

# Learning Subgroups with Maximum Treatment Effects Without Causal Heuristics

Lincen Yang<sup>1</sup>, Zhong Li<sup>\*1, 2</sup>, Matthijs van Leeuwen<sup>1</sup>, Saber Salehkaleybar<sup>1</sup>

<sup>1</sup> Leiden Institute of Advanced Computer Science (LIACS), Leiden University

<sup>2</sup> The Intelligent Computing Research Center, Great Bay University  
{l.yang, z.li, m.van.leeuwen, s.salehkaleybar}@liacs.leidenuniv.nl

## Abstract

Discovering subgroups with the maximum average treatment effect is crucial for targeted decision making in domains such as precision medicine, public policy, and education. While most prior work is formulated in the potential-outcome framework, the corresponding structural causal model (SCM) for this task has been largely overlooked. In practice, two approaches dominate. The first estimates pointwise conditional treatment effects and then fits a tree on those estimates, effectively turning subgroup estimation into the harder problem of accurate pointwise estimation. The second constructs decision trees or rule sets with ad-hoc ‘causal’ heuristics, typically without rigorous justification for why a given heuristic may be used or whether such heuristics are necessary at all.

We address these issues by studying the problem directly under the SCM framework. Under the assumption of a partition-based model, we show that optimal subgroup discovery reduces to recovering the data-generating models and hence a standard supervised learning problem (regression or classification). This allows us to adopt *any* partition-based methods to learn the subgroup from data. We instantiate the approach with CART, arguably one of the most widely used tree-based methods, to learn the subgroup with maximum treatment effect. Finally, on a large collection of synthetic and semi-synthetic datasets, we compare our method against a wide range of baselines and find that our approach, which avoids such causal heuristics, more accurately identifies subgroups with maximum treatment effect.

**Code** — <https://github.com/ylicen/causal-subgroup>

**Extended version** — <https://arxiv.org/pdf/2511.20189>

## Introduction

Subgroup discovery is the task of learning a subgroup, typically described by interpretable rules, from data such that the distribution of a target quantity deviates from that of the full dataset (Atzmueller 2015). In causal inference, the quantity of interest is often the treatment effect, and many works aim to learn subgroups whose effects are enhanced relative to the overall population or a baseline (Su et al. 2009; Wang and Rudin 2022). Such analyses have proved useful in application domains such as healthcare (Lipkovich, Dmitrienko,

and B D’Agostino Sr 2017; Rothwell 2005; Loh, Cao, and Zhou 2019) and education (Athey and Wager 2019).

In this paper we focus on discovering the subgroup with the maximum average treatment effect, which we name as the *maximum-effect subgroup*, not merely a subgroup with an enhanced or above-average effect. Rather than to characterize the heterogeneity of the treatment effects for the whole population in general, we aim to ask for a single optimum, which can answer questions like which subgroup of patients can benefit most from a certain disease treatment plan (Zhang et al. 2017; Goligher et al. 2023; Nagpal et al. 2020). Importantly, this maximum-effect objective also provides a foundation for understanding heterogeneity more generally: one can iteratively remove the discovered subgroup and re-apply the method to the remaining instances—an approach often referred to as sequential covering or divide-and-conquer (Fürnkranz, Gamberger, and Lavrač 2012; Cohen 1995).

Existing methods for learning subgroups with enhanced treatment effects from data largely take the following two approaches. The first approach takes a two-step process (Foster, Taylor, and Ruberg 2011; Huang, Tang, and Kenney 2025): 1) estimating the pointwise conditional treatment effect, and 2) applying an off-the-shelf subgroup discovery method (Lavrač et al. 2004; Van Leeuwen and Knobbe 2012) or fitting a partition-based model, e.g., a classification tree (Breiman et al. 1984) or a rule set (Clark and Niblett 1989), to the estimated pointwise effects to obtain the subgroups with enhanced treatment effects. While flexible, this strategy converts subgroup discovery into the arguably harder task of accurately estimating pointwise conditional treatment effects, and the learned subgroups can be highly sensitive to the pointwise estimation error.

The second approach directly tries to search for subgroups with enhanced treatment effects (Zhou et al. 2024; Dusseldorp and Van Mechelen 2014; Athey and Imbens 2016; Su et al. 2009). It estimates the subgroup treatment effect (i.e., the treatment effect conditioned on the subgroup) and uses that estimate to design the heuristics for searching the subgroups. Algorithmically, the search can be achieved by either a tree-based approach, i.e., grow a tree and pick a subgroup from the leaves (Dusseldorp and Van Mechelen 2014; Su et al. 2009), or a rule-based approach that typically reduces to a combinatorial optimization problem (Zhou et al.

\*corresponding author

2024). In practice, these methods often blend the estimated subgroup treatment effect with regularization terms (based on the number of instances contained in the subgroups). However, the proposed heuristics typically lacked theoretical justification at the time they were introduced, and, when the goal is to find the maximum-effect subgroup, two issues arise: 1) while a large number of different heuristics have been proposed, which heuristic, if any, is theoretically appropriate for identifying the maximizer, and 2) are specialized ‘causal’ heuristics necessary at all?

Thus, these questions hinder the naive approach of using existing methods that discover the subgroups with enhanced treatment effect and then pick the single subgroup within them as the maximum-effect subgroup.

To tackle these shortcomings, we leverage the structural causal model (SCM) framework, and prove that, under the assumption of a partition-based model, the task of learning the maximum-effect subgroup can be reduced to a standard supervised learning task that aims to reveal the data-generating model. We instantiate this approach by training a CART (Breiman et al. 1984) tree, one of the most commonly used methods for supervised learning. We empirically compare against commonly used baseline methods, using both synthetic and semi-synthetic datasets. We demonstrate that our approach, even with the classic yet simple CART algorithm, shows superior performance in the task of maximum-effect subgroup discovery. To our knowledge, this is the first work to ground maximum-effect subgroup discovery within structural causal inference via rigorous theoretical results.

