CAUSAL RULE ENSEMBLE: INTERPRETABLE INFERENCE OF HETEROGENEOUS TREATMENT EFFECTS

KWONSANG LEE[1], FALCO J. BARGAGLI-STOFFI[2], AND FRANCESCA DOMINICI[2]

Abstract. In environmental epidemiology, it is critically important to identify subpopulations that are most vulnerable to the adverse effects of air pollution so we can develop targeted interventions. In recent years, there have been many methodological developments for addressing heterogeneity of treatment effects in causal inference. A common approach is to estimate the conditional average treatment effect (CATE) for a pre-specified covariate set. However, this approach does not provide an easy-to-interpret tool for identifying susceptible subpopulations or discover new subpopulations that are not defined a priori by the researchers. In this paper, we propose a new causal rule ensemble (CRE) method with two features simultaneously: 1) ensuring interpretability by revealing heterogeneous treatment effect structures in terms of decision rules and 2) providing CATE estimates with high statistical precision similar to causal machine learning algorithms. We provide theoretical results that guarantee consistency of the estimated causal effects for the newly discovered causal rules. Furthermore, via simulations, we show that the CRE method has competitive performance on its ability to discover subpopulations and then accurately estimate the causal effects. We also develop a new sensitivity analysis method that examine robustness to unmeasured confounding bias. Lastly, we apply the CRE method to the study of the effects of long-term exposure to air pollution on the 5-year mortality rate of the New England Medicare-enrolled population in United States. Code is available at https://github.com/kwonsang/causal_rule_ensemble.

1. Introduction

There have been many developments in estimating the average treatment effects (ATE). In various fields, the estimation of the ATE provides a central insight on the causal effect of a treatment (e.g., an intervention, an environmental policy, and so on) on an outcome, on average for the whole population. However, in addition to the ATE, it is critically important to identify subpopulations of the population that would benefit the most from a treatment and/or would be most vulnerable to an environmental exposure. In the context of air pollution, it is deemed important to public health to identify the subpopulations that are most vulnerable, so that effective interventions can be put in place to mitigate adverse health effects (Lee et al., 2018).

There is extensive literature on assessing heterogeneity of causal effects that is based on estimating the conditional average treatment effect (CATE). For each combination of covariates $\mathbf{X} = \mathbf{x}$

[1] Department of Statistics, Sungkyunkwan University
[2] Department of Biostatistics, Harvard School of Public Health
Key words and phrases. causal inference, decision rules, interpretability, machine learning, observational study, sensitivity analysis, stability selection.

(i.e., subset of the features space), the CATE can be estimated with the same set of the causal assumptions that are needed for estimating the ATE (Athey and Imbens, 2016). Under the same identification assumptions, earlier works on estimating 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* (Bellman, 1961; Robins and Ritov, 1997). Recently, other nonparametric approaches facilitate machine learning methods such as Random Forest (Breiman, 2001) and Bayesian Additive Regression Tree (BART) (Chipman et al., 2010). These approaches have been successful when the number of features is large. For instance, in their seminal contributions, Foster et al. (2011) and Hill (2011) used forest-based algorithms for the prediction of the missing potential outcomes. In a similar spirit, Hahn et al. (2020) proposed a BART-based approach but with a novel parametrization of the outcome surfaces. 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 asymptotic theory for the conditional treatment effect estimators and valid statistical inference.

Despite the success in accurately estimating the CATE using machine learning methods, these tree ensemble methods offer little guidance about which covariates or, even further, subpopulations (i.e., subsets of the features space defined by multiple covariates) bring about treatment effect heterogeneity. Outputs/results obtained from existing methods are hard to interpret by human experts because parametrizations of the covariate space are complicated. This issue is well-known as *lack of interpretability*. Increasing model interpretability is key to understanding and furthering human knowledge. However, effort to improve interpretability is so far lacking in the current causal inference literature dealing with the study of treatment effect heterogeneity.

In this paper, we propose a novel Causal Rule Ensemble (CRE) method that ensures interpretability, while maintaining a high level of accuracy in estimation. The CRE method uses decision rules obtained from multiple trees, selects a key subset of rules to identify subpopulations contributing to heterogeneous treatment effects, and estimates CATE for each selected rule. Interpretability is usually a qualitative concept, and often defined as the degree to which a human can understand the cause of a decision or consistently predict the results of the model (Miller, 2019; Kim et al.,

2016). Decision rules are ideal for this non-mathematical definition of interpretability. A decision rule consists of simple *if-then* statements regarding several conditions and corresponds to a specific subpopulation. A handful number of decision rules, called *causal rules*, can be chosen by using high performance machine learning techniques, but are still easy to understand.

We achieve the following three main goals: (1) discovering de novo causal rules that lead to heterogeneity of causal effects; (2) providing valid inference and large sample properties about CATE with respect to the newly discovered rules; and (3) assessing sensitivity to unmeasured confounding bias for the rule-specific causal effects. To do so, we follow Athey and Imbens (2016), and rely on a sample-splitting approach that divides the total sample into two smaller subsamples: one for discovering a set of interpretable decision rules that could lead to treatment effect heterogeneity (i.e., discovery sample) and the other for estimating the rule-specific treatment effects (i.e., inference sample). We also tailor a sensitivity analysis method proposed by Zhao et al. (2019) to assess the robustness of the rule-specific treatment effects to unmeasured confounding. Furthermore, the CRE method has several other advantages besides interpretability and estimation accuracy. It allows practitioners to represent treatment effect heterogeneity in a more flexible and stable way. This stability provides higher standards of replicability. Also, it is a powerful tool to disentangle the effect modifiers (namely, drivers of causal effect heterogeneity) from measured confounders.

The reminder of the paper is organized as follows. In Section 2, we introduce the main definitions of CATE and interpretable decision rules. In Section 3, we describe the algorithm used for the discovery of causal rules. Section 4 introduces our innovative ideas regarding estimation and sensitivity analysis. In Section 5, we conduct simulations studies. In Section 6, we apply the proposed method to the Medicare Data. Section 7 discusses the strengths and weaknesses of our proposed approach and areas of future research.

2. Treatment Effect Heterogeneity, Interpretability and Sample-Splitting

2.1. **Causal Treatment Effects.** Suppose there are N subjects. For each subject, let Y_i be an outcome, Z_i be a binary treatment, and $\mathbf{X}_i$ be a K-dimensional vector of covariates. Following the potential outcome framework (Neyman, 1990; Rubin, 1974), we define two potential outcomes $Y_i(1)$ and $Y_i(0)$ as a function of the treatment assigned to each subject i (i.e., $Y_i(Z_i)$). $Y_i(1)$ is the

potential outcome for unit i under treatment, while $Y_i(0)$ is the potential outcome under control. The fundamental problem is that the treatment effect $\tau_i = Y_i(1) - Y_i(0)$ cannot be observed from a given sample $(Y_i, Z_i, \mathbf{X}_i)$ since we can observe only one of the potential outcomes (Holland, 1986). Since either $Z_i = 1$ or $Z_i = 0$ is realized, the corresponding potential outcome $Y_i(Z_i)$ is observed, and the other potential outcome $Y_i(1 - Z_i)$ is *counterfactual* and always missing. For instance, if $Z_i = 1$, we can observe $Y_i(1)$, but cannot observe $Y_i(0)$. What we can observe is the realization of Y_i that can be formalized as $Y_i = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$. It is extremely difficult to estimate τ_i as this quantity is never observed in reality. Instead, we consider the conditional average treatment effect (CATE) $\tau(x)$ defined as $\tau(x) = \mathbb{E}\left[Y_i(1) - Y_i(0) | \mathbf{X}_i = x\right]$ and the average treatment effect (ATE) defined as $\tau = \mathbb{E}_X[\tau(x)]$.

Although $\tau(x)$ cannot be observed, $\tau(x)$ can be estimated under the assumption of no unmeasured confounders (Rosenbaum and Rubin, 1983),

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp Z_i \mid \mathbf{X}_i. \tag{1}$$

This assumption means that the two potential outcomes depend on $\mathbf{X}_i$, but are independent of Z_i conditioning on $\mathbf{X}_i$. By using the propensity score $e(x) = \mathbb{E}[Z_i | \mathbf{X}_i = x]$ (Rosenbaum and Rubin, 1983), the CATE $\tau(x)$ can be identified as

$$\tau(x) = \mathbb{E}\left[\left(\frac{Z_i}{e(x)} - \frac{1 - Z_i}{1 - e(x)}\right) Y_i \mid \mathbf{X}_i = x\right]. \tag{2}$$

However, in practice we do not know whether a considered set $\mathbf{X}_i$ is sufficient for the assumption (1). When there exists a source of unmeasured confounding, this assumption is violated, and the identification results do not hold. Sensitivity analysis provides a useful tool to investigate the impact of unmeasured confounding bias, which will be discussed in Section 4.2.

