Identifying Responders in Randomized Controlled Trials and Observational Studies

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we introduce causal responder detection (CARD), a novel method for responder analysis that identifies treated subjects whose outcomes are statistically different from the control response distribution. Leveraging recent advances in conformal prediction for novelty detection, CARD builds on the AdaDetect framework and, in randomized settings, inherits its finite-sample false discovery rate (FDR) control when coupled with the Benjamini–Hochberg procedure. For observational studies, we propose a propensity score–adjusted extension and establish asymptotic FDR control under standard causal assumptions of ignorability and overlap. Simulation studies and real-data applications demonstrate that CARD effectively detects responders with high power across a range of heterogeneous and distributional treatment-effect scenarios.

1 Introduction

Personalized medicine is expected to advance healthcare in the near future (Vicente et al., 2020). In contrast to a one-size-fits-all approach, personalized medicine advocates for treatments tailored to individual patients based on their clinical characteristics.

Responder analysis in clinical trials is a method used to evaluate the effectiveness of a treatment by identifying and analyzing the subset of participants who respond significantly to the treatment (Henschke et al., 2014). This approach contrasts with the traditional method of evaluating average effects across all participants, which can sometimes obscure the benefits seen in a particular group of responders (Guyatt et al., 1998). When the treatment effect is heterogeneous it is important to identify and understand which individuals respond better or worse to the treatment. Responder analysis is a common practice in clinical trials (Moore et al., 2010; Straube et al., 2010; Chuang et al., 2022). In the context of this manuscript, we focus on statistical evidence that a treated subject's outcome distribution deviates from the control distribution, rather than exceeding a predefined clinically meaningful threshold (Snapinn & Jiang, 2007). We refer to such subjects as responders throughout.

Another approach to understanding heterogeneous treatment effects is estimating the conditional average treatment effect (CATE), which focuses on understanding how the effect of a treatment varies across different segments of the population, rather than only calculating the average treatment effect (ATE). Unlike ATE, which provides a single overall effect, CATE enables the identification of treatment benefits that may be specific to subgroups defined by their unique characteristics (Angus & Chang, 2021).

CATE and responder analysis complement each other. CATE offers insights into the heterogeneous treatment effects in terms of the expected values of certain population segments. Meanwhile, responder analysis provides additional context by indicating the proportion of total responders and extends beyond the expected value. For instance, consider a scenario where a certain treatment increases the variability within a subgroup of the population without necessarily affecting their expected value. Since CATE estimates only the conditional expectation of the treatment effect, it is by construction unable to detect such distributional changes. By contrast, subjects in this subgroup could be identified as responders by a method sensitive to the full outcome distribution.

Individual treatment effects (ITE) provide a comprehensive view of the treatment response. Lei & Candès (2021) and Alaa et al. (2024) proposed methods for constructing confidence intervals for the ITE using split conformal prediction (Papadopoulos, 2008). Lei & Candès (2021) used weighted conformal prediction for combining the confidence intervals for the counterfactual to the ITE confidence interval. By contrast, Alaa et al. (2024) suggested a method based on inductive conformal prediction for specific types of meta-learners.

These confidence intervals offer more information than the CATE point estimate alone. When these confidence intervals are sufficiently powered they can be used for responder analysis, where subjects with confidence intervals that exclude 0 can be determined as responders. However, there are certain settings where they lack the statistical power to detect differences between treated and untreated responses (see Section 4). Duan et al. (2024) suggested an approach to test if the CATE of observations is non-negative (or positive) while controlling the false discovery rate (FDR) over all observations.

CARD (causal responder detection), our proposed responder analysis method, builds on AdaDetect (adaptive detection, Marandon et al. (2024)). AdaDetect is a conformal prediction method for detecting out-of-distribution samples using an adaptive non-conformity score. CARD offers two key innovations over AdaDetect for the task of responder identification. First, we suggest a specialized scorer aimed at discovering responders. This scorer uses recursive partitioning of the feature space, similar to the causal tree algorithm (Athey & Imbens, 2016), but differs by training a classifier to distinguish treated from untreated subjects based on Y within each leaf, rather than estimating the CATE directly. This approach increases the power substantially compared to AdaDetect, which utilizes off-the-shelf classifiers as scorers. Second, we address the potential bias from non-random allocation of treatment by applying a propensity score based adjustment (Rosenbaum & Rubin, 1983; Tibshirani et al., 2019).

The suggested method tests whether each treated observation's (Y_i, X_i) is sampled from the same distribution as the control response distribution, while controlling the FDR, as opposed to providing confidence intervals for individual observations as in Lei & Candès (2021); Alaa et al. (2024). Duan et al. (2024) also provides FDR guarantees, but targets a fundamentally different hypothesis: whether the CATE is non-negative. Because CARD tests for general distributional differences rather than directional mean effects, the two methods address distinct inferential targets, and their statistical power is not directly comparable. When the scientific question concerns the conditional mean, the approach of Duan et al. (2024) is more appropriate; when the interest lies in detecting any departure from the control distribution, CARD is the natural choice.

CARD provides finite-sample FDR control in randomized experiments and asymptotic FDR control in observational settings for identifying treated individuals whose observed outcomes are statistically inconsistent with the control outcome distribution. In contrast to existing FDR-controlled approaches that primarily focus on directional or mean-based hypotheses (e.g., testing whether the CATE is non-negative), CARD is designed to detect fully general distributional deviations, including changes in variance, skewness, or tail behavior.

The rest of Section 1 provides background on the inference task at hand. In Section 2 we give an overview of AdaDetect and causal trees, as they serve as building blocks for our method. In Section 3 we present CARD, discuss its components, and suggest an adjustment to handle cases where the treatment assignment depends on the covariates. In Section 4 we examine the performance of CARD in randomized controlled trials (RCTs) and observational study settings. Finally, we conclude with a discussion and suggested avenues for further research in Section 5.

1.1 Causal Inference

In this paper, we follow the potential outcomes framework (Imbens & Rubin, 2015) with a binary treatment. Given n subjects, we denote by $T_i \in \{0, 1\}$, $(Y_i(1), Y_i(0))$ and X_i, the binary treatment indicator, potential outcomes and the vector of covariates (of length p) of subject i. Using brackets ($[\cdot]$) as a generic notation for a statistical distribution and the tilde ($\sim$) to mean "is distributed as," we assume that the potential outcomes are distributed according to

$$(Y_i(1), Y_i(0), T_i, X_i) \sim [Y(1), Y(0), T, X].$$

Our focus will be on a binary treatment, $T_i \sim \text{Ber}(e(x))$, where $e(x)$ is the propensity score. Under the commonly assumed stable unit treatment value assumption (SUTVA), the observed data set comprises triplets $(Y_{\text{obs},i}, T_i, X_i)$ where $Y_{\text{obs},i} = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0)$.

The ITE τ_i is defined as

$$\tau_i \equiv Y_i(1) - Y_i(0).$$

By definition, for each unit, only one potential outcome is observed while the other remains unobserved. Consequently, the ITEs are unobserved and must be inferred. We also assume strong ignorability throughout the paper:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X.$$

Strong ignorability assumes that there are no unmeasured confounders affecting both the treatment assignment and the potential outcomes. This assumption clearly holds in RCTs. Thus, under this assumption, the treatment assignment is essentially randomized (i.e., independent of potential outcomes) conditional on the covariate values. These assumptions allow us to claim that

$$[Y_i(1)|X = x] \sim [Y_i|X_i = x, T_i = 1],$$

$$\text{and } [Y_i(0)|X = x] \sim [Y_i|X_i = x, T_i = 0],$$

making it possible to identify moments of interest in the distribution of $Y(1) - Y(0)$. Typically, the interest lies in estimating the CATE,

$$\tau(x) = \mathbb{E}(Y(1) - Y(0)|X = x).$$

CATE estimation has garnered much attention in recent years, with multiple methods developed to estimate CATE using machine learning (Künzel et al., 2019; Wager & Athey, 2018; Nie & Wager, 2021). For example, the S-learner consists of training an outcome prediction algorithm $\hat{\mu}(X, T)$, which aims to predict $E(Y|X = x, T = t)$. The CATE estimate is then given by

$$\hat{\tau}(x) = \hat{\mu}(x, T = 1) - \hat{\mu}(x, T = 0).$$

Other parameters of the ITE distribution are also valuable targets. For example, Fort (2016) proposed estimating the conditional quantile treatment effect. CARD's distributional test is sensitive to shifts in any such parameter, not only the conditional mean, which motivates its use when treatment effects may manifest as changes in variance, skewness, or tail behavior.

1.2 Inferential Goal

The objective of responder analysis is to determine whether each treated observation is a responder, $Y_i(1) \neq Y_i(0)$. However, instead of inferring directly on $Y_i(1) - Y_i(0)$ (which requires further assumptions), we test the null hypotheses

$$H_{0,i} : Y_i(1) \sim Y_i(0).$$

Consequently, we can conclude that subjects for whom the null hypothesis is rejected are considered responders. Under the strong ignorability assumption, the above is equivalent to

$$H_{0,i} : (Y_i(1), X_i) \sim P_{Y|X,T=0} \times P_{X|T=1}. \tag{1}$$

One could argue that the objective should be to test $Y_i(1) \sim P_{Y|X=X_i,T=0}$. However, such conditional testing requires assumptions of smoothness and increasingly large sample sizes (Foygel Barber et al., 2021). We do not control the type I error of individual hypotheses, as it depends on the distribution of $X|T = 1$ (see Appendix A). Consequently, we aim to control the FDR across all treated observations.

