc): This is the proportion of final trees that make the correct first split. Only applicable to heterogeneity setting. The first split variable is generally of the greatest contribution to heterogeneity.

Pairwise prediction similarity (PPS)[14]: Measures the accuracy of the subgroup partition, that is,

$$1 - \sum_{i=1}^{N_{\mathrm{te}}}\sum_{j>i}^{N_{\mathrm{te}}} \frac{|\, I_t(i,j) - I_e(i,j)\,|}{\binom{N_{\mathrm{te}}}{2}},$$

where $I_t(i,j)$ and $I_e(i,j)$ indicate whether the observations $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ fall in the same subgroup in the true model and the established tree, respectively.

Mean of square error (MSE): $\theta(\boldsymbol{X})$ is the true treatment effect and $\hat{\theta}(\boldsymbol{X})$ is the estimated treatment effect in test data:

$$\frac{1}{N_{\mathrm{te}}}\sum_{i=1}^{N_{\mathrm{te}}}(\theta(\boldsymbol{X}_i) - \hat{\theta}(\boldsymbol{X}_i))^2.$$

5.3 Implementation details

Code for the simulations is available at https://github.com/Caiwei-Zhang/GenericCausalTree. We implement our GCT method in R based upon the user-written split functions in the R package rpart and retain the main frame of the code by Yang et al.[14]. There are two options considered in our simulations to fit nuisance functions in GCT and CIT. In addition to fitting nuisance functions with the whole dataset before tree building, as we described earlier, the other way is to fit nuisance with the samples in each parent node during tree splitting. The latter enables the capture of the local data characteristics but may also increase the risk of overfitting and slow down tree generation. We will compare the two model-fitting patterns and then select the better one to apply in the

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

data analysis in Section 6. The penalty parameter λ is chosen to be the 95% quantile of $\chi^2(1)$. We set 80% of the data as training data for tree splitting and the remaining 20% as validation data for tree selection. For honest GCT in the second study, we will first take out half of the data for estimation and the remaining half for tree building.

We implement the causal tree in the R package causalTree[①]. The tree splitting and validation criteria need to be specified before running the tree. Referring to Yang et al.[14], two versions of the causal tree are be considered: the original version and the tuning version. The original causal tree (Org.CT) takes the "honest causal tree" criterion as splitting and validation rules. For the tuning version of causal tree (Opt.CT), several alternative splitting and validation rules are provided in the package causalTree, and we will select the optimal combination of rules with the best average performance by grid tuning[②]. In addition, we will perform propensity score weighting in the estimation during the tree-building process to accommodate observational data, as Athey and Imbens suggested[5].

For the ML algorithms used to fit outcome models and propensity score models, we try gradient boosting machine (GBM) and logistic regression model, respectively. All the ML algorithms are implemented with the same parameter settings. We note that the conclusions about the performance of these methods are ordinarily independent of the particular ML algorithm. Many other powerful ML algorithms can be chosen to fit the nuisance functions, such as random forest.

5.4 Simulation results

5.4.1 Comparison of benchmark methods

We summarize the results and provide general remarks on the strengths and drawbacks of the CT, CIT, and GCT algorithms in our simulations. All three algorithms are implemented following Section 5.3. Table 1 reports the average performance metrics in both heterogeneous and homogeneous settings among 1000 repetitions.

Overall, GCT has excellent subgroup identification and estimation performance, followed by CIT-DR and Opt.CT. GCT slightly outperforms CIT in the heterogeneous setting, and both Opt.CT and CIT perform well in the homogeneous setting. In either setting, the estimation error of GCT is the smallest. The performances of CIT under two model-fitting patterns are similar in this experimental setting. However, fitting nuisance functions in parent nodes seems to slightly enhance the performance of GCT. We note that in a more complicated heterogeneous setting, fitting models in parent nodes during splitting could be a wise choice since the response functions are different for distinct parts of covariate space; fitting models in subspaces can greatly aid the detection of local structures.