2.2. **Interpretability and Decision Rules.** Our primary goal is to provide an interpretable structure of $\tau(x)$ (discovery of treatment effect heterogeneity) and then estimate $\tau(x)$ in an efficient and precise manner (estimation of treatment effect heterogeneity). The functional form of $\tau(x)$ is not known, and may have a complicated structure varying across subgroups. A complex model can be considered to get a better estimate of $\tau(x)$, but such model may mask the truly informative structure of $\tau(x)$. Instead, it is possible to "extract" (or "approximate") important subspaces of $\mathbf{X}$

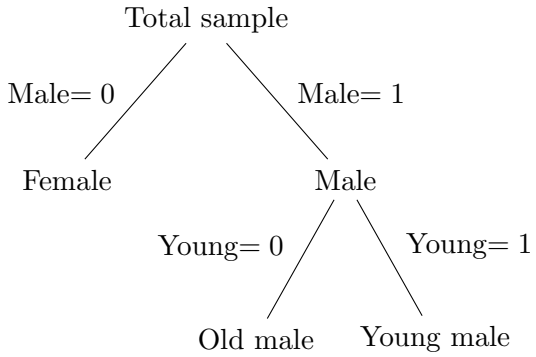

FIGURE 1. An example tree.

that are responsible for the variation of $\tau(x)$, using a sparse representation of the conditional average treatment effect (CATE) defined by parsimonious subsets of the covariates space (Imai et al., 2013). The extracted subspaces may not perfectly describe the treatment effect heterogeneity, but provide a concise and informative summary, thus are more interpretable. This is often referred to as *trade-off* between interpretability and accuracy. More specifically, weighing too much on achieving high accuracy in the estimation of the causal effect for a given covariate-profile generally comes at the cost of compromising interpretability. On the other hand, trying to improve interpretability might lead to too simple structures that are not representative of the true $\tau(x)$ structure, which in turn may reduce the statistical precision. It is important to find a good balance between interpretability and estimation accuracy. Our work focuses on finding this balance and, furthermore, improving interpretability while minimizing the loss of estimation accuracy. To do so, we consider decision rules as base learners, and describe the heterogeneous treatment effect as a linear combination of these learners.

We first introduce decision rules with formal definition. Let S_k be the set of possible values of the kth covariate and $s_{k,m} \subseteq S_k$ be a specific subset corresponding to the mth rule. Then, each decision rule $r_m(x)$ can be represented as

$$r_m(x) = \prod_{k:s_{k,m} \neq S_k} \mathbb{1}(x_k \in s_{k,m}).$$

TABLE 1. Extracting rules from the example tree in Figure 1

Rules	Conditions
r_1	Male= 0
r_2	Male= 1
r_3	Male= 1 & Young= 0
r_4	Male= 1 & Young= 1

Define the covariate space $D = S_1 \times \cdots \times S_K$ as a Cartesian product of K sets. A vector of covariates $\mathbf{X}_i$ must lie in D for all i. Also, we define the subset D_m corresponding to the rule $r_m(x)$, $D_m = s_{1,m} \times \cdots \times s_{K,m}$. Then, $r_m(x) = 1$ if $x \in D_m$ and $r_m(x) = 0$ otherwise.

Decision rules as base learners are easily obtained from decision trees. For example, Figure 1 shows a toy example. The decision tree in this figure consists of several decision rules. Four decision rules can be extracted from this tree. For example, the young male group can be expressed as $r_4 = \mathbb{1}(\text{Male} = 1) \times \mathbb{1}(\text{Young} = 1)$. Other decision rules are listed in Table 1. We note that r_2 represents the internal node of the male group. Each decision rule corresponds to either internal or terminal nodes (leaves) except for the initial node (root). Including rules corresponding to internal nodes may increase the total number of decision rules to consider, but it can be helpful not to miss important decision rules that are essential for describing the variability of $\tau(x)$.

2.3. **Sample Splitting.** Analysis of heterogeneous treatment effects or subgroup analysis are typically conducted for subgroups defined *a priori* to avoid the cherry-picking problem that reports only subgroups with extremely high/low treatment effects (Cook et al., 2004). However, defining subgroups a priori requires a fairly good understanding of the treatment effect, probably from previous literature. On top of that, another problem is that researchers may miss unexpected subgroups. To overcome these limitations, we propose to use data-driven machine learning approaches combined with honest inference (Athey and Imbens, 2016). In particular, we use a sample-splitting approach that divides the total sample into two smaller subsamples (Athey and Imbens, 2016; Lee et al., 2018): (1) discovery and (2) inference subsamples. By using the discovery subsample, decision rules are generated, and among them, a few candidates are selected. These selected rules are regarded as rules given a priori when making statistical inference using the remaining inference subsample.

The main advantage of sample-splitting is to provide *transparency* that is foundational to reproducibility and replicability. The discovery step is performed only using the data in the discovery subsample, which enables researchers to avoid damaging validity of a later inference step. Also, this separated discovery step can be considered as a pre-analysis step to increase efficiency. When causal rules selected by the discovery step represent treatment effect heterogeneity well, in the context of estimation, methods with sample-splitting often perform as good as methods without sample-splitting, or perform better in high-dimensional settings.

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

- Randomly split the total sample into two smaller samples: the discovery and inference subsamples
- The Discovery Step (performed on the discovery subsample):
  (1) Rule generation (Section 3.1)
    (a) Estimate τ_i using existing methods for estimating the CATE
    (b) Use $(\hat{\tau}_i, \mathbf{X}_i)$ to generate a collection of trees through tree-ensemble methods
    (c) From the collection, extract decision rules $r_j$by removing duplicates
  (2) Rule regularization (Section 3.2)
    (a) Generate a new vector $\tilde{\mathbf{X}}_i^*$ whose jth component corresponds to rules r_j
    (b) Apply stability selection to $(\hat{\tau}_i, \tilde{\mathbf{X}}_i^*)$ and select potentially important decision rules $\tilde{\mathbf{X}}$.
- The Inference Step (performed on the inference subsample):
  (1) Estimate the CATE for the selected decision rules represented by $\tilde{\mathbf{X}}$ (Section 4.1)
  (2) Sensitivity analysis for the estimated CATE from the previous step (Section 4.2)

3. Discovery Step: Which Subgroups Are Potentially Important?

This section illustrates a step for discovering potentially important subgroups in terms of decision rules. In particular, we introduce a generic method that creates a set of decision rules and identifies the rule-generated structure. The discovery step consists of two parts: (1) rule generation and (2) rule regularization. For rule-generation, we create base learners that are building blocks to describe the heterogeneous structure of treatment effects. The rule regularization is used to choose *necessary* bulding blocks, and needed for interpretability and stability.

3.1. **Rule Generation.**

We use decision rules as base learners. Decision rules can be externally specified by using prior knowledge. However, we propose to employ a data-driven approach that uses hundreds of trees and extracts decision rules. This approach overcomes two drawbacks that

classical subgroup analysis approaches have: (1) they strongly rely on the subjective decisions on which are the heterogeneous subpopulations to be investigated; and (2) they fail to discover new rules other than the ones that are *a priori* defined by the researchers.

To create base learners for $\tau(x)$, we estimate τ_i first. This estimation is not considered for making inference, but designed for exploring the heterogeneous treatment effect structure. Many different approaches for the estimation of $\tau(x)$ can be considered. There has been methodological advancement in directly estimating $\tau(x)$ by using machine learning methods such as Causal Forest (Wager and Athey, 2018) or Bayesian Causal Forest (BCF) (Hahn et al., 2020). The BCF method directly models the treatment effect such as $\tau_i = m_0(\mathbf{X}_i; \hat{e}) + \alpha(\mathbf{X}_i; \hat{e})Z_i$ using the estimated propensity score $\hat{e}(\mathbf{X}_i)$. Hahn et al. (2020) shows that BCF produces precise estimates of $\tau(x)$. This evidence is consistent with recent works showing an excellent performance of Bayesian machine learning methodologies in causal inference scenarios (Hill, 2011; Hahn et al., 2018; Logan et al., 2019; Bargagli-Stoffi et al., 2019; Starling et al., 2019; Nethery et al., 2019). The BCF method can be applied to the discovery sample to estimate τ_i. We denote such estimates with $\hat{\tau}_i^{BCF}$. Alternative approaches for the estimation of $\tau(x)$ are discussed in Appendix B. Here, we want to highlight that the simulations' results show that the performance of the CRE algorithm in discovering the true causal rules is higher when we use BCF to estimate $\tau(x)$ (additional details on these simulations are provided in Appendix B).