It is noteworthy to mention that our primary contribution is *not* to propose a new, specific heuristic and/or algorithm for discovering subgroups with maximum treatment effects; instead, we provide a theoretical result that offers insights into whether ‘causal’ heuristics are unnecessary for the task of subgroup discovery in this context. Our empirical results are aimed at justifying the general approach of reducing the task of discovering the maximum-effect subgroup to the standard supervised machine learning tasks.

## Related Work

Existing methods for discovering subgroups with maximum/enhanced treatment effects can be categorized into three approaches. The first approach aims to learn a decision tree or a decision rule set, with specifically designed learning criteria and/or heuristics that are tailored for causal inference. These criteria/heuristics often combine an empirical estimate of the subgroup treatment effect—computed by selecting instances that satisfy the rule that describes this subgroup and averaging outcomes within treatment and control—with regularization terms. For instance, the seminal work Causal Tree (Athey and Imbens 2016) estimates the average treatment effect within each node and uses an unbiased proxy for the mean-squared error of the treatment-effect estimator for splitting and pruning, since individual counterfactual outcomes are unobserved. Further, QUINT (Dusseldorp and Van Mechelen 2014) adopts the weighted sum of the subgroup treatment effect of each leaf node plus a term that encourages large subgroup size as the splitting criterion. As QUINT uses a bootstrap procedure for

pruning with an expensive computational cost, the scalability is limited. Similarly, Interaction Tree (Su et al. 2009) uses the difference between the subgroup treatment effects of two children nodes (standardized by the estimated pooled variance) as the splitting criterion.

Beyond trees, SIDES (Lipkovich et al. 2011) directly searches for subgroups without explicitly building a tree. SIDES provides several heuristics to split a subgroup into more refined subgroups, and essentially leaves the choice of which heuristic to use to the users. Last, the more recent rule set methods, including CURLS (Zhou et al. 2024) and MOSIC (Chen et al. 2025), adopt a similar criterion in learning a rule set and directly maximize the subgroup treatment effect, regularized by a model complexity term.

Thus, we conclude that although various criteria have been proposed, there is still no principled guideline for determining which is most appropriate to use in a given practical scenario. Our theoretical analysis and superior empirical experiment results both challenge the necessity of leveraging such learning objectives and/or algorithmic heuristics to learn the maximum-effect subgroup.

The second approach first estimates pointwise conditional treatment effects and then partitions the feature space using those estimates. Specifically, Virtual Twins (Foster, Taylor, and Ruberg 2011) first builds a predictive model to predict the counterfactual outcomes by setting the treatment variable to zero or one. Then, it fits a tree or regressor to produce approximately homogeneous-effect regions. Similarly, the X-learner (Künzel et al. 2019) first splits the dataset into the treatment (i.e., the treatment variable  $T = 1$ ) and the control group ( $T = 0$ ), then separately learns two models to predict the conditional treatment effect of individual points, and finally applies any supervised method to model the pointwise conditional treatment effects. Further, Distill Tree (Huang, Tang, and Kenney 2025) can leverage any off-the-shelf method that predicts pointwise conditional treatment effect, and then use a ‘student model’ to distill the subgroups. The first step here is often referred to as the task of (individual) conditional treatment effect estimation, which can be achieved by various approaches, including double machine learning (Chernozhukov et al. 2018), meta-learners (Nie and Wager 2021; Künzel et al. 2019), (ensemble) tree methods (ATHEY, TIBSHIRANI, and WAGER 2019; Hahn, Murray, and Carvalho 2020), and (deep) representation learning (Curth and Van der Schaar 2021; Shi, Blei, and Veitch 2019; Lee et al. 2025) (although not all these methods have been considered in learning subgroups with the maximum or enhanced treatment effects).

We argue that estimating pointwise conditional treatment effect is a much harder problem than estimating the subgroup treatment effect in practice; hence, the variance of the estimator is in general large and will highly depend on the chosen model.

The third approach is model-based recursive partitioning (Seibold, Zeileis, and Hothorn 2016). Unlike the previous two families, it does not treat the subgroup effect as constant. Instead, it starts with a global predictive model that associates the target variable  $Y$  with the feature variables  $X$  and treatment variable  $T$ . It then iteratively splits the feature

space into hyper-cubes, and refits model parameters within each region. However, as shown by our theoretical results, a subgroup that maximizes the average treatment effect must exhibit homogeneous pointwise treatment effects (which we explain in detail later). By design, model-based approaches allow within-subgroup variation and therefore do not target the maximum-effect subgroup as defined here.

## Theory

We first review the basic concept of structural causal models and the definition of ‘rules’. Then, we show that the maximum-effect subgroup must have a homogeneous pointwise conditional treatment effect. Last, we introduce the partition-based model and present our main theorem, which states that learning the maximum-effect subgroup reduces to the task of learning the underlying data-generating process.

### Preliminaries

**Structural causal models.** A structural causal model (SCM) is a set of functions that fully specifies the data-generating process: given a set of variables  $\{V_i\}_{i \in \{1, \dots, m\}}$ , a SCM can be specified by a set of functions  $V_i = f_i(pa(V_i), U_i)$ , in which  $pa(V_i)$  denotes the parent nodes of  $V_i$  and all  $U_i$  are independent (Pearl 2009). The causal relationship is often represented as a directed acyclic graph (DAG) with nodes  $\{V_i\}$  where there is a directed edge from  $V_j$  to  $V_i$  if  $V_j \in pa(V_i)$ . Given an SCM, we can specify an intervention distribution, denoted as  $P(V_i | do(V_j = v_j))$ , as the conditional probability of  $V_i$  specified by the SCM model after replacing the original equation  $V_j = f_j(pa(V_j), U_j)$  by  $V_j = v_j$ .

Under certain conditions (Pearl 2012), the intervention distribution  $P(V_i | do(V_j = v_j))$  can be computed from expressions written based on the observational distribution. This is often referred to as ‘identifiability’, and can be used to estimate  $P(V_i | do(V_j = v_j))$  from the observational data.