Remark 1. *It is important to note a key distinction between CARD's hypothesis and individual-level treatment effects. CARD tests whether the treated outcome distribution differs from the control distribution, not whether a specific individual's outcome changed. For instance, if $Y(1)$ and $Y(0)$ are identically distributed (e.g., both $N(0, 1)$) but independent, then every individual has a nonzero ITE with probability 1, yet CARD's*

null hypothesis holds. Conversely, CARD can detect cases where the treatment changes the variance or shape of the outcome distribution without altering the mean, scenarios invisible to CATE-based methods. Throughout the paper, we use 'responder' to mean a subject whose outcome is statistically inconsistent with the control distribution, which is a necessary but not sufficient condition for an individual-level treatment effect.

2 Related Work

In this section we provide the relevant background for CARD.

2.1 Causal Tree and Forest

Causal tree (CT) is a type of decision tree specifically designed for estimating CATE (Athey & Imbens, 2016). A tree is a partition, Π, of the covariate space of X. We denote by $l(x;\Pi)$ the partition $x \in l, l \in \Pi$. A partition is obtained by applying a partitioning algorithm, which is used to minimize a loss function on some sample $(y_i, x_i, t_i), i \in \mathcal{I}$. Denote by $\mathcal{I}_j = \{i : t_i = j\}$, $j = 0, 1$, and $n_j = |\mathcal{I}_j|$. Finally, denote $L(\mathcal{I}_j, x, \Pi) = \{i : i \in \mathcal{I}_j, x_i \in l(x;\Pi)\}$. In CTs, at each partition the CATE is estimated by,

$$\begin{aligned}\hat{\tau}(x_i;\Pi) =& \frac{1}{|L(\mathcal{I}_0, x_i, \Pi)|} \sum_{j \in L(\mathcal{I}_0, x_i, \Pi)} y_j \\ &- \frac{1}{|L(\mathcal{I}_1, x_i, \Pi)|} \sum_{v \in L(\mathcal{I}_1, x_i, \Pi)} y_v \ .\end{aligned} \tag{2}$$

The partition, Π is obtained by minimizing the modified CATE mean square error loss,

$$\begin{aligned}Loss(\Pi, \mathcal{I}) =& -\frac{1}{n} \sum_{i \in \mathcal{I}_0 \cup \mathcal{I}_1} \hat{\tau}^2(x_i;\Pi) \\ &+ \frac{2}{n} \sum_{l \in \Pi} \Bigg(\frac{n_0 \hat{\mathrm{V}}(\{y_i : i \in \mathcal{I}_0, x_i \in l\})}{n} \\ &+ \frac{n_1 \hat{\mathrm{V}}(\{y_i : i \in \mathcal{I}_1, x_i \in l\})}{n} \Bigg),\end{aligned} \tag{3}$$

where $n = n_0 + n_1$ is the total number of observations, and $\hat{V}$ is the standard sample variance estimator. To tackle the selection bias in the leaf estimates (as the leaves are selected to maximize this estimate), the authors suggest splitting the data into a training and a test set. The training set is used to find the partitions that minimize Equation (3), and the test-set is used for estimating the CATE within each terminal node (Equation (2)).

The use of sample splitting ensures that the estimates of the CATE in the terminal nodes are unbiased. Thus, any standard inference approach can be used to test the difference in means between the treated and untreated subjects in each terminal leaf using the test set. The trade-off is that there are fewer observations for finding the optimal partition. Furthermore, the method is effective only when generating a single tree. When growing a forest, sample-splitting is not sufficient to ensure valid testing, and different methods for estimating the variance must be applied. These methods have only asymptotic guarantees (Wager & Athey, 2018), even in cases of RCTs.

2.2 Conformal Prediction

Conformal prediction is a framework that provides a method to generate a prediction set that contains the true target value with a $1 - \alpha$ confidence (Angelopoulos et al., 2023). Suppose that a practitioner wishes to learn f, a mapping from X to Y, $f(X) = Y$. We focus on inductive conformal prediction in which the calibration set is independent of the samples used for training $\hat{f}$ (Papadopoulos, 2008).

Suppose there is a predictive model $\hat{f}$, trained on some (X_i, Y_i), where Y_i is a continuous response and X_i are the covariates. Conformal prediction requires a specification of a non-conformity score, $s(x, y)$, (where larger values correspond to worse agreement between x and y) and a calibration set $(X_i, Y_i),\ i = 1, \ldots, k$. The non-conformity score is often learnt based on a training set. To test a new observation, the non-conformity score is computed for the calibration set, $s(\hat{f}(X_1), Y_1), \ldots, s(\hat{f}(X_k), Y_k)$. Assuming that a new observation is from the same distribution as the calibration set, then their non-conformity scores are also distributed identically, thus,

$$\mathbb{P}\left(s(X_{test}, Y_{test}) > q\left((1-\alpha)\frac{k+1}{k}\right)\right) \leq \alpha,$$

where $q(p)$ is the pth empirical quantile based on the calibration set non-conformity scores. This can be used to test $H_0 : Y_{test}|X_{test} \sim Y_i|X_i$, or by inverting the acceptance region generating prediction-intervals (Shafer & Vovk, 2008; Haroush et al., 2021). The choice of the score function is critical for the performance of the conformal predictor. Romano et al. (2019) suggested conformal quantile regression (CQR), an adaptive score function based on quantile regression. This allowed for the conformal interval to vary as function of x, making it adaptive when facing heteroscedasticity. Romano et al. (2019) also generalized the approach and suggested conformal-intervals that are based on the conditional histogram of $Y|X$.

To test $H_0 : Y_{test} \sim P_{Y|X} \times P_X$, using the procedure described above, one must assume that $X_{test} \sim P_X$. When one is unwilling to make such assumption, weighted conformal prediction can be used. Tibshirani et al. (2019) suggested a modification to conformal prediction when facing covariate shift, where the score is weighted according to the likelihood ratio $w(x) = \frac{\mathrm{d}P_{test}(x)}{\mathrm{d}P_0(x)}$, where P_0 is the distribution of the calibration set covariates.

Lei & Candès (2021) proposed a method to obtain confidence intervals for the ITE. In the process, they demonstrated how to construct counterfactual confidence intervals. The method builds on the CQR and utilizes an estimation of the propensity score as weights in weighted conformal prediction to handle distribution shifts.

2.2.1 AdaDetect

Suppose there are two samples, a reference sample $\mathcal{I}_0 = \{i : X_i \sim P_0\}$, and a test sample $\mathcal{I}_1 = \{i : X_i \sim P_i\}$, where P_i can either be P_0 or some other distribution. The interest lies in identifying the specific observations in the test sample who differ from the reference sample, i.e., testing,

$$H_{0,i} : X_i \sim P_0, \forall i \in \mathcal{I}_1. \tag{4}$$

Conformal prediction can be used for this purpose. First, a non-conformity score, s, which measures the degree of disagreement of an observation being sampled from P_0 is applied to $\mathcal{I}_0$ to estimate its empirical distribution under the null hypothesis (Haroush et al., 2021). Then, the same non-conformity is applied to each observation in $\mathcal{I}_1$. A p-value for each observation obtained according to $\mathrm{pv}_i = \frac{1+\sum_{j\in\mathcal{I}_0}\mathbf{1}(s(x_j)>s(x_i))}{|\mathcal{I}_0|+1}$, where $\mathbf{1}$ is the indicator function.

Bates et al. (2023) have shown that the resulting p-values adhere to a specific type of dependency and that the Benjamini-Hochberg (BH) multiplicity adjustment (Benjamini & Hochberg, 1995) can be safely applied to control the FDR. While the procedure is valid, it still remains to decide on s, as different scores will be powerful for different P_1 and P_0.

Recently, Marandon et al. (2024) proposed AdaDetect, an adaptive method for learning the non-conformity score s for such hypotheses. Their non-conformity scores are the probabilities learned by a classifier aimed at discriminating between the reference and test sets. Since the classifier is predicting on the same samples used for training, the probabilities are uncalibrated. Using these probabilities for testing Equation (4) will fail to control the FDR.

AdaDetect tackles this issue by employing "knockoffs," observations designed to be indistinguishable from actual test observations to the classifier, yet for which the null hypothesis (Equation (4)) holds. The non-conformity scores of the knockoffs are distributed similarly to the scores of true test observations under the null hypothesis, allowing estimation of the test-statistic distribution under the null.

The procedure involves sampling k observations from $\mathcal{I}_0$ (denoted $\mathcal{I}_{ko}$) to serve as null calibration points.[1] These are combined with $\mathcal{I}_1$ to form $\mathcal{I}_1^* = \mathcal{I}_{ko} \cup \mathcal{I}_1$, while the remaining control observations form $\mathcal{I}_0^* = \mathcal{I}_0 \backslash \mathcal{I}_{ko}$.

