Causal Rule Ensemble: Interpretable Inference of Heterogeneous Treatment Effects

Kwonsang Lee*[1], Falco J. Bargagli-Stoffi*[2], and Francesca Dominici[2]

[1] *Department of Statistics, Sungkyunkwan University*
[2] *Department of Biostatistics, Harvard School of Public Health*

Abstract

In social and health sciences, it is critically important to identify subgroups of the study population where a treatment (or exposure) has a notably larger or smaller causal effect on an outcome compared to the population average. In recent years, there have been many methodological developments for addressing heterogeneity of causal effects. A common approach is to estimate the conditional average treatment effect (CATE) given a pre-specified set of covariates. However, this approach does not allow to discover new subgroups, but only to estimate causal effects on subgroups that have been specified a priori by the researchers. Recent causal machine learning (ML) approaches estimate the CATE at an individual level in presence of large number of observations and covariates with great accuracy. However, because of their complex parametrization of the feature space, these ML approaches do not provide an interpretable characterization of the heterogeneous subgroups. In this paper, we propose a new Causal Rule Ensemble (CRE) method that: 1) discovers de novo subgroups with significantly heterogeneous treatment effects (i.e., causal rules); 2) ensures interpretability of these subgroups because they are defined in terms of decision rules; and 3) estimates the CATE for each of these newly discovered subgroups with small bias and high statistical precision. We provide theoretical results that guarantee consistency of the estimated causal effects for the newly discovered causal rules. A nice feature of CRE is that it is agnostic to the choices of (i) the ML algorithms that can be used to discover the causal rules, and (ii) the estimation methods for the causal effects within the discovered causal rules. Via simulations, we show that the CRE method has competitive performance as compared to existing approaches while providing enhanced interpretability. We also introduce a new sensitivity analysis to unmeasured confounding bias. We apply the CRE method to discover subgroups that are more vulnerable (or resilient) to the causal effects of long-term exposure to air pollution on mortality.

Keywords: causal inference, decision rules, interpretability, machine learning, observational study, sensitivity analysis, stability selection

We are grateful for helpful feedback from organizers and participants at Harvard University, Sungkyunkwan University, St. Gallen University, KU Leuven, IMT School for Advanced Studies seminars, at the Online Causal Inference seminar (OCIS), the Data Science for Impact Evaluation Webinar (DSIE) and the European Causal Inference Meeting (EUROCIM). This work was supported by EPA grant (83587201-0), NIH grants (R01ES026217, R01MD012769, R01ES028033, 1R01AG060232-01A1, 1R01ES030616, 1R01AG066793-01R01, 1R01ES029950) as well as the Alfred P. Sloan Foundation Grant for the development of "Causal Inference with Complex Treatment Regimes: Design, Identification, Estimation, and Heterogeneity" and the VPR - Harvard University grant on "Climate Change Solutions Fund".

*Co-first authors.

1 Introduction

In the last few decades, many studies have investigated the link between exposure to air pollution and adverse health outcomes, also in the context of causal inference (see Carone et al.; 2020, for a review). Despite strong evidence that exposure to air pollution increases the risk of mortality and morbidity (Schwartz et al.; 2021; Wu et al.; 2020; Nethery et al.; 2020), there is an enhanced sense of urgency to discover *de novo* subgroups – i.e., subsets of the feature space characterized by a given covariate-profile (e.g., female individuals, low-income & male individuals, and so on) – that have the largest (or the smallest) causal effects of air pollution on health (Lee et al.; 2021).

In our motivating application, our goal is to identify subgroups of the Medicare population who are most vulnerable or resilient to long-term exposure to fine particulate matter ($PM_{2.5}$) on mortality. To achieve this goal, we acquired and integrated the data of 1,612,414 Medicare beneficiaries in the New England region of the United States between 2000 and 2006. We consider a binary exposure, indicating whether each individual has been exposed to $PM_{2.5}$ greater than 12 $\mu g/m^3$ or not, which is the current national ambient air quality standard (NAAQS) set by the US environmental protection agency (EPA) for long-term exposure. Based on the two-year annual $PM_{2.5}$ during 2000-2001, we assessed the five-year mortality while adjusting for several potential confounders. More details about the study design are illustrated in Section 5. In the entire Medicare population, Di et al. (2017) found that racial minorities and people with low-income are most vulnerable to air pollution exposure by post-hoc subgroup analysis. Recently, Lee et al. (2021) proposed a single tree approach using sample-splitting, and found several disjoint subgroups with heterogeneous causal effects. However, the discovery can vary with different sample-splitting schemes and could be too simple to describe the heterogeneous structure. We aim to discover de novo subgroups and confirm already discovered subgroups by proposing a new methodology.

There is a rich literature on estimating heterogeneous causal effects by examining conditional average treatment effects (CATE) (Imbens and Wooldridge; 2009; Dominici et al.; 2020). The CATE can be specified at different levels of *granularity*. For instance, at the highest level of granularity, one might want to estimate the individual treatment effect (ITE). At a lower level of granularity, one might want to estimate the average treatment effect for some pre-specified *subgroups* of the population. This latter estimand can also be referred to as the group average treatment effect (GATE) (Jacob; 2019). Both the ITE and GATE are special cases of CATE. Throughout this paper, we will simply use the CATE, rather than the GATE, when referring to the estimated effects in the subgroups detected by the proposed algorithm.

Seminal works on estimating the CATE rely on nearest-neighbor matching and kernel methods (Crump et al.; 2008; Lee; 2009). Wager and Athey (2018) discuss that these approaches may fail in handling a large number of covariates. This issue is often referred to as *curse of dimensionality* (Robins and Ritov; 1997). Recently, other nonparametric machine learning (ML) methods such as the random forest (Breiman; 2001) and Bayesian additive regression tree (BART) (Chipman et al.; 2010) have been proposed to estimate heterogeneity in causal effects. These approaches have been particularly successful in the estimation of the ITE when the number of observations and covariates is large. For instance, Foster et al. (2011) and Hill (2011) used forest-based algorithms for the prediction of the missing potential outcomes and, in turn, estimate the ITE for each individual in the study. In more recent contributions, Wager and Athey (2018) and Athey et al. (2019) developed forest-based methods for the estimation of heterogeneous treatment effects. They also provide an asymptotic theory for the conditional treatment effect estimators and valid statistical inference. In a similar spirit, Hahn et al. (2020) proposed a BART-based approach but with a novel parametrization of the outcome surfaces. Most of the approaches for estimating the CATE require to define covariates values (e.g. subgroups) *a priori* – and can be subject to the *cherry-picking* problem of reporting results only for subgroups with extremely high/low treatment effects (Cook et al.; 2004). Furthermore, defining subgroups a priori requires a fairly good understanding of the treatment effects, possibly from previous literature and may fail to identify unexpected, yet important, heterogeneous subgroups.

Despite the success in accurately estimating the ITE using ML methods, these ensemble methods offer little guidance about which subgroups lead to treatment effect heterogeneity. Results from existing ML approaches are generally hard to interpret because parametrizations of the covariate space are utterly complex. This issue is often referred to as *lack of interpretability*. Increasing model interpretability is particularly important in social and health sciences where the ultimate goal is often to inform policies that affect human beings. For instance, the clear understanding of how a policy works is central to build trust in public interventions and to foster accountability.

In this paper, we propose a novel causal rule ensemble (CRE) method that ensures interpretability, while maintaining a high level of estimation accuracy. We ensure interpretability by means of causal (decision) rules (Lakkaraju et al.; 2016). The *causal rules* are basically decision rules, but with causal meanings. In fact, each causal rule depicts a subgroup with a significantly heterogeneous causal effect (Wang and Rudin; 2017). A causal rule, similarly to a decision rule, consists of simple *if-then* statements regarding several conditions and corresponds to a specific subgroup. The causal rules can be obtained, for instance, by partitioning a certain population through binary splits via binary decision trees (e.g., partitioning individuals based on whether they are Medicaid eligible or not). Interpretability is a non-mathematical concept, yet is often defined as the degree to which a human can understand the cause of a decision or consistently predict the results of the model (Kim et al.; 2016; Miller; 2019). The causal rules fit well to this definition in that they resemble human decision-making processes.

The work proposed in this paper is strongly related to recursive binary partitioning to discover the heterogeneous structure of treatment effects. In fact, one could use the honest causal tree (HCT) algorithm (Athey and Imbens; 2016) to identify causal rules as a form of the subgroup discovery. However, finding causal rules from a single HCT has some drawbacks. It can be prone to over-fitting the training data, so the results can fail to be replicable or generalizable (Strobl et al.; 2009). Also, the discovery is limited to a sub-optimal set of discovered subgroups since the recursive partitioning is often based on a greedy algorithm.

To account for these shortcomings, we propose the CRE method that uses multiple trees rather than a single tree. CRE has two main components: (1) interpretable discovery of subgroups with heterogeneous effects via causal rules; (2) estimation of the causal effects for the newly discovered subgroups. The CRE method uses an ensemble of causal rules obtained from multiple trees; then selects a key subset of important causal rules to identify subpopulations contributing to heterogeneous treatment effects; finally estimates the CATE for each selected rule. CRE allows practitioners to represent treatment effect heterogeneity in a more flexible and stable way. The enhanced stability of CRE as compared to single tree methodologies is obtained from exploiting a large number of trees to discover the causal rules.

The reminder of the paper is organized as follows. In Section 2, we introduce the main definitions of the CATE and interpretable causal rules. In Section 3 we introduce the proposed CRE methodology. In Section 4, we conduct simulations studies. In Section 5, we apply the CRE method to the Medicare Data. Section 6 discusses the strengths and weaknesses of our proposed approach and areas of future research.

2 Treatment Effect Heterogeneity and Interpretability

2.1 Heterogeneous Treatment Effects

Suppose there are N subjects. For each subject, let Y_i be an outcome, Z_i be a binary treatment, X_i be a K-dimensional vector of covariates. Let $\boldsymbol{Y}$ and $\boldsymbol{Z}$ be the corresponding N-dimensional vectors and $\mathbf{X}$ be the $N \times K$ matrix that stacks all the individual observations. We assume that (X_i, Z_i, Y_i) are i.i.d. and sampled from some distribution $\mathcal{P}(\cdot)$. Following the potential outcome framework (Rubin; 1974), let $Y_i(1)$ and $Y_i(0)$ be the two potential outcomes for unit i under treatment and the control, respectively. The fundamental problem is that the ITE for i , $\tau_i = Y_i(1) - Y_i(0)$, cannot be observed from a given sample (X_i, Z_i, Y_i) since we can observe only one of the potential outcomes (Holland; 1986). The CATE $\tau(x)$ can be formally defined as:

$$\tau(x) = E\left[Y_i(1) - Y_i(0) | X_i = x\right], \tag{1}$$

where the average treatment effect (ATE) is $\tau = \mathbb{E}_X[\tau(x)]$. Although the CATE cannot be directly observed, it can be identified under a set of assumptions. In the remainder of this paper, we make the standard assumptions of identifiability in casual inference: 1) unconfoundedness; 2) overlap; 3) Stable Unit Treatment Value Assumption (SUTVA). See the supplementary materials for their mathematical definitions.

2.2 Interpretability and Decision Rules

The main goal of this paper is to identify an interpretable structure of $\tau(x)$ in terms of a few important decision rules and estimate $\tau(x)$ for each discovered rule. Although the functional form of $\tau(x)$ is unknown, it is possible to "extract" a parsimonious set of interpretable rules from $\mathbf{X}$ that are the key drivers of the heterogeneity in the causal effects (Imai et al.; 2013; Ertefaie et al.; 2018). To achieve these, we need to consider the *trade-off* between interpretability and accuracy. More specifically, prioritizing high accuracy in the estimation of $\tau(x)$ for a given subgroup generally comes at the cost of compromising interpretability. On the other hand, prioritizing interpretability might lead to rules that are too few and simple and are not representative of the true heterogeneity in $\tau(x)$. Our work focuses on finding the right balance: improving interpretability while minimizing the loss in precision. To do so, we consider decision rules as base learners, and describe the heterogeneous treatment effect as a linear combination of these learners.

