Computational Statistics and Data Analysis 159 (2021) 107188

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Subgroup causal effect identification and estimation via matching tree☆

Yuyang Zhang [a], Patrick Schnell [a], Chi Song [a], Bin Huang [b], Bo Lu [a,*]

[a] Division of Biostatistics, College of Public Health, The Ohio State University, Columbus, OH, USA
[b] Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, College of Medicine, University of Cincinnati, Cincinnati, OH, USA

ARTICLE INFO

Article history:
Received 27 May 2020
Received in revised form 26 January 2021
Accepted 26 January 2021
Available online 23 February 2021

Keywords:
Potential outcomes
Propensity score
Matched design
Classification and regression tree
Virtual twins

ABSTRACT

Inferring causal effect from observational studies is a central topic in many scientific fields, including social science, health and medicine. The statistical methodology for estimating population average causal effect has been well established. However, the methods for identifying and estimating subpopulation causal effects are relatively less developed. Part of the challenge is that the subgroup structure is usually unknown, therefore, methods working well for population level effect need to be modified to address this. A tree method based on a matched design is proposed to identify subgroups with differential treatment effects. To remove observed confounding, propensity-score-matched pairs are first created. Then the classification and regression tree is applied to the within-pair outcome differences to identify the subgroup structure. This non-parametric approach is robust to model misspecification, which is important because it becomes much harder to specify a parametric outcome model in the presence of subgroup effects. In addition to describing assumptions under which our matching estimator is unbiased, algorithms for identifying subgroup structures are provided. Simulation results indicate that the proposed approach compares favorably, in terms of the percentage of correctly identifying true tree structure, with other competing tree-based methods, including causal trees, causal inference trees and the virtual twins approach. Finally the proposed method is implemented to examine the potential subgroup effect of the timing of Tobramycin use on chronic infection among pediatric Cystic Fibrosis patients.

1. Introduction

Inferring causal effect is a central topic in scientific research. In many fields, including social science, health and medicine, observational studies provide a rich source of data for examining causal effect, when randomized design is not feasible. The statistical methodology for estimating population average causal effect has been well established (Imbens and Rubin, 2015). However, the methods for identifying and estimating subpopulation causal effects are relatively less developed. Heterogeneous subpopulation effects are ubiquitous in many studies, as individuals with different characteristics may respond differently to a certain intervention (Rothwell, 2005). A major challenge in subgroup analysis

☆ Simulation R codes are included as online supplementary materials.

* Correspondence to: 244 Cunz Hall, 1841 Neil Ave., Columbus, OH 43210, USA.
E-mail address: lu.232@osu.edu (B. Lu).

https://doi.org/10.1016/j.csda.2021.107188
0167-9473/

is that the subgroup structure is often unknown in advance. Ad-hoc analysis of subgroup effect may lead to misleading findings and more principled data-driven subgroup analysis strategies are desirable (Lipkovich et al., 2017).

A traditional string of literature on subgroup effects focuses on effect modification as interaction terms in pre-specified models (Kraemer, 2013; Tian et al., 2014). As these parametric or semi-parametric methods often depend on certain structure of the interaction terms, they do not lend themselves well to more complex situations with little knowledge of subgroup structure. Regression tree models are suggested as a good alternative strategy, because of their nonparametric property, natural subgroup interpretation, and ability to accommodate complex data (Loh et al., 2015). If subgroups are formed by covariates stochastically rather than deterministically, latent subgroup modeling approaches can be utilized (Lanza and Rhoades, 2013; Altstein and Li, 2013; Kim et al., 2019).

Identifying subgroup structure in observational data presents one more challenge than randomized trials. In observational studies, patient characteristics could vary substantially across different treatment groups. Strategies for removing this observed confounding include propensity score adjustment, facilitating score adjustment and outcome modeling. Causal tree (CT) combines the propensity score weighting and the regression tree to estimate heterogeneous treatment effects (Athey and Imbens, 2016). It extends the classification and regression tree (CART) (Stone, 1984) by modifying the tree splitting criterion to maximizing the mean squared subgroup treatment effect. The subgroup effects are estimated using inverse propensity score weighting to control confounding (Bang and Robins, 2005). Causal inference tree (CIT) proposes a facilitating score to split the tree, where subjects with similar facilitating scores should have similar propensity scores and treatment effects (Su et al., 2012). Since the facilitating score is related to both treatment assignment and potential outcomes, a parametric likelihood function needs to be specified in the inference. The virtual twins (VT) method aims to find the individual treatment effect then apply the classification tree to identify the subgroup structure (Foster et al., 2011). It predicts the response for the unobserved "twin" for each subject, where the twin refers to the same subject with a different treatment assignment (i.e. the unmeasured copy of the potential outcomes). The outcome prediction can be done with either parametric or nonparametric models.

Matching is another popular way of using propensity score for confounding adjustment in observational studies (Rosenbaum et al., 2010; Lu et al., 2011). First, it is robust to model misspecification as it uses a nonparametric approach to balance covariate distributions. Second, it resembles the randomization design, which is easy to understand by a general audience. Third, it is less prone to model manipulation as the causal effect estimation is conducted only after good matches are created and the outcome variable is never used in the matching process. Fourth, it avoids the instable estimation of weighting method when propensity scores are close to 0 or 1, which might be more likely to occur with many subgroups. But the use of matching in subpopulation causal effect estimation is quite limited in the literature. Hsu et al. (2013) use matching design and the classification tree to examine effect modification with a focus on design sensitivity for potential unmeasured confounding. One reason of limited use of matching design in subgroup analysis is that it is hard to ensure the matching quality in every subgroup in practice, especially when there are many subgroups. Dong et al. (2020) propose a subgroup balancing propensity score and discuss its use in matching and weighting. As long as the treated and control group distributions are reasonably overlapped, matching estimator can be used to estimate the subpopulation effects, referred to as subgroup average treatment effects for the treated (subgroup ATTs).

In this paper, instead of focusing on balancing covariate distributions in known subgroups, we develop a novel approach, i.e. matching tree (MT), as a principled data-driven strategy for subgroup identification, when subgroup indicators are unknown. Individual subjects from different treatment groups are first matched to remove confounding. The tree-based method is then applied to within-pair outcome differences to identify the subgroup structure, with trimming strategy to prevent overfitting. The matching based estimator is unbiased for subgroup treatment effect estimation among people with similar characteristics as the treated subjects, which is also robust to model misspecification. The resulting subgroup structure may be further utilized by content expert to improve the intervention strategies. The paper is organized as follows. Section 2 provides justifications for matching based subgroup estimator and details for matching tree algorithms. Section 3 presents two simulation studies. Section 4 applies the proposed method to study the potential subgroup effect of the timing of Tobramycin use on chronic infection among pediatric CF patients. Section 5 ends the paper with discussions on limitation and potential extensions.

2. Matching tree for subpopulation effects

In this section, we first show that matching on the propensity score produces unbiased estimator for subpopulation causal effects of interest, then present the matching tree algorithm for identifying subpopulations with differential causal effects.

2.1. Unbiased estimation for subpopulation causal effect

To establish the unbiased subpopulation effect estimation, we follow the potential outcomes framework (Imbens and Rubin, 2015). With a dichotomous treatment option, T ($T = 1$ for treated and $T = 0$ for control or standard care), we denote (Y^1, Y^0) to be the potential outcomes under treatment and control respectively. Let $\mathbf{X}$ be a matrix that contains the pretreatment covariates and Z be the indicator of subgroups. Our primary interest is the subpopulation average causal effect for the treated, $E(Y^1 - Y^0|Z)$, where the expectation integrates over the distribution of treated subjects (Dong et al., 2020). The estimation hereafter is carried out in such a context.

Because we cannot observe both potential outcomes for the same subject in a real study, the causal effect identification needs more assumptions than what are usually required for estimating the effects' association. Two assumptions are commonly used (Imbens and Rubin, 2015): (1) Stable Unit Treatment Value Assumption (SUTVA), which dictates no interference between units and no variation of a specified treatment level; (2) strongly ignorable treatment assignment assumption, which implies that the probability of receiving treatment should be strictly between 0 and 1 for each subject, and the treatment assignment is unconfounded given observed covariates. For well executed randomized studies, the ignorability assumption is guaranteed to hold by design. For observational studies, it is a crucial assumption as the treatment reception might be compromised by subject level characteristics. In practice, researchers tend to identify and collect data based on a long list of pre-treatment covariates and make adjustment on them in the analysis to ensure treatment assignment ignorability.