The non-conformity score is learned by training a classifier aimed at discriminating between $\mathcal{I}_0^*$ and $\mathcal{I}_1^*$, where higher values indicate the sample is more likely to be from $\mathcal{I}_1^*$. The prediction will also serve as our non-conformity score. Intuitively, if $X_i \sim P_0$ then the $s(X_i), i \in \mathcal{I}_1$ are exchangeable with $s(X_j), j \in \mathcal{I}_{ko}$. Thus, a conformal p-value can be obtained as

$$\text{pv}_i = \frac{1 + \sum_{j=1}^{k} \mathbf{1}\left(s(x_j) > s(x_i)\right)}{k+1}, \tag{5}$$

which quantifies how likely it is that the non-conformity score of a given observation is sampled from the non-responders' non-conformity score distribution. Due to the exchangeability of the knockoff scores and the tested observation score, each conformal p-value is marginally super-uniform under the null (i.e., $P(\text{pv}_i \leq \alpha | H_0) \leq \alpha$). The use of a classifier generates an adaptive non-conformity score, which was shown to asymptotically achieve the optimal Neyman-Pearson test power (Marandon et al., 2024).

3 CARD

We propose CARD, a targeted approach that uses a specialized scorer to enhance the power in detecting changes in the response distribution (compared to the control) among treated subjects. Similar to AdaDetect, CARD employs a classifier to score observations of interest and uses knockoffs to control the FDR in finite samples. However, CARD differs by utilizing a scorer specifically tailored to the problem's structure, substantially improving its power compared to a direct application of AdaDetect with generic classifiers.

We initially discuss the suggested method in the RCT setting, which naturally avoids confounding due to differences in the distributions of $X|T = 1$ and $X|T = 0$ due to the randomization to treatment. In Section 3.3, we also discuss approaches for addressing such differences in non-randomized settings.

Let $\mathcal{I}_0$ and $\mathcal{I}_1$ denote the indices for untreated and treated observations, respectively. Additionally, define the set of treated observations indices where (Equation (1)) holds as $\mathcal{H}_0 = \{i \in \mathcal{I}_1 : (Y_i(1), X_i) \sim (Y_i(0), X_i)\}$.

3.1 Inference in RCTs

To test the hypotheses Equation (1), we split the untreated subjects into two groups, $\mathcal{I}_0 = \mathcal{I}_{ko} \cup \mathcal{I}_0^*$ and train a classifier to differentiate between $\mathcal{I}_1^* = \mathcal{I}_1 \cup \mathcal{I}_{ko}$ and $\mathcal{I}_0^*$. Let $s(x_i, y_i)$ be the resulting score according to the classifier. We then obtain p-values as in Equation (5),

$$\text{pv}_i = \frac{1 + \sum_{j \in \mathcal{I}_{ko}} \mathbf{1}\left(s(x_j, y_j) > s(x_i, y_i)\right)}{k+1}.$$

In RCTs, randomization ensures that $X|T = 1 \sim X|T = 0$. If the null hypothesis (Equation (1)) is true for observation i, then $Y_i|T_i = 1, X_i \sim Y_i|T_i = 0, X_i$. Thus, $\{(X_i, Y_i)\}_{i \in \mathcal{I}_{ko}}$ are exchangeable with $\{(X_j, Y_j)\}_{j \in \mathcal{I}_1 \cap \mathcal{H}_0}$. If the classifier is invariant to the ordering of its training inputs (a property satisfied by most practical classifiers; see eq. (8) in Marandon et al. (2024)), then by Lemma 3.2 and Theorem 3.3 of Marandon et al. (2024), the BH procedure applied to the resulting p-values will control the FDR.

To handle the setting of observational studies, where $X|T = 1 \not\sim X|T = 0$, we suggest a weighted version of AdaDetect based on weighted conformal prediction (Tibshirani et al., 2019); see Section 3.3. A graphical illustration of the procedure is provided in Section 3.1.

[1] We refer to these as "knockoffs" throughout, though this usage differs from the model-X knockoffs framework of Candes et al. (2018), where knockoffs are synthetic copies of covariates constructed for variable selection. Here, knockoffs are real control observations reassigned to the test set.

3.2 Scorer

To increase power beyond that of AdaDetect, we suggest learning the non-conformity scores using a responder tree. The responder tree mirrors the principles of CTs, wherein the data is recursively partitioned on X to minimize the overall Loss_y across partitions until certain stopping criteria, such as reaching a maximum tree depth or convergence of loss are met. While the splits are done according to X the evaluation in the terminal leaves is based solely on Y and T.

The splits the responder tree makes, are driven by the loss on the terminal nodes (where only Y and T are considered). The aim of the tree is to best classify observations by their treatment assignment (treated or not). The loss the responder tree attempts to minimize is

$$\text{Loss}(\Pi, \mathcal{I}_0, \mathcal{I}_1) = \sum_{l \in \Pi} c_l \, \text{Loss}_y(\{(y_i, t_i) : i \in \mathcal{I}_0 \cup \mathcal{I}_1, \; x_i \in l\}), \tag{6}$$

where Loss_y is a classification loss (such as Gini impurity or cross-entropy) computed on the depth-2 tree that classifies T from Y within leaf l, and

$$c_l = \frac{|\{i \in \mathcal{I}_0 \cup \mathcal{I}_1 : x_i \in l\}|}{|\mathcal{I}_0 \cup \mathcal{I}_1|}.$$

We minimize the specified loss (Equation (6)) using recursive partitioning, a standard approach to optimize classification trees (Rokach & Maimon, 2005). The primary partitions are based on X. For each subset of

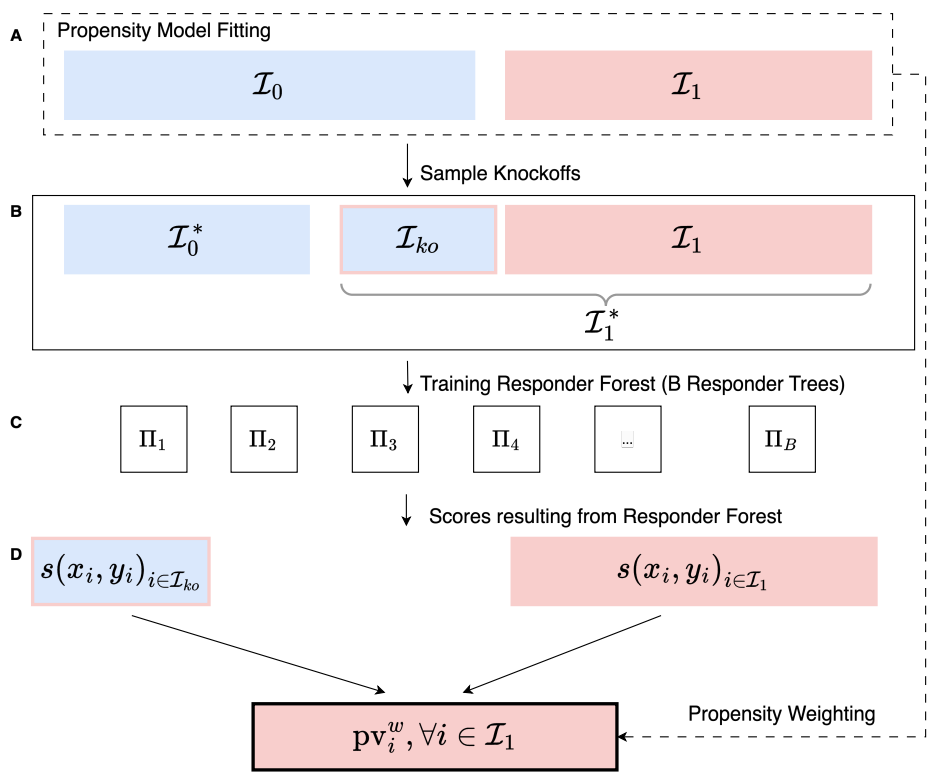

Figure 1: An illustration of CARD is provided where red fill indicates treated samples, and blue fill represents untreated samples. Knockoffs are depicted with a red frame. In step A, the propensity model is fit, and knockoffs are sampled from the untreated observations. Step B involves preparing the data for the classifier by combining the knockoffs with the treated observations. In step C, the responder forest is trained to classify between $\mathcal{I}_0^*$ and $\mathcal{I}_1^*$. Finally, in step D, the predictions from the responder forest are used as non-conformity scores. These scores are then weighted according to the propensity score to calculate p-values

observations defined by these partitions, we fit a shallow (depth 2) classification tree, where Y is used to classify T. The loss associated with each partition is calculated based on the weighted classification loss of the shallow trees applied to each subset.

Intuitively, the responder tree focuses the classification task on detecting distributional discrepancies in Y within covariate-adaptive regions of X. By partitioning on X but evaluating splits using treatment-label predictability from Y, the scorer focuses on outcome shifts rather than generic covariate differences.

While a single responder tree is described, we can also consider a forest of scorers, where in a similar vein to Random Forest (RF), a bootstrap sample of observations and a sample of features are taken to grow each tree (Hastie et al., 2009). The score is the average prediction across the different trees as standard in classification random-forest.

3.3 Propensity Adjusted CARD

The propensity score is the probability of treatment assignment conditional on observed baseline characteristics, i.e., $T = 1|X$. If the propensity is a function of X and $Y(0) \not\perp X$, then the knockoffs are no longer exchangeable with the treated observations under the null. To overcome this issue, we suggest a propensity adjusted p-value,

$$\mathrm{pv}_i^w = w_i^* + \sum_{j \in \mathcal{I}_{ko}} w_j^* \cdot \mathbf{1}\left(s(x_j, y_j) > s(x_i, y_i)\right), \tag{7}$$

where the weights are defined by $w_j = \frac{e(x_j)}{1-e(x_j)}$, $j \in \{i\} \cup \mathcal{I}_{ko}$ and $w_j^* = \frac{w_j}{\sum_{l \in \{i\} \cup \mathcal{I}_{ko}} w_l}$. This is similar to the weighting used in (Lei & Candès, 2021).