Once the unit level treatment effect $\hat{\tau}_i$ is obtained, one can fit a decision tree using the data $(\hat{\tau}_i, \mathbf{X}_i)$. After fitting a tree, decision rules can be extracted as discussed above. However, using a single tree is neither an efficient nor a stable way to find important decision rules. This is due to two main factors, the *greedy* nature of tree-based algorithms and their lack of flexibility. Binary trees are greedy algorithms as they do not subdivide the population based on the overall best splits (the set of splits that would lead to the minimization of the overall criterion function), but they pick the best split at each step (the one that minimizes the criterion function at that particular step)[1]. Moreover, binary trees may not spot *simultaneous* patterns of heterogeneity in the data because of their binary nature. Imagine that the treatment effect differs by sex and high school diploma. The tree-based algorithm may be able to spot only one of the two drivers of heterogeneity

[1]Optimal trees (Bertsimas and Dunn, 2017) accommodate for this shortcoming at the cost of an extremely higher computational burden.

if one affects the treatment effect more than the other. Even when both the heterogeneity drivers are discovered they are spotted in a suboptimal way as an interaction between the two variables (i.e., women with a high school diploma, men with no diploma, etc).

To accommodate for these shortcomings of single trees, tree ensemble methods such as Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001) can be applied to obtain a collection of trees. Nalenz and Villani (2018) discuss that boosting and Random Forest are different in nature, thus using both approaches can generate a wider set of decision rules. Even after removing duplicate rules, this will lead to a more diverse set of candidate rules, thus will increases the probability to capture important rules. We follow Friedman and Popescu (2008) and Nalenz and Villani (2018) to use the same settings of the tuning parameters for gradient boosting and Random Forest. Also, we note that one should avoid using too lengthy (i.e., many conditions) or too many decision rules, which results in reducing interpretability.

3.2. **Rule Regularization and Stability Selection.** Denote $r_m(x)$ as the generated rules from Section 3.1, with $m = 1, \ldots, M^*$. Since each rule $r_m(x)$ indicates whether x satisfies the rule or not, it can take a value 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 $\tau(x)$ is. Although the original dataset $\mathbf{X}$ is not high-dimensional, $\tilde{\mathbf{X}}^*$ can be high-dimensional.

The set of the generated rules 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. Then, rule selection is applied to distinguish between few important rules and many insignificant ones. This regularization step improves understanding of the heterogeneity and increases efficiency. Such step improves interpretability: if too many rules are selected, this information may be too complex to understand. We consider the following linear regression model of the form,

$$\tau(x) = \beta_0 + \sum_{m=1}^{M^*} \beta_m r_m(x) + \epsilon \tag{3}$$

Since a linear model is considered, the model (3) lends a familiar interpretation of the coefficients $\{\beta_m\}_0^{M^*}$. Using this linear model, one can employ the following penalized regression to select

important rules:

$$\underset{\beta_m}{\operatorname{argmin}}\left\{\beta_0 + \sum_{m=1}^{M^*} \beta_m r_m(x)\right\} \text{ subject to } ||\beta_m||_p \leq \lambda \tag{4}$$

where λ is the regularization parameter and $||\cdot||_p$ is the l_p-norm. 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 (4) by employing the l_1-norm ($\sum_{m=1}^{M^*} |\beta_m|$). 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). Using a regularization parameter λ, we can see that the estimate $\hat{\boldsymbol{\beta}}$ shrinks toward to zero as λ increases, which provides a sparse structure close to the true model. Consistency of LASSO variable selection has been studied in Zhao and Yu (2006) and others. However, the biggest challenge is to choose a proper value of λ for consistent selection. Cross-validation is usually accompanied to choose λ. However, it may fail for high-dimensional data (Meinshausen and Bühlmann, 2006). Stability selection (Meinshausen and Bühlmann, 2010) can be used to enhance the performance of the LASSO estimator. Roughly speaking, by using a subsampling scheme, selection probabilities of variables (decision rules in our paper) can be estimated. Given two parameters (i.e., cut-off threshold and the average number of selected variables), variable selection can be done in a comparatively robust and transparent way. Also, Meinshausen and Bühlmann (2010) discussed that the solution of stability selection depends little on the initial regularization chosen, which is a desirable feature when selecting decision rules.

Among initially generated M^* decision rules, we assume that M (with $M \ll M^*$) decision rules are selected as an output from the discovery procedure. Then, we can define $\tilde{\mathbf{X}}$ as the matrix containing only the selected decision rules. Figure 2 depicts the intuition behind these steps of rules discovery and selection. The Figure shows a simple forest composed of just five trees. Each node of each tree (with the exclusion of the roots) represents a causal rule (*rules' discovery*), while the nodes highlighted in red represent the causal rules that are selected by the stability selection methodology (*rules' selection*).

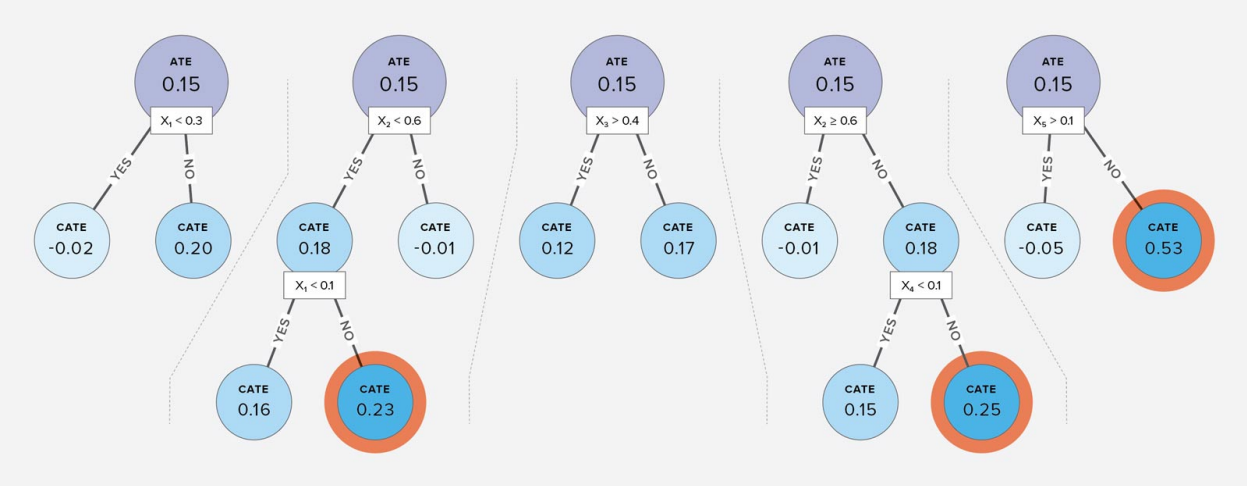

FIGURE 2. Rules' discovery and selection in a simple forest.

4. Inference Step: Which Subgroups Are Really Different?

Using the rules that were discovered and then selected in the discovery step, the remaining inference sample is used to make inference. We propose a generic approach to estimate the rule-specific treatment effect, and compare it with existing approaches. Furthermore, we propose a new sensitivity analysis method to assess the impact of unmeasured confounding bias on causal conclusion.

4.1. Estimating the subgroup-specific treatment effect.

Decision rules are selected through the discovery step by using the linear model in (3),

$$\boldsymbol{\tau} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with $\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{x}) = 0$ and $\text{var}(\boldsymbol{\epsilon}|\mathbf{x}) = \sigma^2\boldsymbol{I}$. The main goal of this inference step is to estimate $\boldsymbol{\beta}$ that represents rule-specific treatment effects. We estimate these CATE for the subpopulations corresponding to the selected rules. We use ordinary least squares (OLS) to estimate $\boldsymbol{\beta}$. However, this estimator cannot be obtained as the unit level treatment effect $\boldsymbol{\tau}$ is not known. Hence, we define a new vector $\boldsymbol{\tau}^* = (\tau_1^*, \ldots, \tau_N^*)^T$ that is an estimate of $\boldsymbol{\tau}$. The newly defined $\boldsymbol{\tau}^*$ is just an intermediate value used in the inference step. Although $\boldsymbol{\tau}^*$ is an estimate, to distinguish this with the fitted value obtained by using the linear model, we save hat notation.