In terms of tree splitting in the heterogeneous setting, GCT builds the most correct tree models during 1000 repetitions under both model-fitting patterns. Meanwhile, the first split variable is generally of the greatest contribution to heterogeneity, obtaining the highest accuracy of the first split shows that GCT has a strong ability to detect heterogeneity. Additionally, GCT shows good power in subgroup identification

Table 1. The summary of average performance metrics during 1000 repetitions for three tree-based algorithms. The upper panel "Hetero" corresponds to the results in heterogeneous setting; the lower panel "Homo" corresponds to the results in homogeneous setting. The optimal metrics are marked in bold text.

Setting	Metrics	CT		CIT-DR		GCT	
		Org.CT	Opt.CT	fitBe	fitIn	fitBe	fitIn
Hetero	corr.tree	0.002	0.459	0.803	0.806	0.822	**0.862**
	num.leaf	31.108	2.294	1.890	1.872	2.388	**2.046**
	num.noise	22.859	0.490	0.025	**0.015**	0.309	0.070
	fir.splt.acc	0.622	0.677	0.837	0.836	**0.970**	0.939
	PPS	0.621	0.779	0.896	0.895	**0.948**	0.943
	MSE	9.107	1.653	0.560	0.553	**0.268**	0.288
	time	0.577	**0.504**	5.315	4.514	1.079	2.401
Homo	corr.tree	0.037	**0.987**	0.975	0.976	0.956	0.956
	num.leaf	9.324	**1.023**	1.032	1.037	1.055	1.059
	num.noise	8.324	**0.023**	0.032	0.037	0.055	0.059
	PPS	0.259	**0.995**	0.990	0.989	0.980	0.980
	MSE	4.557	0.138	0.052	0.051	**0.033**	0.034
	time	0.512	**0.501**	2.593	3.278	0.483	1.561

with the best PPS.

We notice that the average number of leaf nodes of CIT-DR is less than 2 and the number of noise splits is close to 0, indicating that the splitting of CIT is relatively conservative in the heterogeneous case. This may be attributed to the method of treatment effect estimation. The treatment effect estimation of the CIT-DR algorithm is based on the propensity score weighed counterfactual prediction. This method pools the treatment A and covariates $\boldsymbol{X}$ together to fit the counterfactual functions $E(Y|\boldsymbol{X},A)$, where the treatment is considered in the same position as the other covariates, which may weaken the effect of treatment. When fitting this outcome model, the response surfaces corresponding to $A=1$ and $A=0$ interfere, blurring the difference between them and negatively affecting the prediction of $E(Y|\boldsymbol{X},A)$. Since the two response surfaces are more different in heterogeneous settings than in homogeneous settings, we find that the CIT's performance in homogeneous settings is always better than that in heterogeneous settings in the following simulations. At the same time, the pooled data increase the difficulty of fitting outcome model $E(Y|\boldsymbol{X},A)$, which may place higher demands on the ML algorithm used. The results in Supporting information S.2.1 also demonstrate this point: CIT is relatively sensitive to different ML algorithms. In contrast, for GCT, we do not directly conduct the prediction involving A on Y when fitting outcome models. We fit $E(Y|\boldsymbol{X})$, the effect of $\boldsymbol{X}$ on Y, and then remove this part of the effect from Y by partialing out. This method can alleviate the interference

① github.com/susanthey/causalTree

② According to the tuning result, the best combination of parameters for Opt.CT is split.Rule = "tstats", split. Honest = "TRUE", cv.option = "matching" in heterogeneous setting, and split.Rule = "tstats", split.Honest = "FALSE", cv.option = "matching" in a homogeneous setting.

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

caused by obvious differences between response surfaces. This may be why GCT outperforms CIT in this heterogeneous setting.

In addition, the performance of Opt.CT is far better than that of Org.CT, indicating that the causal tree is extremely dependent on parameter tuning. The Org.CT performs poorly in both settings since it tends to generate very large trees. As we noted, the causal tree criterion guides tree-splitting according to expected mean square error (EMSE), and it is prone to introduce split even when the two branches have the same treatment effect, as long as the variances of outcome in the treated group and control group decrease. Splitting on the noise variables that only have an impact on the outcome but not on the treatment will also lower the EMSE.

In terms of estimation, the MSE shown in Table 1 and the boxplot in Fig. 1 highlight the promise of the GCT algorithm for accurate estimations of treatment effects. GCT yields the lowest MSE among these methods in both settings.