**Subgroups and rules.** A subgroup is often described by a rule, which is a logical conjunction of literals (Van Leeuwen and Knobbe 2012; Zhou et al. 2024): given a dataset with feature variables  $X_j, j \in \{1, \dots, m\}$ , a literal is in the form of  $X_j \in R_j$ , in which  $X_j$  is a single feature variable and  $R_j$  represents a subset of the domain of  $X_j$ ; hence, a rule can be denoted as a conjunction of several literals  $\bigwedge X_j \in R_j$ .

### Causal subgroup discovery

We first consider the general form of structural causal model (SCM) for estimating treatment effect, i.e.,

$$\begin{aligned} X &= N_X, T = f_T(X, N_T), Y = f_Y(T, X, N_Y), \\ N_X, N_T, N_Y &\text{ are independent exogenous noises,} \end{aligned}$$

with the DAG as shown in Fig. 1.

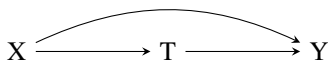


Figure 1: The DAG for causal subgroup discovery.

$X$  can be a multivariate random variable with its domain denoted as  $\mathcal{X}$ . Further, we also assume that there are no hidden confounders that can affect both the treatment variable  $T$  and the target variable  $Y$ .<sup>1</sup>

To simplify notations, we assume that  $X$  is discrete and  $Y$  is a binary target variable *when developing our theory*. However, our theoretical results can be directly extended to continuous  $X$  (by replacing the  $\sum$  with the integral sign in our derivations) and to numeric targets (by simply replacing the  $P(\cdot)$  with  $E(\cdot)$ ). Our experiments do contain datasets with continuous  $X$  and  $Y$ .

Our goal is to find a subgroup such that the treatment effect of the subgroup is maximized. Specifically, we define the *subgroup treatment effect* as follows.

**Definition 1** (Subgroup treatment effect). *Given a  $Q \subseteq \mathcal{X}$ , in which  $\mathcal{X}$  is the domain of  $X$ , we say  $Q$  is a subgroup, and meanwhile we define the subgroup treatment effect as*

$$\begin{aligned} A(Q) &:= P(Y = 1 | do(T := 1), X \in Q) - \\ &P(Y = 1 | do(T := 0), X \in Q). \end{aligned} \quad (1)$$

We hence formally state our problem as finding the maximum-effect subgroup

$$Q^* = \arg \max_Q A(Q). \quad (2)$$

Further, under the assumption that there exists no hidden confounder and  $P(X \in Q) > 0$  (i.e., subgroups with probability zero are not interesting in practice),  $A(Q)$  is *identifiable* from the observational distribution:

**Proposition 1** (Identifiable).  $A(Q) = E_{X|X \in Q}[P(Y = 1 | T = 1, X) - P(Y = 1 | T = 0, X)]$  under the assumptions stated above.

The proof is deferred to the supplementary materials of the extended version. Next, we can further prove that an maximum-effect subgroup  $Q^*$  must have homogeneous treatment effect, which we formally define as follows:

**Definition 2** (Subgroup with a homogeneous treatment effect). *A subgroup  $Q \subset \mathcal{X}$  has homogeneous treatment effect if  $\forall Q' \subset Q, A(Q) = A(Q')$ . Equivalently, as  $Q'$  can be a subset that contains only one single point, i.e.,  $Q' = \{x'\}$  for some  $x' \in \mathcal{X}$ , we also have  $A(Q) = A(X = x), \forall x \in Q$ , in which  $A(X = x) := P(Y | do(T := 1), X = x) - P(Y | do(T := 0), X = x)$ .*

The next theorem justifies the approach of finding the maximum-effect subgroup through subgroups with homogeneous treatment effects.

**Theorem 1.** *For a subset  $Q' \subset Q$ , if  $A(Q') \leq A(Q)$  then we must have  $A(Q \setminus Q') \geq A(Q)$ , and the equality  $A(Q \setminus Q') = A(Q)$  only holds when  $A(Q') = A(Q)$ .*

<sup>1</sup>Previous methods for the task of subgroup discovery for enhanced/maximum treatment effects (Athey and Imbens 2016; Dusseldorp and Van Mechelen 2014; Su et al. 2009; Wang and Rudin 2022; Zhou et al. 2024) often leverage the potential outcome framework (Rubin 1974), and they often have the ‘‘exchangeability’’ (Rosenbaum and Rubin 1983) assumption, which is equivalent to the no hidden-confounder assumption.

*Proof.*

$$\begin{aligned}
A(Q) &= \sum_{x \in Q} (P(Y = 1 | do(T := 1), X = x) - \\
&\quad P(Y = 1 | do(T := 0), X = x)) P(X = x | X \in Q) \\
&= \sum_{x \in Q'} (P(Y = 1 | do(T := 1), X = x) - \\
&\quad P(Y = 1 | do(T := 0), X = x)) P(X = x | X \in Q) + \\
&\quad \sum_{x \in Q \setminus Q'} (P(Y = 1 | do(T := 1), X = x) - \\
&\quad P(Y = 1 | do(T := 0), X = x)) P(X = x | X \in Q) \quad (3)
\end{aligned}$$

Note that  $\forall x \in Q'$ , we have  $P(X = x) = P(X = x, X \in Q')$ , and as a result

$$\begin{aligned}
P(X = x | X \in Q) &= \frac{P(X = x, X \in Q)}{P(X \in Q)} = \frac{P(X = x)}{P(X \in Q)} \\
&= \frac{P(X = x) P(X \in Q')}{P(X \in Q) P(X \in Q')} = P(X = x | X \in Q') \frac{P(X \in Q')}{P(X \in Q)};
\end{aligned}$$

Similarly, we can also show that

$$P(X = x | X \in Q) = P(X = x | X \in Q \setminus Q') \frac{P(X \in Q \setminus Q')}{P(X \in Q)}.$$

Thus, by substituting  $P(X = x | X \in Q)$  in Eq. 3, we have

$$A(Q) = \frac{P(X \in Q')}{P(X \in Q)} A(Q') + \frac{P(X \in Q \setminus Q')}{P(X \in Q)} A(Q \setminus Q').$$

Hence

$$\min\{A(Q'), Q \setminus Q'\} \leq A(Q) \leq \max\{A(Q'), Q \setminus Q'\}; \quad (4)$$

and the equality only holds when  $A(Q') = A(Q \setminus Q')$ .  $\square$

Thus, an maximum-effect subgroup  $Q^*$  must have homogeneous treatment effect (otherwise the subgroup can be further split and one of the subsets will have a higher average treatment effect).