The estimate τ_i^* can be viewed as an observable quantity with some sampling error although τ_i^* is an estimated value. The quantity τ_i^* can be represented by

$$\tau_i^* = \tau_i + u_i \quad \text{where} \quad \mathbb{E}(u_i|\mathbf{x}_i) = 0 \text{ and } \text{var}(u_i|\mathbf{x}_i) = w_i. \tag{5}$$

In general, the variance w_i is not constant across all individuals. By combining it with the model (3), the modified linear model is obtained as

$$\tau_i^* = \beta_0 + \sum_{j=1}^{M} \beta_j \tilde{X}_{ij} + \nu_i = \tilde{\mathbf{X}}_i \boldsymbol{\beta} + \nu_i, \tag{6}$$

where $\tilde{\mathbf{X}}_i$ is the ith row of $\tilde{\mathbf{X}}$ and $\nu_i = \epsilon_i + u_i$. Lewis and Linzer (2005) considered a similar linear model like the model (6) and found via simulation studies that the OLS estimator performs well in many cases. Also, they found that since the error ν_i is not homoscedastic, considering White or Efron's heteroscedastic-consistent standard error can provide an efficient estimate. Based on these findings, we also consider the OLS estimator for $\boldsymbol{\beta}$ that can be defined as

$$\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\tau}^*. \tag{7}$$

Also, the fitted value $\hat{\boldsymbol{\tau}}$ is defined as $\hat{\boldsymbol{\tau}} = \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\boldsymbol{\tau}^*$.

The intermediate vector $\boldsymbol{\tau}^*$ can be chosen in various ways. However, to validate our estimator (7), each τ_i^* has to be unbiased with finite variance. Given that the matrix $\tilde{\mathbf{X}}$ is fixed, we can prove that the estimator $\hat{\beta}_j, j = 1, \ldots, M$ is a consistent estimator of β_j that is the average treatment effect for the subgroups defined by the decision rule r_j if r_j is not included in another rule $r_{j'}$.

Theorem 1. *If τ_i^* satisfies the model (5) (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 $\boldsymbol{\beta}$.*

An example is to use the inverse probability weighting (IPW) or stabilized IPW (SIPW) approaches. Both $\tau_i^* = \hat{\tau}_i^{IPW}$ and $\tau_i^* = \hat{\tau}_i^{SIPW}$ satisfy the model (5). 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 $\tau_i^* = \hat{\tau}_i^{(S)IPW}$ 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*

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

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 $\hat{\nu}_i = \tau_i^* - \tilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}$. This estimator is robust and often called the White's estimator (White, 1980). There are other approaches to obtain a heteroscedasticity consistent covariance matrix, which is 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 τ_i^*, then feasible generalized least squares estimators (Lewis and Linzer, 2005) can be considered.

Instead of estimation, hypothesis testing for identifying true decision rules can be considered. As we will see in a simulation study, true rules representing treatment effect heterogeneity are discovered and selected with high probability. However, at the same time, non-true rules are selected with high probability. If one wants to know which rules are really important for describing treatment effect heterogeneity, variable selection can be used to choose a final model for treatment effect heterogeneity. For instance, the null hypothesis $H_0 : \boldsymbol{\beta} = 0$ can be tested by using a Wald-type test statistic $T_n = n\hat{\boldsymbol{\beta}}^T\hat{\mathbf{V}}_n^{-1}\hat{\boldsymbol{\beta}}$. If $T_n > \chi^2_{M,1-\alpha}$, H_0 is rejected.

4.2. **Sensitivity Analysis.** The validity and consistency of the estimator $\hat{\boldsymbol{\beta}}$ rely on the assumption of no unmeasured confounders and correct specification of the propensity score model. However,

in reality, there is no guarantee that these assumptions are satisfied. Also, such assumptions are not directly testable. Without further knowledge about unmeasured confounding, it is extremely difficult to quantify how much bias can occur. Instead of attempting to quantify the degree of unmeasured confounding in the given dataset, it is more realistic to see how our causal conclusion will change with respect to various degrees of such bias. In this subsection, we propose a sensitivity analysis to examine the impact of potential violations of the no unmeasured confounding assumption. We introduced a generic approach for estimating $\boldsymbol{\beta}$ in Section 4.1. However, in our sensitivity analysis, we consider the special case where τ_i^* is estimated as $\hat{\tau}_i^{SIPW}$. Define $\mathbf{W} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$ that is a $M \times N$ 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

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

We consider the marginal sensitivity model that was introduced by Tan (2006) and Zhao et al. (2019). Let the true propensity probability $e_0(\mathbf{x}, y; a) = \mathrm{P}_0(Z = 1|\mathbf{X} = \mathbf{x}, Y(a) = y)$ for $a \in \{0, 1\}$. If the assumption of no unmeasured confounders holds, this probability would be the same as $e_0(\mathbf{x}) = \mathrm{P}_0(Z = 1|\mathbf{X} = \mathbf{x})$ that is identifiable from the data. Unfortunately, this assumption cannot be tested since $e_0(\mathbf{x}, y; a)$ is generally not identifiable from the data. For each sensitivity parameter Λ that will be introduced in detail later, the maximum deviation of $e_0(\mathbf{x}, y; a)$ from the identifiable quantity $e_0(\mathbf{x})$ is restricted, and sensitivity analysis is conducted for each Λ to see if there is any qualitative change of our conclusion. In addition to this non-identifiability of $e_0(\mathbf{x}, y; a)$, 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.

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

(1) Generate a matrix $\mathbf{W} = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$
(2) In the ℓth of L iterations:
   (a) Generate a bootstrapped sample: $(Z_i^{(\ell)}, Y_i^{(\ell)}, \mathbf{X}_i^{(\ell)})_{i=1,\ldots,N}$.
   (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 $\tilde{Y}_{j,N_1+1} \geq \ldots \geq \tilde{Y}_{j,N}$
   (d) Compute an estimate $\hat{\gamma}^{(\ell)}$ by fitting the logistic regression with $(Z_i^{(\ell)}, \mathbf{X}_i^{(\ell)})$
   (e) Solve the following optimization problems:

$$\min \text{ or } \max \frac{\sum_{i=1}^{N_1} \tilde{Y}_{ji}^{(\ell)}[1 + q_i \exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=1}^{N_1}[1 + q_i \exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]} - \frac{\sum_{i=N_1+1}^{N} \tilde{Y}_{ji}^{(\ell)}[1 + q_i \exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=N_1+1}^{N}[1 + q_i \exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}$$

   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 $U_j = Q_{1-\alpha/2}\left(U_j^{(\ell)}\right), \ell = 1, \ldots, L$

Our sensitivity model assumes that the true propensity probability $e_0(\mathbf{x}, y; a) = \mathrm{P}_0(Z = 1|\mathbf{X} = \mathbf{x}, Y(a) = y)$ satisfies:

$$e_0(\mathbf{x}, y; a) \in \mathcal{E}_{\gamma_0}(\Lambda) = \{0 < e(\mathbf{x}, y; a) < 1 : 1/\Lambda \leq OR\{e(\mathbf{x}, y; a), e_{\gamma_0}(\mathbf{x})\} \leq \Lambda\} \quad \text{for } a \in \{0, 1\} \quad (8)$$

where $e_{\gamma_0}(\mathbf{x}) = \mathrm{P}_{\gamma_0}(Z = 1|\mathbf{X} = \mathbf{x})$ and $OR\{e(\mathbf{x}, y; a), e_{\gamma_0}(\mathbf{x})\} = \{(1-e(\mathbf{x}, y; a)) \cdot e_{\gamma_0}(\mathbf{x})\}/\{e(\mathbf{x}, y; a) \cdot (1-e_{\gamma_0}(\mathbf{x}))\}$. The deviation of $e_0(\mathbf{x}, y; a)$ 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 $\Lambda \geq 1$. When $\Lambda = 1$, $e_0(\mathbf{x}, y; a) = e_{\gamma_0}(\mathbf{x})$ for all a, which implies that there is no violations of the assumptions. If the propensity score model is correctly specified, $e_{\gamma_0}(\mathbf{x}) = e_0(\mathbf{x})$. Also, if there is no unmeasured confounder, $e_0(\mathbf{x}, y; a) = e_0(\mathbf{x})$. Therefore, under the assumptions of correct propensity score model and no unmeasured confounder, $e_0(\mathbf{x}, y; a) = e_{\gamma_0}(\mathbf{x})$. Our sensitivity model considers violations of both assumptions. This sensitivity analysis 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).