We define the mth decision rule (e.g., being female & being eligible to Medicaid) as $r_m(x) = \prod_{k:s_{k,m}\neq S_k} \mathbb{1}(x_k \in s_{k,m})$ where S_k is the set of all possible values of the kth covariate and $s_{k,m} \subseteq S_k$. Hence, the covariate space $\mathcal{D}$ can be defined as a Cartesian product of K sets: $\mathcal{D} = S_1 \times \cdots \times S_K$. A vector of covariates X_i must lie in $\mathcal{D}$ for all i. Also, we define the subset $\mathcal{D}_m$ corresponding to the rule $r_m(x)$, $\mathcal{D}_m = s_{1,m} \times \cdots \times s_{K,m}$. Then, $r_m(x)$ is 1 if $x \in \mathcal{D}_m$ and 0 otherwise.

Base learners that are decision rules can be easily obtained from decision trees. Consider a toy example shown in Figure 1. The illustrated decision tree consists of four decision rules, where each decision rule corresponds to a certain internal or external node (subgroup) in the decision tree. For instance, the young female group can be expressed as $r_4 = \mathbb{1}(\text{Female} = 1) \times \mathbb{1}(\text{Young} = 1)$. Other decision rules are listed in Table 1. In addition to the external nodes, the base learners contain the internal nodes such as r_2 representing the male group. By including the internal and external nodes, the CRE method that we propose can flexibly choose appropriate decision rules. The advantages of using both nodes are discussed in Nalenz and Villani (2018).

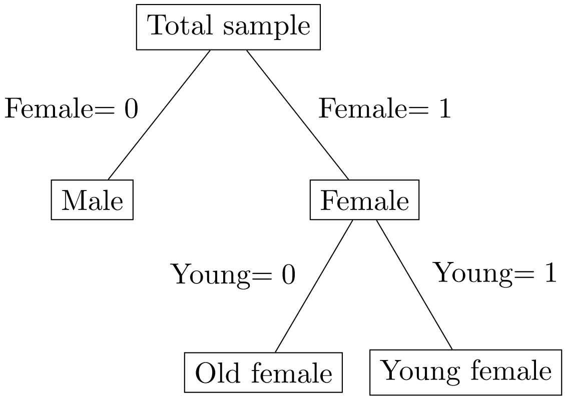

Figure 1: An example tree.

Rule	Condition	Subgroup
r_1	Female= 0	Male
r_2	Female= 1	Female
r_3	Female= 1 & Young= 0	Old female
r_4	Female= 1 & Young= 1	Young female

Table 1: The decisioncorresponding rules and their corresponding subgroups.

3 Methodology

In this section we propose the CRE method for discovering and estimating the treatment effect heterogeneity. Algorithm 1 illustrates the main steps of the proposed methodology. The CRE method is agnostic to the choice of the methods used at the different steps: multiple causal ML methodologies could be employed to perform the causal rule-generation, regularization and estimation.

3.1 Discovery Step: Which Subgroups Are Potentially Important?

We start by using a sample-splitting approach that divides the total sample into two smaller subsamples (Athey and Imbens; 2016; Lee et al.; 2021): (1) discovery subsample ($\mathcal{I}^{dis}$) and (2) inference subsample ($\mathcal{I}^{inf}$). This section provides details on the discovery step which consists of two parts: rule-generation and rule-regularization. For rule-generation, we create base learners that are building blocks to describe the heterogeneous structure of $\tau(x)$. The rule-regularization is used to create only the *necessary* building blocks among the ones previously generated, and is needed to enhance interpretability and stability.

3.1.1 Rule-Generation

We propose a new rule-generation approach in two steps. In the first step, we obtain an estimate of the ITE, $\hat{\tau}_i$, using any suitable method. In the second step, we detect the heterogeneity (and,

Algorithm 1 Overview of the Causal Rule Ensemble (CRE) Method

Inputs: N units i (X_i, Z_i, Y_i), where X_i is the feature vector, Z_i is the treatment indicator, and Y_i is the observed response.
Outputs: : (1) a set of interpretable causal rules, (2) a set CATE estimates for the corresponding causal rules, and (3) sensitivity analysis for the CATE estimates.
Procedure:

1. **Honest Splitting**: randomly split the total sample into two smaller samples: discovery subsample ($\mathcal{I}^{dis}$) and inference subsample ($\mathcal{I}^{inf}$).
2. **The Discovery Step** (performed on $\mathcal{I}^{dis}$):
   (a) Rule-generation (Section 3.1.1):
      (i) Obtain an estimate for the ITE, $\hat{\tau}_i^{dis}$, using one of the existing methods – e.g., Bayesian causal forest, causal forest, inverse propensity weighting, stabilized inverse propensity weighting, and so on.
      (ii) Fit tree ensemble methods such as random forest and gradient boosting on $\hat{\tau}_i^{dis}$ to generate a number of trees representing the heterogeneity in $\hat{\tau}_i^{dis}$. Each of the q distinct binary trees is denoted by $\mathcal{T}_j$, where $\mathcal{T}$ represents the entire tree: its structure, its internal nodes and its terminal nodes (leaves).
      (iii) From each tree $\mathcal{T}_j, j = 1, \ldots, q$, extract decision rules (i.e., the nodes of each tree) $r_m(x), m = 1, \ldots, M^*$. Each decision rule corresponds to an interpretable subgroup with potentially heterogeneous causal effects.
   (b) Rule-regularization (Section 3.1.2):
      (i) Generate $\tilde{\mathbf{X}}^*$ as a new matrix whose columns are the decision rules: since each rule $r_m(x)$ indicates whether x satisfies the rule or not, it can be either 0 or 1.
      (ii) Apply penalized regression (e.g., LASSO, stability selection) to select important decision rules using a linear model of the form:
$$\underset{\alpha_m}{\text{minimize}}\left\{\sum_{i\in\mathcal{I}^{dis}}\left(\hat{\tau}_i^{dis} - \alpha_0 - \sum_{m=1}^{M^*}\alpha_m\tilde{X}_{im}^*\right)^2 + \lambda\sum_{m=1}^{M^*}|\alpha_m|\right\}.$$
      (iii) The selected M (with $M \ll M^*$) rules, say $r_m(x), m = 1, \ldots, M$, are representative of the heterogeneity in the causal effects and we call the rules the **causal rules**.
3. The Inference Step (performed on $\mathcal{I}^{inf}$):
   (a) Causal effect estimation for each causal rule (Section 3.2.1):
      (i) Obtain an estimate for the ITE, $\hat{\tau}_i^{inf}$, using existing methods using existing methods – e.g., Bayesian causal forest, causal forest, inverse propensity weighting, stabilized inverse propensity weighting, and so on.
      (ii) Generate the 0-1 matrix $\tilde{\mathbf{X}}$ corresponding to the causal rules r_m and define the modified linear model using the causal rules:
$$\hat{\tau}_i^{inf} = \beta_0 + \sum_{j=1}^{M}\beta_j\tilde{X}_{ij} + \nu_i = \tilde{\mathbf{X}}_i\boldsymbol{\beta} + \nu_i$$
      where $\tilde{\mathbf{X}}_i$ is the ith row of $\tilde{\mathbf{X}}$.
      (iii) Consider the OLS estimator for $\boldsymbol{\beta}$: $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\hat{\boldsymbol{\tau}}^{inf}$ to estimate the CATE for each causal rule. Note that if the method chosen to estimate ITE in (3a) provides consistent estimators of $\hat{\tau}_i^{inf}$, then $\hat{\boldsymbol{\beta}}$ is consistent (Theorem 1) and asymptotically normal (Theorem 2).
   (b) Sensitivity analysis of unmeasured confounding for the CATE estimates from the previous step (Section 3.2.2): See Algorithm 2 for a detailed description of the algorithm that we developed for sensitivity analysis.

in turn, the causal rules) in the ITE by fitting tree-ensemble methodologies on the estimated ITE. As shown in the supplementary materials, this two-step procedure for rule-generation enhances the stability and precision in the detection of causal rules as compared to single step approaches (e.g., direct rule-discovery via causal trees).

The estimation of the ITE in this first step of rule-generation is not used for making inference, but is designed for exploring the heterogeneous treatment effect structure only. Many different approaches for the ITE estimation can be considered, such as inverse probability weighting (Robins and Ritov; 1997), BART (Hill; 2011), causal forest (Wager and Athey; 2018), or Bayesian causal forest (BCF) (Hahn et al.; 2020). In the supplementary materials, we show with simulations that the BCF performs well when applying the CRE method as compared to alternative approaches.

Once the ITE estimate on the discovery sample – say $\hat{\tau}_i^{dis}$ – is obtained, we implement the *fit-the-fit* approach to discovery of causal rules. More specifically, we fit tree ensemble methods such as random forest (Breiman; 2001) and gradient boosting (Friedman; 2001) to discover the decision rules such as $\hat{\tau}_i^{dis} \approx \mathcal{T}_1(\mathbf{X}_i^{dis}) + ... + \mathcal{T}_q(\mathbf{X}_i^{dis})$ where each of the q distinct binary trees is denoted by $\mathcal{T}_j$ and $\mathcal{T}$ represents the entire tree: its structure, its internal and terminal nodes (leaves). The maximal complexity of these trees (i.e., their depth) can be set by the researchers to allow for the discovery of simpler (i.e., less lengthy) rules. We follow Friedman and Popescu (2008) and Nalenz and Villani (2018) for the settings of the tuning parameters for gradient boosting and random forest. From each tree $\mathcal{T}_j, j = 1, \ldots, q$, we extract decision rules. Then, we generate a pooled set of the decision rules containing $r_m(x), m = 1, \ldots, M^*$. Since each rule $r_m(x)$ indicates whether x satisfies the rule or not, it can be either 0 or 1. Define $\tilde{\mathbf{X}}^*$ as a new matrix whose columns are the decision rules. The number of rules, M^*, is usually larger than K and depends on how heterogeneous τ_i^{dis} is. Although the original data set is not high-dimensional, $\tilde{\mathbf{X}}^*$ can be high-dimensional. The set of $r_m(x)$ is a set of potentially important rules. It can contain actually important decision rules that describe the true heterogeneity in the treatment effects. However, insignificant rules may be contained in the set as well. Therefore, regularization is needed to find such important decision rules.

3.1.2 Rule-Regularization, Stability Selection and Causal Rules

A step of rule-regularization can improve interpretability by removing redundant rules while setting aside necessary decision rules. We consider the following linear regression model of the form,

$$\hat{\tau}_i^{dis} = \alpha_0 + \sum_{m=1}^{M^*} \alpha_m \tilde{X}_{im}^* + \epsilon_i \tag{2}$$

Since a linear model is considered, the model (2) lends a familiar interpretation of the coefficients $\{\alpha_m\}_0^{M^*}$. Using this linear model, one can employ the following penalized regression to select important rules:

$$\underset{\alpha_m}{\text{minimize}} \left\{ \sum_{i \in \mathcal{I}^{dis}} \left(\hat{\tau}_i^{dis} - \alpha_0 - \sum_{m=1}^{M^*} \alpha_m \tilde{X}_{im}^* \right)^2 + \lambda \sum_{m=1}^{M^*} |\alpha_m| \right\} \tag{3}$$

where λ is the regularization parameter. However, variable selection has been known as a notoriously difficult problem.

The least absolute shrinkage and selection operator (LASSO) estimator (Tibshirani; 1996) has been popular and widely used over the past two decades in order to solve the problem in (3). The usefulness of this estimator among other penalization regression methods is demonstrated in various applications (Su et al.; 2016; Belloni et al.; 2016; Chernozhukov et al.; 2016, 2017). When the data is high-dimensional, selecting λ can be challenging. A stability selection approach proposed by Meinshausen and Bühlmann (2010) can be considered. The stability selection can improve the performance of a rule-selection algorithm using subsamples (see supplementary materials for details). Among initially generated M^* decision rules, we assume that M (with $M \ll M^*$) decision rules are selected as an output from the stability selection procedure. We call these selected rules the *causal rules*. Using these rules, we can define $\tilde{\mathbf{X}}$, which is a sub-matrix of $\tilde{\mathbf{X}}^*$, as the matrix containing only the selected causal rules.