As shown in Rosenbaum and Rubin (1983), the strongly ignorable treatment assignment for the entire population can be achieved through adjusting for a scalar propensity score. Next, we establish the same property in the context of subpopulation causal effects in Lemma 1, following the idea of Rosenbaum and Rubin (the proof is given in Appendix A.1).

Lemma 1. *Denote the propensity score as* $e(\mathbf{X}, Z) = P(T = 1|\mathbf{X}, Z)$. *If the treatment assignment is strongly ignorable in subpopulations given a vector of covariates,* $\mathbf{X}$, *i.e.*

$$(Y^1, Y^0) \perp\!\!\!\perp T|\mathbf{X}, Z \ \ and \ \ 0 < P(T = 1|\mathbf{X}, Z) < 1,$$

then the treatment assignment is strongly ignorable in subpopulations given the propensity score, i.e.

$$(Y^1, Y^0) \perp\!\!\!\perp T|e(\mathbf{X}, Z), Z \ \ and \ \ 0 < P(T = 1|e(\mathbf{X}, Z), Z) < 1.$$

Given Lemma 1, we can construct unbiased subpopulation causal effect estimation via the propensity score (the proof of Proposition 1 is given in Appendix A.2).

Proposition 1. *Suppose the subpopulation strongly ignorable treatment assignment holds given observed covariates* $\mathbf{X}$ *and subgroup indicator Z. Denote the propensity score as* $e(\mathbf{X},Z)$. *Then the propensity-score-adjusted subgroup effect estimator is unbiased for estimating the subgroup average treatment effect.*

In practice, the adjustment of the propensity score can be done in the form of matching, stratification, weighting or regression modeling. We also note it is not necessary to always include Z in the propensity score model. The propensity score only needs to include the true confounders, as shown in the following corollary (the proof is given in Appendix A.3). But when researchers are not certain about whether Z is a confounder, the general advice is to include it in the propensity score.

Corollary 1. *With subpopulation strongly ignorable treatment assignment assumption in Lemma 1, if Z is the subpopulation indicator but not a confounder, then the subgroup effect estimator adjusted for the propensity score* $e(\mathbf{X})$ *is unbiased for estimating the subgroup average treatment effect.*

2.2. Matching tree algorithms

When the subgroup indicators are not known in advance, we need to first find out the subgroup composition before we can estimate the causal effects. Most existing methods for subgroup identification involve potential outcome modeling, either parametrically or semiparametrically. The major drawback is that they tend to suffer from model misspecification. To reduce the reliance on parametric modeling, some Virtual Twin methods use tree-based models, e.g. BART or random forest, to predict the missing potential outcomes for each subject (Chipman et al., 2010). Causal tree approach provides a different perspective by taking advantage of the propensity score weighting. Since the propensity score possesses a critical piece of information in causal inference, we propose to combine propensity score matching and classification tree as an alternative to Virtual Twin methods, as it does not require explicit modeling of the potential outcomes. Within each pair, the observed outcome of the matched control serves as a natural candidate for the missing potential outcome of the treated subject.

Specifically, our method takes following steps. First, we estimate the propensity score of treatment reception and identify a list of possible subgroup defining variables (SDVs). Usually, the propensity score is estimated via a logistic regression model and the subgroup defining variables are determined by the substantive experts. The covariate design matrix $\mathbf{X}$ contains all relevant covariates and subgroup defining variables. Each column of $\mathbf{X}$ presents one variable and each row of the $\mathbf{X}$ presents one subject. Second, we create matched pairs between treated and control groups based on the estimated propensity scores and the subgroup defining variables. This step is different from the conventional propensity score matching as we also need to match on subgroup defining variables for subgroup identification. Third, we calculate the outcome differences within each pair and regress such differences over subgroup defining variables using CART. Lastly, we prune the tree by examining the Least Square Means (LSmeans) simultaneous confidence intervals (Lin et al., 2019). The LSmeans method is a general way of estimating multiple subgroup effects with control of overall Familywise Error

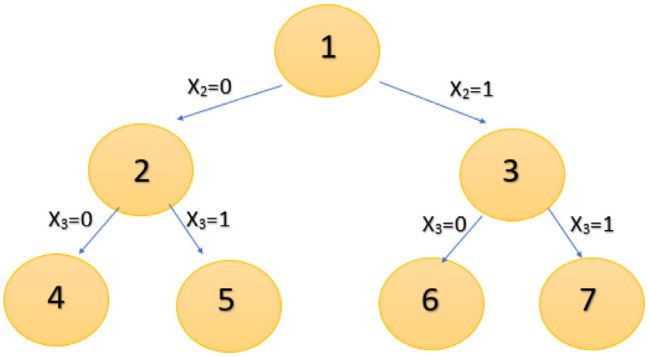


Fig. 1. LSmeans prune.

Table 1
LSmeans prune for nodes 2, 4, 5.

	LSmeans	S.D.	α	Lower bound	Upper bound
$X_3 = 0$	−0.29	0.161	0.0125	−0.65	0.07
$X_3 = 1$	−0.482	0.127	0.0125	−0.77	−0.2

Table 2
LSmeans prune for node 3, 6, 7.

	LSmeans	S.D.	α	Lower bound	Upper bound
$X_3 = 0$	2.08	0.195	0.0125	1.64	2.52
$X_3 = 1$	4.43	0.154	0.0125	4.08	4.77

Rate. Since we only consider binary trees, for each pruning step, the multiplicity adjustment is basically a Bonferroni correction with two groups. For two nodes under the same parent, if their confidence intervals overlap, it indicates a non-significant difference and we prune these two nodes to the parent node. We keep this pruning process till there is only root node or no confidence interval overlap between the nodes under the same parent. Section 2.2.1 provides a simulated example for illustrative purpose.

2.2.1. A simulated example for illustrating the tree pruning

In this subsection, we show how to use the LSmeans method to prune a small tree. First, we generate a dataset of 2000 subjects, with three covariates, X_1 (standard normal distribution), X_2 and X_3 (Bernoulli distributions with probability 0.4 and 0.6), respectively. The treatment status is generated from a Bernoulli distribution with probability 0.5. The outcome is generated from the model $Y = -3 + 5 \times T + 5 \times X_1 + 5 \times X_2 \times T + 5 \times X_2 \times X_3 \times T + \epsilon$, where ϵ follow the standard normal distribution. The initial tree structure is presented in Fig. 1.

There are four terminal nodes determined by X_2 and X_3, named nodes 4 to 7. The nodes 4 and 5 share the same parent — node 2. The nodes 6 and 7 share the same parent — node 3. For tree pruning, first, we consider the group of nodes 2, 4, and 5. Table 1 shows the LSmeans estimates with the simultaneous 95% confidence interval bounds. The α value is 0.0125 (=0.05/4) since there are two terminal nodes and two-sided tests are used. Because the confidence intervals for two groups determined by X_3 overlap, we prune nodes 4 and 5 to the parent node 2.

Next, we apply the same pruning procedure to the group of nodes 3, 6, and 7. Based on result shown in Table 2, since two confidence intervals do not overlap, we keep the structure and do not prune. Therefore, the final tree structure indicates three subgroups — subjects in three terminal nodes 2, 6 and 7.

2.2.2. Algorithms for binary and continuous subgroup indicators

From an implementation perspective, matching on additional variables on top of the propensity score requires modifications to existing software packages. Matching on continuous variables presents another layer of complexity than matching on binary variables, because it is almost impossible to match exactly on a continuous variable. Practically, we have to use a caliper to specify a range of values for acceptable matchings. Therefore, we present the algorithms separately for binary and continuous variables. The existing optimal matching package can accommodate exact matching with binary variables, but not the caliper matching with continuous variables. So, to be consistent, we implement nearest neighbor matching for both binary and continuous variables.

Algorithm 1 delineates the steps for implementing matching tree with possible binary subgroup indicators. Algorithm 2 delineates the steps for implementing matching tree with possible continuous subgroup indicators. The difference is

that caliper matching for continuous variables is used in Algorithm 2 and subsequently, the covariate values from treated subjects are used in CART.