### Our proposed model

We propose a partition-based model, which we formally define by the following structural causal model (SCM):

**Definition 3.** *The partition-based model is defined as*

$$\begin{aligned}
X &= N_X, T = f_T(X, N_T), \\
Y &= f_Y(T, \sum_i^I i \cdot \mathbf{1}_{K_i}(X), N_Y)
\end{aligned}$$

$$K_i \subseteq \mathcal{X}, \forall i, j \in I, K_i \cap K_j = \emptyset,$$

$N_X, N_T, N_Y$  are independent exogenous noises,

in which  $\mathbf{1}_{K_i}(\cdot)$  is the indicator function which is equal to one if  $X \in K_i$  (otherwise, it is zero), and  $I = \{1, 2, \dots, |I|\}$  is an index set. The causal relationship among variables can be represented by Figure 1 as well.

In the partition-based model, we assume the feature space can be partitioned into small subsets, within each the conditional probability  $P(Y|X, do(T = t))$ ,  $t \in \{0, 1\}$  becomes homogeneous; i.e.,  $\forall x \in K_i, P(Y|X = x, do(T = t)) = P(Y|X \in K_i, do(T = t))$ .

Although most subgroup discovery methods adopt the potential outcome framework (Rubin 1974), we argue that they often implicitly assume this partition-based model. Specifically, when predicting the subgroup treatment effects, all instances in the same subgroup (or a tree leaf node for tree-based methods) are assigned with a single predicted outcome for  $T = t$ , where  $t \in \{0, 1\}$  (Athey and Imbens 2016).

We next show that under the condition that the data is generated according to the partition-based model (Def. 3), the subgroup that maximizes the treatment effect, defined in Eq. 1, is one of the  $K_i, i \in I$ .

**Theorem 2.** *Denote  $S = \sum_{i \in I} i \cdot \mathbf{1}_{K_i}(X)$ ,  $K_i \subseteq \mathcal{X}, \forall i, j \in I, K_i \cap K_j = \emptyset$ . Denote  $A(S = i) = P(Y = 1 | S = i, do(T := 1)) - P(Y = 1 | S = i, do(T := 0))$ , and without loss of generality, assume that*

$$A(S = 1) \geq A(S = i), \forall i \in I \quad (5)$$

Then,  $\forall Q \subseteq \mathcal{X}$ , we have

$$\begin{aligned}
A(S = 1) &\geq P(Y = 1 | X \in Q, do(T := 1)) \\
&\quad - P(Y = 1 | X \in Q, do(T := 0)) \quad (6)
\end{aligned}$$

*Proof.* Without loss of generality, we assume that  $S \neq 0$  no matter what value  $X$  takes. This can be achieved by making the last subgroup  $K_n = \mathcal{X} \setminus (\cup_{i \in \{1, \dots, n-1\}} K_i)$ . The right hand side (RHS) of Eq. 6 is

$$\begin{aligned}
\text{RHS} &\stackrel{(a)}{=} \frac{P(Y = 1, X \in Q | do(T := 1))}{P(X \in Q | do(T := 1))} - \\
&\quad \frac{P(Y = 1, X \in Q | do(T := 0))}{P(X \in Q | do(T := 0))} \\
&\stackrel{(b)}{=} \frac{\sum_i P(Y = 1, X \in Q \cap K_i | do(T := 1))}{P(X \in Q | do(T := 1))} - \\
&\quad \frac{\sum_i P(Y = 1, X \in Q \cap K_i | do(T := 0))}{P(X \in Q | do(T := 0))} \\
&= \sum_i \left( \frac{P(X \in Q \cap K_i | do(T := 1))}{P(X \in Q | do(T := 1))} \right. \\
&\quad \left. P(Y = 1 | X \in Q \cap K_i, do(T := 1)) \right) - \\
&\quad \sum_i \left( \frac{P(X \in Q \cap K_i | do(T := 0))}{P(X \in Q | do(T := 0))} \right. \\
&\quad \left. P(Y = 1 | X \in Q \cap K_i, do(T := 0)) \right) \\
&\stackrel{(c)}{=} \sum_i \left( P(Y = 1 | X \in Q \cap K_i, do(T := 1)) - \right. \\
&\quad \left. P(Y = 1 | X \in Q \cap K_i, do(T := 0)) \right) \frac{P(X \in Q \cap K_i)}{P(X \in Q)}
\end{aligned}$$

$$\begin{aligned}
&\stackrel{(d)}{=} \sum_i \frac{P(X \in Q \cap K_i)}{P(X \in Q)} (P(Y = 1|S = i, do(T := 1)) - \\
&P(Y = 1|S = i, do(T := 0))) \\
&\stackrel{(e)}{\leq} \sum_i \frac{P(X \in Q \cap K_i)}{P(X \in Q)} (P(Y = 1|S = 1, do(T := 1)) - \\
&P(Y = 1|S = 1, do(T := 0))) \\
&= \text{LHS of Eq. 6,}
\end{aligned}$$

where in: (a) we use the Bayes rule; (b) we partition the event  $\{X \in Q\}$  into subsets  $\{X \in Q \cap K_i\}_{i \in I}$ ; (c) we use the rule of ‘deletion of intervention’ of do-calculus (Pearl 2009); (d) we leverage our partition-based model property that since  $Y = f_Y(T, \sum_i^I i \cdot \mathbf{1}_{K_i}(X), N_Y)$ ,  $P(Y|S = i, do(T = t)) = P(Y|X \in K_i, do(T = t)) = P(Y|X \in K_i \cap Q, do(T = t))$ ; (e) we use the assumption Eq. 5.  $\square$

Theorem 2 states that, under a partition-based model that divides the feature space into subsets  $\{K_i\}_{i \in I}$ , the subgroup  $Q^*$  that maximizes the subgroup treatment effect must be one of the partition subset, i.e.,  $Q^* = \arg \max_{i \in I} A(K_i)$ . Consequently, searching for the maximum-effect subgroup reduces to two steps: 1) recovering the data-generating partition, e.g., by maximum likelihood (equivalently, minimizing cross-entropy) or a surrogate such as the Gini index used by CART (Breiman et al. 1984), and 2) evaluating the learned subsets and selecting the one with the largest estimated treatment effect. In practice, the learned partition typically contains far fewer subsets than the sample size (to avoid overfitting), so scanning all subsets is computationally lightweight.