Figure 2 depicts the intuition behind these steps of rule-generation and selection. This figure shows a simple forest composed of just five trees. Each tree represent the heterogeneity in the causal effects and can be though as coming from the fit-the-fit procedure illustrated in the previous

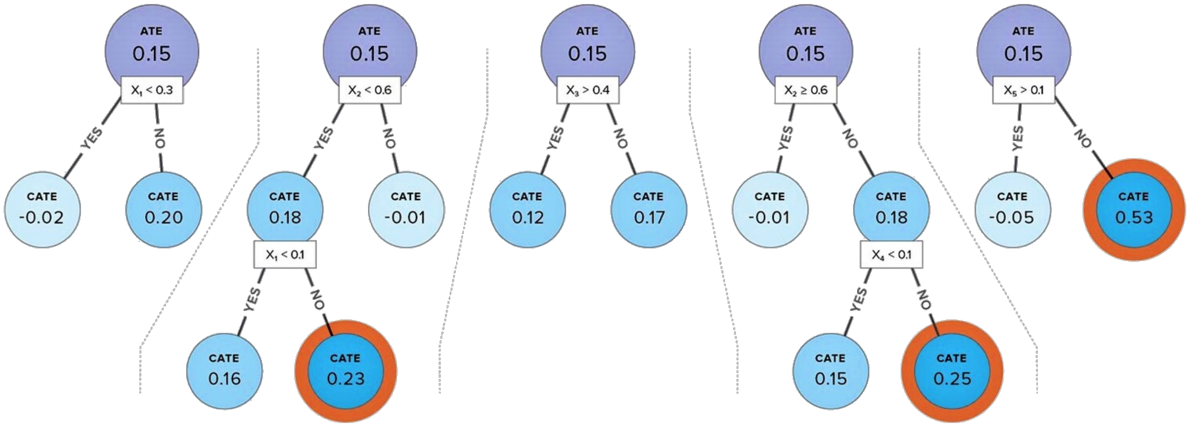

Figure 2: Rule generation and selection in a simple forest. The five trees represent the heterogeneity in the causal effect that can be obtained from the fit-the-fit procedure. Each node of each tree represents a decision rule and the bluer the shade of the node, the larger the causal effect for the associated subgroup. The nodes highlighted in red represent the causal rules that are selected by the stability selection methodology.

Section. Each node of each tree (excluding the initial nodes) represents a decision rule constructed from the *rule-generation* step. The bluer the shade of the node, the larger the causal effect for the decision rule. The nodes highlighted in red represent the causal rules that are selected by the stability selection methodology (*rule-selection*). These causal rules are the ones that are most informative on the heterogeneity of the causal effects as they represent the subgroups with the larger effect variation.

3.2 Inference Step: Which Subgroups Are Really Different?

We propose a general approach to estimate the rule-specific treatment effect. To estimate the CATE for the causal rules obtained from the discovery step, we use the three identification assumptions (unconfoundedness, overlap and SUTVA). We want to emphasize here that we can use any approach for estimating the CATE. However, among the approaches, when either the inverse probability weighting (IPW) or the stabilized IPW (SIPW) estimator is used, we are able to assess the impact of unmeasured confounding bias on the causal rule-specific effects. This particular situation is discussed and a new sensitivity analysis method is introduced in Section 3.2.2.

3.2.1 Estimating the subgroup-specific treatment effect

Once we have discovered the set of the causal rules from the discovery sample, the remaining inference sample ($\mathcal{I}^{inf}$) can be used to make inference. Suppose there are N^{inf} units in the inference sample. Define a new vector $\hat{\boldsymbol{\tau}}^{inf} = (\hat{\tau}_1^{inf}, \ldots, \hat{\tau}_{N^{inf}}^{inf})^T$ that is an estimate of $\boldsymbol{\tau}$ on the inference sample. Again, we highlight that any estimation method for the ITE could be employed. Since $\hat{\tau}_i^{dis}$ and $\hat{\tau}_i^{inf}$ are estimated on independent samples and in separate steps, one can estimate them using different statistical methodologies. From simulation studies in Section 4, we find evidence that estimation of both $\hat{\tau}_i^{dis}$ and $\hat{\tau}_i^{inf}$ through BCF produces the best comparative performance.

The estimated value $\hat{\tau}_i^{inf}$ can be represented as:

$$\hat{\tau}_i^{inf} = \tau_i^{inf} + u_i \quad \text{where} \quad \mathbb{E}(u_i|X_i) = 0 \text{ and } \text{var}(u_i|X_i) = w_i. \tag{4}$$

By reworking the model (2), the modified linear model using the causal rules is obtained as:

$$\hat{\tau}_i^{inf} = \beta_0 + \sum_{j=1}^{M} \beta_j \tilde{X}_{ij} + \nu_i = \tilde{\mathbf{X}}_i \boldsymbol{\beta} + \nu_i, \tag{5}$$

where $\tilde{\mathbf{X}}_i$ is the ith row of $\tilde{\mathbf{X}}$. We consider the OLS estimator for $\boldsymbol{\beta}$ that can be defined as:

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \hat{\boldsymbol{\tau}}^{inf}. \tag{6}$$

We show that the proposed OLS estimator in (6) is consistent (Theorem 1) and asymptotically normal (Theorem 2) in the supplementary materials. The variance of $\hat{\boldsymbol{\beta}}$ may be of interest. One can estimate it through the well-known sandwich formula. We refer to the supplementary materials for more details on this estimator.

3.2.2 Sensitivity Analysis

The estimator $\hat{\boldsymbol{\beta}}$ compactly represents the treatment effect heterogeneity. The validity and consistency of $\hat{\boldsymbol{\beta}}$ rely on the assumption of no unmeasured confounders and correct specification of the propensity score model. However, in practice, we do not know whether a considered set X_i

is sufficient for the unconfounded assumption to hold. When there exists a source of unmeasured confounding, this assumption is violated, and the identification results do not hold. Sensitivity analysis can be a useful tool to investigate the impact of unmeasured confounding bias.

We propose a new sensitivity analysis method that can assess the robustness of our conclusion. This method does not attempt to quantify the degree of unmeasured confounding bias in a given data set. Instead, it is more realistic to see how our causal conclusion will change with respect to various degrees of such bias. As mentioned above, we consider the special case where $\hat{\tau}_i^{inf}$ is the SIPW estimate $\hat{\tau}_i^{SIPW}$ from Hirano et al. (2003).

Define $\mathbf{W} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$ that is a $M \times N^{inf}$ matrix. Also, let W_j be the jth row of $\mathbf{W}$ and W_{ji} be the (j,i) element of $\mathbf{W}$. Our estimator $\hat{\beta}_j$ is explicitly represented by:

$$\begin{aligned}
\hat{\beta}_j &= \hat{\beta}_j(1) - \hat{\beta}_j(0) \quad \text{where} \\
\hat{\beta}_j(1) &= \left[\frac{1}{N}\sum_{i=1}^{N}\frac{Z_i}{\hat{e}(X_i)}\right]^{-1}\left[\sum_{i=1}^{N}\frac{W_{ji}Y_iZ_i}{\hat{e}(X_i)}\right] \\
\hat{\beta}_j(0) &= \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1-Z_i}{1-\hat{e}(X_i)}\right]^{-1}\left[\sum_{i=1}^{N}\frac{W_{ji}Y_i(1-Z_i)}{1-\hat{e}(X_i)}\right].
\end{aligned}$$

We consider the marginal sensitivity model that was introduced by Tan (2006) and Zhao et al. (2019). Let the true propensity probability $e_0(\mathbf{x}, y; z) = \mathrm{P}_0(Z = 1|\mathbf{X} = \mathbf{x}, Y(z) = y)$ for $z \in \{0, 1\}$. If the assumption of no unmeasured confounders holds, this probability would be the same as $e_0(\mathbf{x}) = \mathrm{P}_0(Z = 1|\mathbf{X} = \mathbf{x})$ (i.e., $e_0(\mathbf{x}, y; z) = e_0(\mathbf{x})$) that is identifiable from the data. Unfortunately, this assumption cannot be tested since $e_0(\mathbf{x}, y; z)$ is generally not identifiable from the data. In addition to this non-identifiability of $e_0(\mathbf{x}, y; z)$, there is another difficulty in obtaining $e_0(\mathbf{x})$ non-parametrically when $\mathbf{X}$ is high-dimensional. In practice, $e_0(\mathbf{x})$ is estimated by a parametric logistic model in the form of $e_\gamma(\mathbf{x}) = \exp(\gamma'\mathbf{x})/\{1 + \exp(\gamma'\mathbf{x})\}$ where $e_{\gamma_0}(\mathbf{x})$ can be considered as the best parametric approximation of $e_0(\mathbf{x})$, and used for sensitivity analysis.

Our sensitivity model uses the sensitivity parameter $\Lambda \geq 1$ that restricts the maximum deviation of $e_0(\mathbf{x}, y; z)$ from the identifiable quantity $e_{\gamma_0}(\mathbf{x}) = \mathrm{P}_{\gamma_0}(Z = 1|\mathbf{X} = \mathbf{x})$. Sensitivity analysis is conducted for each Λ to see if there is any qualitative change of our conclusion. More formally, define the set $\mathcal{E}_{\gamma_0}(\Lambda)$and assume $e_0(\mathbf{x}) \in \mathcal{E}_{\gamma_0}(\Lambda)$,

$$e_0(\mathbf{x}, y; z) \in \mathcal{E}_{\gamma_0}(\Lambda) = \left\{ 0 < e(\mathbf{x}, y; z) < 1 : \frac{1}{\Lambda} \leq \frac{(1 - e(\mathbf{x}, y; z)) \cdot e_{\gamma_0}(\mathbf{x})}{e(\mathbf{x}, y; z) \cdot (1 - e_{\gamma_0}(\mathbf{x}))} \leq \Lambda \right\} \quad \text{for } z \in \{0, 1\}. \tag{7}$$

The deviation of $e_0(\mathbf{x}, y; z)$ is symmetric with respect to the parametrically identifiable quantity $e_{\gamma_0}(\mathbf{x})$, and the degree of the deviation is governed by the sensitivity parameter Λ. When $\Lambda = 1$, $e_0(\mathbf{x}, y; z) = e_{\gamma_0}(\mathbf{x})$ for all z, which implies no violations of the following assumptions; the correctly specified propensity score model assumption (that is, $e_{\gamma_0}(\mathbf{x}) = e_0(\mathbf{x})$) and no unmeasured confounder assumption (that is, $e_0(\mathbf{x}, y; z) = e_0(\mathbf{x})$). For $\Lambda > 1$, the proposed sensitivity model considers violations of both assumptions. This model resembles the model proposed by Rosenbaum (2002). The connection between the two models is illustrated in Section 7.1 in Zhao et al. (2019).

If we denote $W_{ji}Y_i$ by $\tilde{Y}_i^j$ and treat $\tilde{Y}_i^j$ as if it is an observed outcome, the $100(1-\alpha)\%$ confidence interval of β_j can be constructed by using the percentile bootstrap through the procedure in Algorithm 2. Confidence intervals have at least $100(1-\alpha)$ % coverage probability even in the presence of unmeasured confounding. The validity of the percentile bootstrap confidence interval $[L_j, U_j]$ is proved and an efficient way to find this interval is discussed in Zhao et al. (2019).

4 Simulations

In this section, we introduce two simulation studies to assess the performance of the CRE method. First, we assess the performance of CRE in discovering the true causal rules during the discovery step. Second, we assess the overall performance of both the discovery and inference steps in terms of estimation accuracy of the rule-specific causal effects.