Lemma 1. *Under the null hypothesis (Equation* (1)*), and assumptions A1-3, the weighted p-values (Equation* (7)*) control the type I error at level α,*

$$\mathbb{P}(\mathrm{pv}_i^w < \alpha) \leq \alpha.$$

A.1. ***SUTVA (Stable Unit Treatment Value Assumption)****: The observed outcome equals the potential outcome under the received treatment, i.e.,*

$$Y_{obs,i} = T_i \cdot Y_i(1) + (1 - T_i) \cdot Y_i(0).$$

A.2. ***Ignorability (Unconfoundedness)****: Conditional on covariates X, treatment assignment is independent of the potential outcomes,*

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i \mid X_i.$$

A.3. ***Overlap (Positivity)****: Each treatment level is possible for all values of X,*

$$0 < \mathbb{P}(T_i = t \mid X_i = x) < 1 \quad \textit{for all } t \in \{0, 1\}, \textit{ and for all } x \textit{ such that } \mathbb{P}(X_i = x) > 0.$$

The proof is given in Appendix D. It is important to note that the resulting p-values do not have a theoretical guarantee for controlling the FDR in finite samples, particularly in the context of non-randomized experiments. However, in our experiments, these p-values effectively control the FDR, even in smaller sample sizes, despite the presence of non-trivial unknown propensity. Furthermore, it is shown that as $n \to \infty$, the BH procedure using the weighted p-values recovers the theoretical control of the FDR (Jin & Candès, 2023).

Although results are typically presented with oracle propensity, in practice, the propensity score is estimated from the same data we intend to analyze. To reduce over-fitting, we fit the propensity estimator using m-fold fitting (Jacob, 2020), training on $m - 1$ folds and estimating the propensity on the remaining fold. We examine the empirical validity of this adjustment in Section 4.2.

4 Simulations

4.1 RCT Simulations

We begin by evaluating the performance of CARD in RCT settings using a simulation study. We use a variant of the data generation process that was used by Lei & Candès (2021) and Wager & Athey (2018). The data generation process is outlined as follows: We sample $X^* \sim N(0, \Sigma_{p\times p})$, where $\Sigma_{i,j} = \rho^{|i-j|}$. To obtain X we transform $X_i = \Phi(X_i^*)$. The potential outcomes are generated according to,

$$Y(0) = 4 \cdot (X_1 + X_2) + \epsilon,$$
$$Y(1) = 4 \cdot (X_1 + X_2) + rf(X_1) \cdot f(X_2) + \epsilon,$$

where $\epsilon \sim N(0, \sigma(x)^2)$ and

$$f(x) = \mathbf{1}(x > 0.5) \cdot \frac{2}{1 + \exp(-12 \cdot (x - 0.5))}.$$

That is, the null hypothesis (Equation (1)) is false for $X_1 > 0.5, X_2 > 0.5$. Treatment assignments are randomly determined following a Bernoulli distribution, Ber(0.5). We explore several variations in the parameters: $r \in \{-1, +1\}$ to assess the effect of signal sign, $\sigma(x) \in \{1, -2 \cdot \ln(x)\}$ to compare homoscedastic and heteroscedastic settings, $\rho \in \{0, 0.9\}$ to distinguish between independent and dependent settings, and $p \in \{10, 100\}$ to evaluate the impact of the number of features. We also vary the number of observations, $n \in \{250, 500, 1000, 2000\}$. Each setting is repeated 100 times, and the averages are reported.

We compare CARD (using 20% of the untreated subjects as knockoffs) with several other methods of detecting responders.

1. Global Testing: Testing if $Y_i, T_i = 1 \sim Y(0)$, which is also valid for testing Equation (1). The p-value is obtained on the basis of the $Y(0)$ eCDF (see Appendix C).
2. AdaDetect (Marandon et al., 2024): Applying the original AdaDetect method with RF as the classifier. Again, 20% of the untreated observations are used as knockoffs.
3. CQR (Romano et al., 2019): We apply the CQR non-conformity score to obtain p-values (see Appendix B), equivalent to counterfactual confidence intervals as suggested by Lei & Candès (2021). Boosting is used for the estimation of the conditional quantiles, with 80% of the untreated sample is used for estimation and 20% for testing.

All methods are valid for testing Equation (1). To handle multiplicity, we adjust the p-values of all methods using the BH procedure. The FDR is expected to be controlled at level 0.1. Other CATE estimation methods can also be used for testing for responders such as CF, BART and bootstrap with Meta-Learners (Wager & Athey, 2018; Chipman et al., 2012; Künzel et al., 2019). However, Lei & Candès (2021) showed they lack finite coverage control, and thus we omit them from formal comparisons.

The power results are presented in Figure 2, with additional FDR and power analyses provided in Appendix E. The use of AdaDetect with RF as the classifier yields poor results, which is unsurprising as RF attempts to detect differences across all features, unlike competing methods focusing on Y or $Y|X$.

CARD is powerful in all scenarios, significantly outperforming competing methods in most cases, except in homoscedastic scenarios with positive signals where they are surpassed by the CQR and global methods for $n < 2000$. Its power increases the most as the sample size increases.

As implied by its name, the global method performs well when responders are discernible solely by considering Y, particularly in scenarios with positive signal sign and homoscedastic noise. Because the global method does not split the data, it significantly outperforms other methods in this specific scenario. However, when the differences in distributions are only discernible upon conditioning on X (see Appendix C for an illustrative example), the method has virtually no power.

Unlike CARD, the CQR non-conformity score is learnt only on the basis of the untreated distribution. This leads to much better performance of CARD across all settings except for positive signal and homoscedastic noise.

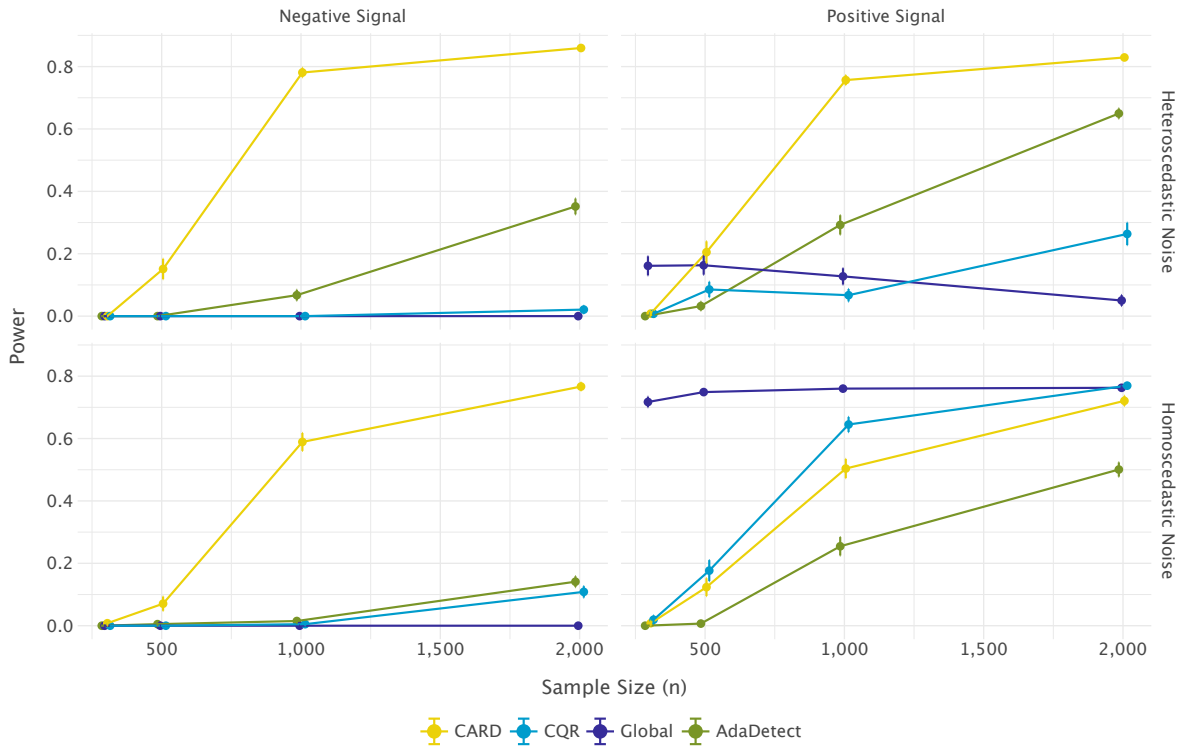

Figure 2: Power analysis of various methods for different sample sizes. The plots display the power as a function of sample size (n) in the $p = 10$ scenario. The power of CARD is higher in all scenarios except for the homoscedastic noise and positive signal, where it reaches parity when n is sufficiently large. Intervals indicate ± 1 s.e. around the estimate.

4.2 Observational Study Simulation

The asymptotic control of the FDR in the weighted CARD procedure is theoretically assured. In the following section, we examine the impact of weighting on CARD in finite sample settings.

We employ the same data generation protocol as outlined in Section 4, changing only the propensity function to mimic an observational study with confounders, $e = \frac{1}{1+\exp(1.75 \cdot X_1 - 0.825)}$. The propensity function ensures that the probability of receiving treatment is 0.5. We present results for the $p = 10$ and positive signal scenario, as the results are consistent across scenarios (see Appendix E). Furthermore, we restrict our simulations to CARD.