Algorithm 1 Matching Tree: Binary Subgroup Indicator

Input Outcome Y, Propensity Score, Binary Covariates Design Matrix $\mathbf{X}$

Output Identified Subgroup Indicators

1: Do $1-1$ nearest neighbor matching on both propensity score and binary covariates $\mathbf{X}$, with all columns of $\mathbf{X}$ exactly matched.
2: Obtain the matched set S, which contains n pairs.
3: Calculate the difference on the outcome within each pair, denoted as ΔY_i, where $i = 1, 2, \ldots, n$.
4: Regress ΔY_i versus $\mathbf{X}$ using CART.
5: **while** LSmeans confidence intervals overlap between nodes **do**
6: Prune the tree
7: **end while**

Algorithm 2 Matching Tree: Continuous Subgroup Indicator

Input Outcome Y, Propensity Score, Continuous Covariates Design Matrix $\mathbf{X}$

Output Identified Subgroup Indicators and Corresponding Cut-off Values

1: Do $1-1$ nearest neighbor matching both on the propensity score and continuous covariates $\mathbf{X}$, with columns matched on the predefined caliper.
2: Obtain the matched set S, which contains n pairs.
3: Calculate the difference on the outcome within each pair, denote as ΔY_i, where $i = 1, 2, \ldots, n$.
4: Regress ΔY_i versus $\mathbf{X}$ using CART, where covariate values are taken from the treated subjects.
5: **while** LSmeans confidence intervals overlap between nodes **do**
6: Prune the tree
7: **end while**

When analyzing real data, we would separate potential binary subgroup indicators from potential continuous subgroup indicators, then run the algorithm for each group of variables. Based on the pruned trees, we could identify candidates for subgroup indicators. If there are both binary and continuous subgroup indicators, we would combine them to re-run the matching tree algorithm to see if the tree can be further pruned. For categorical variables with more than two levels, we may define multiple dummy variables and include them as binary variables. To further simplify the process, we may also categorize the continuous variables into several meaningful groups, then we only need to run the algorithm for binary variables. Simulation studies are conducted in the next section to evaluate the empirical performance of our proposed method, in comparison to several competing approaches.

3. Simulation

In this section, we present two simulation studies: a smaller study for examining the unbiasedness of matching based subpopulation causal effect estimators, and a larger study for comparing matching tree with causal inference tree, causal tree, and virtual twins method, in terms of subpopulation structure identification.

3.1. A small simulation for unbiased subpopulation causal effect estimation

To examine whether the matching based subpopulation causal effect estimators are unbiased or not, we first generate four covariates (two continuous ones and two binary ones) and based on the covariates, we generate the treatment

Table 3
Continuous outcome, subpopulation ignorability.

	Subgroup 1			Subgroup 2		
	%Bias	Coverage	SD	%Bias	Coverage	SD
Scenario CC1	0.75	95.0	0.09	1.82	94.3	0.11
Scenario CC2	1.91	97.0	0.09	1.37	96.5	0.09
Scenario CC3	0.88	93.1	0.11	1.56	94.5	0.09
Scenario CC4	0.99	93.9	0.09	1.56	97.2	0.11

reception probability via a logistic regression model. Then we generate the outcome variable via either a linear model (for continuous outcome) or a logistic regression model (for binary outcomes).

Each simulated dataset consists of 2000 observations and the simulation is repeated for 1000 times. The details of simulation setup are presented as follows:

- Generate X_1, X_2 independently from the standard Normal distribution $N(0, 1)$.
- Generate X_3, X_4 from the Bernoulli distribution with probability 0.4, 0.6, respectively.
- Generate the treatment indicator with $logit(P(T = 1)) = -1.3 + log(2) \times X_1 + log(2) \times X_3$, and the coefficients are chosen to ensure the sample size ratio of 1 : 3 between the treated and control group, approximately.

We consider four scenarios for the continuous outcome and the outcome generating model is $Y \sim N(\mu, 1)$, where $\mu = -3.85 + 5 \times T + 5 \times X_1 + X_3 + EMF$ (EMF stands for effect modifier function):

- Scenario CC1: binary non-confounding effect modifier:
  $EMF = -10 \times X_4 \times T$.
- Scenario CC2: binary confounding effect modifier:
  $EMF = -10 \times X_3 \times T$.
- Scenario CC3: continuous non-confounding effect modifier:
  $EMF = -10 \times I(X_2 > 0.2) \times T$.
- Scenario CC4: continuous confounding effect modifier:
  $EMF = -10 \times I(X_1 > 0.2) \times T$.

First, we conduct matching in subgroups defined by the true effect modifier to obtain exact match on the subgroup indicator. Second, we run a linear regression within each subgroup separately and the model only includes the treatment indicator to see if matching can remove confounding due to covariates. For example, in scenario CC1, we conduct 1 : 1 optimal propensity score matching without replacement in subgroups $X_4 = 1$ and $X_4 = 0$ separately. The propensity score model is a logistic regression model including X_1 to X_4. Within each matched subgroup dataset, we run a simple linear regression using Y as the outcome and T as the predictor. The percentage bias (%Bias), i.e., $(\hat{\beta} - 5)/5 \times 100\%$, the 95% confidence interval coverage (Coverage), and the average standard deviation estimates (SD) are reported in Table 3:

The biases under all scenarios are small (less than 2%). The 95% CI coverages are close to the nominal level. This implies the matching based method is practically unbiased for estimating subpopulation effects.

We also conduct simulations for four binary outcome scenarios and the findings are similar to the continuous case. Details are included in the Appendix A (Appendix A.4).

3.2. A large simulation for subpopulation structure identification

Next, we examine the empirical performance of matching tree (MT) in comparison to several competing methods. The other tree-based methods being considered are causal tree (CT), causal inference tree (CIT), virtual twins (VT) with Bayesian additive regression tree (BART) and virtual twins (VT) with causal forest (CF). Among the five methods, Matching Tree is based on matched data, while virtual twins with BART (VT.BART) and virtual twins with causal forest (VT.CF) need to use full data. Causal tree and causal inference tree have no restrictions regarding the data structure, so we run both matched data and full data with them.

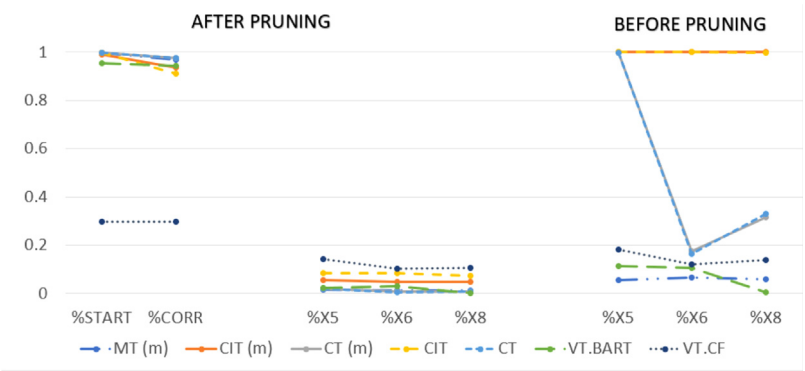
3.2.1. Continuous outcome

To obtain the continuous outcome, first we generate X_1, X_2, X_3, X_7 from the Normal distributions, with means 3, −5, 0, −1, respectively, standard deviation 1. Next we generate X_4, X_5, X_6, X_8 from the Bernoulli distributions with success probabilities 0.6, 0.5, 0.4, 0.55, respectively.

The treatment assignment is determined by the following logistic regression model:

$$logit(P(T = 1)) = (0.6 + 0.4 \times log(|X_1|) + 1.4 \times \sqrt{|X_2|} - 3.2X_4X_5).$$

The outcome is generated as Normal distribution with a standard Normal error term, and there are two scenarios for the mean term:



Legend:

%CORR: The proportion of the trees that correctly specify the subgroup structure, i.e. only partition by X_4

%START: The proportion of the trees that first partition by the true effect modifier, i.e. first partitioning by X_4

%X_k: The proportion of the trees that partition by the incorrect variable X_k, i.e. X_k is not the true effect modifier.

MT (m) : Matching Tree that runs with the matched data.

CIT (m): Causal Inference Tree that runs with the matched data.

CT (m): Causal Tree that runs with the matched data.

CIT: Causal Inference Tree that runs with the full data.