To investigate the performance of these methods as the sample size grows, we gradually increase the sample size from 1000 to 20000 and report the trend of MSE as shown in Fig. 2. The MSE of GCT and Opt.CT declines steadily as the sample size increases. In contrast, the MSE by CIT algorithms cannot converge under the given sample sizes, and its splitting performance does not improve with the increased sample size. The complete performance metrics are presented in Supporting information S.2.2 because of limited space, with a brief analysis of tree splitting and estimation.

Furthermore, we conduct additional experiments to evaluate the performance of these methods in different scenarios more comprehensively: (Ⅰ) We have mentioned above that changing the ML algorithm used to fit outcome models will not influence the ranking of these tree-based methods. To illustrate this, we try to fit the outcome models of GCT and CIT by random forest (RF) and generalized linear model (GLM) in addition to GBM. The performance metrics are summarized in Supporting information S.2.1. It demonstrates that the choice of ML algorithms will affect the performance of trees but could hardly change the ranking of trees: GCT has better performance regardless of ML methods. It also indicates that CIT is relatively sensitive to ML methods with

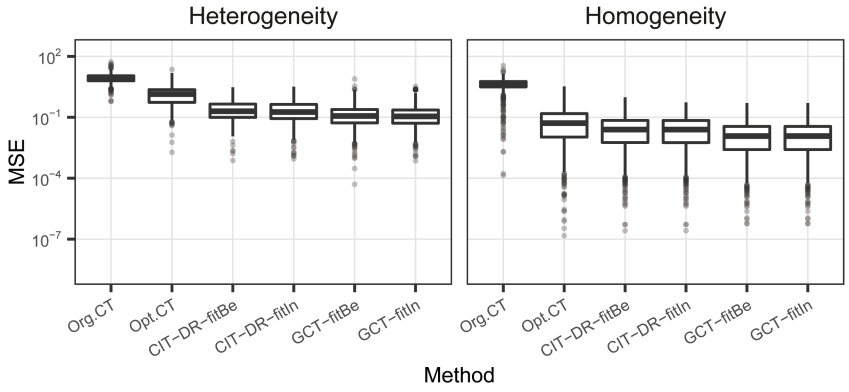


Fig. 1. The distribution of MSE over 1000 simulated datasets in heterogeneous (left) and homogeneous settings (right).

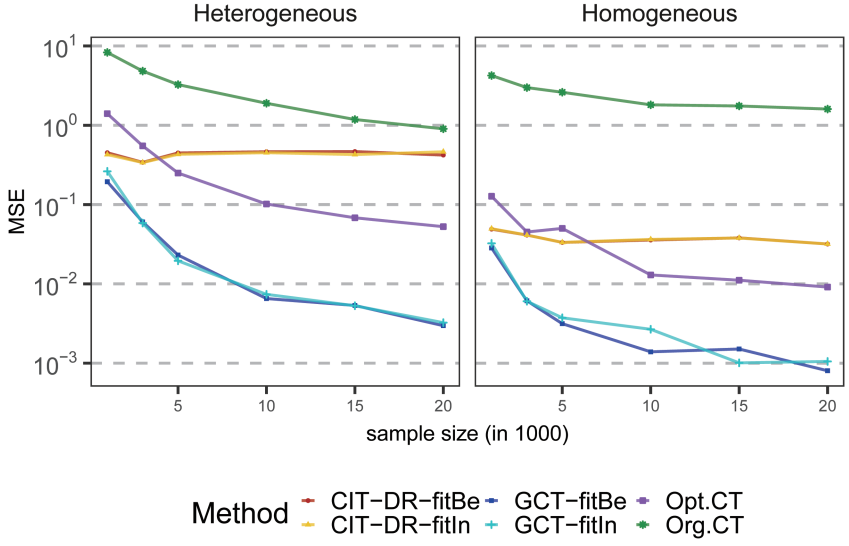


Fig. 2. The MSE curves of three tree-based algorithms with sample sizes increased from 1000 to 20000. The curves of CIT-DR-fitBe and CIT-DR-fitIn almost overlap in a homogeneous setting.

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

larger fluctuations in the subgroup identification performance. (Ⅱ) The results show that all three methods are negatively impacted by unobserved confounding. In terms of tree splitting, the performance of CIT drops the most, followed by GCT and Opt.CT drops the least. However, GCT still performs best. For more details, see the results in Supporting information S.2.3.