4.1 Simulation study: Rule-Discovery

We start by evaluating the performance of CRE by assessing how many times CRE can identify the true underlying causal rules. Then, we compare the performance of CRE with respect to the honest causal tree (HCT) method (Athey and Imbens; 2016). The HCT method is a causal decision tree algorithm and discovers disjoint decision rules through a single tree. HCT has been extensively

Algorithm 2 Constructing the confidence interval of β_j for each sensitivity parameter Λ

```
Inputs: The inference subsample of $N^{inf}$ units with $(X_i, Z_i, Y_i)$, $i \in \mathcal{I}^{inf}$ and the causal rules $r_m(x), m = 1, \ldots, M$ obtained from the discovery step
Outputs: The estimates of the causal rule-specific treatment effect $\beta_j, j = 1, \ldots, M$
Procedure:
  1. Generate a 0-1 matrix $\tilde{\mathbf{X}}$ and a matrix $\mathbf{W} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$.
  2. In the $\ell$th of $L$ iterations:
      (a) Generate a bootstrapped sample: $(Z_i^{(\ell)}, Y_i^{(\ell)}, X_i^{(\ell)})_{i=1,\ldots,N^{inf}}$.
      (b) Generate transformed outcomes $\tilde{Y}_{ji}^{(\ell)} = W_{ji}Y_i^{(\ell)}$ for all $i$, where $W_{ji}$ is the $(j, i)$ element of $\mathbf{W}$.
      (c) Reorder the index such that the first $N_1 = \sum_{i=1}^{N} Z_i^{(\ell)}$ units are treated with $\tilde{Y}_{j1} \geq \ldots \geq \tilde{Y}_{j,N_1}$ and the rest are control with $\tilde{Y}_{j,N_1+1} \geq \ldots \geq \tilde{Y}_{j,N^{inf}}$
      (d) Compute an estimate $\hat{\gamma}^{(\ell)}$ by fitting the logistic regression with $(Z_i^{(\ell)}, X_i^{(\ell)})$
      (e) Solve the following optimization problems:
            $\text{min or max} \frac{\sum_{i=1}^{N_1} \tilde{Y}_{ji}^{(\ell)}[1 + q_i \exp\{-\hat{\gamma}^{(\ell)} X_i^{(\ell)}\}]}{\sum_{i=1}^{N_1}[1 + q_i \exp\{-\hat{\gamma}^{(\ell)} X_i^{(\ell)}\}]} - \frac{\sum_{i=N_1+1}^{N} \tilde{Y}_{ji}^{(\ell)}[1 + q_i \exp\{\hat{\gamma}^{(\ell)} X_i^{(\ell)}\}]}{\sum_{i=N_1+1}^{N}[1 + q_i \exp\{\hat{\gamma}^{(\ell)} X_i^{(\ell)}\}]}$
          subject to $1/\Lambda \leq q_i \leq \Lambda$, for $1 \leq i \leq N$, and denote the minimum as $L_j^{(\ell)}$ and the maximum as $U_j^{(\ell)}$
  3. Construct the $(1 - \alpha)$-coverage confidence interval $[L_j, U_j]$ where $L_j = Q_{\alpha/2}\left(L_j^{(\ell)}\right)$ and $U_j = Q_{1-\alpha/2}\left(U_j^{(\ell)}\right), \ell = 1, \ldots, L$
```

employed in the literature and it was the building block of other methodologies for interpretable discovery of causal rules such as the one proposed by Lee et al. (2021).

As we discussed in Section 3.1, to apply CRE, many approaches can be considered to estimate τ_i. We have found via a separate simulation study, discussed in the supplementary materials, that the BCF approach has better performance than other approaches such as IPW and BART-based outcome regressions, and slightly outperforms BART. Thus, in this simulation study, we use BCF to estimate τ_i. We call this version of the CRE method *CRE-BCF*. For the data-generating process, we generate the covariate matrix $\mathbf{X}$ with 10 binary covariates from X_{i1} to $X_{i,10}$. The binary treatment indicator Z_i is drawn from a binomial distribution, $Z_i \sim \textit{Binom}(\pi_i)$ where $\pi_i = \text{logit}(-1 + X_{i1} - X_{i2} + X_{i3})$. We consider three factors: (i) the number of decision rules, (ii) the effect size k from 0 to 2, and (iii) sample size $N = 1000$ or 2000.

To produce a more meaningful comparison between CRE-BCF and HCT, we restricted the form of causal rules that can be obtained through a binary tree. In particular, the potential outcomes are generated by $Y_i(0) \sim N(X_{i1} + 0.5X_{i2} + X_{i3}, 1)$ and $Y_i(1) = Y_i(0) + \tau(X_i)$. The observed outcome is $Y_i = Y_i(0)(1 - Z_i) + Y_i(1)Z_i$. For the case of two causal rules, $\tau = \tau(X_i) = k$ if $X_{i1} = 0, X_{i2} = 0$, $\tau = -k$ if $X_{i1} = 1, X_{i2} = 1$, and $\tau = 0$ otherwise. While in the case of four causal rules we have that $\tau = k$ if $(X_{i1}, X_{i2}, X_{i3}) = (0, 0, 1)$, $\tau = 2k$ if $(X_{i1}, X_{i2}, X_{i3}) = (0, 0, 0)$, $\tau = -k$ if $(X_{i1}, X_{i2}, X_{i3}) = (0, 1, 0)$, $\tau = -2k$ if $(X_{i1}, X_{i2}, X_{i3}) = (0, 1, 1)$ and $\tau = 0$ otherwise. This scenario was chosen because it is the most favourable to the HCT algorithm as these causal rules can be discovered by a single tree. Figure 3 shows the results when (X_{i1}, X_{i2}, X_{i3}) are both the confounders and effect modifiers. Moreover, we introduce variations in the set of covariates that are used to define the causal rules (we refer to these variables as effect modifiers) by switching (X_1, X_2, X_3) with (X_8, X_9, X_{10}). This change represents the case where effect modifiers are different from confounders. Investigating this scenario is important since confounders may affect the ability of the algorithm to spot the correct causal rules. Figure 4 shows the results when a set of the confounders is different that of the effect modifiers.

We found that CRE-BCF outperforms HCT in the detection of the true causal rules, especially when the number of rules is larger than 2 and when the the same variables are used as confounders and effect modifiers. Both these scenarios are expected in real-world applications. Also, it is important to note that CRE provides a smaller number of detected rules (4 to 7) as compared to HCT (10 to 84).

4.2 Simulation study: Rule-specific Effect Estimation

In this subsection, we evaluate the overall performance of the CRE method including both the discovery and inference steps. In this simulation study, we use the BCF approach for both discovery and inference steps when applying the CRE method. In particular, in the discovery step, $(X_i, \hat{\tau}_i^{BCF})$ is used to discover and select causal rules. During the inference step, $\hat{\tau}_i^{inf}$ is estimated by using the BCF approach (CRE-BCF).

We also consider a BCF approach that does not use the sample-splitting technique, and we call this *original-BCF*. The original BCF provides the estimate $\hat{\tau}_i$, but does not provide an interpretable form of the heterogeneous treatment effects. On the contrary, the CRE-BCF not only provides a

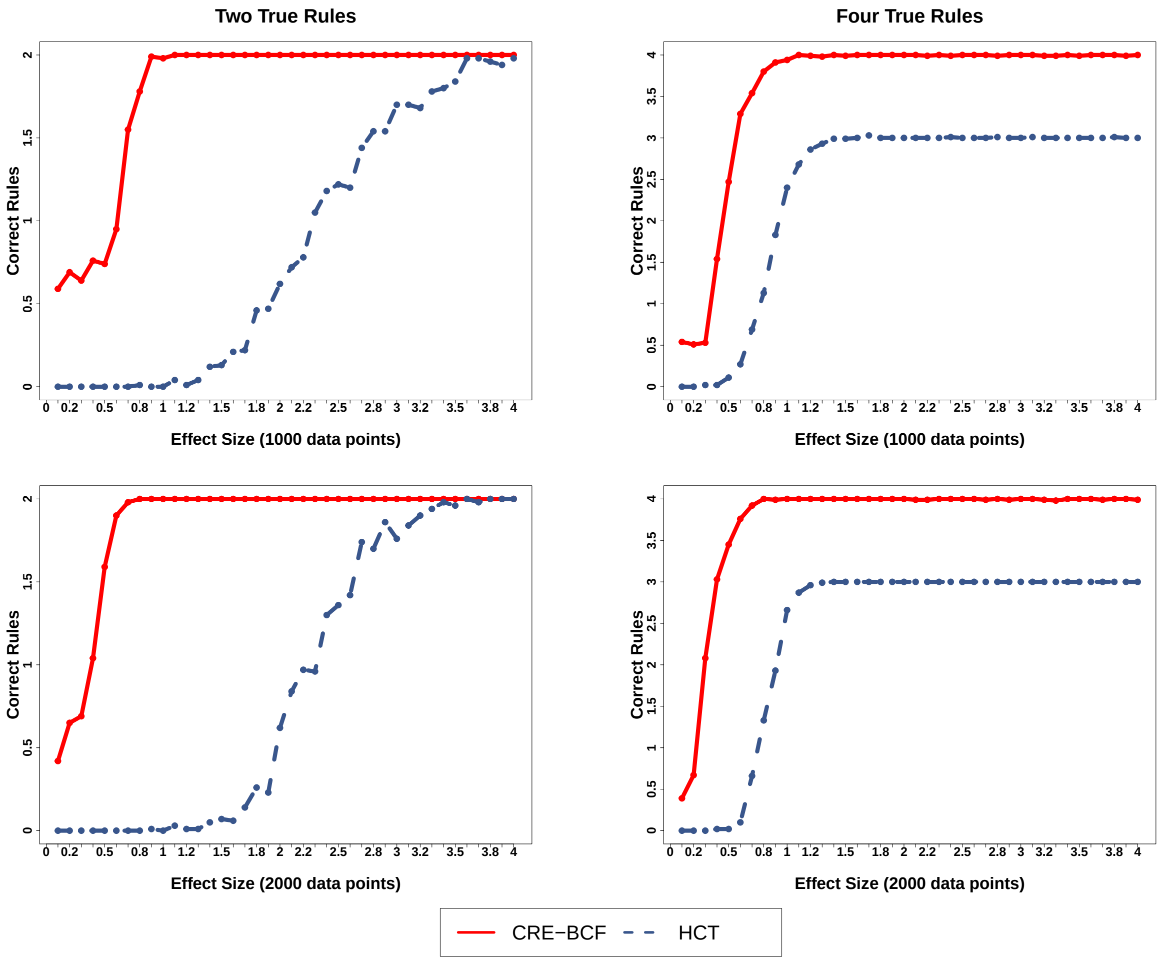

Figure 3: Average number of correctly discovered rules in the case where the same variables are used as confounders and effect modifiers. The first column depicts the case of two true rules while the second column the case of four true rules. In the first row the sample size is 1,000 while in the second it is 2,000.

set of causal rules that significantly increase interpretability of findings, but also provides the estimate with respect to the discovered structure. It is difficult to compare the two methods in terms of interpretability, so in this simulation study, we compare them in terms of estimation accuracy. Athey and Imbens (2016) recommends to use a (50%, 50%) ratio between the discovery and inference samples for sample-splitting, but Lee et al. (2021) shows through simulation studies that (25%, 75%) has better performance than (50%, 50%). We investigate six different ratios from (10%, 90%) to (50%, 50%). Note that the original-BCF can be considered as an extreme case of