CT: Causal Tree that runs with the full data.

VT.BART: Virtual Twins with Bayesian additive regression tree runs with the full data.

VT.CF: Virtual Twins with Causal Forest runs with the full data.

Fig. 2. Scenario C1: Subgroup identification.

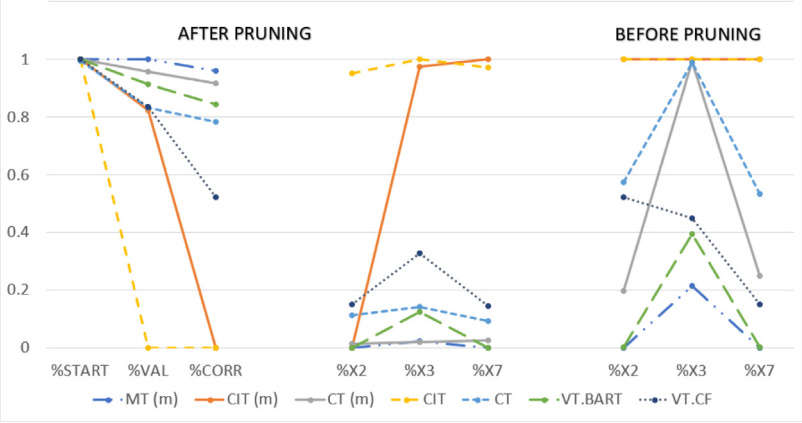
- Scenario C1 (Binary effect modifier and confounder):
  $\mu = -3.85 + 0.5 \times sin(X_1) + 2 \times exp(X_3) + 2 \times X_4 + 1.5 \times X_6 + T \times (2 + 5 \times X_4)$.
- Scenario C2 (Continuous effect modifier and confounder):
  $\mu = -3.85 + 0.5 \times sin(X_1) + 2 \times exp(X_3) + 2 \times X_4 + 1.5 \times X_6 + T \times (2 + 5 \times I_{(X_1 > 2.4)})$.

There are 2000 observations in each simulated dataset. The simulation is repeated for 1000 times for all scenarios. We have also considered cases where the effect modifier is not a confounder (details are presented in Appendix A.5).

To summarize the performance of each method for identifying the subgroup structure, several metrics are reported in the following two figures:

Fig. 2 presents the results for scenario C1, where the binary subgroup indicator is also a confounder (X_4). The x-axis presents different metrics as defined in the figure legend and the y-axis shows the performance of each metric in proportions. The left five columns present the final tree after the LSmeans pruning, and the right three columns present results without pruning. The first two columns show the overall performance of the method (higher proportion indicates better performance). The remaining six columns are the proportions of each individual variable that should not be selected (lower proportion indicates better accuracy). Comparing all seven methods, the performance of accuracy are not very different, except virtual twins with causal forest. The matching tree and two causal tree methods show better results with accuracy around 97% or higher. The before-pruning columns indicate some overfitting issues with causal inference tree and causal tree methods, as they tend to partition with extra variables before pruning. Matching tree and virtual twins methods do not seem to overfit much.

Fig. 3 presents the results for scenario C2, where the continuous subgroup indicator is also a confounder (X_1). We use a small interval [2.3, 2.5] as an acceptable region for cut-off point identification, since the true cut-off value for X_1 is 2.4. The overall performance, shown in the left two columns, indicates that all methods become much worse in comparison to scenario C1, except matching tree, which remains a high accuracy of 96%. Both causal inference tree methods are very bad with little chance of identifying the true subgroup structure. Excluding causal inference tree methods, methods using matched data tend to perform better. Among methods using full data, virtual twins with BART produces the best result. Overfitting becomes a bigger issue as shown in columns before pruning.



Legend:

%CORR: The proportion of the trees that correctly specify the subgroup structure, i.e. only partition by X_1.

%START: The proportion of the trees that first partition by the true effect modifier, i.e. first partitioning by X_1.

%VAL: For the continuous effect modifier, the proportion of the trees that partition by the right pre-defined cut-off point.

$\%X_k$: The proportion of the trees that partition by the incorrect variable X_k, i.e. X_k is not the true effect modifier.

MT (m) : Matching Tree that runs with the matched data.

CIT (m): Causal Inference Tree that runs with the matched data.

CT (m): Causal Tree that runs with the matched data.

CIT: Causal Inference Tree that runs with the full data.

CT: Causal Tree that runs with the full data.

VT.BART: Virtual Twins with Bayesian additive regression tree runs with the full data.

VT.CF: Virtual Twins with Causal Forest runs with the full data.

Fig. 3. Scenario C2: Subgroup identification.

3.2.2. Binary outcome

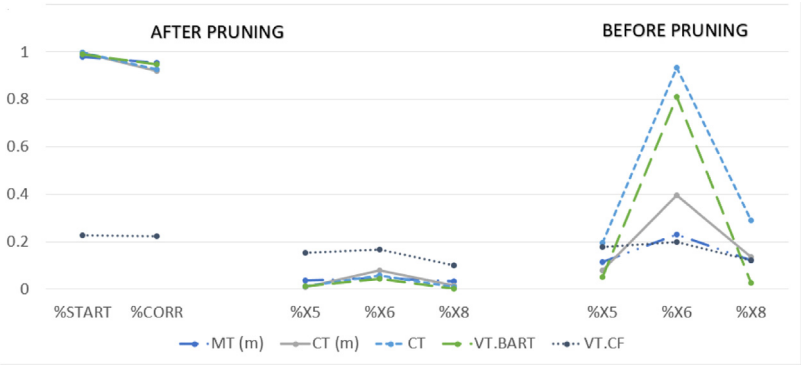
We simulate binary outcomes using the same set of covariates. The true outcome is generated based on the following models:

- Scenario B1 (Binary effect modifier and confounder):
  $logit(P(Y=1)) = -1 + log(1.15) \times X_1 + log(1.15) \times X_3 + log(1.5) \times X_4 + log(1.5) \times X_6 + T \times (log(2) + log(10) \times X_4)$.
- Scenario B2 (Continuous effect modifier and confounder):
  $logit(P(Y=1)) = -1 + log(1.15) \times X_1 + log(1.15) \times X_3 + log(1.5) \times X_4 + log(1.5) \times X_6 + T \times (log(2) + log(10) \times I_{(X_1 > 2.4)})$.

Since causal inference tree method are based on Normal likelihood and it cannot be used for modeling binary outcomes, we remove them from comparison in this subsection. We have also considered cases where the effect modifier is not a confounder (details are presented in Appendix A.6).

Fig. 4 presents the tree identification results for scenario B1, where the binary subgroup indicator is also a confounder (X_4). Consistent with the findings in scenario C1, all methods produce similar results with more than 90% accuracy, except virtual twins with causal forest. Matching tree and virtual twins with BART are slightly better. Overfitting seems to be an issue for some methods, especially causal tree with full data, as shown in columns before pruning.

Fig. 5 presents the tree identification results for scenario B2, where the continuous subgroup indicator is also a confounder (X_1). Similar to scenario C2, we use [2.3, 2.5] as an acceptable region for correctly identifying the cut-off point. All methods show more than 70% accuracy with matching tree and virtual twin with BART above 80%. The before-pruning results indicate that overfitting is still an issue for most methods.



Legend:

%CORR: The proportion of the trees that correctly specify the subgroup structure, i.e. only partition by X_4.

%START: The proportion of the trees that first partition by the true effect modifier, i.e. first partitioning by X_4.

$\%X_k$: The proportion of the trees that partition by the incorrect variable X_k, i.e. X_k is not the true effect modifier.

MT (m) : Matching Tree that runs with the matched data.

CT (m): Causal Tree that runs with the matched data.

CT: Causal Tree that runs with the full data.

VT.BART: Virtual Twins with Bayesian additive regression tree runs with the full data.

VT.CF: Virtual Twins with Causal Forest runs with the full data.

Fig. 4. Scenario B1: Subgroup identification.

Overall, matching tree is the top choice for matched data and virtual twins with BART is the top choice for full data based on our simulations. In the next section, we apply matching tree method to a real data example to study the timing effect of Tobramycin use on chronic infections among pediatric Cystic Fibrosis patients.