We consider four methods of handling non-trivial propensity: 1) None, ignoring propensity and applying CARD without any adjustment; 2) oracle adjustment, where weighting is based on true propensity; 3) RF classifier propensity adjustment; and 4) logistic regression propensity adjustment. The propensity models are fit using 10-fold cross-fitting.

The results, shown in Figure 3, indicate that all methods except “None" control the FDR at the expected level of 0.1. The estimating methods exhibit higher power than the oracle, a result attributed to cross-fitting and errors in propensity estimation. However, since the method is generally conservative, it controls the FDR at the expected level, which is promising for real-world applications. An example analysis using a subset of the ACIC 2016 data (Dorie et al., 2019) is presented in Appendix F. An additional real-world data example is provided in Appendix G.

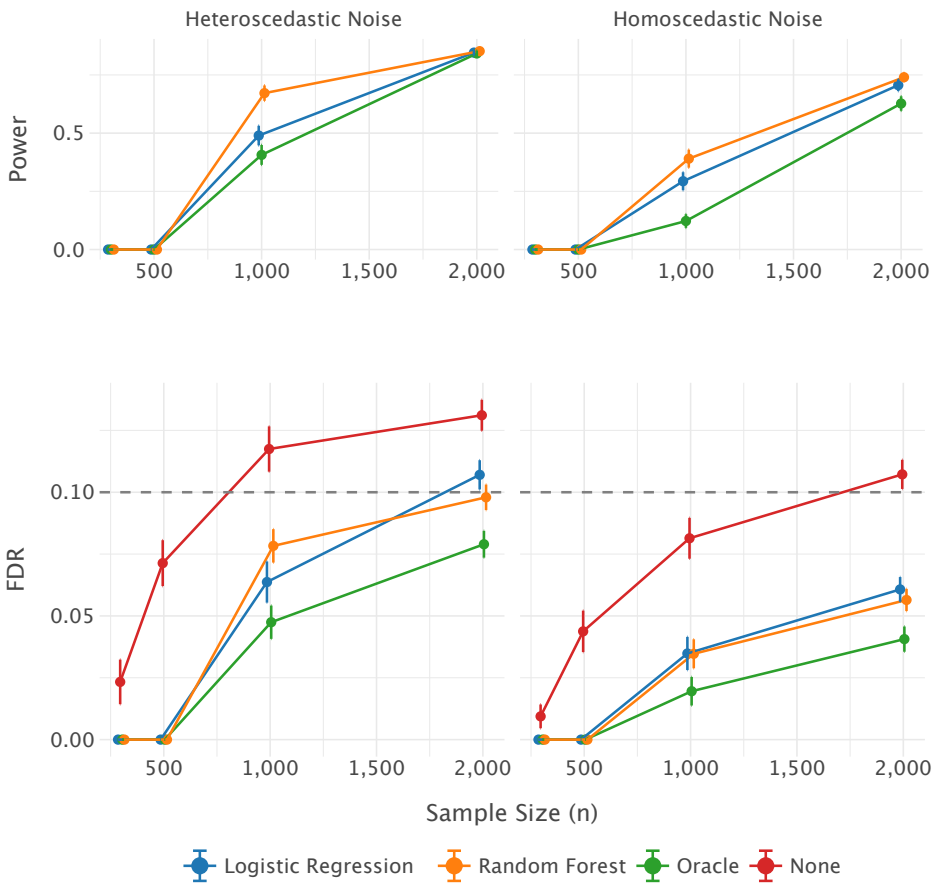

Figure 3: Power and FDR of the various methods of handling propensity. The None method is removed from the power (upper row) plots, since it fails to control the FDR at the expected level (bottom row). Intervals indicate ± 1 s.e around the estimate.

5 Discussion

CARD is a powerful method for identifying responders within a treatment group, offering finite-sample FDR control in RCTs and applicability to observational studies under standard causal assumptions. The detection of responders complements CATE analysis: while CATE measures the average treatment effect within population segments, CARD quantifies the proportion of individuals whose outcome distributions differ from control. For example, a subgroup may exhibit substantial CATE, yet responder analysis may reveal that only a small fraction of individuals within that subgroup truly experience a distributional shift under treatment. Unlike mean-based approaches, CARD tests for differences in the full outcome distribution, allowing detection of variance, skewness, or tail changes even when the conditional mean remains unchanged.

The main drawback of CARD is that it offers only marginal control over the Type I error, and not conditional on the covariates, X. This results from the finite sample control provided by the method of Foygel Barber et al. (2021), a limitation also shared by counter-factual ITE confidence intervals (Lei & Candès, 2021). Extending the method to test conditional hypotheses remains an area for future research.

While CARD is presented as a method for detecting responders, it can also be applied to other use-cases:

1. A global null test for treatment effectiveness, allowing for quick screening of various datasets.

2. Identifying subspaces of X in which the treatment and non-treatment response distributions differ. A single responder tree tries to find a subspace of X where the distributions $Y(0)|X \in A$ and $Y(1)|X \in A$ are different. Testing the difference between the two distributions using only the knockoffs in the identified subspace, $\mathcal{I}_{ko} \cap A$ provides a valid test for distribution differences given A.

Similar to Duan et al. (2024), CARD can also be configured to test only for positive (or negative) responders. Another issue with CARD (and any other split-conformal prediction method) is its inherent randomness due to the split (or selection of knockoffs). Fortunately, by pairing CARD with the method suggested by Bashari et al. (2024), most of this randomness can be removed while maintaining similar power. Finally, combining this method with various techniques for constructing valid confidence intervals for the true discovery proportion (Hemerik & Goeman, 2018; Millstein et al., 2022) can offer further insights into the proportion of responders.

References

Ahmed M Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. *Advances in Neural Information Processing Systems*, 36, 2024.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Derek C Angus and Chung-Chou H Chang. Heterogeneity of treatment effect: estimating how the effects of interventions vary across individuals. *JAMA*, 326(22):2312–2313, 2021.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Meshi Bashari, Amir Epstein, Yaniv Romano, and Matteo Sesia. Derandomized novelty detection with fdr control via conformal e-values. *Advances in Neural Information Processing Systems*, 36, 2024.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1):289–300, 1995.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *Annals of Applied Statistics*, 6(1):266–298, 2012.

Chien-Chia Chuang, Isabelle Guillemin, Claus Bachert, Stella E Lee, Peter W Hellings, Wytske J Fokkens, Nicolas Duverger, Chunpeng Fan, Nadia Daizadeh, Nikhil Amin, et al. Dupilumab in crswnp: responder analysis using clinically meaningful efficacy outcome thresholds. *The Laryngoscope*, 132(2):259–264, 2022.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1): 43–68, 2019.

Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Interactive identification of individuals with positive treatment effect while controlling false discoveries. *Journal of Causal Inference*, 12(1):20230059, 2024.

Margherita Fort. Unconditional and conditional quantile treatment effect: Identification strategies and interpretations. *Topics in Theoretical and Applied Statistics*, pp. 15–24, 2016.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Hamlet Gasoyan, Elizabeth R Pfoh, Rebecca Schulte, Phuc Le, W Scott Butsch, and Michael B Rothberg. One-year weight reduction with semaglutide or liraglutide in clinical practice. *JAMA Network Open*, 7(9): e2433326–e2433326, 2024.

Gordon H Guyatt, Elizabeth F Juniper, Stephen D Walter, Lauren E Griffith, and Roger S Goldstein. Interpreting treatment effects in randomised trials. *British Medical Journal*, 316(7132):690–693, 1998.

Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A statistical framework for efficient out of distribution detection in deep neural networks. In *International Conference on Learning Representations*, 2021.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Random forests. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, pp. 587–604, 2009.

Jesse Hemerik and Jelle J Goeman. False discovery proportion estimation by permutations: confidence for significance analysis of microarrays. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):137–155, 2018.

Nicholas Henschke, Annefloor van Enst, Robert Froud, and Raymond W Ostelo. Responder analyses in randomised controlled trials for chronic low back pain: an overview of currently used methods. *European Spine Journal*, 23:772–778, 2014.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Daniel Jacob. Cross-fitting and averaging for machine learning estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2007.02852*, 2020.

Ying Jin and Emmanuel J Candès. Model-free selective inference under covariate shift via weighted conformal p-values. *arXiv preprint arXiv:2307.09291*, 2023.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10): 4156–4165, 2019.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183, 2024.

Joshua Millstein, Francesca Battaglin, Hiroyuki Arai, Wu Zhang, Priya Jayachandran, Shivani Soni, Aparna R Parikh, Christoph Mancao, and Heinz-Josef Lenz. Fcrci: Fdr confidence interval selection and adjustment for large-scale hypothesis testing. *Bioinformatics Advances*, 2(1):vbac047, 2022.

R Andrew Moore, Owen A Moore, Sheena Derry, Paul M Peloso, Arnold R Gammaitoni, and Hongwei Wang. Responder analysis for pain relief and numbers needed to treat in a meta-analysis of etoricoxib osteoarthritis trials: bridging a gap between clinical trials and clinical practice. *Annals of the Rheumatic Diseases*, 69(2):374–379, 2010.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 2021.

Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in Artificial Intelligence*. Citeseer, 2008.

Lior Rokach and Oded Maimon. Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(4):476–487, 2005.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Steven M Snapinn and Qi Jiang. Responder analyses and the assessment of a clinically relevant treatment effect. *Trials*, 8:1–6, 2007.