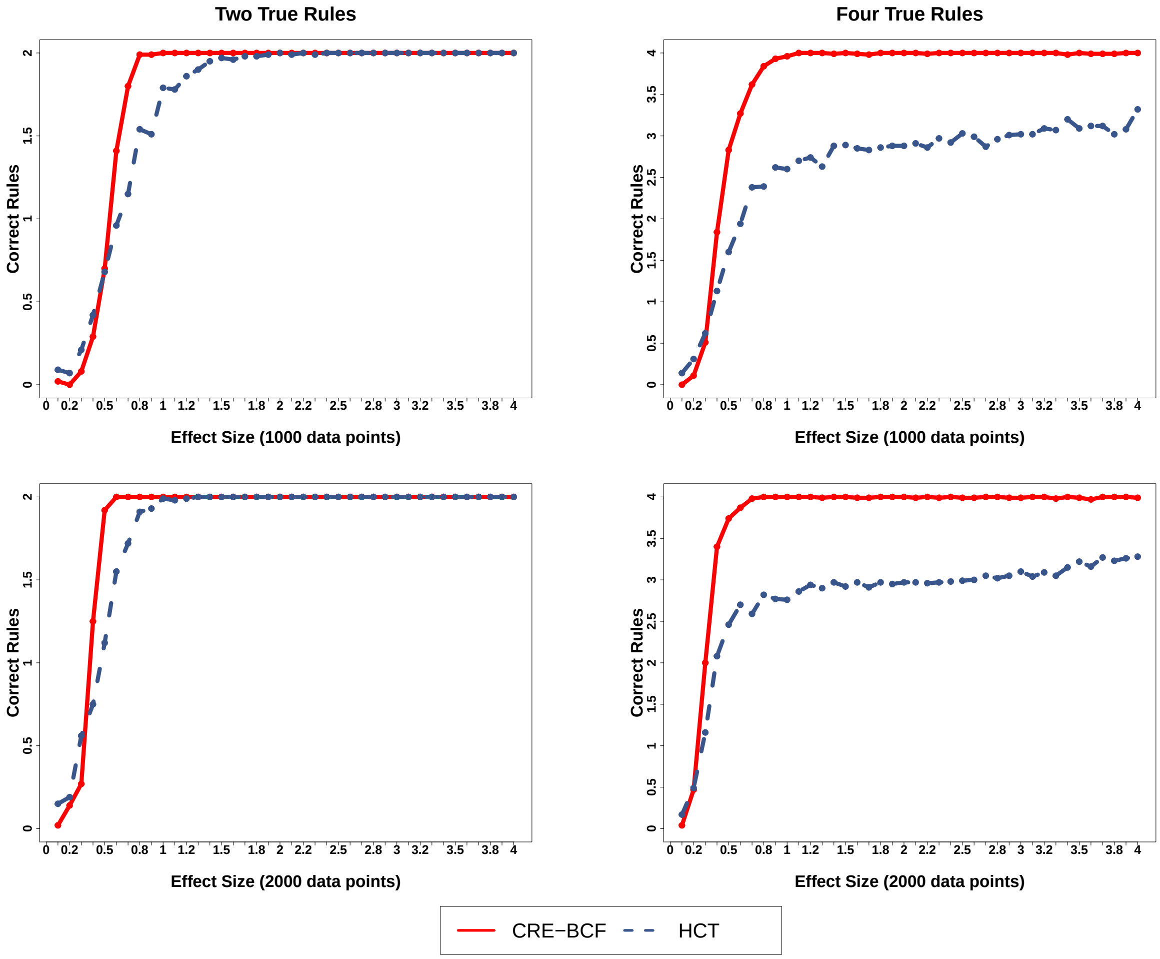

Figure 4: Average number of correctly discovered rules in the case where the confounders are different from the effect modifiers. The first column depicts the case of two true rules while the second column the case of four true rules. In the first row the sample size is 1,000 while in the second it is 2,000.

(0%, 100%).

For the data generating process, covariates, treatment, and potential outcomes are generated in the same way as in the previous simulation study. In this simulation, the only difference is that we assume that there are two true underlying decision rules: (1) $X_{i1} = 0, X_{i2} = 0$ and (2) $X_{i1} = 1, X_{i2} = 1$. The treatment effect τ_i is defined as $\tau_i = 1$ if $X_{i1} = 0, X_{i2} = 0$, $\tau_i = -1$ if $X_{i1} = 1, X_{i2} = 1$, and $\tau_i = 0$ otherwise. We consider four samples sizes: $N = 500, 1000, 1500$ and 2000. We consider the root mean squared error (RMSE) to compare the two methods. Table 2

Table 2: RMSE comparison between the CRE and BCF methods

	Method						
N	CRE50	CRE40	CRE30	CRE25	CRE20	CRE10	BCF
500	0.568	0.520	0.486	**0.478**	0.498	0.598	0.391
1000	0.399	0.366	0.347	**0.343**	0.344	0.406	0.355
1500	0.326	0.299	0.279	**0.276**	0.278	0.320	0.309
2000	0.283	0.258	0.239	**0.231**	0.234	0.262	0.237

shows the performance comparison using RMSE. We consider 1000 simulated data sets for each sample size and provide the average of 1000 RMSE values. Among the considered ratios, (25%, 75%) provides the least RMSE for every sample size. The RMSE value decreases as the proportion of the discovery sample increases up to 25%, and it starts to increase after 25%. When the sample size is small (i.e., $N = 500$), the RMSE for CRE25 is higher than that for BCF, however, it is lower when N is moderately large. This simulation result shows that even though, for instance, the CRE25 method uses 75% of the total sample for inference, it is as efficient as the original BCF that uses 100% of the total sample for inference.

5 Application to the Medicare data

We apply the proposed CRE method to the Medicare data to study the heterogeneous effects of long-term exposure to $\text{PM}_{2.5}$ on 5-year mortality. We consider the same data of 110,911 matched pairs that are used in Lee et al. (2021). The population consists of Medicare beneficiaries in New England regions in the United States between 2000 and 2006. The treatment is whether the two-year (2000-2001) average of exposure to $\text{PM}_{2.5}$ is greater than 12 $\mu g/m^3$. The outcome is five-year mortality measured between 2002-2006. For an individual, the outcome is 1 if he/she died before the end of 2006 and 0 otherwise. There are four individual level covariates: sex (female, male), age (65-70, 71-75, 76-80, 81-85, 86+), race (white, non-white), and Medicaid eligibility (eligible, non-eligible). Medicaid eligibility is considered as a proxy for socioeconomic status: if an individual is eligible for Medicaid, it is highly likely that he/she has lower household income. In the matched data, these four variables are exactly matched. There are also 8 ZIP code-level or 2 county-level covariates (body mass index, smoker rate, Hispanic population rate, Black population rate, median household income, median value of housing, % of people below the poverty level, % of people below high school education, % of owner occupied housing and population density), and they are balanced between the treated and control groups, see Lee et al. (2021) for more detailed discussion about the covariates and their balance. The CRE method using a (25%, 75%) discovery/inference ratio is considered, which leads to 27,500 and 82,591 matched pairs respectively. We ignore the matched structure and use $27,500 \times 2 = 55,000$ individuals in the discovery sample as if they were obtained independently. Since the outcome variable is binary, we use the logistic BART approach to estimate $\hat{\tau}_i^{dis}$ and discover the causal rules. Note that for a continuous outcome, as shown in the simulation study, the BCF approach can be considered. Also, note that both BCF and BART have shown the best comparative performance in the rule discovery step (see supplementary materials).

In the discovery step with only four individual-level covariates, the CRE method discovers four causal rules, r_1, r_2, r_3, r_4, shown in Table 3. Compared to the single-tree discovery from Lee et al. (2021) all the causal rules from CRE represent subgroups with heterogeneous effects while the single-tree essentially contains uninformative subgroups. The baseline subgroup that does not correspond to any causal rule $r_j, j = 1, 2, 3, 4$ represents the combined subgroup of (1) Medicaid eligible white people aged between 81-85 and (2) white people aged above 85. This subgroup is similar to the single-tree finding with the highest treatment effects. From this baseline subgroup, each of the four causal rules indicates a certain subgroup. Another distinctive feature of the CRE method compared to the single-tree method is that the causal rules can overlap and their treatment effects can be linearly added up. Therefore, the four causal rules compactly summarizes the heterogeneous effect structure.

Furthermore, we extend the CRE method to the four individual-level and eight ZIP code-level covariates. The ZIP code-level covariates could not be considered in Lee et al. (2021) because pairs were matched exactly only on individual-level covariates. When applying the CRE method, five rules are discovered including the three previous rules, r_2, r_3, r_4 and two additional rules, r_5, r_6. The additional rules are defined by Hispanic population (10% above or not), Education (did not complete high school 30% above or not), and Population density (above the average or not). The rule r_5 identifies a subgroup of people living in areas where the proportion of hispanic is below

10% and the proportion of people who did not complete high school is above 30%. The rule r_6 is included in r_5, and identifies a subgroup of r_5 with the additional condition that the population density is below the average.

Next, once the causal rules are discovered, we use the remaining 82,591 pairs in the inference sample to estimate the rule-specific treatment effects. The rules r_j induce the model $\tau(x) = \beta_0 + \sum_j \beta_j r_j(x)$. For the first rule set $\{r_1, \ldots, r_4\}$, the corresponding coefficients are estimated and reported in Table 3. For the inference step, we consider $\hat{\tau}_i^{inf} = \hat{\tau}_i^{SIPW}$ to claim the asymptotic normality, obtain the 95% confidence intervals using the asymptotic distributions, and perform sensitivity analysis on the estimated CATE. On the first part of the right column of Table 3, the intercept β_0 that is the coefficient of the baseline group shows the positive value of the treatment effect 0.07. This implies that there is a seven percentage point increase of the 5-year mortality rate. However, other causal rules have negative β_j, which shows a decrease in the mortality rate if a subject belongs to one of the four subgroups. For instance, consider a 70 years old white male who is not Medicaid eligible. He corresponds to r_2, r_3, r_4, but not r_1. Therefore, the estimated treatment effect for him is $\hat{\beta}_0 - \hat{\beta}_2 - \hat{\beta}_3 - \hat{\beta}_4 = 0.070 - 0.012 - 0.027 - 0.033 = -0.002$. Note that a subgroup represented by a linear combination of the causal rules can have the null effect. Though the estimates for r_1 and r_2 are negative, they are not statistically significant at a significance level $\alpha = 0.05$. However, the estimates for r_3, r_4 are significant, which means that people below 80 (i.e., r_3) and people below 85 not being eligible for Medicaid (i.e., r_4) are significantly less vulnerable than the baseline subgroup. When including ZIP code-level covariates, all the estimates for the discovered rules $\{r_2, r_3, r_4, r_5, r_6\}$ are negative. Similarly, r_2 is not significantly different. For the newly discovered $\{r_5, r_6\}$, only r_6 is statistically significant.

Finally, to evaluate the robustness of the above finding about the treatment effect heterogeneity, we conduct sensitivity analysis. Under the sensitivity model in (7), we obtain several sets of 95% CIs for the coefficients for each Λ. Table 4 shows the 95% CIs for Λ from 1.01 to 1.05. As Λ increases all the CIs get wider. When $\Lambda = 1.04$, all the coefficients contain zero and there is no evidence for the heterogeneity. In particular, $\Lambda = 1.04$ means that if there is an unmeasured confounder that can make the estimated propensity score deviated from the true score by 1.04 in terms of the odds

Table 3: Discovering decision rules and estimating the coefficients for the decision rules

	Rules	Covariates			
		Individual		Indiv. + ZIP code	
#	Description	Est.	95% CI	Est.	95% CI
	Intercept	0.070	(0.054, 0.087)	0.077	(0.061, 0.094)
r_1	$\mathbb{1}(\text{white} = 0)$	-0.008	(-0.027, 0.011)		
r_2	$\mathbb{1}(65 \leq \text{age} \leq 75)$	-0.012	(-0.024, 0.000)	-0.009	(-0.021, 0.002)
r_3	$\mathbb{1}(65 \leq \text{age} \leq 80)$	-0.027	(-0.045, -0.010)	-0.027	(-0.045, -0.010)
r_4	$\mathbb{1}(65 \leq \text{age} \leq 85) \cdot \mathbb{1}(\text{Medicaid} = 0)$	-0.033	(-0.050, -0.016)	-0.031	(-0.048, -0.015)
r_5	$\mathbb{1}(\text{hispanic } \% = 0) \cdot \mathbb{1}(\text{education} = 1)$			-0.019	(-0.038, 0.000)
r_6	$\mathbb{1}(\text{hispanic } \% = 0) \cdot \mathbb{1}(\text{education} = 1)$ $\cdot \mathbb{1}(\text{population density} = 0)$			-0.045	(-0.067, -0.022)

Table 4: Sensitivity analysis for the treatment effect heterogeneity by using the percentile bootstrap

Rules	Sensitivity Parameter Λ				
	1.01	1.02	1.03	1.04	1.05
Inter.	(0.054, 0.100)	(0.045, 0.110)	(0.035, 0.121)	(0.024, 0.130)	(0.014, 0.140)
r_2	(-0.026, 0.007)	(-0.031, 0.012)	(-0.036, 0.017)	(-0.041, 0.022)	(-0.047, 0.028)
r_3	(-0.051, -0.003)	(-0.060, 0.004)	(-0.069, 0.014)	(-0.077, 0.023)	(-0.086, 0.033)
r_4	(-0.054, -0.009)	(-0.062, 0.000)	(-0.071, 0.009)	(-0.078, 0.017)	(-0.087, 0.024)
r_5	(-0.040, 0.004)	(-0.046, 0.012)	(-0.053, 0.016)	(-0.059, 0.020)	(-0.066, 0.026)
r_6	(-0.072, -0.017)	(-0.080, -0.011)	(-0.085, -0.004)	(-0.091, 0.002)	(-0.098, 0.008)

ratio scale, then our finding about the heterogeneity can be explained by this unmeasured bias. Even if the heterogeneity can be explained by an unmeasured bias at $\Lambda = 1.04$, the treatment effect of the baseline subgroup (i.e., intercept) still remains significant.