The $(1-\alpha)$-coverage confidence interval of β_j can be constructed by using the percentile bootstrap. First, we denote $W_{ji}Y_i$ by $\tilde{Y}_i^j$ and treat $\tilde{Y}_i^j$ as if it is an observed outcome, then the confidence interval for each β_j can be constructed 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 $[L_j, U_j]$ can be proved by using Theorem 1 in Zhao et al. (2019). In Step (2e) of Algorithm 2, the optimization problem can be efficiently solved by separating two simpler optimization problems. The minimum is obtained when the first part $\frac{\sum_{i=1}^{N_1}\tilde{Y}_{ji}^{(\ell)}[1+q_i\exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=1}^{N_1}[1+q_i\exp\{-\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}$ is minimized and the second part $\frac{\sum_{i=N_1+1}^{N}\tilde{Y}_{ji}^{(\ell)}[1+q_i\exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}{\sum_{i=N_1+1}^{N}[1+q_i\exp\{\hat{\gamma}^{(\ell)}\mathbf{X}_i^{(\ell)}\}]}$ is maximized. For instance, the minimization of the first part is achieved at $\{q_i : q_i = 1/\Lambda \text{ for } i = 1,\ldots,b, \text{ and } q_i = \Lambda \text{ for } i = b+1,\ldots,N_1\}$ for some b. To find the minimum, it is required to check every possible value of b, and the computational complexity is $O(N_1)$. See Proposition 2 in Zhao et al. (2019) for more details.

5. Simulation

In this section, we introduce two simulation studies to assess the performance of the CRE method. In the first simulation study, we evaluate the discovery step in terms of how well the method performs in discovering the true underlying decision rules. In the second simulation study, we evaluate the overall performance of both the discovery and inference steps in terms of estimation accuracy.

5.1. **Simulation study: Rules Discovery.**

In order to evaluate the ability of CRE in the discovery step we run a series of simulations in which we assess how many times CRE can spot the true underlying decision rules. First, we assess the *absolute* performance of CRE; then we compare its performance with the one of 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. In order to evaluate the ability of discovering decision rules we consider, among the discovered rules, how many times true rules are captured.

As we discussed in Section 3, to apply the CRE method, many approaches can be considered to estimate the individual treatment effect τ_i. We have found via simulation studies that the BCF approach has better performance than other approaches such as $\hat{\tau}_i^{IPW}$ or BART (in Appendix B,

we show the comparative performances of BCF, BART, IPW and outcome regression). Thus, in this simulation study, we implement the BCF approach for the CRE method. We call this version of the CRE method *CRE-BCF*. Also, 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, $Z_i \sim Binom(\pi_i)$ where $\pi_i = \text{logit}(-1 + X_{i1} - X_{i2} + X_{i3})$. Finally, the output is generated 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_{i1}, X_{i2} and 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.

In order to produce a more meaningful comparison between CRE-BCF and HCT, we restricted the data generating process to causal rules that are representable 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 $Y_i(1) = Y_i(0) + \tau(\mathbf{X}_i)$. For the case of two causal rules, $\tau = \tau(\mathbf{X}_i) = k$ 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. 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 the effect modifiers are different from the confounders. Investigating this scenario is important since confounders may affect the ability of the algorithm to spot the correct causal rules. Figures 3 and 4 depict the results in the case with the same variables for both confounders and effect modifiers and in the case of different variables, respectively.

When (X_1, X_2, X_3) are both confounders and effect modifiers, CRE-BCF consistently outperforms HCT, while in the other scenario CRE-BCF is still performing better but the performance gap is narrower especially in the case with two true causal rules. In both the scenarios, the gap is wider as the number of true rules is four. This shows that HCT is consistently unable to detect all the four true rules. These simulations show that the usage of CRE-BCF results in an increase in the detection of the true causal rules especially when there is overlap between the confounders and effect modifiers. This is a very interesting feature of CRE-BCF as similar scenarios are very

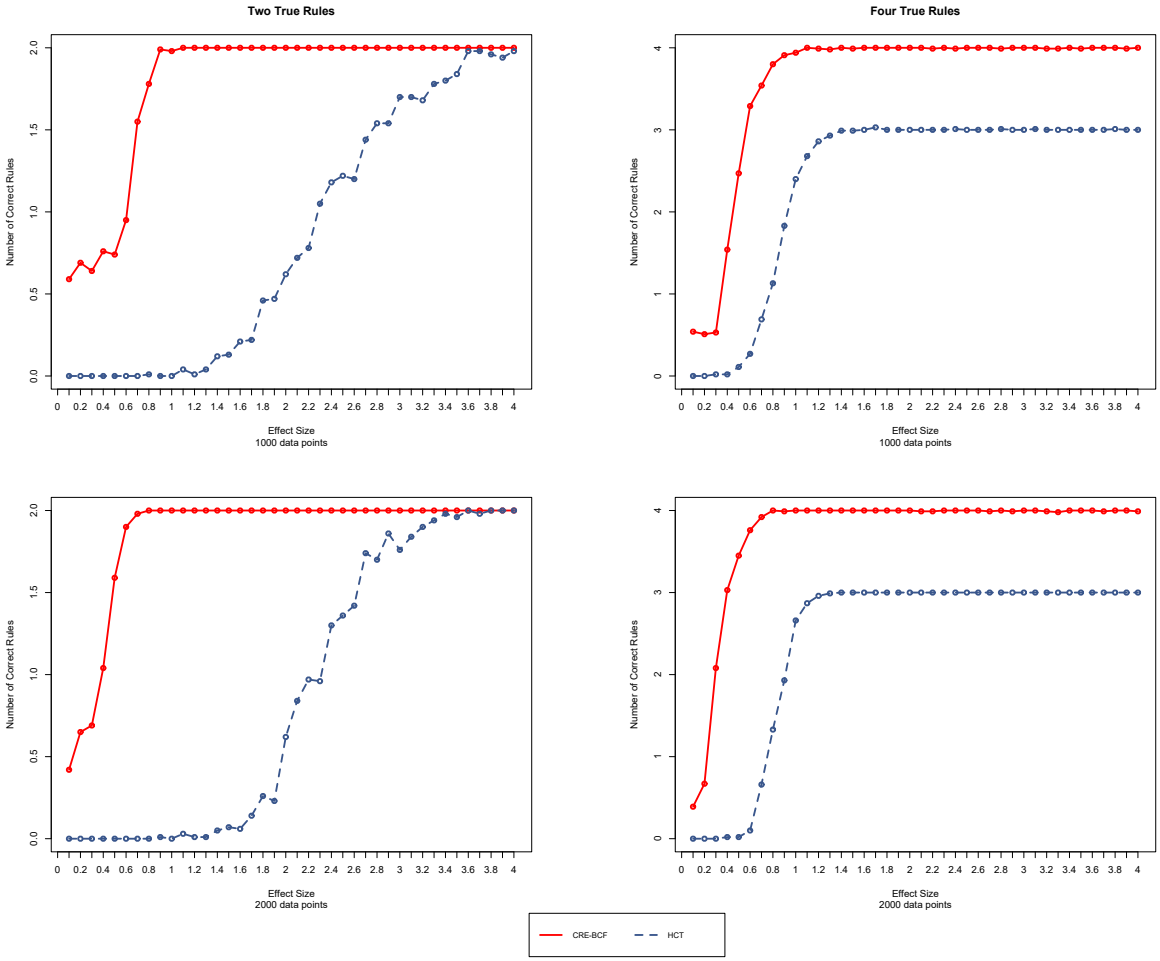

FIGURE 3. Average number of correctly discovered rules in the case where the same variables are used as confounders and effect modifiers. The first column depicts the case of two true rules while the second column the case of four true rules. In the first row the sample size is 1,000 while in the second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

likely in 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. 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).