Sebastian Straube, Sheena Derry, R Andrew Moore, Jocelyn Paine, and Henry J McQuay. Pregabalin in fibromyalgia-responder analysis from individual patient data. *BMC Musculoskeletal Disorders*, 11:1–8, 2010.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Astrid M Vicente, Wolfgang Ballensiefen, and Jan-Ingvar Jönsson. How personalised medicine will transform healthcare by 2030: the icpermed vision. *Journal of Translational Medicine*, 18:1–4, 2020.

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

A Marginal vs Conditional Responder Testing

The test proposed for $H_{0,i} : (X_i, Y_i) \sim (X_0, Y_0)$, is a test of the joint distributions, as opposed to a test of the conditional distributions $H_{0,i} : Y_i(1)|X_i = x_i \sim Y_i(0)|X_i = x_i$. This implies, that the type I error is not controlled conditional on X, but rather marginalized over X, i.e.,

$$\begin{aligned} &\mathbb{P}_{(X_0,Y_0)|T=1}(\mathrm{PV}_i \leq \alpha) = \\ &\int \mathbb{P}_{X|T=1}(X = x)\mathbb{P}_{Y(0)|X=x}\left(\mathrm{PV}_i \leq \alpha\right)) \, dx \leq \alpha. \end{aligned} \tag{8}$$

In most practical examples, the CARD's Type I error rate when testing conditional hypotheses is negligibly above the nominal level. However, it is possible to construct scenarios where the CARDs test's Type I error rate reaches up to 1 when testing these hypotheses.

Consider the case where $Y(0)|X = 0 + \epsilon$, $Y(1)|X = \mathbf{1}(X > 0.01) + \epsilon$, and $\epsilon \sim N(0, 10^{-7})$. Furthermore, let $T \sim Ber(0.5)$ and $X \sim U[0, 1]$. For a large number of observations, the discriminators will make the correct split on $X_1 > 0.01$, with a following split on $Y > 10^{-5}$ (or any other value which separates the distributions well). Due to these splits all of the null conditional hypotheses would be rejected (all observations for which $X_1 < 0.01$), yet, the joint type I error will be controlled at the expected level.

B Conformal Quantile Regression (CQR)

To produce CQR, we split the data into a training set, $\mathcal{I}_{tr}$, and a calibration set, $\mathcal{I}_{cal}$. First, the training set is used for fitting a quantile regression model (such as linear quantile regression, quantile boosting, quantile trees, etc.). The model is trained to predict the conditional quantiles $q_{\alpha_l}(x)$ and $q_{\alpha_h}(x)$, where $\alpha_l + \alpha_h = \alpha$, and $1 - \alpha$ is the coverage probability. Then, the calibration set is used to estimate the non-conformity score,

$$s(X_i, Y_i) = \max\left(q_{\alpha_l}(X_i) - Y_i, Y_i - q_{\alpha_h}(X_i)\right).$$

To construct the CI, the $1 - \alpha$ quantile of the non-conformity scores are computed on the calibration set, denoted by $Q_{1-\alpha}$. Finally, the confidence interval is,

$$C(X_{n+1}) = [q_{\alpha_l}(X_i) - Q_{1-\alpha}, q_{\alpha_h}(X_i) + Q_{1-\alpha}].$$

To obtain p-values, we use the following,

$$\mathrm{pv}_i = \frac{1 + \sum_{x_j, y_j \in \mathcal{I}_{cal}} \mathbf{1}\left(s(x_j, y_j) \geq s(x_i, y_i)\right)}{|\mathcal{I}_{cal}| + 1}. \tag{9}$$

C Global Testing

The task of detecting responders is essentially a task of detecting out-of-distribution observation based on the response. A simple approach would be to calculate the empirical p-values of $Y_i, T_i = 1$ observations based on the null sample, the p-values are,

$$\mathrm{pv}_i^{\mathrm{global}} = 2 \min\left(\frac{1 + \sum_{j=1}^{n_0} \mathbf{1}\left(y_j \geq y_i\right)}{n_0 + 1}, \frac{1 + \sum_{j=1}^{n_0} \mathbf{1}\left(y_j \leq y_i\right)}{n_0 + 1}\right).$$

A major drawback of the method is its inability to handle subjects which are responders conditional on some X. To illustrate the issue, we will use a somewhat contrived example, which exaggerates what otherwise is a very common phenomena. Suppose that $X \sim U[0, 1]$,

$$Y(0) = -5 \cdot \mathbf{1}(x < 0.5) + 5 \cdot \mathbf{1}(x > 0.5) + \epsilon,$$

and

$$Y(1) = -5 \cdot \mathbf{1}(x < 0.25) + 5 \cdot \mathbf{1}(x > 0.5) + \epsilon,$$

where $\epsilon \sim N(0,1)$.

The global testing approach has no power, since the responder do not appear extreme considering Y (see Figure 4 A), while when considering $0.25 < X < 0.75$, it is clear that the observation centered around $Y = 0$ are anomalies (see Figure 4 B). Indeed, in this example the power of the CARD is 1, while the global testing has a power of 0.

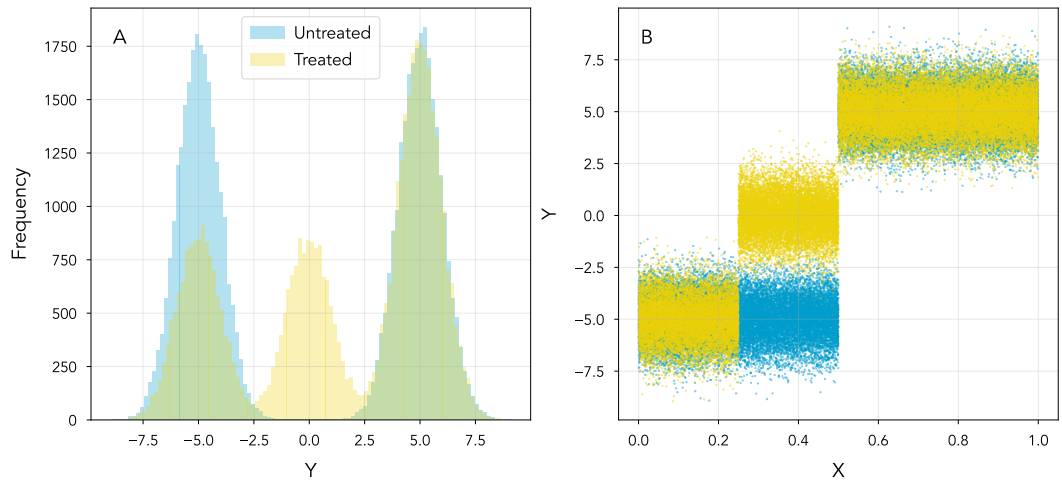

Figure 4: A. Histogram of Y, the responders are located between the two modalities of $Y(0)$. B. Scatter plot of Y X, it is clear to see the responders around $Y = 0$ and $0.25 < X < 0.5$.

D Proof of the Lemma in Lemma 1

Proof. For simplicity, we assume that $\mathcal{I}_{ko} = \{1, \ldots, k\}$ and $i = k+1$.

Denote $Z_i = (X_i, Y_i)$, $z_i = (x_i, y_i)$. Let $c = \frac{\mathbb{P}(T=0)}{\mathbb{P}(T=1)}$, and let

$$\tilde{w}_i(x,y) = \begin{cases} 1 & 1 \leq i \leq k \\ c\frac{e(x)}{1-e(x)} & i = k+1 \end{cases}.$$

Under the null hypothesis, it holds that for every x, y, $1 \leq i_0 \leq k$

$$\mathbb{P}((X,Y) = (x,y)|T=0)\tilde{w}_{k+1}(x,y) = \mathbb{P}((X,Y) = (x,y)|T=1)\tilde{w}_{i_0}(x,y). \tag{10}$$

Indeed, the left hand side of Equation (10) is

$$\begin{aligned}
&\frac{\mathbb{P}((X,Y) = (x,y)|T=0)\mathbb{P}(T=0)P(T=1|X=x)}{\mathbb{P}(T=1)P(T=0|X=x)} = \\
&\frac{\mathbb{P}(T=0|(X,Y) = (x,y))\mathbb{P}((X,Y) = (x,y))\mathbb{P}(T=1|X=x)}{\mathbb{P}(T=1)\mathbb{P}(T=0|X=x)} = \\
&\frac{\mathbb{P}((X,Y) = (x,y))\mathbb{P}(T=1|X=x)}{\mathbb{P}(T=1)} = \\
&\mathbb{P}((X,Y) = (x,y)|T=1).
\end{aligned}$$

The first equality is by Bayes rule, the second equality is due to the null hypotheses ($\mathbb{P}(Y=y, T=t|X=x) = \mathbb{P}(Y=y|X=x)\mathbb{P}(T=t|X=x)$) and the last equality is by definition.