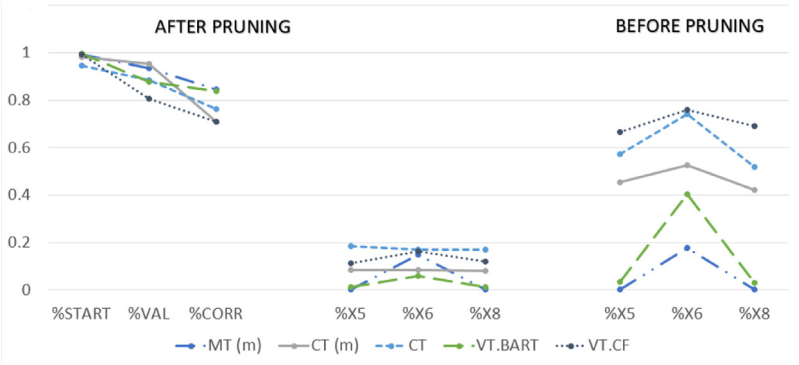
4. Effect of the timing of Tobramycin use on chronic infections among pediatric CF patients

Cystic Fibrosis (CF) is one of the most common, life-shortening genetic diseases among children (Foundation, 2013). At least 30,000 people are affected by CF in the United States, and about 1000 new cases are diagnosed every year. The primary cause of CF is mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene. This mutation leads to dysfunction of the CFTR protein, which affects the function of lung, causing difficulty breathing, lung infections, and subsequent death. Currently, there is no known cure for CF.

The primary cause of death among CF patients is respiratory failure due to chronic airway infection and loss of lung function. Pseudomonas aeruginosa (PA) is the most common type of infections in CF patients. Since PA is strongly associated with increased inflammation, higher probability of pulmonary exacerbations, and higher mortality, treating chronic PA is crucial for treating CF patients (Kosorok et al., 2001). According to the CF Pulmonary Guidelines, Tobramycin (TOBI) is an efficacious drug for treating chronic PA infection in CF patients (Ramsey et al., 1999).

Many studies have examined the use of TOBI among CF patients. Forced expiratory volume in 1 s (FEV1) is a common measure of lung function, which can be collected from patients six years or older. In a randomized clinical trial setting, Konstan et al. (2014) used the linear mixed-effects model to show that the use of TOBI in CF patients is associated with improvement in FEV1. VanDyke et al. (2013) discussed whether TOBI is effective in treating chronic PA infection among CF patients using observational data. They used a two-stage least square instrumental variable approach, where center-level prescription was considered as the instrumental variable. The model results showed that the patients with chronic PA infection who took TOBI had better FEV1 than those not taking TOBI.

Since more than 75% CF patients are diagnosed before age 2, it is very important to develop treatment plan at early ages. In our study, we want to investigate whether the use of TOBI can treat the infections in young CF patients, especially those younger than one year old. We choose to use body weight increase rate as the outcome since FEV1 can only be measured after the age of six.



Legend:

%CORR: The proportion of the trees that correctly specify the subgroup structure, i.e. only partition by X_4.

%START: The proportion of the trees that first partition by the true effect modifier, i.e. first partitioning by X_4.

$\%X_k$: The proportion of the trees that partition by the incorrect variable X_k, i.e. X_k is not the true effect modifier.

MT (m) : Matching Tree that runs with the matched data.

CT (m): Causal Tree that runs with the matched data.

CT: Causal Tree that runs with the full data.

VT.BART: Virtual Twins with Bayesian additive regression tree runs with the full data.

VT.CF: Virtual Twins with Causal Forest runs with the full data.

Fig. 5. Scenario B2: Subgroup identification.

4.1. US CF foundation patient registry dataset

The CF Foundation Patient Registry is a comprehensive database containing information on the health status of CF patients who received care in CF Foundation-accredited care centers and agreed to participate in the Registry. This information is used to create CF care guidelines and to assist researchers in studying CF treatments and outcomes and designing CF related clinical studies. We use the registry data from January 1, 1996, to December 31, 2015 to evaluate two treatment strategies: taking TOBI at the first occurrence of PA infection (early treatment) versus taking TOBI at the second or later PA infection occurrences (late treatment).

We consider the patients enrolled in the registry under one year of age, including 5657 records. After removing cases with missing PA infection or weight information, there are 5218 patients in the final analytical file, including 1330 early treatment and 3888 late treatment patients. The continuous outcome is the weight increase rate within one year after the first chronic PA infection, defined as

$$\frac{Wt\_Percentile\_PA1\_1y - Wt\_Percentile\_PA1}{Wt\_Percentile\_PA1},$$

where $Wt\_Percentile\_PA1$ denotes the weight percentile at the first PA infection, and $Wt\_Percentile\_PA1\_1y$ denotes the patient's weight percentile at one year after the first PA infection.

To apply our subgroup identification algorithm, we are particularly interested in if there is one or more subgroups that present different treatment effects. Therefore, different treatment strategies can be targeted to different subgroups. Given the observational nature of the registry data, we first use propensity score matching to balance covariate distributions, then apply matching tree algorithm to identify potential subgroups.

Logistic regression is used to estimate the propensity of TOBI timing status with covariates identified based on clinical knowledge (Konstan et al., 2007), including sex, insurance coverage, other drugs related to CF, baseline measures for age, number of bacterial tests, year entered in registry, weight-for-age percentile, and height-for-age percentile. Summary statistics for covariates by treatment groups are displayed in Table 4.

Table 4
Summary statistics.

	All patients	Early TOBI patients	Late TOBI patients
N	5218	1330	3888
Female, %(n)	45.9 (2399)	46.2 (614)	45.9 (1785)
TOBI_alt, %(n)	6.70 (350)	6.99 (93)	6.61 (257)
Insurance, %(n)			
Medicaid	40.2 (2098)	40.1 (533)	40.2 (1565)
Non-Medicate	57.8 (3018)	57.3 (762)	58.0 (2256)
Unknown	1.95 (102)	1.80 (24)	2.00 (78)
Age, mean(SD)	2.39 (2.62)	2.6 (2.84)	2.32 (2.53)
Bacterial, mean (SD)	8.00 (8.27)	9.10 (9.79)	7.61 (7.62)
Year, mean (SD)	2006 (4.87)	2006 (4.85)	2006 (4.88)
Weight percentile, mean (SD)	33.02 (27.66)	34.13 (27.67)	32.63 (27.65)
Height percentile, mean (SD)	35.21 (27.15)	34.39 (26.45)	35.49 (27.39)

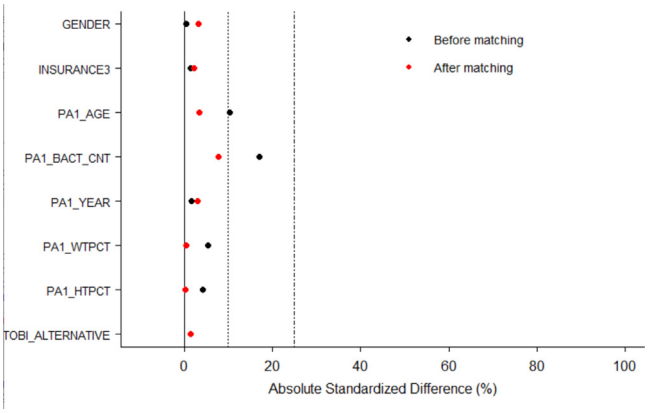


Fig. 6. ASD plot.

4.2. Subgroup identification

To investigate whether the early use of TOBI can benefit some subgroup of CF patients, we apply the matching tree method to identify subgroups. Following the previous literature (Dasenbrook et al., 2008), we consider the covariates in Table 4 as potential effect modifiers. Nearest neighbor algorithm is used to implement matching and post-matching balance is examined by checking the absolute standardized differences (ASD) for each covariate. Fig. 6 shows good overall balance as no ASD is more than 10%. With well-balanced covariates, we can proceed with subgroup identification and treatment effect estimation.