5.4.2 The inference by honest GCT

For the second study, we test the effectiveness of honest GCT in terms of inference. We report the coverage rate of 95% confidence intervals on each subgroup in the correct tree models. We further show the simultaneous coverage rate of 95% simultaneous confidence intervals on both subgroups, following the approach of Hothorn et al.[27], which allows for simultaneous inference procedures in semiparametric models, widening confidence intervals by controlling the type I error for each null hypothesis lower than $\alpha^* = 1-(1-0.05)^{1/G}$. As we can obtain the confidence intervals of all the nodes in each tree by GCT, we wonder if the confidence interval at each node in a tree could cover its true average treatment effect, even if the tree is incorrectly built. To this end, we report the coverage proportion (Cov.Prop), which shows the proportion of confidence intervals covering the true average treatment effects of nodes in a tree.

The results are shown in Table 2. We notice two points from the results. First, when $N = 2000$, nonhonest GCT performs better than honest GCT. Focused on the heterogeneous setting, the coverage rates of the single confidence interval and simultaneous confidence interval by nonhonest GCT are closer to the nominal coverage rate than honest GCT, indicating that halving the sample indeed sacrifices the accuracy of both partition and estimation for honest GCT. However, as N increases, the performance discrepancies in subgroup identification and estimation between the two versions of GCT narrow, such as the proportion of correct trees and MSE, are almost equal to those of nonhonest GCT. In general, both versions of GCT can give valid conditional coverage. Meanwhile, the average percentage of confidence intervals covering the true average treatment effects on the tree nodes is almost 90%. Thus, we can be more confident in inferring the average treatment effect of each node in the tree by GCT.

In Supporting information S.2.4, we further implement the honest GCT and honest CIT with sample size $N = 2000$ to compare their inference performance. The results show that although CIT has good performance in splitting and estimation as before, considering the inference performance, CIT lacks validity, and its coverage rate of 95% confidence intervals is much lower than the nominal coverage rate. Even in the homogeneous case, CIT's coverage rate of confidence intervals is not ideal.

Table 2. Summary of the performance metrics for honest GCT and nonhonest GCT. On the basis of the correct trees, we report the coverage rate of their confidence intervals. In the heterogeneous setting, Cov1 and Cov2 denote the proportion of times that confidence intervals cover the true treatment effect on each subgroup during 1000 repetitions, and "Covsim" reports the proportion of times that confidence intervals simultaneously cover the true treatment effects on both subgroups during 1000 repetitions. In a homogeneous setting, "Cov" denotes the proportion of times that confidence intervals cover the true treatment effect on the root node. The abbreviations for other metrics are described in Table 1.

Metrics	$N = 2000$				$N = 4000$			
	honest GCT		nonhonest GCT		honest GCT		nonhonest GCT	
	fitBe	fitIn	fitBe	fitIn	fitBe	fitIn	fitBe	fitIn
Hetero								
corr.tree	0.767	0.873	0.849	0.910	0.857	0.916	0.857	0.920
num.leaf	2.315	2.068	2.568	2.156	2.580	2.156	2.934	2.160
num.noise	0.292	0.075	0.422	0.120	0.443	0.117	0.735	0.118
fir.splt.acc	0.915	0.953	0.999	0.996	0.994	1.000	1.000	1.000
PPS	0.924	0.951	0.975	0.982	0.972	0.983	0.976	0.989
MSE	0.288	0.206	0.070	0.063	0.059	0.045	0.039	0.025
Cov1	0.939	0.935	0.939	0.948	0.940	0.937	0.956	0.952
Cov2	0.932	0.940	0.947	0.953	0.949	0.948	0.946	0.949
Cov.sim	0.934	0.924	0.958	0.956	0.935	0.932	0.953	0.952
Cov.prop	0.945	0.947	0.898	0.929	0.946	0.944	0.885	0.912
time	1.317	2.873	5.190	9.654	5.376	9.055	25.433	31.502
Homo								
corr.tree	0.950	0.948	0.953	0.942	0.946	0.936	0.963	0.937
num.leaf	1.063	1.060	1.077	1.089	1.068	1.104	1.046	1.106
PPS	0.978	0.979	0.979	0.976	0.980	0.971	0.989	0.976
MSE	0.016	0.015	0.011	0.011	0.006	0.006	0.004	0.006
Cov	0.954	0.961	0.956	0.968	0.962	0.959	0.963	0.963
Cov.prop	0.955	0.959	0.941	0.949	0.962	0.959	0.950	0.943
time	0.696	1.956	3.114	7.520	3.311	7.175	18.210	27.948