Equation (10) implies that $(Z_1, ..., Z_{k+1})$ are weighted exchangeable ((Tibshirani et al., 2019, Definition 1)) – for every permutation σ,

$$\frac{\mathbb{P}(Z_1 = v_1, \ldots, Z_{k+1} = v_{k+1})}{\tilde{w}_1(v_1) \cdots \tilde{w}_{k+1}(v_{k+1})} = \frac{\mathbb{P}(Z_1 = v_{\sigma(1)}, \ldots, Z_{k+1} = v_{\sigma(k+1)})}{\tilde{w}_1(v_{\sigma(1)}) \cdots \tilde{w}_{k+1}(v_{\sigma(k+1)})}.$$

Now, we let E_z be the event that $\{Z_1, ..., Z_{k+1}\} = \{z_1, ..., z_{k+1}\}$. By the above, there is a constant C_z such that

$$\begin{aligned}\mathbb{P}(Z_1 = v_{\sigma(1)}, \ldots, Z_{k+1} v_{\sigma(k+1)}) &= C_z \tilde{w}_1(v_1) \ldots \tilde{w}_{k+1}(v_{k+1}) \\ &= C_z \tilde{w}_{k+1}(v_{\sigma(k+1)}).\end{aligned}$$

Assuming for simplicity that all the values are different, using the same steps as in the proof of (Tibshirani et al., 2019, Lemma 3), it holds that

$$\begin{aligned}\mathbb{P}(Z_{k+1} = z_j | E_z) =& \frac{\sum_{\sigma:\sigma(k+1)=j} \mathbb{P}(Z_1 = v_{\sigma(1)}, \ldots, Z_{k+1} = v_{\sigma(k+1)})}{\sum_{\sigma} \mathbb{P}(Z_1 = v_{\sigma(1)}, \ldots, Z_{k+1} = v_{\sigma(k+1)})} \\ =& \frac{k! \tilde{w}_{k+1}(z_j)}{\sum_{l=1}^{k+1} k! \tilde{w}_{k+1}(z_l)} = w_j^*.\end{aligned}$$

Finally, notice that the score s is constant given E_z. This implies together with the above

$$\mathbb{P}(\mathrm{pv}_i^w < \alpha | E_z) \leq \alpha.$$

And after marginalizing we get the required result. □

E Additional RCT Simulations

E.1 Power Results

Here we offer additional power results for the RCT setting simulations (Section 4.1). The plots shown in Figure 5 illustrate the power as a function of the number of observations. The conclusions from Section 4.1 are similar across the various scenarios. The CARD method outperforms all methods except in the positive signal and homoscedastic noise scenario.

E.2 FDR Results

Across all simulations, all methods control the FDR at the expected level of 0.1.

F ACIC 2016 Example

We present an illustrative example of an analysis performed using CARD, utilizing the first dataset from the ACIC 2016 challenge (available in the following repository and described in Dorie et al. (2019)). This dataset comprises 4,802 observations and 58 covariates. After applying one-hot encoding to the categorical variables, we retained 82 covariates in total.

The dataset includes information on the two potential outcomes and the treatment assignment for each observation. We analyzed both the full dataset and a sub-sample consisting of 1,000 observations. In our analysis, we applied CARD with propensity adjustment to the original treatment allocation, treating it as an observational study. To simulate a RCT, we reassigned treatment by sampling from a $Ber(0.5)$ distribution. In the RCT setting we applied CARD without a propensity adjustment. In the original treatment allocation, 858 observations were treated, while in the simulated RCT, 2,368 received treatment.

CARD was applied with 20% of the untreated samples serving as knockoffs. To control the FDR, we applied the BH procedure, setting the FDR threshold at 0.1. When necessary, propensity scores were estimated using a random forest classifier.

n	RCT	Observational Study
1000	0.53	0.73
4802	0.62	0.74

Table 1: Proportion of rejected hypotheses for the different scenarios for the ACIC 2016 data.

We report the proportion rejected hypotheses in each setting (Table 1). Surprisingly, CARD seems to perform better in the observational study setting. This is likely due to CARD benefiting from a large untreated sample, where a large number of knockoffs can be sampled, without compromising the scorer performance. This can also be seen in Figure 7, where the p-values in the observational setting are smaller due to the larger number of knockoffs used.

To further analyze the data, we tested for differences between responders and non-responders across each covariate. A Welch's t-test was employed, and we applied a Bonferroni correction (chosen for simplicity in this exploratory post-hoc analysis) to account for multiple comparisons. This analysis was conducted solely under the full RCT setting. The results revealed significant differences in several covariates, most notably for covariate X20, with a p-value of 10^{-72}. More sophisticated methods, such as logistic regression, can be

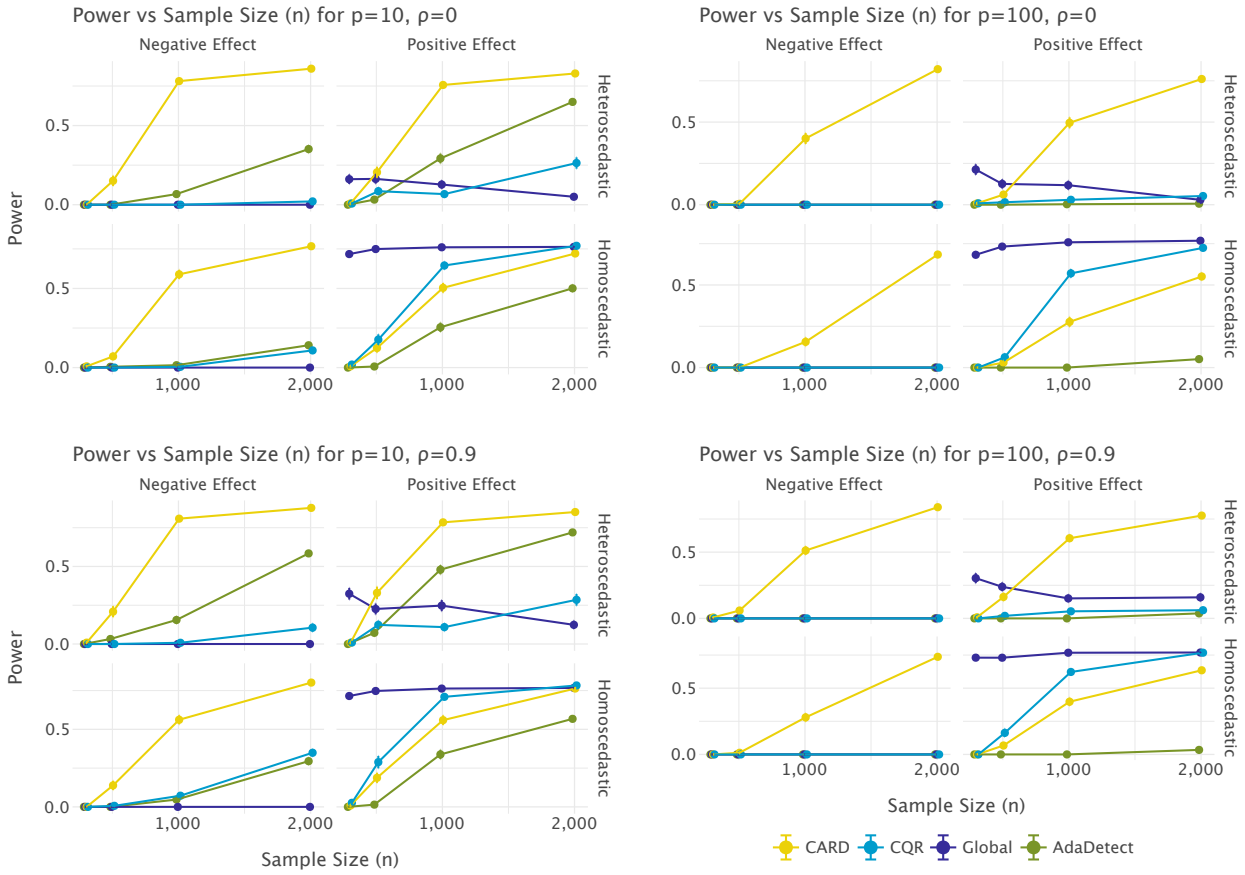

Figure 5: Power analysis of various methods for different sample sizes. The plots display the power as a function of sample size (n) in the various scenarios. Due to the correlations the AdaDetect method performs better compared to its performance when the covariates are independent from one another, but still the best performing method is CARD except in the positive signal and homoscedastic noise, where the CQR and Global methods achieve better power for $n < 2000$.

employed to account for dependencies between covariates, and give further insight on the factors driving the response.

G Exploratory Real-World Data Example: Semaglutide in Obstructive Sleep Apnea Patients

We conclude with an illustrative real-world application of CARD to an observational dataset of patients diagnosed with obstructive sleep apnea (OSA). The purpose of this analysis is to demonstrate the empirical behavior of CARD in a high-dimensional, confounded clinical setting, rather than to provide confirmatory inference. In particular, this example highlights how CARD can be used as a discovery-oriented tool for identifying individual-level treatment response signals in real-world data.

The dataset consists of $23,999$ patients with a diagnosis of OSA and complete baseline covariate information. Patients were followed for two years: a one-year screening period followed by a one-year weight-loss tracking period. Treatment corresponds to the initiation of semaglutide, while the control group comprises patients receiving no pharmacological weight-loss intervention. For the untreated cohort, a weight-loss start date was assigned randomly to avoid immortal time bias. The outcome of interest is the relative weight change from the start of the tracking period to the follow-up.