In sum, our theorem questions the necessity of using bespoke ‘causal’ heuristics and/or learning criteria for discovering maximum-effect subgroups. Instead, Theorem 2 indicates that we can use any off-the-shelf decision-tree-based or rule-based methods for classification (for nominal  $Y$ ) or regression (for continuous  $Y$ ) for such task.

## Learning the Subgroup from Data

As our Theorem 2 does not restrict us to use any specific learning method, we instantiate our approach with CART (Breiman et al. 1984), a widely used tree method for classification and regression. Notably, rather than proposing a specific, new algorithm for learning maximum-effect subgroups, we aim to empirically validate the general approach of learning such subgroups by reducing it to a standard classification or regression task. Hence, other advanced tree/rule-based methods can be adopted here as well (Yang and van Leeuwen 2022, 2024; Hu, Rudin, and Seltzer 2019; Briša, van der Linden, and Demirović 2025).

Specifically, for a discrete target variable  $Y$ , we fit a CART decision tree, together with the standard criterion Gini-index. Gini-index is a surrogate loss of cross-entropy loss, and hence an approximation for maximizing the likelihood. Meanwhile, for a continuous target variable, we fit a CART regression tree, with the standard mean squared error (MSE) as the learning criterion (Breiman et al. 1984). For numeric targets, optimizing MSE corresponds to maximum

likelihood under the assumption that the conditional distribution of the target variable is Gaussian. To mitigate overfitting, we apply the common approach of cost-complexity tree pruning with the help of cross-validation.

After the decision (regression) tree is learned from data, we next describe how to find the subgroup with the maximum treatment effect. Given any leaf node of the learned tree, the path from the root node to this leaf node can be considered a single rule  $R$ . Hence, each rule defines a subgroup (hyper-rectangle) of the feature space  $\mathcal{X}$  which we denote as  $Q$ . Notably, if rule  $R$  contains a condition of  $T = t$ , ( $t \in \{0, 1\}$ ), the subgroup is obtained by ignoring the internal node that contains the condition for the treatment variable  $T$ .

To estimate the subgroup treatment effect of the subgroup  $Q$ , we adopt ‘honest’ inference that is proposed in the previous work of Causal Tree (Athey and Imbens 2016). Before we train the decision (regression) tree, we randomly split the dataset into the training and test sets. We first train the decision (regression) tree model with the training set; then, with the model learned from it, we estimate the treatment effect for each subgroup with the test set. Precisely, given the instances of the test set that is contained in the subgroup  $Q$ , we further split it into two subsets, one with  $T = 0$  and the other with  $T = 1$ . With these two subsets, we can calculate the empirical conditional probability  $\hat{P}(Y|T = t, X \in Q)$ ,  $t \in \{0, 1\}$ .

Next, since CART aims to partition the feature space such that each leaf node contains instances with approximately homogeneous conditional probabilities, its estimate of  $P(Y|X, T)$  for all instances reaching a particular leaf  $Q$  is given by the empirical average of the target values of those instances: e.g., for  $T = 1$  and a binary target  $Y = 1$ ,  $\hat{P}(Y = 1|X = x, T = 1) = \frac{|\{(x, t, y) \in D: y=1, x \in Q, t=1\}|}{|\{(x, t, y) \in D: x \in Q, t=1\}|} =: \hat{P}(Y = 1|X \in Q, T = 1)$ , in which  $D$  denotes the dataset and  $|\cdot|$  denotes set cardinality. By substituting these empirical estimates into the expression in Proposition 1, we obtain the proposed estimator for  $A(Q)$ :

$$\hat{A}(Q) := \hat{P}(Y = 1|X \in Q, T = 1) - \hat{P}(Y = 1|X \in Q, T = 0).$$

Finally, we go over all subgroups and pick the subgroup with the highest estimated treatment effect as our maximum-effect subgroup learned from data, which we denote as  $\hat{Q}$ .

## The advantage over ‘causal’ heuristics

Guided by Theorem 1, one might be tempted to maximize the empirical subgroup treatment effects and use it as a heuristic when learning the tree-based partition model. For instance, when splitting a tree node  $Q$  into two children  $Q_L$  and  $Q_R$ , one could regard each tree node as a subgroup (as previously described), and build the heuristics in the form of  $\max\{\hat{A}(Q_L), \hat{A}(Q_R)\}$ , or  $|\hat{A}(Q_L) - \hat{A}(Q_R)|$ . These heuristics are commonly used in previous works (Dusseldorp and Van Mechelen 2014; Athey and Imbens 2016; Su et al. 2009) as specifically designed ‘causal’ heuristics.

However, these heuristics are fragile because  $\hat{A}(Q)$  requires accurate estimation of *both* terms in the definition of  $A(Q)$  (Eq. 1). During tree growth, imbalances between the

treatment ( $T = 1$ ) and control ( $T = 0$ ) groups are common. For instance, the extreme case would be that there are many data points satisfying the condition  $\{T = 1 \wedge X \in Q\}$  yet few points satisfying the condition  $\{T = 0 \wedge X \in Q\}$ . Under such circumstances, one might prefer to shrink  $Q$  to potentially increase  $\hat{P}(Y = 1|T = 1, X \in Q)$ , while also being inclined to expand  $Q$  such that it contains more instances, in order to have a more reliable estimation for  $\hat{P}(Y = 1|T = 0, X \in Q)$ . These objectives are inherently conflicting and cannot be achieved simultaneously.