We apply algorithm 1 to identify subgroups defined by binary covariates and apply algorithm 2 to identify subgroups defined by continuous covariates. With algorithm 1, gender, insurance and TOBI_alt are considered as potential subgroup defining variables. We first conduct 1-1 matching on both estimated propensity scores and these three variables, with the three variables exactly matched. With matched data, we calculate the difference between the weight increase rate within each pair and regress this difference with gender, insurance, and TOBI_alt, using CART. After tree pruning, it turns out none of them is picked up, so no binary subgroup defining variables are considered further. With algorithm 2, weight percentile and height percentile are considered as potential subgroup defining variables. The matching is similarly conducted and the only difference is that we need to use caliper matching for continuous variables, instead of exact matching. Algorithm 2 finds a tree structure as presented in Fig. 7, where the n values indicate the number of pairs in each subgroup, i.e., the leftmost leaf node includes 84 pairs of patients (168 individuals). Specifically, the following four subgroups are identified by the matching tree method:

- height percentile $< 4.4\%$ and weight percentile $< 3.5\%$
- height percentile $< 4.4\%$ and weight percentile $\geq 3.5\%$
- height percentile $\geq 4.4\%$ and weight percentile $< 4.3\%$
- height percentile $\geq 4.4\%$ and weight percentile $\geq 4.3\%$

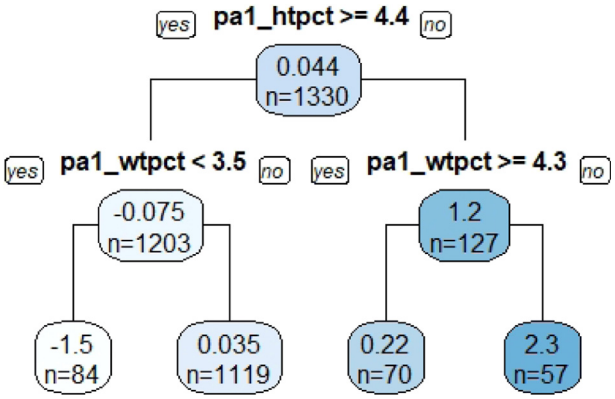


Fig. 7. Subgroup identification.

Table 5
Real data analysis result.

	Treatment effect (CI)	SD	P-value
Hpt $<$ 4.4% & Wpt $<$ 4.3%	0.123 (0.073, 0.173)	0.020	$<$0.0001*
Hpt $<$ 4.4% & Wpt $\geq$ 4.3%	0.018 ($-$0.027, 0.063)	0.018	0.33
Hpt $\geq$ 4.4% & Wpt $<$ 3.5%	$-$0.02 ($-$0.063, 0.023)	0.017	0.28
Hpt $\geq$ 4.4% & Wpt $\geq$ 3.5%	$-$0.0003 ($-$0.013, 0.012)	0.005	0.998

4.3. Subgroup effects estimation

With the identified subgroup structure, we estimate the treatment effect in each subgroup by LSmeans method, which accounts for the multiplicity in multiple group mean estimation. Lin et al. (2019) discussed the general strategy to construct simultaneous confidence intervals for both subgroup effects and the overall population effect. Our application is a special case of their method, which corresponds to a Bonferroni correction, since our four subgroups are independent. Since the outcome, one-year weight increase rate, is highly skewed, the logarithm transformation is used to make it closer to a normal distribution. Before the log transformation, we shift individual weight increase rate by adding 20 percentage points to make sure all values are positive. We use the transformed outcome as Y in the following linear model with T being the timing of TOBI, A, B, C, D being the four subgroups:

$$\begin{aligned} Y &= \beta_0 + \beta_T \times T + \beta_B \times I(B) + \beta_C \times I(C) + \beta_D \times I(D) \\ &\quad + \beta_{TB} \times T \times I(B) + \beta_{TC} \times T \times I(C) + \beta_{TD} \times T \times I(D) + \epsilon, \end{aligned}$$

where I(B), I(C) and I(D) are indicator variables for groups B, C, and D. The point estimators are β_T, $\beta_T + \beta_{TB}$, $\beta_T + \beta_{TC}$, $\beta_T + \beta_{TD}$ for subgroups A, B, C and D, respectively. We use the R function "glht" (in the package "multcomp") with two pairwise commands (group and treatment assignment) to produce estimates displayed in Table 5 ("Hpt" for the height-for-age percentile variable and "Wpt" for the weight-for-age percentile variable). The confidence intervals (CI) are 95% simultaneous confidence intervals calculated by the package with multiplicity adjustment.

The treatment effect is significant in the first subgroup, i.e., small weight and small height babies, showing a beneficial effect of taking TOBI early. Since the estimation is based on matched data, the interpretation of the treatment effect should be for those who actually received TOBI at their first PA infection. Among babies who actually were treated with TOBI at first PA infection, with height percentile less than 4.4% and weight percentile less than 4.3%, the average treatment effect of log weight increase rate is 0.123 (95% CI, 0.071 to 0.175). To get a better idea about the treatment effect in the original scale, we also calculated the mean weight increase rate difference between treated and control subjects in the subgroup showing significant effect, which is 2.16 percentage points. However, for the other three subgroups, no significant effects for TOBI timing are present. In other words, there is no evidence for benefit of early TOBI treatment for bigger babies (in terms of either height or weight).

5. Discussion

Discovering subgroups with heterogeneous effects is challenging in observational studies, as subjects in different treatment groups may have different characteristics. We propose a matching design to first remove confounding, then

apply tree-based methods to identify subgroup structure using within pair outcome differences. It is similar to virtual twins method in the sense that subjects in each pair resemble twins with different treatment assignment. But the key difference is that we use matching design to find the outcomes of the twins, rather than using predictive modeling, which may be vulnerable to model misspecification. As a result of the matched design, our estimator should be interpreted as average subpopulation treatment effect for the treated, as the group who actually receive the treatment is used as the reference population (Dong et al., 2020). If the average subpopulation effect for all is of primary interest, we may consider using the propensity score weighting based tree method, i.e. causal tree.

One practical issue with matching implementation is that we might not be able to match exactly on all possible effect modifiers in practice. This is unlikely to be an issue when sample size is large enough and the size ratio between treated and control is adequate. However, when the sample size is small to moderate and there are many effect modifiers, exact matching may not be possible on all covariates. We propose a strategy of splitting the covariates into several groups and running Matching tree separately for each of them. The practical issue is that these trees might be constructed with different matched sets. Consequently, the identification of subgroup structure may not be optimal. We view this as a trade-off between nonparametric approaches like matching and more parametric predictive modelings. If sample size is limited, matching tree may not be able to identify all subgroups. On the other hand, finding subgroups via modeling may also be questionable, as it relies too much on model structural assumptions.

Another important topic on causal inference in observation data is how to handle unmeasured confounding, as the strongly ignorable treatment assignment assumption may not always hold. A common strategy in population level causal effect evaluation is to use sensitivity analyses, which assess the impact of the potential unobserved confounder from a quantitative perspective (Gastwirth et al., 1998; Rosenbaum, 2005; Nattino and Lu, 2018). It is very likely that unmeasured confounding may exist at subgroup level, future work should gear towards developing appropriate sensitivity analysis strategies for tree-based subgroup analysis methods. A good starting point is Hsu and colleague's work on effect modification and design sensitivity in observational studies (Hsu et al., 2013).

Acknowledgments

Lu's work was partially supported by grant DMS-2015552 from National Science Foundation, USA. Huang's work was supported by the CF Statistical Network and Expertise Award (StatNet SZCZES18Y7), and by a Patient Centered Outcomes Research Institute Method Award (BH ME1408-19894). The researchers thank CFF for use of the registry data and the patients, care providers, and clinic coordinators at US CF centers for their contributions to the registry. The authors also thank the associate editor and two anonymous reviewers for their insightful comments, which lead to substantial improvements in the presentation of the methodology.

Appendix A

This proofs of Lemma 1 and Proposition 1 follow the idea of Rosenbaum and Rubin (1983) where they proved the strong ignorability for the general population. We extend their results to subpopulations using a similar approach. The proofs are included here for completeness.