6 Data analysis

To illustrate the feasibility of the GCT algorithm proposed in the proceeding sections, we conduct subgroup identification in an observational study that evaluates the effect of race on access to opioid use disorder (OUD) treatment in the United States. These data come from the Treatment Episode Data Set: Admissions 2015 (TEDS-A-2015), a program that collected information on admissions to substance abuse treatment in 2015, which is available on the website of Substance Abuse & Mental Health Data Archive (SAMHDA)①. We only use the data of Maryland in the following analysis. The original dataset of Maryland contains information on 107509 participants with 62 variables. After removing the redundant variables and filtering out invalid levels of some variables, the final dataset contains 45396 observations with 26 variables. The processed data provide demographic characteristics such as age, gender, race, marital status, education, pregnancy at the time of admission, and veteran status, as well as socio economic characteristics such as employment status, living arrangement, the source of income, health insurance, and primary source of payment. Substance abuse characteristics, such as whether the patient receives medication-assisted opioid therapy, the number of prior treatment episodes, the

① https: //www.datafiles.samhsa.gov/

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

primary substance abuse problem and the usual route of administration, are also included. In related studies[28, 29], the disparities between African Americans and Whites are mainly discussed. Here we set the treatment variable A as an indicator for being African American or White, where $A = 1$ denotes Whites (29901 participants) and $A = 0$ denotes African Americans (15495 participants). The outcome variable Y is the number of waiting days to receive OUD treatment. We truncate the extremely large values of outcome Y as the 99% quantile. A detailed description of all variables is provided in Supporting information S.3.1.

We apply the honest GCT algorithm to the processed dataset. GBM is used to fit both the propensity score and conditional mean of the outcome before tree building. In the implementation of honest GCT, the ratio of training data and estimation data is 1∶1; and among training data, we extract 20% as validation data to perform tree selection.

6.1 Analysis

The optimal tree by honest GCT is shown in Fig. 3. The final tree consists of four terminal nodes, splitting at whether or not medication-assisted opioid therapy (MAT) was received (methuse), the primary source of payment (primpay), and the number of prior treatment episodes (noprior). Thus, receiving MAT is the most vital variable for racial disparities. Personal financial status and previous substance abuse situations are also critical factors that affect access to OUD treatment. Contrary to general knowledge, the treatment effect estimator in the whole population (root node) is 0.042, indicating that the average waiting time for Whites is slightly longer than that for African Americans overall. However, the situation will be different in subgroups. We find that after the first split, based on whether a patient received MAT, the waiting days for Whites is 0.18 shorter than African Americans in those who received MAT (72% of the whole population). For another branch that did not receive MAT (28% of the population), African Americans have a priority of 0.61 day over Whites in receiving OUD treatment. The second split is at the primary sources of payment. Based on the first split, when the primary sources of payment of patients fall in self-pay, health insurance companies, medicare, medicaid, or other government payment, which have a proportion of 70%, Whites still have a slight priority of 0.078 day in access to OUD treatment. It is worth noting that those who have not received MAT and whose primary source of payment is other, experience the most significant racial disparities in OUD treatment, reaching 3.3 days. Whites have overwhelming superiority to OUD treatment in this subgroup.

The tree splits by the number of previous treatment episodes at the third level. In addition to MAT and the primary source of payment, if participants had no prior treatment episodes, which cover 20% of the whole population, the waiting time for Whites is longer than that for African Americans by 0.29 day. On the other 50% of participants who have had at least one prior treatment episode, Whites have a priority of 0.21 day compared to African Americans. Although it is habitually believed that African Americans are generally at a disadvantage over Whites in access to OUD treatment, the results reveal that African Americans have priority in receiving OUD treatment in almost half of the participants in Maryland state. Strongest racial inferiority for African Americans only appears in approximately 2% of the population who have received MAT and pay by other primary sources for this treatment episode.