Because treatment assignment is non-random, we estimated the propensity score $e(x) = \mathbb{P}(T = 1 \mid X = x)$ using a random forest classifier with 5-fold cross-fitting. The model achieved strong discrimination (ROC AUC $= 0.851$), indicating substantial selection into treatment based on prior anti-diabetic medication use,

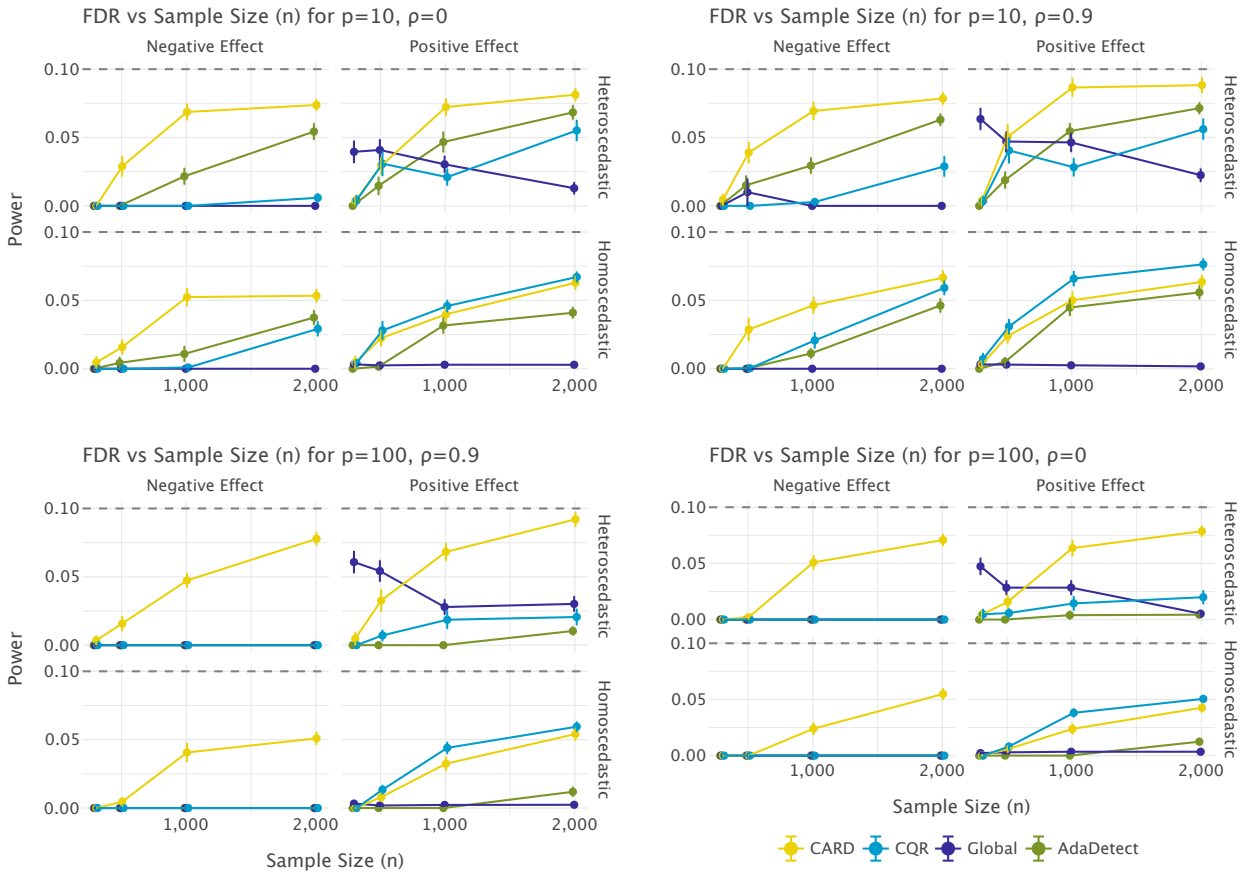

Figure 6: FDR of various methods for different sample sizes. The plots display the FDR as a function of sample size (n).

Table 2: CARD Diagnostics and Empirical Results (Semaglutide Cohort). The analysis demonstrates strong signal separation compared to null settings. Despite strong selection bias (AUC 0.851), the placebo test confirms that the discovery rate in the treated group significantly exceeds background noise.

Category	Metric	Value
Model Diagnostics	Propensity Score ROC AUC	0.851
	Common Support (Overlap) [a]	51.3%
	Placebo Discovery Rate [b]	3.1%
Responder Statistics	Total Treated Population (N)	2,530
	Identified Responders (n)	585
	Proportion of Cohort	23.1%
Clinical Outcomes	Mean Weight Change (Responders)	−8.04%
	Mean Weight Change (Non-Responders)	−2.67%
	Standardized Mean Difference (SMD)	−0.931
Top Predictors	1. Prior Anti-Diabetic Meds	(Imp: 0.28)
	2. Age at Enrollment	(Imp: 0.16)
	3. Baseline Year	(Imp: 0.11)

[a] Percentage of patients with propensity scores $e(x) \in [0.05, 0.95]$.
[b] Discovery rate when CARD is applied to a randomly partitioned untreated population (null setting) using $\alpha = 0.1$.

age, and baseline BMI. While this strong selection limited the common support to 51.3% of the cohort (see Table 2), there remained sufficient overlap to identify a large treated sub-population for whom valid counterfactual comparisons could be made using untreated knockoffs.

We applied CARD using a causal classification forest as the non-conformity scorer. Under formal FDR control using the BH procedure at $\alpha = 0.1$, CARD yielded no discoveries. This lack of power likely reflects the severe confounding (propensity $AUC = 0.851$) and limited overlap (51.3%). To explore whether CARD's scores nonetheless capture treatment-related signal, we conducted an exploratory analysis using nominal (unadjusted) p-values at the same threshold. To validate these exploratory results, we conducted a placebo test by applying CARD to the untreated population with randomized treatment labels (using the same parameters as in the main analysis). This null analysis yielded a discovery rate of only 3.1%, whereas the

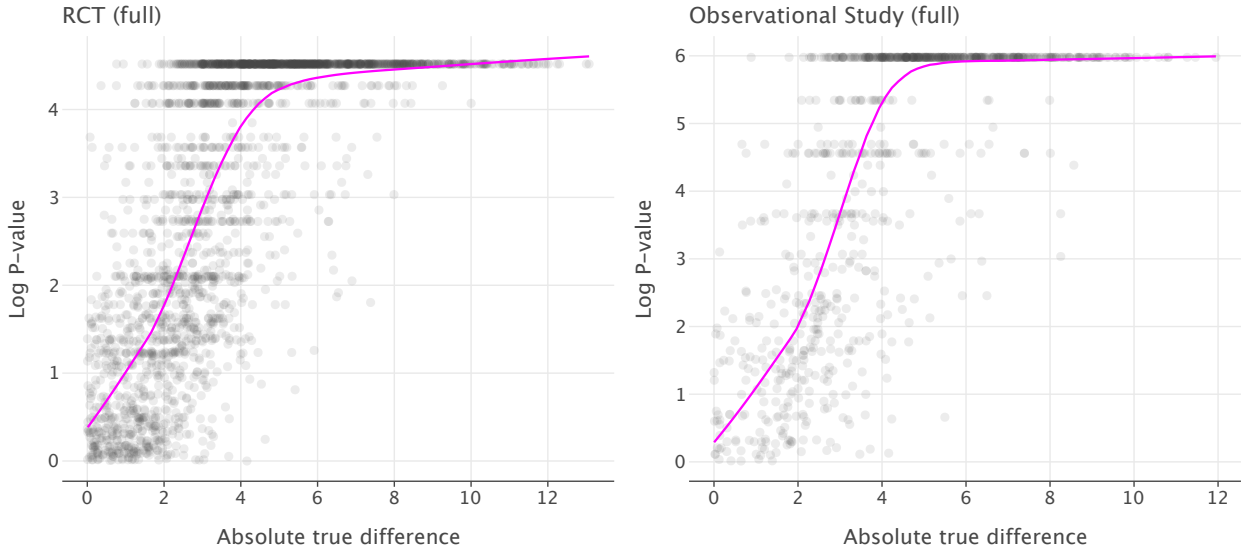

Figure 7: Log-p-values as a function of the true difference, measured by the Conditional Average Treatment Effect (CATE), for the ACIC 2016 dataset. A smooth LOWESS curve (shown in purple) illustrates the trend. As CATE increases, the p-values decrease as expected, with smaller p-values observed in the observational setting, due to the higher number of knockoffs used.

true treated group yielded a 23.1% responder rate, although it is important to keep in mind that in the placebo case the distribution of covariates of the treated and untreated are similar. The large difference in responders rates suggests that CARD is identifying a treatment-specific signal and not only noise.

Using this nominal threshold, CARD identified 585 responders among $2,530$ treated patients. These responders exhibited a mean weight reduction of -8.04%, compared to only -2.67% among non-responders. The magnitude of this separation is reflected in a large standardized mean difference (SMD $= -0.931$), providing strong evidence that the CARD scores capture clinically meaningful heterogeneity.

To further characterize the responder phenotype, we trained a secondary random forest classifier to discriminate between CARD-identified responders and non-responders using baseline covariates. Permutation importance scores identified prior anti-diabetic medication use, younger age, and higher baseline BMI as the primary drivers of treatment response. These findings align with recent real-world evidence suggesting that patients with established metabolic dysfunction and higher baseline weight often experience the most significant relative weight reduction upon initiation of high-potency GLP-1 receptor agonists (Gasoyan et al., 2024).

Overall, this example illustrates both the potential and the limitations of CARD in severely confounded real-world settings. While CARD's scores appear to capture treatment-related signal, the inability to achieve formal discoveries under BH correction underscores the importance of sufficient overlap and moderate confounding for achieving sufficient power. By providing a clear separation between responders and non-responders that is robust to placebo-style testing, CARD allows researchers to prioritize specific subgroups for downstream confirmatory studies.