A.1. Proof of Lemma 1

Proof. First, we can show that $P(T=1|e(\mathbf{X},Z),Z)=e(\mathbf{X},Z)$. This is because

$$\begin{aligned} &P(T=1|e(\mathbf{X},Z),Z)=E(T|e(\mathbf{X},Z),Z)=E[E(T|\mathbf{X},Z)|e(,\mathbf{X}Z),Z] \\ &=E[e(\mathbf{X},Z)|e(\mathbf{X},Z),Z]=e(\mathbf{X},Z). \end{aligned}$$

Therefore, the positivity inequality holds. Next, we will prove the conditional independence.

$$\begin{aligned} &P(T=1|Y^1,Y^0,Z,e(\mathbf{X},Z))=E(T|Y^1,Y^0,Z,e(\mathbf{X},Z)) \\ &=E(E(T|Y^1,Y^0,Z,\mathbf{X})|Y^1,Y^0,Z,e(\mathbf{X},Z))=E(E(T|\mathbf{X},Z)|Y^1,Y^0,Z,e(\mathbf{X},Z)) \\ &=E(e(\mathbf{X},Z)|Y^1,Y^0,Z,e(\mathbf{X},Z))=e(\mathbf{X},Z). \end{aligned}$$

Similarly, we have:

$$\begin{aligned} &P(T=1|Z,e(\mathbf{X},Z))=E(T|Z,e(\mathbf{X},Z))=E(E(T|\mathbf{X},Z)|Z,e(\mathbf{X},Z)) \\ &=E(e(\mathbf{X},Z)|Z,e(\mathbf{X},Z))=e(\mathbf{X},Z). \end{aligned}$$

Therefore, $P(T=1|Y^1,Y^0,Z,e(\mathbf{X},Z))=P(T=1|Z,e(\mathbf{X},Z))$. □

A.2. Proof of Proposition 1

Proof. By Lemma 1, we have $(Y^1, Y^0) \perp\!\!\!\perp T | e(\mathbf{X}, Z), Z$. Then,

$$\begin{aligned}
&E_{e(\mathbf{X},Z)|Z}(E(Y|e(\mathbf{X},Z),T=1,Z) - E(Y|e(\mathbf{X},Z),T=0,Z)) \\
&= E_{e(\mathbf{X},Z)|Z}(E(Y^1|e(\mathbf{X},Z),T=1,Z) - E(Y^0|e(\mathbf{X},Z),T=0,Z)) \\
&= E_{e(\mathbf{X},Z)|Z}(E(Y^1|e(\mathbf{X},Z),Z) - E(Y^0|e(\mathbf{X},Z),Z)) \\
&= E(Y^1|Z) - E(Y^0|Z).
\end{aligned}$$

where $E_{e(\mathbf{X},Z)|Z}$ is the expectation of the distribution of propensity score in the subpopulation defined by Z. □

A.3. Proof of Corollary 1

Proof. First, for the positively inequality, it is easy to show that $P(T=1|e(\mathbf{X}),Z) = e(\mathbf{X})$ because,

$$\begin{aligned}
&P(T=1|e(\mathbf{X}),Z) = E(T|e(\mathbf{X}),Z) = E[E(T|\mathbf{X})|e(\mathbf{X}),Z] \\
&= E[e(\mathbf{X})|e(\mathbf{X}),Z] = e(\mathbf{X}).
\end{aligned}$$

Then, for the conditional independence.

$$\begin{aligned}
&P(T=1|Y^1,Y^0,Z,e(\mathbf{X})) = E(T|Y^1,Y^0,Z,e(\mathbf{X})) \\
&= E(E(T|Y^1,Y^0,Z,\mathbf{X})|Y^1,Y^0,Z,e(\mathbf{X})) = E(E(T|\mathbf{X})|Y^1,Y^0,Z,e(\mathbf{X})) \\
&= E(e(\mathbf{X})|Y^1,Y^0,Z,e(\mathbf{X})) = e(\mathbf{X}).
\end{aligned}$$

Thus,

$$\begin{aligned}
P(T=1|Y^1,Y^0,Z,e(\mathbf{X})) &= P(T=1|Z,e(\mathbf{X})). \\
(Y^1,Y^0) &\perp\!\!\!\perp |e(\mathbf{X}),Z.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E_{e(\mathbf{X})|Z}(E(Y|e(\mathbf{X}),T=1,Z) - E(Y|e(\mathbf{X}),T=0,Z)) \\
&= E_{e(\mathbf{X})|Z}(E(Y^1|e(\mathbf{X}),T=1,Z) - E(Y^0|e(\mathbf{X}),T=0,Z)) \\
&= E_{e(\mathbf{X})|Z}(E(Y^1|e(\mathbf{X}),Z) - E(Y^0|e(\mathbf{X}),Z)) \\
&= E(Y^1|Z) - E(Y^0|Z).
\end{aligned}$$

where $E_{e(\mathbf{X})|Z}$ is the expectation of the distribution of propensity score in the subpopulation defined by Z. □

A.4. Simulation for unbiased subpopulation causal effect estimation – Binary outcome

The baseline covariates are generated same as the continuous outcome. The binary outcome generating models is

$$logit(P(Y=1)) = log(3) + 2 \times T + log(1.5) \times X_1 + log(1.5) \times X_3 + \text{EMF}.$$

where EMF indicates effect modifier function.

- Scenario BB1: Binary Outcome, Binary non-confounder, effect modifier.
  $(\text{EF} = 3 \times X_4 \times T)$
- Scenario BB2: Binary Outcome, Binary confounder, effect modifier. Binary Outcome, Binary non-confounder, effect modifier.
  $(\text{EF} = 3 \times X_3 \times T)$
- Scenario BB3: Binary Outcome, Continuous non-confounder, effect modifier.
  $(\text{EF} = 3 \times I(X_2 > 0.2) \times T)$
- Scenario BB4: Binary Outcome, Continuous confounder, effect modifier.
  $(\text{EF} = 3 \times I(X_1 > 0.2) \times T)$ (see Table A.6).

A.5. Simulation for subpopulation structure identification: Continuous outcome, non-confounding effect modifier

See Tables A.7 and A.8.

A.6. Simulation for subpopulation structure identification: Binary outcome, non-confounding effect modifier

See Tables A.9 and A.10.

Table A.6
Binary outcome, subpopulation ignorability.

RD	Subgroup 1			Subgroup 2		
	%Bias	Coverage	SD	%Bias	Coverage	SD
Scenario BB1	1.16	94.2	0.02	1.21	95.3	0.02
Scenario BB2	1.47	95.1	0.02	0.47	95.2	0.03
Scenario BB3	1.57	94.9	0.03	0.76	95.2	0.02
Scenario BB4	1.91	95.0	0.02	0.22	95.3	0.03
RR	%Bias	Coverage	SD	%Bias	Coverage	SD
Scenario BB1	1.37	100	0.08	1.46	99.8	0.09
Scenario BB2	1.69	99.1	0.08	0.41	98.6	0.09
Scenario BB3	1.89	99.2	0.10	0.95	99.9	0.08
Scenario BB4	2.16	99.2	0.08	0.49	99.2	0.10
OR	%Bias	Coverage	SD	%Bias	Coverage	SD
Scenario BB1	1.12	95.5	0.93	2.27	96.6	0.43
Scenario BB2	2.09	98.2	0.96	0.63	95.7	0.35
Scenario BB3	1.08	94.6	0.96	1.06	96.2	0.35
Scenario BB4	1.81	94.9	0.95	1.17	95.5	0.36

Table A.7
Identification-Continuous outcome, binary effect modifier, non-confounder.

	Tree after prune					Tree before prune		
	%START	%CORR	$\%X_4$	$\%X_5$	$\%X_8$	$\%X_4$	$\%X_5$	$\%X_8$
MT (m)	99.9%	97.6%	1.2%	0.7%	0.8%	5.9%	4.1%	5.2%
CIT (m)	99.3%	93.4%	6.1%	5.3%	5.2%	100%	100%	100%
CT (m)	100%	95.6%	4.4%	1.8%	2.5%	100%	45.5%	45.5%
CIT	96.5%	87.1%	12.5%	10.5%	8.7%	100%	100%	99.8%
CT	100%	94.2%	5.8%	2.4%	3.2%	100%	45%	44.9%
VT.BART	99.6%	98.7%	1.1%	0%	0%	38.5%	3.1%	0.2%
VT.CF	76.9%	76.5%	1.2%	4.8%	4.6%	3.2%	12.9%	9.8%

Note: Methods that run with the matched data are denoted with "(m)"; otherwise, the methods run with the full data.

Table A.8
Identification–Continuous outcome, continuous effect modifier, non-confounder.