Looking at the most affected subgroup in Fig. 4, Whites also show a delay in receiving treatment compared with other

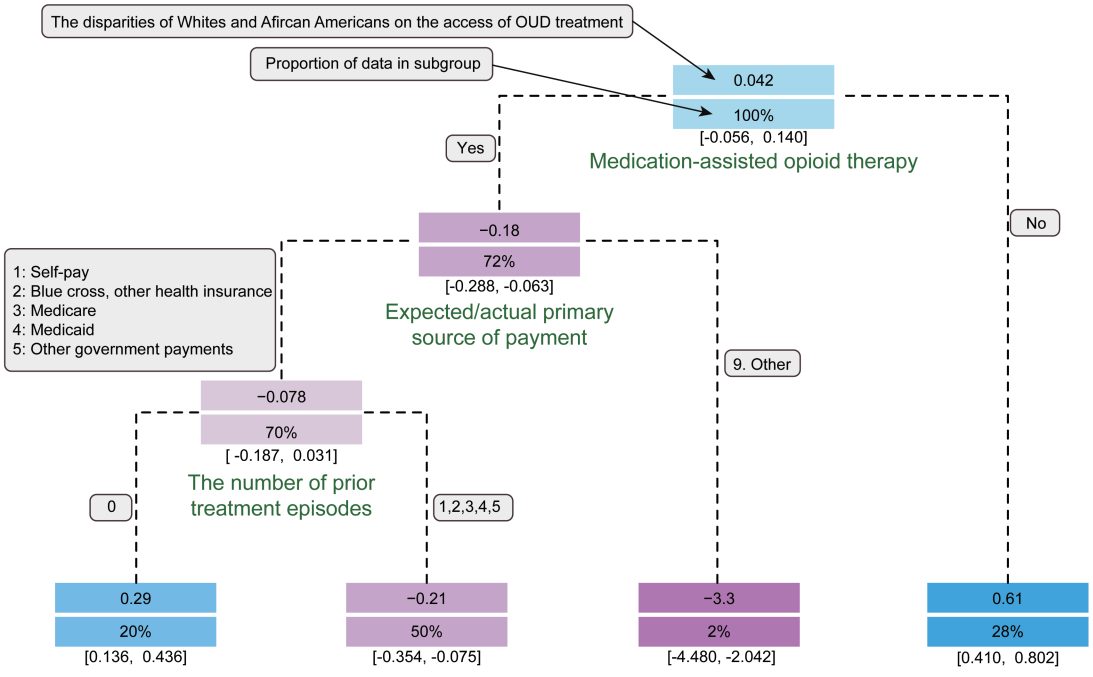


Fig. 3. The final tree model for TEDS-A (Maryland, 2015). The square bracket under each node is the confidence interval for this subgroup.