6 Discussion

In this paper, we introduce a new method for studying treatment effect heterogeneity that notably improves interpretability in terms of causal rules. The proposed CRE methodology accommodates for well-known shortcomings of binary trees by providing a more stable methodology to discover and estimate heterogeneous effects while maintaining high levels of interpretability. Indeed, CRE is stable to sample-to-sample variations, leading to more reproducible results, and its flexibility allows for the discovery of a wider set of causal rules. Also, CRE provides robust results for the detection of causal rules in the presence of overlap between confounders and effect modifiers.

The CRE method is a general method that is completely compatible with existing methods for estimating the CATE. The performance of CRE may vary with respect to the choice of existing methods to generate base decision rules in the discovery sample and intermediate values in the inference sample. Therefore, the CRE method can be thought of as a *refinement* process of the outputs produced by existing methods. If an estimation method for the CATE has great precision, then it is highly likely that it detects the treatment effect heterogeneity during the estimation procedure. When the CRE method is accompanied with this estimation method, the CRE method discovers the underlying treatment effect structure with high probability and represents this structure in an easy-to-interpret form.

The maximal number and complexity of the rules can be set by researchers or practitioners. Indeed, a few simple (i.e. not lengthy) rules are utterly important for public policy implications, where policy guidelines need to be as simple and as general as possible. However, when it comes to precision medicine, discovering a possibly lengthy rule that is specific to a patient could be of interest. Also the choice of how many causal rules to discover in the discovery step may depend on the questions that practitioners want to answer. For example, policymakers generally want to discover a short list of risk factors. A few important subgroups defined by the risk factors are usually easy-to-understand, and further foster focused discussions about the assessments of potential risks and benefits of policy actions. Also, due to the restriction of resources, public health can be promoted efficiently when prioritized subgroups are available. Conversely, in precision medicine a comparatively larger set of causal rules can be chosen. Indeed, an important goal is to identify patient subgroups that respond to treatment at a much higher (or lower) rate than the average (Loh et al.; 2019). Also, identifying a subgroup that must avoid the treatment due to its excessive side effects can be valuable information. However, discovering only a few subgroups is likely to miss this extreme subgroup.

From simulations we showed that, CRE has a good performance both in the discovery and inference steps. Interestingly, CRE has an enhanced comparative performance in scenarios in which there is overlap between confounders and effect modifiers. This is an important quality for real-world applications. For instance, it can be the case in pollution studies, that income is a confounder as poorer people live in neighbours with higher levels of pollution, and also an effect modifier as poorer people may have worst living conditions and, in turn, experience higher negative effects from pollution. In this scenarios, discovery of causal rules through CRE could greatly improve over other methodologies.

A number of extensions of the CRE method can be possible. First, the CRE method maintains the benefits that existing methods have – i.e., asymptotic normality and unbiasedness. If an existing method can produce unbiased point estimates for $\tau(x)$ with valid confidence intervals, the CRE method can also produce unbiased estimates for $\boldsymbol{\beta}$ with valid confidence intervals. Bayesian methods such as BART or BCF can be also used, and it is empirically shown that they perform really well. However, the validity of Bayesian inference such as constructing credible intervals remains as a future research question. Second, the discovery step of the CRE method can be considered as a dimension reduction procedure. We used a set of decision rules as a basis, but it may be possible to use other forms to characterize the treatment effect heterogeneity. Finally, we proposed an approach for sensitivity analysis of unmeasured confounding bias based on the inverse probability of treatment weighting estimator. However, a more general approach for sensitivity analysis that can be compatible with a larger class of estimation methods would be helpful. Future research is needed for developing such sensitivity analysis.

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., Wager, S. et al. (2019). Generalized random forests, *The Annals of Statistics* **47**(2): 1148–1178.

Bargagli-Stoffi, F. J., De-Witte, K. and Gnecco, G. (2019). Heterogeneous causal effects with imperfect compliance: a novel bayesian machine learning approach, *arXiv preprint arXiv:1905.12707* .

Bargagli-Stoffi, F. J., Tortù, C. and Forastiere, L. (2020). Heterogeneous treatment and spillover effects under clustered network interference, *arXiv preprint arXiv:2008.00707* .

Belloni, A., Chernozhukov, V., Hansen, C. and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control, *Journal of Business & Economic Statistics* **34**(4): 590–605.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Carone, M., Dominici, F. and Sheppard, L. (2020). In pursuit of evidence in air pollution epidemiology: the role of causally driven data science, *Epidemiology* **31**(1): 1–6.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects, *American Economic Review* **107**(5): 261–65.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. and Robins, J. M. (2016). Locally robust semiparametric estimation, *arXiv preprint arXiv:1608.00033* .

Chipman, H. A., George, E. I. and McCulloch, R. E. (2010). BART: Bayesian additive regression trees, *The Annals of Applied Statistics* **4**(1): 266–298.

Cook, D. I., Gebski, V. J. and Keech, A. C. (2004). Subgroup analysis in clinical trials, *Medical Journal of Australia* **180**(6): 289–291.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity, *The Review of Economics and Statistics* **90**(3): 389–405.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F. and Schwartz, J. D. (2017). Air pollution and mortality in the Medicare population, *New England Journal of Medicine* **376**(26): 2513–2522.

Dominici, F., Bargagli-Stoffi, F. J. and Mealli, F. (2020). From controlled to undisciplined data: estimating causal effects in the era of data science using a potential outcome framework, *arXiv preprint arXiv:2012.06865* .

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*, Vol. 38, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Ertefaie, A., Hsu, J. Y., Page, L. C. and Small, D. S. (2018). Discovering treatment effect heterogeneity through post-treatment variables with application to the effect of class size on mathematics scores, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **67**(4): 917–938.

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.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *The Annals of Statistics* **29**(9): 1189–1232.

Friedman, J. H. and Popescu, B. E. (2008). Predictive learning via rule ensembles, *The Annals of Applied Statistics* **2**(3): 916–954.

Hahn, P. R., Murray, J. S., Carvalho, C. M. et al. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects, *Bayesian Analysis* .

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference, *Journal of Computational and Graphical Statistics* **20**(1): 217–240.

Hirano, K., Imbens, G. W. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica* **71**(4): 1161–1189.

Holland, P. W. (1986). Statistics and causal inference, *Journal of the American Statistical Association* **81**(396): 945–960.

Imai, K., Ratkovic, M. et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation, *The Annals of Applied Statistics* **7**(1): 443–470.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation, *Journal of Economic Literature* **47**(1): 5–86.

Jacob, D. (2019). Group average treatment effects for observational studies, *arXiv preprint arXiv:1911.02688* .

Kim, B., Khanna, R. and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability, *Advances in Neural Information Processing Systems*, pp. 2280–2288.

Lakkaraju, H., Bach, S. H. and Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684.

Lee, K., Small, D. S. and Dominici, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies, *Journal of the American Statistical Association* pp. 1–33.

Lee, M.-j. (2009). Non-parametric tests for distributional treatment effect for randomly censored responses, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(1): 243–264.

Lewis, J. B. and Linzer, D. A. (2005). Estimating regression models in which the dependent variable is based on estimates, *Political Analysis* **13**(4): 345–364.

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.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model, *The American Statistician* **54**(3): 217–224.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**(4): 417–473.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* **267**: 1–38.

Nalenz, M. and Villani, M. (2018). Tree ensembles with rule structured horseshoe regularization, *The Annals of Applied Statistics* **12**(4): 2379–2408.

Nethery, R. C., Mealli, F., Sacks, J. D. and Dominici, F. (2020). Evaluation of the health impacts of the 1990 clean air act amendments using causal inference and machine learning, *Journal of the American Statistical Association* pp. 1–12.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models, *Statistics in medicine* **16**(3): 285–319.

Rosenbaum, P. R. (2002). *Observational studies*, New York: Springer.

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.

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

Schwartz, J., Wei, Y., Di, Q., Dominici, F., Zanobetti, A. et al. (2021). A national difference in differences analysis of the effect of pm2. 5 on annual death rates, *Environmental Research* **194**: 110649.

Strobl, C., Malley, J. and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests., *Psychological Methods* **14**(4): 323.

Su, L., Shi, Z. and Phillips, P. C. (2016). Identifying latent structures in panel data, *Econometrica* **84**(6): 2215–2264.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores, *Journal of the American Statistical Association* **101**(476): 1619–1637.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**(1): 267–288.

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

Wang, T. and Rudin, C. (2017). Causal rule sets for identifying subgroups with enhanced treatment effect, *arXiv preprint arXiv:1710.05426* .

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**(4): 817–838.

Wu, X., Braun, D., Schwartz, J., Kioumourtzoglou, M. and Dominici, F. (2020). Evaluating the impact of long-term exposure to fine particulate matter on mortality among the elderly, *Science Advances* **6**(29): eaba5692.

Zhao, Q., Small, D. S. and Bhattacharya, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**(4): 735–761.

Online Appendix

A Assumptions

Assumption 1 (Unconfoundedness).

$$(Y(1), Y(0)) \perp\!\!\!\perp Z \mid X_i.$$

This assumption means that the two potential outcomes depend on $\mathbf{X}$, but are independent of Z conditioning on $\mathbf{X}$.

Assumption 2 (Overlap).

$$0 < e(x) < 1 \;\forall\, x \in \mathcal{D},$$

where $\mathcal{D}$ is the features' space and $e(x) = \mathbb{E}[Z|\mathbf{X} = x]$ is the propensity score (Rosenbaum and Rubin; 1983).

Assumption 3 (Stable Unit Treatment Value Assumption (SUTVA)).

$$Y_i(Z_i) = Y_i,$$
$$Y_i(Z_i) = Y(Z_1, Z_2, \cdots, Z_i, \cdots, Z_N).$$

Assuming (1-3) to hold and by using the propensity score, CATE $\tau(x)$ can be identified.

B Simulations for Two-Steps Rule-Discovery Procedure

We have conducted simulations to compare the causal rules discovered by the Honest Causal Tree (Athey and Imbens; 2016) analysis (HCT) with the rules discovered using a simplified version of our proposed CRE approach described in the paper, in which we employed BCF to obtain ITEs and then applied CART to discover an interpretable tree structure (BCF+CART).

We simulated two true causal rules. We generated the covariate matrix $\mathbf{X}$ with 10 binary covariates and two designs: Design 1 had no confounding while Design 2 introduced confounding in the data generating process. For Design 1, we assumed $Z_i \sim Bern(0.5)$. The potential outcomes were generated by $Y_i(0) \sim N(0,1)$ and $Y_i(1) = Y_i(0) + \tau$, and the observed outcomes were created as $y = y_0 \cdot (1 - Z_i) + y_1 \cdot Z_i$. For Design 2, we assumed $Z_i \sim Bern(\pi_i)$, where $\pi_i = \text{logit}(-1 + x_{j3} - x_{j4} + x_{j5})$. The potential outcomes were generated by $Y_i(0) \sim N(x_{j3} + 0.5x_{j4} + x_{j5}, 1)$ and $Y_i(1) = Y_i(0) + \tau$. Finally, the observed outcomes were created as $y = y_0 \cdot (1 - Z_i) + y_1 \cdot Z_i + f(\mathbf{x})$ where $f(\mathbf{x})$ is a linear function of the confounders x_{j3}, x_{j4} and x_{j5}. For both designs we generated two causal rules.

Table 5: Simulations results for two-steps rule-discovery.

		Sample Size					
Effect Size	Method	500	1,000	5,000	500	1,000	5,000
		Design 1			Design 2		
0.50	CT	0.18	0.25	0.40	0.02	0.03	0.05
	BCF+CART	0.27	0.39	1.00	0.03	0.20	0.99
1.00	CT	0.44	0.52	0.84	0.07	0.08	0.14
	BCF+CART	1.00	1.00	1.00	0.84	0.99	1.00

The two causal rules were: 1) $x_{j1} = 0, x_{j2} = 0$ then $\tau = a$, 2) $x_{j1} = 1, x_{j2} = 1$, then $\tau = -a$, where $a \in \{0.5, 1\}$. Results were obtained by aggregating over 500 simulated dataset for each sample size and design. In each setting, BCF+CART achieved better subgroup identification than HCT.

The intuition behind the superior performance of double ML approaches, such as BCF+CART, is that they exploit more information on the heterogeneous subgroups than single learner algorithms while simultaneously controlling better (through the BCF estimation of CATE) for measured confounding. In this simplified version causal rules are discovered through a single tree instead that a forest of trees as in the proposed CRE approach.