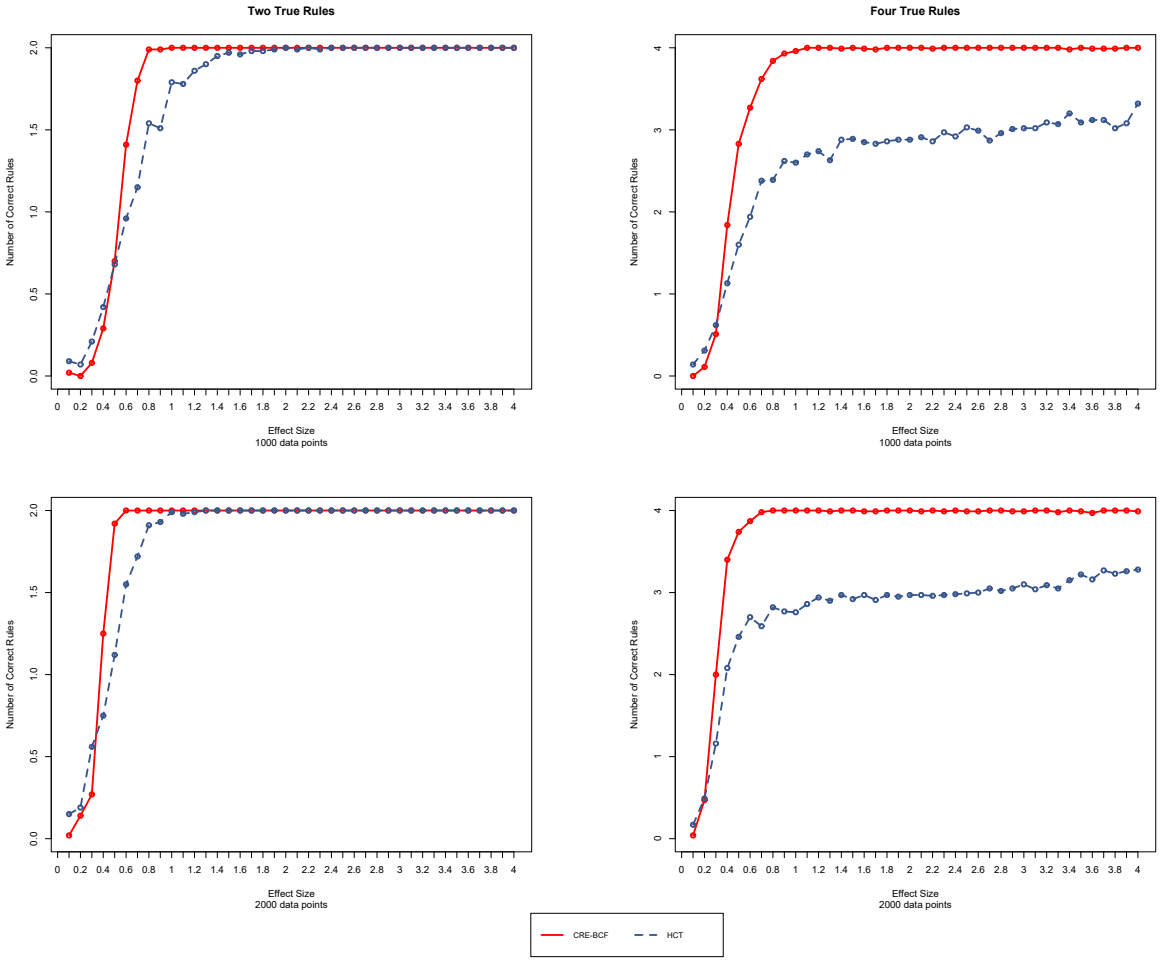

FIGURE 4. Average number of correctly discovered rules in the case where the confounders are different from the effect modifiers. The first column depicts the case of two true rules while the second column the case of four true rules. In the first row the sample size is 1,000 while in the second it is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

5.2. **Simulation study: Rule-specific Effects Estimation.** In this subsection, we evaluate the overall performance of the CRE method including both the discovery and inference steps. In the previous simulation study, we found that the BCF approach shows a great performance in discovering underlying decision rules. Also, it has been shown that the BCF approach estimates τ_i with great accuracy heuristically. Thus, 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,

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

$(\mathbf{X}_i, \hat{\tau}_i^{BCF})$ is used in order to discover and select important decision rules. During the inference step, the intermediate variable τ_i^* is estimated by using the BCF approach (CRE-BCF).

To compare, we 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 set of decision 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. (2018) 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 (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, 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 shows the performance comparison using RMSE. We consider 1000 simulated datasets 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.

6. Application to the Medicare data

We apply the proposed CRE method to the Medicare data in order to study the effect of long-term exposure to fine particulate matter ($PM_{2.5}$) on 5-year mortality. Lee et al. (2018) studied the treatment effect of exposure to $PM_{2.5}$ with 110,091 matched pairs and discovered treatment effect heterogeneity in a single tree structure by using the HCT approach. However, as pointed out in their discussion, a single tree unavoidably contains subgroups that are not informative for describing treatment effect heterogeneity due to the nature of tree algorithms. In particular, they found six disjoint subgroups, but only three of them are informative to describe the treatment effect heterogeneity. Also, as we discussed and showed in our simulation study, the HCT approach can be affected by sample-to-sample variations.

For a brief overview of the matched Medicare data, it contains 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 $PM_{2.5}$ is greater than 12 $\mu g/m^3$. The outcome is five-year mortality measured between 2002-2006. For an individual, the outcome is 0 if he/she was died before the end of 2006 and 1 otherwise. There are four individual level covariates - sex (male, female), 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 variable indicating socioeconomic status. If an individual is eligible for Medicaid, it is highly likely that he/she has lower household income, thus we use this variable as a proxy for low income. In the matched data, these four variables are exactly matched. There are also 8 ZIP code-level or 2 county-level covariates, and they are fairly balanced between the treated and control groups, see Table 2 in Lee et al. (2018) for the covariate balance. Namely, the additional variables used are body-mass-index (BMI), 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

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)

occupied housing and population density. Also, see Di et al. (2016) for general description about the Medicare data.

We apply the CRE method to the same discovery and inference samples that are split by using a (25%, 75%) ratio and contain 27,500 and 82,591 matched pairs respectively. Since two individuals in a matched pair share the same covariates, but experience different treatment values, the observed outcomes can be considered as two potential outcomes for a hypothetical individual that represents the corresponding matched pair. The treated-minus-control difference can be considered as an estimate of τ_i for matched pair i. However, since our outcome is binary, the estimate $\hat{\tau}_i^{Matching}$ can take only one of three possible values $\{-1, 0, 1\}$. This discrete feature is undesirable under the linear regression model (3). Instead, we ignore the matched structure and use $27500 \times 2 = 55000$ individuals in the discovery sample as if they were obtained independently. We use the logistic BART approach to estimate potential outcome functions $m_z(x) = \mathbb{E}[Y_i(z)|\mathbf{X}_i = x]$, $z = 0, 1$. Then, the estimate $\hat{\tau}_i^{BART} = \hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)$ is obtained. Note that for a continuous outcome, as shown in the simulation study, the BCF approach can be considered.

In the discovery step with the estimate $\hat{\tau}_i^{BART}$, we first apply the CRE method only with four individual-level covariates. Four decision rules, r_1, r_2, r_3, r_4, are discovered. Table 3 shows the descriptions for these rules on the left column. Based on this finding, the model $\tau(x) = \beta_0 + \sum_{j=1}^{4} \beta_j r_j(x)$ is considered for the later inference step. The first rule is for non-white people that is only 7% of the total population. The next three rules are defined by age, and r_2 is included in r_3. The intercept β_0 represents the treatment effect for the subgroup of Medicaid eligible white people

aged between 81-85 and white people aged above 85. This subgroup corresponds to the single tree finding in Lee et al. (2018).

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. (2018) 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 means 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 means a subgroup of r_5 with the additional condition that the population density is below the average.

Before making inference with the discovered rules, we want to emphasize two aspects in the discovery step of the CRE method. First, since the CRE method discovers decision rules instead of a whole tree, only important subgroups (decision rules) are selected, and other subgroups are represented by the intercept. For example, Lee et al. (2018) discovered six subgroups, but only a few of them are informative for describing the treatment effect heterogeneity. Second, the discovered rules are stable. Being stable means that if another discovery sample is chosen, the discovered rules are hardly changed while a discovered tree varies with its size and terminal nodes. This robustness to sample-to-sample variation makes findings replicable. This feature is extremely significant because higher standards of reproducibility are deemed important in the context of open science.

Next, with the discovered decision rules, we use the remaining 82,591 pairs in the inference sample to estimate the rule-specific treatment effects. For the first rule set $\{r_1, \ldots, r_4\}$, the corresponding coefficients are estimated and reported in Table 3. To obtain the estimates and 95% confidence intervals, we consider $\tau_i^* = \hat{\tau}_i^{SIPW}$ to claim the asymptotic normality and obtain the 95% confidence intervals using the asymptotic distributions. Causal Forest method (Athey et al., 2019) can be also considered since it guarantees the asymptotic normal distribution for $\boldsymbol{\beta}$. Note that the logistic BART approach can be considered as the discovery step, but the asymptotic properties are not guaranteed. On the first part of the right column of Table 3, all the coefficients except for the

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)

intercept have the negative sign. The intercept indicates that individuals who do not belong to the discovered rules $\{r_1, r_2, r_3, r_4\}$ are significantly affected by exposure to air pollution. There is a 7 percentage point increase of mortality rate. Though the estimates for r_1, 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 others. 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.