Thus, leveraging  $\hat{A}(Q)$  as splitting heuristics either leads to large variance in the training phase and the risk of overfitting, or a compromise between the two subsets of instances with  $T = 1$  and  $T = 0$ , respectively. We avoid this pitfall by optimizing the Gini index, a well-behaved surrogate for cross-entropy (i.e., log-likelihood maximization). This objective remains stable under imbalance and defers the treatment effect estimation to an ‘honest’ stage where both subsets with ( $T = 1$ ) and ( $T = 0$ ) are adequately supported.

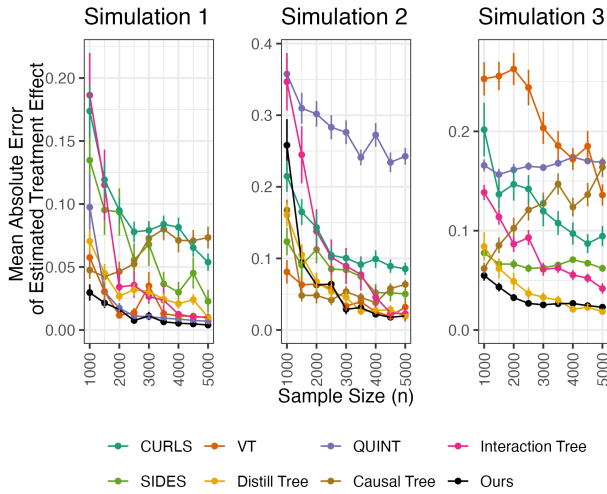


Figure 2: The sample sizes versus the absolute value of the difference between the estimated treatment effect of the learned subgroup  $\hat{Q}$  and that of the ground-truth subgroup  $Q_{gt}$  (lower is better). All simulations are repeated 50 times, and we use the error bar to represent the standard error.

## Experiment

### Implementation and baseline methods

We consider seven baseline methods in total. Five use specifically designed causal heuristics to learn the subgroups, including SIDES (Lipkovich et al. 2011; Lipkovich, Dmitrienko, and B D’Agostino Sr 2017), Interaction Trees (Su et al. 2009), QUINT (Dusseldorp and Van Mechelen 2014), Causal Tree (Athey and Imbens 2016), and the recently proposed CURLS (Zhou et al. 2024). The remaining two adopt the two-step approach, i.e., first estimating the pointwise conditional treatment effect and then fitting a model to the pointwise estimates to learn the subgroups. These include Virtual Twins (VT) (Foster, Taylor, and Ru-

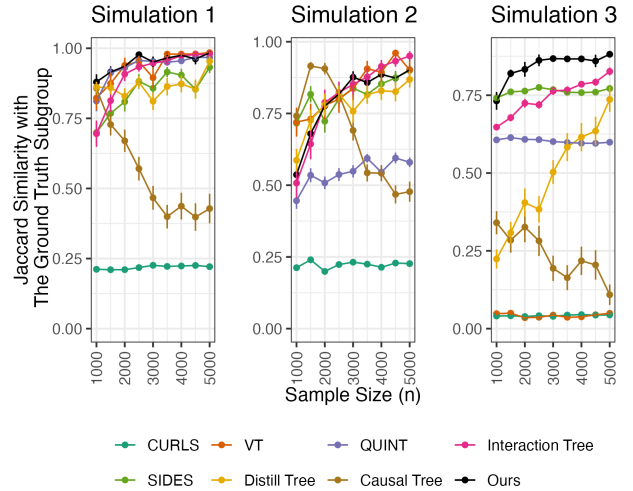


Figure 3: The sample sizes versus the Jaccard similarity of the learned subgroup  $\hat{Q}$  and the ground-truth subgroup  $Q_{gt}$  (higher is better). All simulations are repeated 50 times, and we use the error bar to represent the standard error.

berg 2011) (only suitable for binary targets), and the recently proposed Distill Tree (Huang, Tang, and Kenney 2025).

To ensure a fair comparison, we use the same ‘honest’ inference protocol for all methods (i.e., to train the model on a training split, and to estimate the subgroup treatment effects on a held-out test split). All hyperparameters are tuned as in the original baseline work; when unspecified, we follow the authors’ tuning procedures or use the default values from their code. The full implementation details are included in the supplementary materials.

### Synthetic Datasets

We consider the following simulations: 1) **(Simulation 1) when features and treatment are independent**, i.e., the features  $X_1, X_2 \sim N(0, 1)$ , the treatment  $T \sim Ber(0.5)$ , and the target  $Y|X_1 > 1 \wedge T = 1 \sim Ber(0.8)$ ,  $Y|X_1 < -1 \wedge T = 0 \sim Ber(0.75)$ ,  $Y|otherwise \sim Ber(0.2)$ , in which  $Ber(\cdot)$  represents the Bernoulli distribution; 2) **(Simulation 2) when features and treatment are dependent**, i.e., the features  $X_1, X_2 \sim N(0, 1)$  and the treatment  $T|X_1 \geq 0 \sim Ber(0.8)$ ,  $T|X_1 < 0 \sim Ber(0.2)$ , and the target  $Y$  is generated as in the previous case; 3) **(Simulation 3) Rule list simulator**. The features  $X_1, \dots, X_5 \sim N(0, 1)$  and the treatment  $T \sim Ber(0.5)$ . Define rule<sub>1</sub> :  $X_1 > -1 \& X_2 > -1 \& X_3 > -1$ , rule<sub>2</sub> :  $X_1 > -1 \& X_2 > -1 \& \neg rule_1$ , rule<sub>3</sub> :  $X_1 > -1 \& \neg rule_1 \& \neg rule_2$ . Then  $Y|T = 1, rule_1 \sim Ber(0.8)$ ,  $Y|T = 1, rule_2 \sim Ber(0.6)$ ,  $Y|T = 1, rule_3 \sim Ber(0.4)$ , and  $Y|otherwise \sim Ber(0.2)$ .