	Tree after prune						Tree before prune		
	%START	%VAL	%CORR	$\%X_2$	$\%X_3$	$\%X_7$	$\%X_2$	$\%X_3$	$\%X_7$
MT (m)	100%	98.9%	90.9%	0%	3.4%	0%	0%	22.3%	0%
CIT (m)	100%	84.9%	0%	98.2%	100%	99.1%	100%	100%	100%
CT (m)	100%	78.7%	68.9%	3.8%	2.3%	5.8%	26.8%	53.9%	30.3%
CIT	100%	70.1%	0%	100%	100%	99.2%	100%	100%	100%
CT	99.7%	59.9%	30.1%	11.7%	49.3%	9.9%	52.5%	99.9%	45.8%
VT.BART	100%	99.9%	66.2%	0.1%	0.3%	0%	0.3%	2.5%	0.2%
VT.CF	100%	79.0%	69.8%	12.4%	14.1%	13.7%	65.1%	66.5%	66.3%

Note: Methods that run with the matched data are denoted with "(m)"; otherwise, the methods run with the full data.

Table A.9
Identification – Binary outcome, binary effect modifier, non-confounder.

	Tree after prune					Tree before prune		
	%START	%CORR	$\%X_4$	$\%X_5$	$\%X_8$	$\%X_4$	$\%X_5$	$\%X_8$
MT (m)	98.5%	93.5%	6.5%	3.2%	2.7%	9.8%	7.9%	5.6%
CT (m)	99.5%	89.8%	5.4%	4%	3.4%	27.3%	30.4%	30.7%
CT	98.6%	96.8%	0.6%	1.8%	0.7%	10.1%	54.1%	33.8%
VT.BART	99.8%	96.2%	3%	0.8%	0.3%	85.2%	3.7%	1.4%
VT.CF	64.4%	63.9%	5.2%	8.3%	6.1%	12%	16.5%	11.9%

Note: Methods that run with the matched data are denoted with "(m)"; otherwise, the methods run with the full data.

Appendix B. Simulation *R* codes

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2021.107188.

Table A.10
Identification – Binary outcome, continuous effect modifier, non-confounder.

	Tree after prune						Tree before prune		
	%START	%VAL	%CORR	$\%X_2$	$\%X_3$	$\%X_7$	$\%X_2$	$\%X_3$	$\%X_7$
MT (m)	92.6%	82.7%	73%	0%	15.7%	0%	0%	17.5%	0%
CT (m)	68.2%	50%	30.2%	16.3%	15.2%	15.1%	36.9%	38.9%	34.5%
CT	96.3%	83.7%	56.7%	15%	11.3%	15.3%	50.1%	36.1%	53.1%
VT.BART	87.3%	55.4%	47.2%	11.8%	33.8%	11.5%	15.9%	48.2%	14.3%
VT.CF	97.7%	67.2%	52%	23.9%	26.2%	26.6%	73.2%	71.8%	73.7%

Note: Methods that run with the matched data are denoted with "(m)"; otherwise, the methods run with the full data.

References

Altstein, L., Li, G., 2013. Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model. Biometrics 69 (1), 52–61.

Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proc. Natl. Acad. Sci. 113 (27), 7353–7360.

Bang, H., Robins, J.M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61 (4), 962–973.

Chipman, H.A., George, E.I., McCulloch, R.E., et al., 2010. Bart: Bayesian additive regression trees. Ann. Appl. Stat. 4 (1), 266–298.

Dasenbrook, E.C., Merlo, C.A., Diener-West, M., Lechtzin, N., Boyle, M.P., 2008. Persistent methicillin-resistant staphylococcus aureus and rate of fev1 decline in cystic fibrosis. Amer. J. Respir. Crit. Care Med. 178 (8), 814–821.

Dong, J., Zhang, J.L., Zeng, S., Li, F., 2020. Subgroup balancing propensity score. Stat. Methods Med. Res. 29 (3), 659–676.

Foster, J.C., Taylor, J.M., Ruberg, S.J., 2011. Subgroup identification from randomized clinical trial data. Stat. Med. 30 (24), 2867–2880.

Foundation, C.F., 2013. Cystic Fibrosis Foundation Patient Registry 2012 Annual Data Report. Cystic Fibrosis Foundation Bethesda, MD.

Gastwirth, J.L., Krieger, A.M., Rosenbaum, P.R., 1998. Dual and simultaneous sensitivity analysis for matched pairs. Biometrika 85 (4), 907–920.

Hsu, J.Y., Small, D.S., Rosenbaum, P.R., 2013. Effect modification and design sensitivity in observational studies. J. Amer. Statist. Assoc. 108 (501), 135–148.

Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.

Kim, H.J., Lu, B., Nehus, E.J., Kim, M.-O., 2019. Estimating heterogeneous treatment effects for latent subgroups in observational studies. Stat. Med. 38 (3), 339–353.

Konstan, M.W., Morgan, W.J., Butler, S.M., Pasta, D.J., Craib, M.L., Silva, S.J., Stokes, D.C., Wohl, M.E.B., Wagener, J.S., Regelmann, W.E., et al., 2007. Risk factors for rate of decline in forced expiratory volume in one second in children and adolescents with cystic fibrosis. J. Pediatr. 151 (2), 134–139.

Konstan, M.W., Wagener, J.S., Pasta, D.J., Millar, S.J., Morgan, W.J., 2014. Clinical use of tobramycin inhalation solution (tobi®) shows sustained improvement in fev1 in cystic fibrosis. Pediatr. Pulmonol. 49 (6), 529–536.

Kosorok, M.R., Zeng, L., West, S.E., Rock, M.J., Splaingard, M.L., Laxova, A., Green, C.G., Collins, J., Farrell, P.M., 2001. Acceleration of lung disease in children with cystic fibrosis after Pseudomonas aeruginosa acquisition. Pediatr. Pulmonol. 32 (4), 277–287.

Kraemer, H.C., 2013. Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. Stat. Med. 32 (11), 1964–1973.

Lanza, S.T., Rhoades, B.L., 2013. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. Prev. Sci. 14 (2), 157–168.

Lin, H.-M., Xu, H., Ding, Y., Hsu, J.C., 2019. Correct and logical inference on efficacy in subgroups and their mixture for binary outcomes. Biom. J. 61 (1), 8–26.

Lipkovich, I., Dmitrienko, A., B D'Agostino Sr, R., 2017. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat. Med. 36 (1), 136–196.

Loh, W.-Y., He, X., Man, M., 2015. A regression tree approach to identifying subgroups with differential treatment effects. Stat. Med. 34 (11), 1818–1833.

Lu, B., Greevy, R., Xu, X., Beck, C., 2011. Optimal nonbipartite matching and its statistical applications. Amer. Statist. 65 (1), 21–30.

Nattino, G., Lu, B., 2018. Model assisted sensitivity analyses for hidden bias with binary outcomes. Biometrics 74 (4), 1141–1149.

Ramsey, B., Pepe, M., Quan, J., Otto, K., Montgomery, A., Williams-Warren, J., Vasiljev, K., Borowitz, D., Bowman, C., Marshall, B., et al., 1999. Cystic fibrosis inhaled tobramycin study group: intermittent administration of inhaled tobramycin in patients with cystic fibrosis. New Engl. J. Med. 340 (1), 23–30.

Rosenbaum, P.R., 2005. Sensitivity analysis in observational studies. Encyclopedia Statist. Behav. Sci. 1809–1814.

Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

Rosenbaum, P.R., et al., 2010. Design of Observational Studies, Vol. 10. Springer.

Rothwell, P.M., 2005. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 365 (9454), 176–186.

Stone, C.J., 1984. Classification and regression trees. Wadsworth Int. Group 8, 452–456.

Su, X., Kang, J., Fan, J., Levine, R.A., Yan, X., 2012. Facilitating score and causal inference trees for large observational studies. J. Mach. Learn. Res. 13 (Oct), 2955–2994.

Tian, L., Alizadeh, A.A., Gentles, A.J., Tibshirani, R., 2014. A simple method for estimating interactions between a treatment and a large number of covariates. J. Amer. Statist. Assoc. 109 (508), 1517–1532.

VanDyke, R.D., McPhail, G.L., Huang, B., Fenchel, M.C., Amin, R.S., Carle, A.C., Chini, B.A., Seid, M., 2013. Inhaled tobramycin effectively reduces fev1 decline in cystic fibrosis. an instrumental variables analysis. Ann. Amer. Thorac. Soc. 10 (3), 205–212.