Also, we want to emphasize the interpretation of the coefficients in the inference step. A non-significant estimate does not mean that the corresponding subgroup has the null treatment effect. In this context, non-significance means that the decision rule is no longer important for describing treatment effect heterogeneity. For instance, consider a subgroup of Black people above 85. This groups belongs to r_1 only. The estimate $\hat{\boldsymbol{\beta}}_1$ is shown as not significant. However, the treatment effect for this subgroup is $\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 = 0.062$ with the 95% CI (0.041, 0.084) meaning that this subgroup is significantly affected by air pollution. For another example, a subgroup of Black people below 75 has the effect $\hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\beta}}_3 = 0.023$ (95% CI: (0.003, 0.043)), which is still significant. Since we have the asymptotic distribution for $\boldsymbol{\beta}$, any subgroup's treatment effect can be examined.

Finally, to evaluate the robustness of the above finding about the treatment effect heterogeneity, we conduct sensitivity analysis in the inference step by setting $\tau_i^* = \hat{\tau}_i^{SIPW}$. We use the following model for sensitivity analysis, $\tau_i = \beta_0 + \sum_{j=2}^{6} \beta_j r_j$. Under the sensitivity model (8), for each Λ, we obtain several sets of 95% CIs for the coefficients. 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 ratio scale, then our finding about the heterogeneity can be explained by this unmeasured bias. One further thing to note is that 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) is significant.

7. Discussion

This paper proposes a new data-driven method for studying treatment effect heterogeneity that notably improves interpretability and provides helpful guidance about subgroups with heterogeneous effects in terms of decision rules. Moreover, the proposed CRE methodology accommodates for well-known shortcomings of binary trees by providing a more stable, flexible and robust methodology to discover and estimate heterogeneous effects. 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.

Though the CRE method makes inference using a smaller inference subsample due to sample-splitting, it maintains estimation precision at a similar level as other existing methods while providing an interpretable form of the treatment effect heterogeneity. The CRE method is a generic method that is completely compatible with existing methods for estimating 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 τ_i^* 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. Indeed, a few simple rules are utterly important for public policy implications. However, when it comes

to precision medicine, discovering a possibly lengthy rule that is specific to a patient could be of interest.

The proposed CRE method requires a researcher to specify some of the so-called tuning parameters. In particular, one should choose a number of trees that are generated for extracting decision rules, and the cut-off threshold in stability selection during the discovery step. Previous studies show that the performance is not much affected by specification of the parameters. Also, the studies provide a general guidance of how to specify the parameters. However, the optimal choice of the splitting ratio between the discovery and inference samples is not known yet. Even though the ratio (25%, 75%) is shown to have the best performance through simulation studies, we do not know whether this ratio works best for all real-world datasets. It may be possible to require a larger proportion of the discovery sample if data is sparse. On the contrary, a smaller proportion is needed when the underlying effect structure is simple.

Regardless of the splitting ratio, it is more important to choose a proper number of decision rules during the discovery step. The choice may depend on the questions that practitioners want to answer. For example, public policy makers 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. Instead, a comparatively larger set of decision rules can be chosen, for instance, in precision medicine. 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.

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. A 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). Resursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*.

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*.

Bellman, R. E. (1961). *Adaptive control processes: a guided tour*. Princeton University Press.

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.

Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, 106(7):1039–1082.

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

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., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., and Schwartz, J. (2016). Assessing PM2.5 exposures with high spatiotemporal resolution across the continental United States. *Environmental science & technology*, 50(9):4712–4721.

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

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., Dorie, V., and Murray, J. S. (2018). Atlantic causal inference conference (acic) data analysis challenge 2017.

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.

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

Lee, K., Small, D. S., and Dominici, F. (2018). Discovering effect modification and randomization inference in air pollution studies. *arXiv preprint arXiv:1802.06710*.

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.

Logan, B. R., Sparapani, R., McCulloch, R. E., and Laud, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using bayesian additive regression trees. *Statistical Methods in Medical Research*, 28(4):1079–1093.

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. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

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., Dominici, F., et al. (2019). Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The Annals of Applied Statistics*, 13(2):1242–1267.

Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments, section 9. *Roczniki Nauk Roiniczych*, X:1–51. reprinted in Statistical Science, 1990, 5, 463–485.

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.

Starling, J. E., Murray, J. S., Lohr, P. A., Aiken, A. R., Carvalho, C. M., and Scott, J. G. (2019). Targeted smooth bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation. *arXiv preprint arXiv:1905.09405*.

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.

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

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

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

Appendix A. Proofs

A.1. **Proof of Theorem 1.** By multiplying $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$ on the both sides of model (6), 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, $\text{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$.

A.2. **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 $N(0, \mathbf{Q}^{-1}\Omega\mathbf{Q}^{-1})$.

Appendix B. Comparison between BCF, BART, IPW, OR for rule discovery

As we discussed in Section 3, 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

$$\hat{\tau}_i^{IPW} = \left(\frac{Z_i}{\hat{e}(\mathbf{X}_i)} - \frac{1 - Z_i}{1 - \hat{e}(\mathbf{X}_i)}\right) Y_i \tag{9}$$

where $\hat{e}(\mathbf{X}_i)$ is the estimate of the propensity score $e(x)$ at $\mathbf{X}_i$. The estimate $\hat{e}(\mathbf{X}_i)$ can be obtained by fitting a logistic regression on $(Z_i, \mathbf{X}_i)$. The estimator is validated based on the identification result (2). 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 $\hat{e}(\mathbf{X}_i)$ 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),

$$\hat{\tau}_i^{SIPW} = \left\{ \left(\frac{1}{N}\sum_{i=1}^N \frac{Z_i}{\hat{e}(\mathbf{X}_i)}\right)^{-1} \frac{Z_i}{\hat{e}(\mathbf{X}_i)} - \left(\frac{1}{N}\sum_{i=1}^N \frac{1 - Z_i}{1 - \hat{e}(\mathbf{X}_i)}\right)^{-1} \frac{1 - Z_i}{1 - \hat{e}(\mathbf{X}_i)} \right\} Y_i. \tag{10}$$

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 $\hat{m}_1(\mathbf{X}_i)$. The unit level treatment effect can be estimated by either subtracting the observed outcome and imputed counterfactual, i.e. $\tau_i^{OR} = Y_i^{obs} - (\hat{m}_1(\mathbf{X}_i) \cdot (1 - Z_i) + \hat{m}_0(\mathbf{X}_i) \cdot Z_i)$, or by subtracting the imputed potential outcomes, i.e., $\tau_i^{BART} = \hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)$. 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): (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 Subsection 3.2 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.