We vary the sample sizes from 1000 to 5000, and repeat the simulation 50 times. We first apply our method to learn the maximum-effect subgroup  $\hat{Q}$  from data, and obtain the estimated subgroup treatment effect denoted as  $\hat{A}(\hat{Q})$ . Next, as we know the ground-truth maximum-effect subgroup, which we denote as  $Q_{gt}$ , we also investigate its es-

timated subgroup treatment effect, denoted as  $\hat{A}(Q_{gt})$ . We report the difference between them  $|\hat{A}(\hat{Q}) - \hat{A}(Q_{gt})|$  (lower is better) in Figure 2, and meanwhile report the Jaccard similarity between the instances contained in  $Q_{gt}$  and those contained in  $\hat{Q}$  in Figure 3 (higher is better).

As shown in Figures 2–3, our method performs competitive in Simulation 1 and 2 where the ground-truth partition contains only two subgroups. Notably, several heuristics-driven causal methods already begin to fail in these simple cases, either showing suboptimal effect estimates or poor subgroup recovery. In Simulation 3, our method clearly outperforms the five heuristics-driven baselines in both metrics with smaller absolute effect error and higher Jaccard similarity. Further, although the two-step approach algorithm Distill Tree is a close competitor in effect error (Figure 2), it performs substantially worse in Jaccard similarity (Figure 3).

| Method          | Mean ground-truth subgroup treatment effect | Proportions of ranked first among all 77 datasets |
|-----------------|---|---|
| <b>Ours</b>     | <b>10.540</b>                               | <b>0.519</b>                                      |
| CURLS           | 7.410                                       | 0.180   |
| CausalTree      | 7.843                                       | 0.143   |
| DistillTree     | 7.451                                       | 0.130   |
| InteractionTree | 6.280                                       | 0.039   |
| QUINT           | 5.135                                       | 0.000   |
| SIDES           | 4.622                                       | 0.013   |

Table 1: Ground-truth subgroup treatment effect for the learned subgroups of the 77 semi-synthetic datasets (higher is better). Our method ranks first in 51.9% of all datasets.

### Semi-synthetic Datasets

We consider the commonly used semi-synthetic simulator ACIC-2016 (Dorie et al. 2019), which can generate semi-synthetic datasets based on the Infant Health and Development Program (IHDP) (Louizos et al. 2017). We used 77 semi-synthetic datasets generated by this simulator, and we again adopt the honest estimation as previously described.

As the outcome is given for both  $T = 1$  and  $T = 0$ , the ground-truth treatment effect is known for each single instance. We take the average treatment effects for all instances in our learned subgroup and report this ground-truth subgroup average treatment effects for each dataset (higher is better). As reported in Table 1, our learned subgroups have the highest ground-truth subgroup treatment effect.

Further, we demonstrate the ground-truth average treatment effect for the learned subgroups for each individual dataset in Figure 4. The results show that, for a significant proportion of all 77 datasets, we have identified subgroups with substantially larger average treatment effects than those of other baselines. We finally emphasize that these semi-synthetic datasets are *not* simulated based on a partition-based model; thus, it demonstrates that our method is empirically robust and can generalize to more general cases.

### Conclusion and Discussion

We studied the problem of learning the subgroup with the maximum treatment effect under the the structural

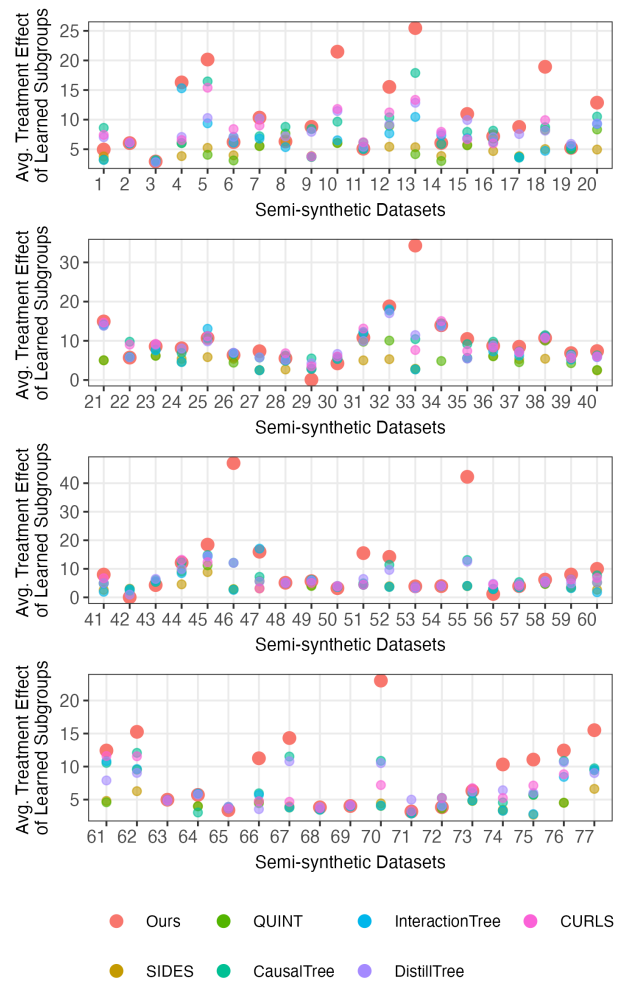


Figure 4: Ground-truth average treatment effect of the learned subgroup (higher is better).

causal model framework. We first showed that any maximizer must exhibit homogeneous pointwise treatment effects, which motivated us to consider the partition-based model. Our main theorem then established that discovering the maximum-effect subgroup reduces to a standard regression/classification problem under a partition-based model, and hence challenges the necessity of the bespoke ‘causal’ heuristics. We instantiated the approach with CART and paired it with the ‘honest’ estimation. We compared against several baselines with both synthetic and semi-synthetic datasets and demonstrated that our method has superior performance in discovering maximum-effect subgroups.