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

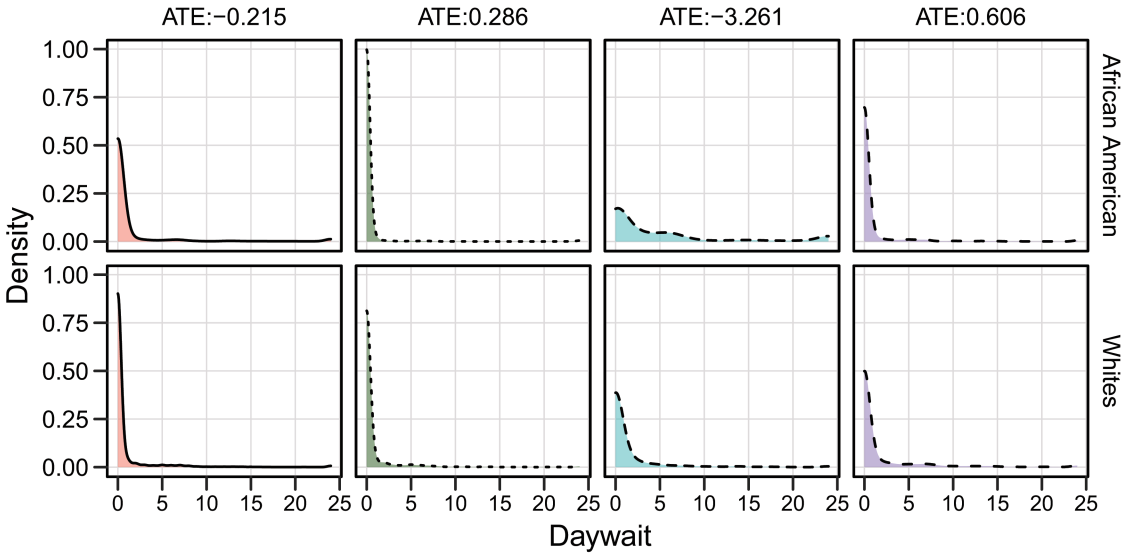


Fig. 4. The distribution of days waiting to OUD treatment in four subgroups.

subgroups. However, African Americans are more affected, so the waiting time still presents a substantial racial discrepancy. In this subgroup, the distribution of days waiting of African Americans is dispersed: Nearly half of the African Americans wait for more than two days to enter OUD treatment, while a similar waiting period is rare in other subgroups. Therefore, more attention should be given to patients who have received MAT and use other sources of payment, especially African Americans.

To further compare with the literature on the effect of racial disparities on access to OUD treatment, we extract the baseline of valid trees (BVT), which is the most in-depth tree among the trees that have positive heterogeneity and complexities. The BVT is presented in Supporting information S.3.2. It contains eleven leaf nodes with different treatment effects. In addition to the first three splits shown in Fig. 3, the BVT also splits at the service setting at admissions, the frequency of use of the primary substance, education, the number of substances, the source of income, and pregnancy. Several of these variables are considered as the most effective variables for racial disparities[29], demonstrating the rationality of GCT in practical applications.

7 Conclusions

We believe that causal inference at the subgroup level will be more intelligible for decision-making in large-scale data analysis. We gather specific units with similar characteristics and treatment effects for subgroup identification and approximate their individual treatment effects by an average value. We improve the existing tree-based method, called GCT, that identifies subgroups with significant heterogeneity and allows for valid inference of the subgroup treatment effects. As discussed previously, the method of estimating treatment effects on nodes contributes greatly to how trees behave. A semiparametric framework is embedded into GCT for treatment effect estimation and allows us to provide explicit asymptotic properties of the STE estimators by trees. We can further conduct valid conditional inference for the tree estimators combined with honest estimation in observational studies, which has not been achieved before.

We conduct several experiments to compare GCT with the other two benchmark tree-based methods. The GCT algorithm performs better in heterogeneous settings than the other two, showing good power in detecting heterogeneity. Another strength of GCT lies in its estimation accuracy; the idea of partialling out can better address the interference of confounding on estimation. Moreover, the asymptotic properties of tree estimators provide us with theoretical guarantees when using GCT to conduct inference. Through the simulations, GCT shows good convergence in subgroup identification and estimation and reaches the nominal coverage rate, which coincides with our tree estimator theorems.

Depending on the scenarios and goals of data analysis, subgroup identification has continued application prospects. Perhaps it makes practical sense to adapt the model to more complex data types, such as extremely imbalanced data. At the same time, it is also valuable to further improve the operations and relevant theories of statistical inference of tree treatment effect estimators.

Supporting information

The supporting information for this article can be found online at https://doi.org/10.52396/JUSTC-2022-0054.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (72071187, 11671374, 71731010, 71921001) and Fundamental Research Funds for the Central Universities (WK3470000017 , WK2040000027).

Conflict of interest

The authors declare that they have no conflict of interest.

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102

Biographies

Caiwei Zhang is a graduate student under the supervision of Prof. Zemin Zheng at the University of Science and Technology of China. Her research focus on causal inference based on machine learning.

Zemin Zheng is a Full Professor at the Department of Management, University of Science and Technology of China (USTC). He received his B.S. degree from the USTC in 2010 and Ph.D. degree from the University of Southern California in 2015. Afterwards, he has been working in the Department of Management, USTC. His research mainly focuses on high dimensional statistical inference and big data problems. He has published some articles in top journals of Statistics, including *Journal of Royal Statistical Society Series B*, *The Annals of Statistics, Operations Research*, and *Journal of Machine Learning Research*. In addition, he has also presided over several scientific research projects, including the Youth Project of the National Natural Foundation of China (NSFC) and General Projects of the NSFC.