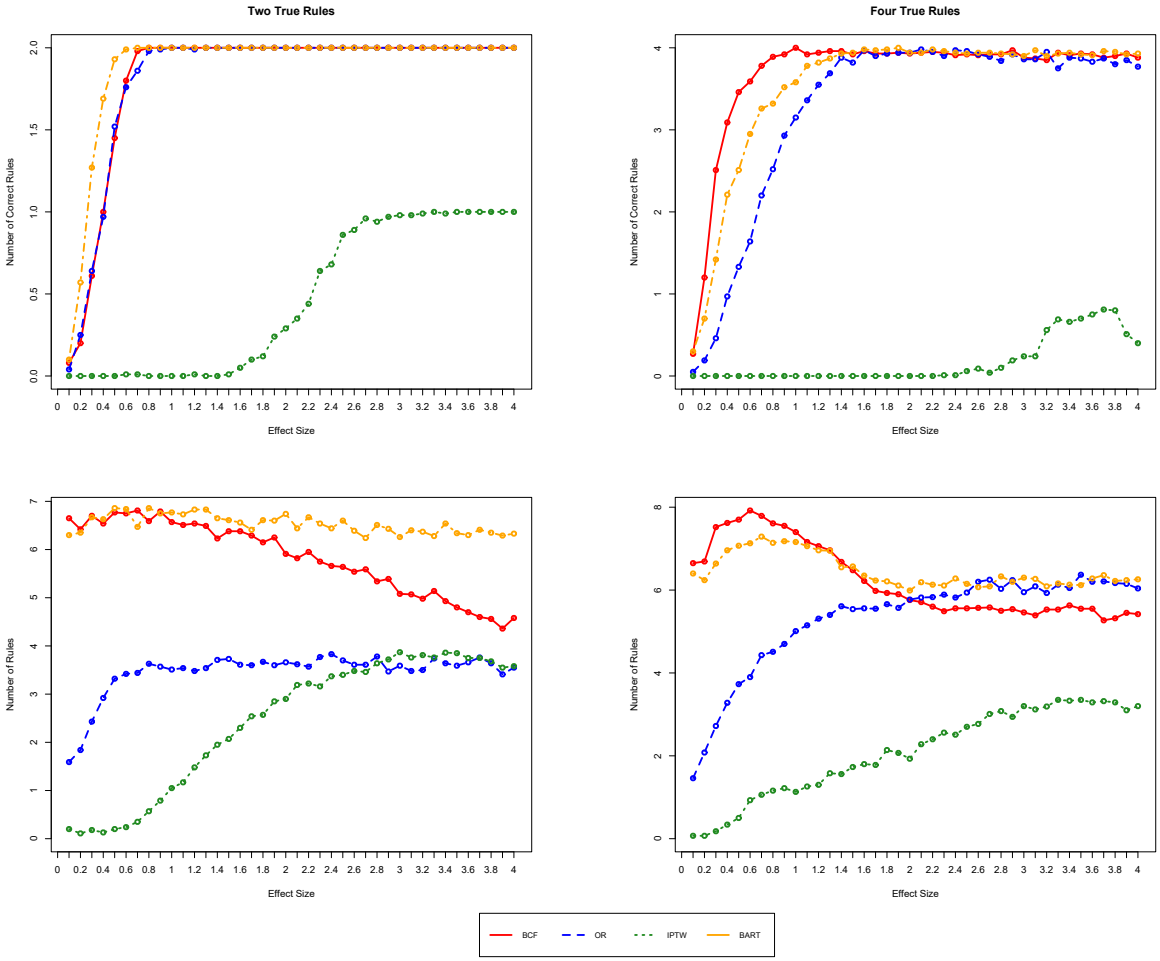

FIGURE 5. Comparison between BCF, BART, IPW, OR for rules' discovery. The plots show the variation in the number of correct rules (first row) and the number of detected rules (second row) as the effect size increases, in the cases of two true rules (first column) and four true rules (second column).

APPENDIX C. 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

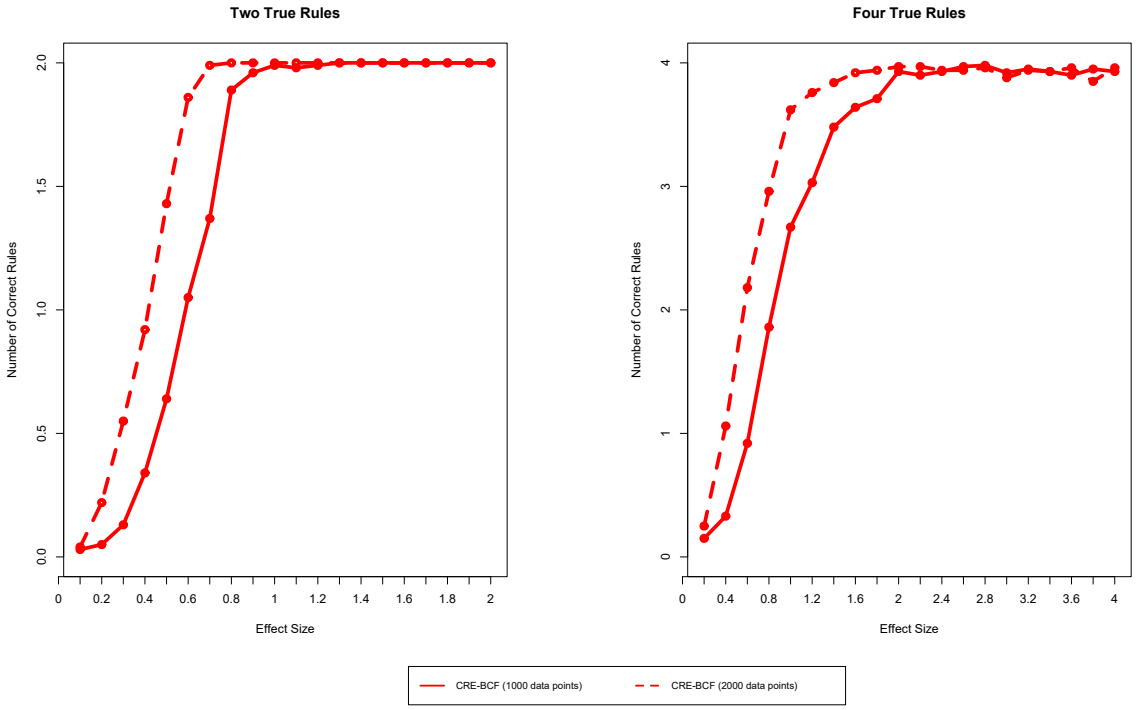

FIGURE 6. Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).

not sizable as the performance of CRE-BCF is excellent also with the smaller sample size. Table 5 depicts the performance of CRE-BCF for the simulation scenario introduced in 5.1. 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 D, 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.

APPENDIX D. 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 5. 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 3.2 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.

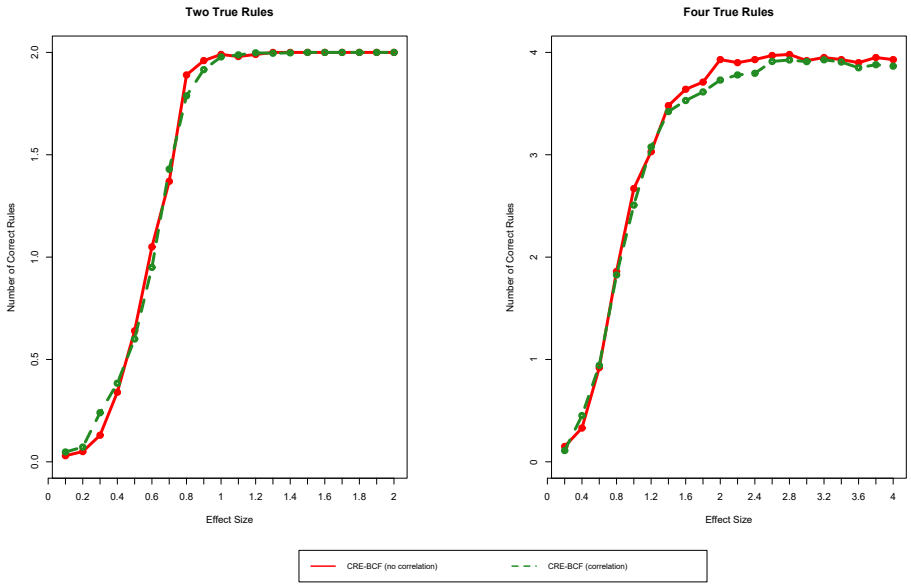

FIGURE 7. Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).

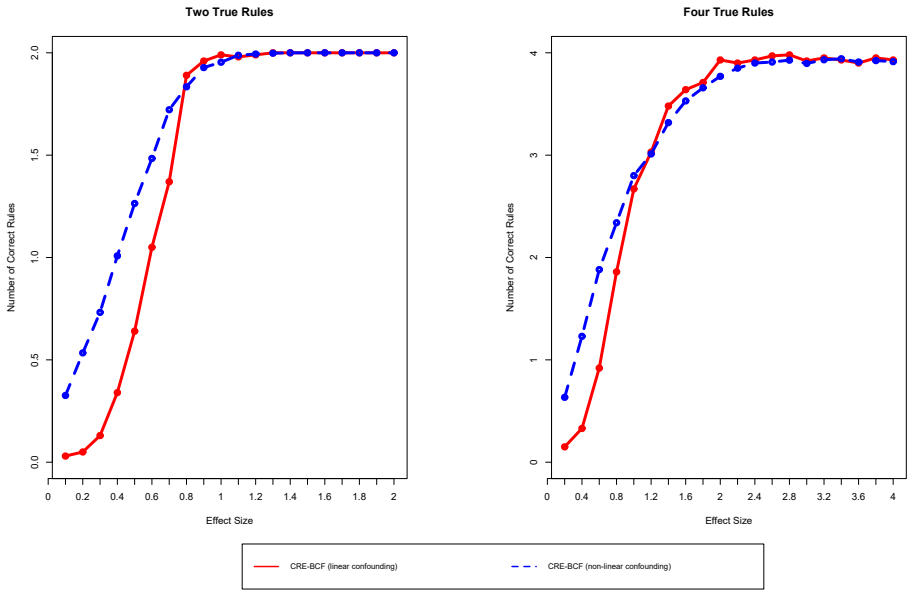

FIGURE 8. Average number of correctly discovered rules for CRE-BCF in the case of two true rules (left panel) and four true rules (right panel).