Overall, our results support a simple and general recipe: learn a supervised partition, then estimate effects and select the subgroup. The limitations of this work may be the no-hidden-confounder assumption and the exact-partition assumption. Future research directions might include relaxing these assumptions and exploring other (advanced) algorithms for learning partition-based models.

## References

- Athey, S.; and Imbens, G. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360.
- ATHEY, S.; TIBSHIRANI, J.; and WAGER, S. 2019. GENERALIZED RANDOM FORESTS. *The Annals of Statistics*, 47(2): 1148–1178.
- Athey, S.; and Wager, S. 2019. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2): 37–51.
- Atzmueller, M. 2015. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1): 35–49.
- Breiman, L.; Friedman, J.; Stone, C. J.; and Olshen, R. A. 1984. *Classification and regression trees*. CRC press.
- Brīta, C. E.; van der Linden, J. G.; and Demirović, E. 2025. Optimal Classification Trees for Continuous Feature Data Using Dynamic Programming with Branch-and-Bound. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11131–11139.
- Chen, W.; Pan, W.; Gan, K.; and Wang, F. 2025. MOSIC: Model-Agnostic Optimal Subgroup Identification with Multi-Constraint for Improved Reliability. *arXiv preprint arXiv:2504.20908*.
- Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; and Robins, J. 2018. Double/debiased machine learning for treatment and structural parameters.
- Clark, P.; and Niblett, T. 1989. The CN2 induction algorithm. *Machine learning*, 3(4): 261–283.
- Cohen, W. W. 1995. Fast effective rule induction. In *Machine learning proceedings 1995*, 115–123. Elsevier.
- Curth, A.; and Van der Schaar, M. 2021. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34: 15883–15894.
- Dorie, V.; Hill, J.; Shalit, U.; Scott, M.; and Cervone, D. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1): 43–68.
- Dusseldorp, E.; and Van Mechelen, I. 2014. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine*, 33(2): 219–237.
- Foster, J. C.; Taylor, J. M.; and Ruberg, S. J. 2011. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24): 2867–2880.
- Fürnkranz, J.; Gamberger, D.; and Lavrač, N. 2012. *Foundations of rule learning*. Springer Science & Business Media.
- Goligher, E. C.; Lawler, P. R.; Jensen, T. P.; Talisa, V.; Berry, L. R.; Lorenzi, E.; McVerry, B. J.; Chang, C.-C. H.; Leifer, E.; Bradbury, C.; et al. 2023. Heterogeneous treatment effects of therapeutic-dose heparin in patients hospitalized for COVID-19. *Jama*, 329(13): 1066–1077.
- Hahn, P. R.; Murray, J. S.; and Carvalho, C. M. 2020. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3): 965–1056.
- Hu, X.; Rudin, C.; and Seltzer, M. 2019. Optimal sparse decision trees. *Advances in Neural Information Processing Systems*, 32.
- Huang, M.; Tang, T. M.; and Kenney, A. M. 2025. Distilling heterogeneous treatment effects: Stable subgroup estimation in causal inference. *arXiv preprint arXiv:2502.07275*.
- Künzel, S. R.; Sekhon, J. S.; Bickel, P. J.; and Yu, B. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165.
- Lavrač, N.; Kavšek, B.; Flach, P.; and Todorovski, L. 2004. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5(Feb): 153–188.
- Lee, S.; Liu, R.; Song, W.; Li, L.; and Zhang, P. 2025. SubgroupTE: Advancing Treatment Effect Estimation with Subgroup Identification. *ACM transactions on intelligent systems and technology*, 16(3): 1–23.
- Lipkovich, I.; Dmitrienko, A.; and B D’Agostino Sr, R. 2017. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine*, 36(1): 136–196.
- Lipkovich, I.; Dmitrienko, A.; Denne, J.; and Enas, G. 2011. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine*, 30(21): 2601–2621.
- Loh, W.-Y.; Cao, L.; and Zhou, P. 2019. Subgroup identification for precision medicine: A comparative review of 13 methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(5): e1326.
- Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.
- Nagpal, C.; Wei, D.; Vinzamuri, B.; Shekhar, M.; Berger, S. E.; Das, S.; and Varshney, K. R. 2020. Interpretable subgroup discovery in treatment effect estimation with application to opioid prescribing guidelines. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, 19–29.
- Nie, X.; and Wager, S. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319.
- Pearl, J. 2009. *Causality*. Cambridge university press.
- Pearl, J. 2012. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*.
- Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.
- Rothwell, P. M. 2005. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454): 176–186.

- Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5): 688–701.
- Seibold, H.; Zeileis, A.; and Hothorn, T. 2016. Model-based recursive partitioning for subgroup analyses. *The international journal of biostatistics*, 12(1): 45–63.
- Shi, C.; Blei, D.; and Veitch, V. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.
- Su, X.; Tsai, C.-L.; Wang, H.; Li, B.; et al. 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(2).
- Van Leeuwen, M.; and Knobbe, A. 2012. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, 25(2): 208–242.
- Wang, T.; and Rudin, C. 2022. Causal rule sets for identifying subgroups with enhanced treatment effects. *INFORMS journal on computing*, 34(3): 1626–1643.
- Yang, L.; and van Leeuwen, M. 2022. Truly unordered probabilistic rule sets for multi-class classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 87–103. Springer.
- Yang, L.; and van Leeuwen, M. 2024. Conditional density estimation with histogram trees. *Advances in Neural Information Processing Systems*, 37: 117315–117339.
- Zhang, W.; Le, T. D.; Liu, L.; Zhou, Z.-H.; and Li, J. 2017. Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics*, 33(15): 2372–2378.
- Zhou, J.; Yang, L.; Liu, X.; Gu, X.; Sun, L.; and Chen, W. 2024. CURLS: Causal Rule Learning for Subgroups with Significant Treatment Effect. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4619–4630.