C Theoretical Properties

Given that the matrix $\tilde{\mathbf{X}}$ is fixed, we can prove that the estimator $\hat{\beta}_i, j = 1, \ldots, M$ is a consistent estimator of β_i, that is, the average treatment effect for the subgroups defined by the decision rule r_i.

Theorem 1. *If $\hat{\tau}_i^{inf}$ satisfies the model* (4) *(Condition 1) and $\mathbb{E}(\tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i) = \mathbf{Q}$ is a finite positive definite matrix (Condition 2), then the estimator $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \boldsymbol{\tau}^*$ is a consistent estimator for $\boldsymbol{\beta}$.*

Proof of Theorem 1. By multiplying $(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T$ on the both sides of model (5), we have $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \boldsymbol{\nu}$. From the assumptions, $\text{plim} \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i = \mathbf{Q}$ and $\text{plim} \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{X}}_i^T \nu_i = \text{plim} \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{X}}_i^T (\epsilon_i + u_i) = 0$. By Slutsky's theorem, plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

For instance, one could use the inverse probability weighting (IPW) or stabilized IPW (SIPW) approaches as both $\hat{\tau}_i^{inf} = \hat{\tau}_i^{IPW}$ and $\hat{\tau}_i^{inf} = \hat{\tau}_i^{SIPW}$ satisfy the model in (4). Therefore, the following corollary can be obtained from Theorem 1:

Corollary 1. *The estimator $\hat{\boldsymbol{\beta}}^{(S)IPW} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \boldsymbol{\tau}^*$ where $\hat{\tau}_i^{inf} = \hat{\tau}_i^{(S)IPW}$ is consistent.*

We need additional assumptions to prove asymptotic normality of $\hat{\boldsymbol{\beta}}$. For a general covariate matrix, the following three Conditions are required: (3) $\mathbb{E}(\tilde{\mathbf{X}}_{ij}^4) < \infty$, (4) $\mathbb{E}(\nu_i^4) < \infty$; (5) $\mathbb{E}(\nu_i^2 \tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i) = \boldsymbol{\Omega}$ is a positive definite matrix. Since $\tilde{\mathbf{X}}_{ij}$ is either 0 or 1 in our setup, Condition (3) is satisfied by design. The following theorem represents the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.

Theorem 2. *If Conditions (1)-(5) hold, then*

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}) \quad \text{as} \quad N \to \infty$$

where $\mathbf{V} = \mathbf{Q}^{-1} \boldsymbol{\Omega} \mathbf{Q}^{-1}$.

The variance $\mathbf{V}$ usually has to be estimated. The variance-covariance matrix estimator $\hat{\mathbf{V}}_n = \hat{\mathbf{Q}}^{-1} \hat{\boldsymbol{\Omega}} \hat{\mathbf{Q}}^{-1}$ can be obtained by the sandwich formula where $\hat{\mathbf{Q}} = n^{-1} \sum_{i=1}^{N} \tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i$, $\hat{\boldsymbol{\Omega}} = n^{-1} \sum_{i=1}^{N} \hat{\nu}_i^2 \tilde{\mathbf{X}}_i^T \tilde{\mathbf{X}}_i$ and $\hat{\nu}_i = \hat{\tau}_i^{inf} - \tilde{\mathbf{X}}_i \hat{\boldsymbol{\beta}}$. This estimator is robust and often referred to as the

White's estimator (White; 1980). There are other approaches to obtain a heteroscedasticity consistent covariance matrix as discussed in Long and Ervin (2000). For small samples, Efron's estimator (Efron; 1982), known as HC3 estimator, can be considered alternatively. Also, if the variance w_i is known from the large sample properties of existing methods for obtaining $\hat{\tau}_i^{inf}$, then feasible generalized least squares estimators (Lewis and Linzer; 2005) can be considered.

Proof of Theorem 2. To prove normality, we use the expression $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\nu}$. By rewriting this, we have $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (N^{-1}\sum_{i=1}^N \tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i)^{-1} \times N^{-1/2}\sum_{i=1}^N \tilde{\mathbf{X}}_i^T\nu_i$. From the assumptions, we also have $\mathbb{E}(\tilde{\mathbf{X}}_i^T\nu_i) = 0$ and $\text{var}(\tilde{\mathbf{X}}_i^T\nu_i) = \mathbb{E}(\nu_i^2\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i) < \infty$. Then, by the central limit theorem, $N^{-1/2}\sum_{i=1}^N \tilde{\mathbf{X}}_i^T\nu_i$ converges in distribution to $N(0,\Omega)$. By Slutsky's theorem and Cramer-Wold theorem, $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to $N(0,\mathbf{Q}^{-1}\Omega\mathbf{Q}^{-1})$.

D Comparison between BCF, BART, IPW, OR for Rule Discovery

As we discussed in Section 3.1, one of the attractive features of the CRE method is that we can consider many approaches to estimate the individual treatment effect τ_i. Among many existing methods, one approach to note is the inverse probability weighting (IPW) estimator

$$\hat{\tau}_i^{IPW} = \left(\frac{Z_i}{\hat{e}(X_i)} - \frac{1-Z_i}{1-\hat{e}(X_i)}\right)Y_i \tag{8}$$

where $\hat{e}(X_i)$ is the estimate of the propensity score $e(x)$ at X_i. The estimate $\hat{e}(X_i)$ can be obtained by fitting a logistic regression on (Z_i, X_i). Although $\hat{\tau}_i^{IPW}$ is an unbiased estimator of $\tau(x)$ (i.e., $\mathbb{E}[\hat{\tau}_i^{IPW}|X = x] = \tau(x)$), the transformed value $\hat{\tau}_i^{IPW}$ can be highly fluctuating when $\hat{e}(X_i)$ is close to 0 or 1. To avoid extreme values of $\hat{\tau}_i^{IPW}$, we can instead use the stabilized version $\hat{\tau}_i^{SIPW}$ (Hirano et al.; 2003),

$$\hat{\tau}_i^{SIPW} = \left\{\left(\frac{1}{N}\sum_{i=1}^N \frac{Z_i}{\hat{e}(X_i)}\right)^{-1}\frac{Z_i}{\hat{e}(X_i)} - \left(\frac{1}{N}\sum_{i=1}^N \frac{1-Z_i}{1-\hat{e}(X_i)}\right)^{-1}\frac{1-Z_i}{1-\hat{e}(X_i)}\right\}Y_i. \tag{9}$$

Although $\hat{\tau}_i^{IPW}$ or $\hat{\tau}_i^{SIPW}$ is enough to use as an estimate of τ_i, another approach of imputing missing potential outcomes can be considered. Without estimating the propensity score, functions for two potential outcomes, say $m_1(x) = \mathbb{E}[Y|Z = 1, \mathbf{X} = x]$ and $m_0(x) = \mathbb{E}[Y|Z = 0, \mathbf{X} = x]$, are estimated. Then missing potential outcomes are imputed by the estimated functions $\hat{m}_0$ and $\hat{m}_1$. For instance, if $Z_i = 0$, $Y_i(0)$ is observed as Y_i and $Y_i(1)$ is imputed by $\hat{m}_1(X_i)$. The unit level treatment effect can be estimated by either subtracting the observed outcome and imputed counterfactual, i.e. $\tau_i^{OR} = Y_i^{obs} - (\hat{m}_1(X_i) \cdot (1 - Z_i) + \hat{m}_0(X_i) \cdot Z_i)$, or by subtracting the imputed potential outcomes, i.e., $\tau_i^{BART} = \hat{m}_1(X_i) - \hat{m}_0(X_i)$. We refer to the former methodology as outcome regression (OR) and to the latter as BART imputation (Hill; 2011). Indeed, we implement both these methods for the estimation of the unit level treatment effect using the Bayesian Additive Regression Tree (BART) algorithm (Chipman et al.; 2010).

Here, we show that the BCF method for the estimation of τ_i has the better performance as compared to other approaches discussed above. In order to evaluate the ability of discovering decision rules, two factors are considered in line with similar simulations scenarios (Bargagli-Stoffi et al.; 2019, 2020): (1) how many rules are discovered and (2) among the discovered rules, how many times true rules are captured. We implement the simulation scenario introduced in Section 3.1.2 with uncorrelated covariates, linear confounding and 2,000 data points. The obtained results are depicted in Figure 5. The four different plots in this figure show the variation in the number of correct rules (first row) and the number of detected rules (second row) as the effect size increases. The plots in the first column depict the results in the case of two true causal rules, while in the second column depict the results in the case of four true causal rules. In the case of two true causal rules, BCF, BART and OR similarly perform with respect to their ability to identify the true causal rules, while in the case of four true causal rules there is a clear advantage in using BART/BCF over OR. Also, IPW consistently underperform in both the scenarios. With respect to the number of rules detected, OR and IPW are more "conservative" (i.e., smaller number of detected rules) while BART is less conservative. As BCF shows the best performance in terms of correctly detected rules, while it detects consistently less rules than the others. Also, BCF requires a lower computational cost than BART.

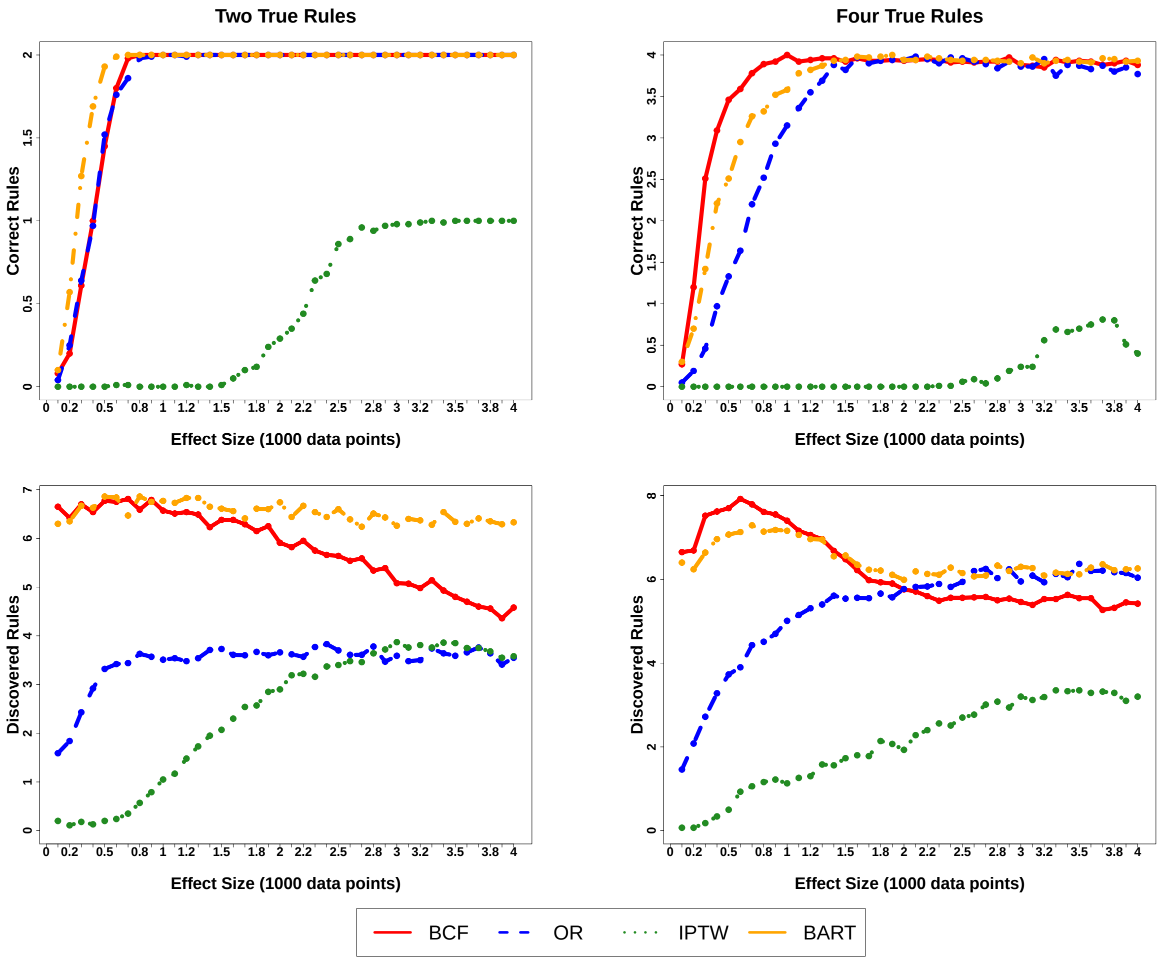

Figure 5: Comparison between BCF, BART, IPW, OR for rules' discovery. The plots show the variation in the number of correct rules (first row) and the number of detected rules (second row) as the effect size increases, in the cases of two true rules (first column) and four true rules (second column).