References

[1] Altman N, Krzywinski M. Association, correlation and causation. *Nature Methods,* **2015**, *12* (10): 899–900.

[2] Zhang L, Zou H, Zhao Y, et al. Association between blood circulating vitamin D and colorectal cancer risk in Asian countries: A systematic review and dose-response meta-analysis. *BMJ Open,* **2019**, *9* (12): e030513.

[3] Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics,* **2019**, *47* (2): 1148–1178.

[4] Künzel S R, Sekhon J S, Bickel P J, et al. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences,* **2019**, *116* (10): 4156–4165.

[5] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences,* **2016**, *113* (27): 7353–7360.

[6] Robinson P M. Root-*N*-consistent semiparametric regression. *Econometrica,* **1988**: 931–954.

[7] Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association,* **2018**, *113* (523): 1228–1242.

[8] Fan Y, Lv J, Wang J. DNN: A two-scale distributional tale of heterogeneous treatment effect inference. SSRN 3238897, **2018**.

[9] Johansson F, Shalit U, Sontag D. Learning representations for counterfactual inference. In: Proceedings of the 33rd International Conference on Machine Learning. New York: PMLR, **2016**: 3020–3029.

[10] Shalit U, Johansson F D, Sontag D. Estimating individual treatment effect: Generalization bounds and algorithms. In: Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, **2017**: 3076–3085.

[11] Zhang Z, Lan Q, Ding L, et al. Reducing selection bias in counterfactual reasoning for individual treatment effects estimation. arXiv: 1912.09040, **2019**.

[12] Atan O, Jordon J, van der Schaar M. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. In: Thirty-Second AAAI Conference on Artificial Intelligence. Palo Alto, CA: Association for the Advancement of Artificial Intelligence, **2018**: 2071–2078.

[13] Su X, Tsai C L, Wang H, et al. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research,* **2009**, *10*: 141–158.

[14] Yang J, Dahabreh I J, Steingrimsson J A. Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics,* **2022**, *78* (2): 624–635.

[15] Foster J C, Taylor J M, Ruberg S J. Subgroup identification from randomized clinical trial data. *Statistics in Medicine,* **2011**, *30* (24): 2867–2880.

[16] Breiman L, Friedman J, Olshen R, et al. Classification and regression trees. Belmont, CA: Wadsworth International Group, 1984, 37(15): 237–251.

[17] Chernozhukov V, Demirer M, Duflo E, et al. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. Cambridge, MA: National Bureau of Economic Research, **2018**.

[18] Park C, Kang H. A groupwise approach for inferring heterogeneous treatment effects in causal inference. arXiv: 1908.04427, **2019**.

[19] Rubin D B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology,* **1974**, *66* (5): 688–701.

[20] Imbens G W, Rubin D B. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge, UK: Cambridge University Press, **2015**.

[21] Hernán M A, Robins J M. Causal Inference: What If. Boca Raton, FL: Chapman & Hall/CRC, **2020**.

[22] Chernozhukov V, Chetverikov D, Demirer M, et al. Double/debiased/neyman machine learning of treatment effects. *American Economic Review,* **2017**, *107* (5): 261–65.

[23] Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika,* **2021**, *108* (2): 299–319.

[24] Berk R, Brown L, Buja A, et al. Valid post-selection inference. *The Annals of Statistics,* **2013**, *41* (2): 802–837.

[25] Lee J D, Sun D L, Sun Y, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics,* **2016**, *44* (3): 907–927.

[26] Fithian W, Sun D, Taylor J. Optimal inference after model selection. arXiv: 1410.2597, **2014**.

[27] Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. *Biometrical Journal,* **2008**, *50* (3): 346–363.

[28] Guerrero E G. Enhancing access and retention in substance abuse treatment: the role of medicaid payment acceptance and cultural competence. *Drug and Alcohol Dependence,* **2013**, *132* (3): 555–561.

[29] Kong Y, Zhou J, Zheng Z, et al. Using machine learning to advance disparities research: Subgroup analyses of access to opioid treatment. *Health Services Research,* **2022**, *57* (2): 411–421.

DOI: 10.52396/JUSTC-2022-0054
JUSTC, **2023**, 53(11): 1102