E Detailed Simulation Results

Figure 6 depicts the simulation results for two scenarios: two true rules (left panel) and four true rules (right panel). We consider 100 simulated datasets for each effect size and we report the average number of correctly discovered rules (CDR). We report the results for both $N = 1000$ (red solid line) and $N = 2000$ (red dashed line) data points. We find that in the scenario with 2,000 data points CRE-BCF is faster in discovering the actual rules. However, the difference in

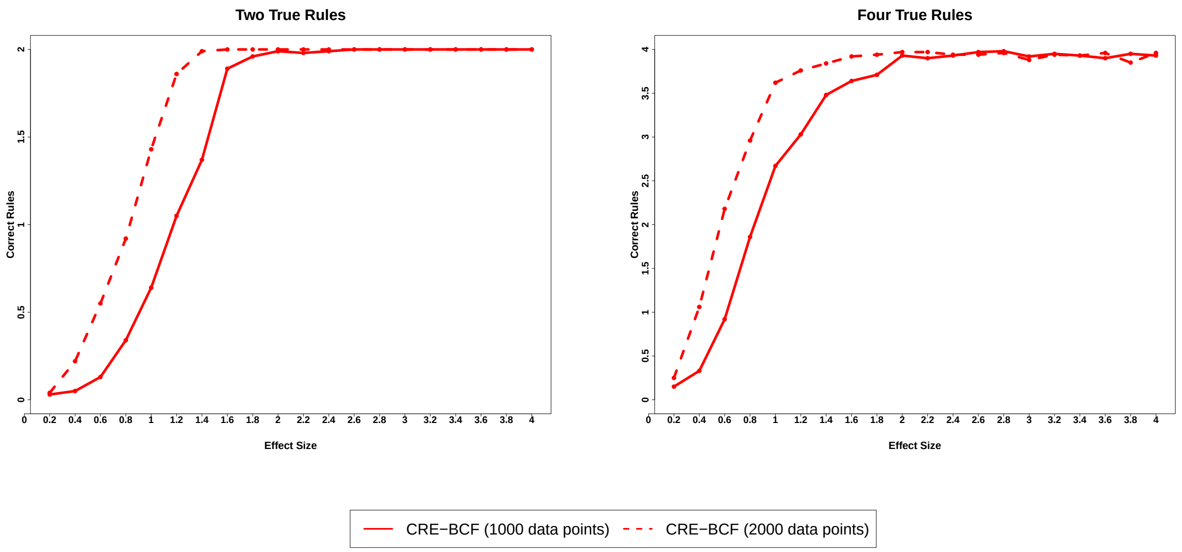

Figure 6: Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).

not sizable as the performance of CRE-BCF is excellent also with the smaller sample size. Table 6 depicts the performance of CRE-BCF for the simulation scenario introduced in 4.1. We provide both the average number of correctly discovered rules (CDR), the proportion of times when the correct rules are discovered (π) and the average number of discovered rules (DR). As the effect size k increases CRE-BCF is always able to spot the true causal rules. Even for smaller effect sizes that are not significantly different than the null effect, CRE performs well in the both cases. In Appendix F, we run a series of simulations introducing additional variations in the correlation between the covariates and the functional form of $f(\mathbf{X})$. Here, it is worth highlighting that none of these variations in the data generating process decreases the ability of CRE to correctly spot the true underlying causal rules.

F Additional Simulations

In this section, we additionally consider correlation between covariates in $\mathbf{X}$. Such correlations were introduced to investigate whether or not correlated covariates negatively affect the ability

Table 6: Performance of the CRE-BCF method in discovering the true underlying causal rules

Linear Scenario												
	Two Rules						Four Rules					
	1,000			2,000			1,000			2,000		
k	CDR	π	DR	CDR	π	DR	CDR	π	DR	CDR	π	DR
0.1	0.03	0.00	6.53	0.04	0.01	6.41	0.15	0.01	6.54	0.25	0.01	6.57
0.2	0.05	0.01	6.54	0.22	0.05	6.73	0.33	0.04	6.51	1.06	0.14	6.88
0.3	0.13	0.02	6.50	0.55	0.15	6.57	0.92	0.09	6.66	2.18	0.35	6.82
0.4	0.34	0.09	6.54	0.92	0.33	6.60	1.86	0.29	7.04	2.96	0.48	7.55
0.5	0.64	0.17	6.64	1.43	0.61	6.82	2.67	0.40	7.44	3.62	0.83	7.68
0.6	1.05	0.42	6.59	1.86	0.90	6.80	3.03	0.56	7.48	3.76	0.83	7.66
0.7	1.37	0.59	6.84	1.99	0.99	6.67	3.48	0.71	7.59	3.84	0.89	7.60
0.8	1.89	0.90	6.61	2.00	1.00	6.65	3.64	0.79	7.61	3.92	0.94	7.51
0.9	1.96	0.96	6.77	2.00	1.00	6.66	3.71	0.82	7.66	3.94	0.95	7.39
1.0	1.99	0.99	6.68	2.00	1.00	6.55	3.93	0.96	7.71	3.97	0.97	7.35
1.1	1.98	0.98	6.79	2.00	1.00	6.51	3.90	0.92	7.51	3.97	0.98	7.24
1.2	1.99	0.99	6.63	2.00	1.00	6.36	3.93	0.95	7.62	3.94	0.96	6.88
1.3	2.00	1.00	6.65	2.00	1.00	6.67	3.97	0.97	7.64	3.94	0.94	6.68
1.4	2.00	1.00	6.76	2.00	1.00	6.30	3.98	0.98	7.59	3.96	0.96	6.82
1.5	2.00	1.00	6.65	2.00	1.00	6.31	3.92	0.95	7.38	3.88	0.89	6.52
1.6	2.00	1.00	6.65	2.00	1.00	6.29	3.95	0.95	7.18	3.94	0.94	6.23
1.7	2.00	1.00	6.61	2.00	1.00	6.26	3.93	0.93	7.16	3.93	0.93	6.04
1.8	2.00	1.00	6.47	2.00	1.00	6.01	3.90	0.90	7.16	3.96	0.96	6.01
1.9	2.00	1.00	6.50	2.00	1.00	6.03	3.95	0.95	6.94	3.85	0.85	5.84
2.0	2.00	1.00	6.24	2.00	1.00	5.93	3.93	0.93	6.90	3.96	0.97	5.84

of CRE-BCF to discover the true causal rules. It can be possible that CRE-BCF faces harder times in correctly picking the variables that are responsible for the heterogeneous effects, as all the variables are correlated with each other. Figure 7 depicts the simulation results with 0.3 correlation in the case of 1000 data points and 2 and 4 true causal rules, respectively. This figure shows that correlation does not affect the performance of CRE.

Moreover, we generate non-linearities in the confounding $f(\mathbf{X})$. In particular, the data generating process introduced in Section 3.1.2 is reworked as follows: $y = y_0 \cdot (1 - z_i) + y_1 \cdot z_i + exp\{X_1 - X_2 \cdot X_3\}$. This non-linear $f(\mathbf{X})$ is introduced in order to check the robustness of the CRE-BCF model to non-linear confounding. We can argue that, in many real-world applications, confounders can interact with each other and can have non-linear associations with the output. Again, we generate two different scenarios with 2 and 4 true causal rules and 1000 data points. Figure 8 shows that this kind of non-linear confounding is not harmful, rather BCF can discover the true causal rules for small effect sizes. This is due to the fact that both BCF (Hahn et al.; 2020) and CRE are able to deal with non-linearities in a excellent way.

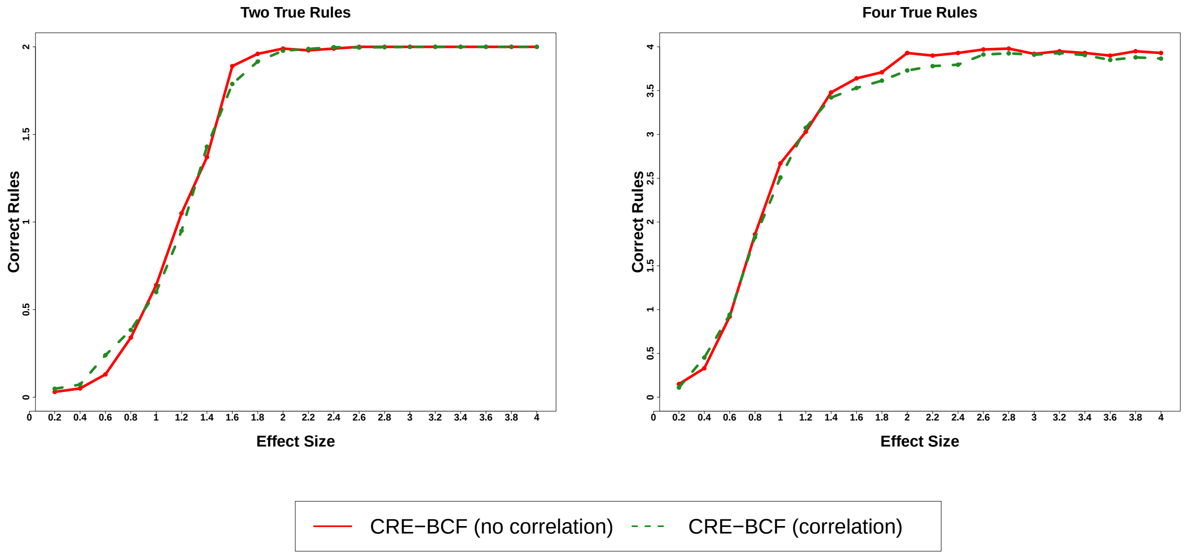

Figure 7: Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).

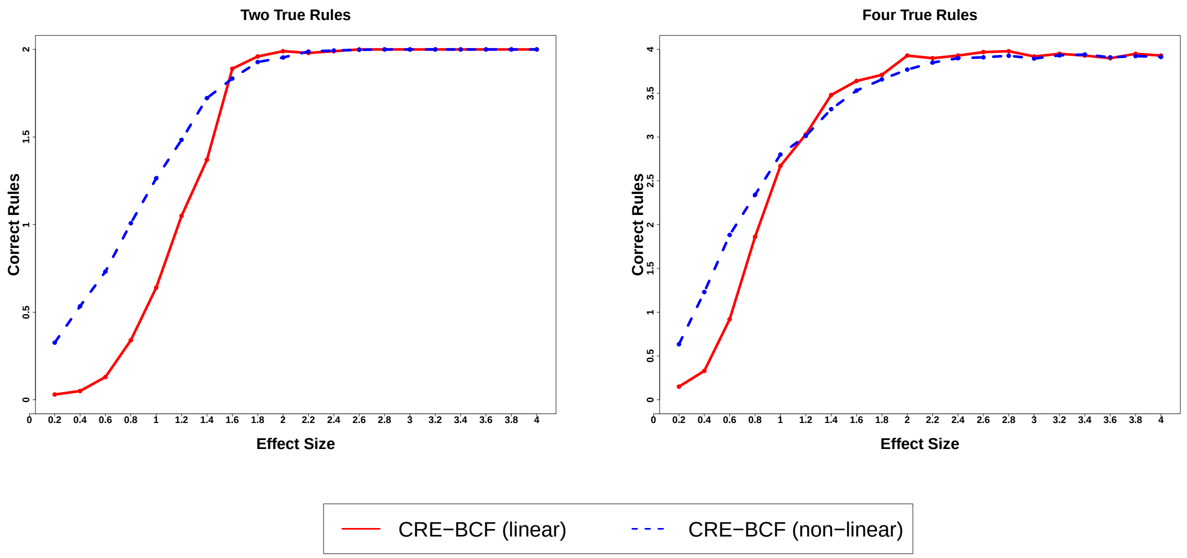

Figure 8: Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).