Research Article

Boyan Duan*, Larry Wasserman, and Aaditya Ramdas

Interactive identification of individuals with positive treatment effect while controlling false discoveries

https://doi.org/10.1515/jci-2023-0059
received September 08, 2023; accepted May 10, 2024

Abstract: Out of the participants in a randomized experiment with anticipated heterogeneous treatment effects, is it possible to identify which subjects have a positive treatment effect? While subgroup analysis has received attention, claims about individual participants are much more challenging. We frame the problem in terms of multiple hypothesis testing: each individual has a null hypothesis (stating that the potential outcomes are equal, for example), and we aim to identify those for whom the null is false (the treatment potential outcome stochastically dominates the control one, for example). We develop a novel algorithm that identifies such a subset, with nonasymptotic control of the false discovery rate (FDR). Our algorithm allows for interaction – a human data scientist (or a computer program) may adaptively guide the algorithm in a data-dependent manner to gain power. We show how to extend the methods to observational settings and achieve a type of doubly robust FDR control. We also propose several extensions: (a) relaxing the null to nonpositive effects, (b) moving from unpaired to paired samples, and (c) subgroup identification. We demonstrate via numerical experiments and theoretical analysis that the proposed method has valid FDR control in finite samples and reasonably high identification power.

Keywords: individual treatment effect, heterogeneous treatment effect, multiple testing, subgroup discovery

MSC 2020: 62M07

1 Introduction

Subgroup identification – or identifying subgroups of the population that have some positive response to a treatment – has been a major topic in the clinical trial community and the causal literature (see Lipkovich et al. [1], Powers et al. [2], Loh et al. [3], and the references therein). Typically, the treatment effect in the investigated population varies with the subject's gender, age, and other covariates. Identifying subjects with positive effects can help guide follow-up research and provide medical guidance. However, most existing methods do not have an error control guarantee at the level of the individual – it is possible that most subjects in the identified subgroup do not have positive effects. For example, an identified subgroup could be "female subjects younger than 40," which may typically mean that the average treatment effect is positive in this subgroup, but it is possible that only 10% of them with age between 18 and 20 years may truly have a positive treatment effect.

*** Corresponding author: Boyan Duan,** Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, 15213, United States of America, now at Google, e-mail: boyand@alumni.cmu.edu
Larry Wasserman: Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, 15213, United States of America, e-mail: larry@stat.cmu.edu
Aaditya Ramdas: Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA, 15213, United States of America, e-mail: aramdas@stat.cmu.edu

Open Access. © 2024 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

This article considers the identification of positive effects at an individual level: among the participants in a trial, which ones have the treatment potential outcome larger than control potential outcome (we call this the subject's treatment effect)? This is a hard question to answer in general, because we only observe one or the other potential outcome. But we will give a nontrivial answer that may be powerless in the worst case, but powerful in cases where the covariates are informative about the treatment effects, and never making too many false claims.

For each subject, the available data we have include its treatment assignment, covariates, and observed outcome. We will identify a set of subjects as having positive effects, with an error control on false identification, and without any modeling assumption on how the (treated/control) outcomes vary with covariates. Below is an example to visualize the problem and prepare for our solution.

1.1 An illustrative example

Consider an experiment involving 500 subjects, each is assigned to treatment or control independently with probability 1/2. Suppose each subject is associated with two simple covariates, which are independently and uniformly distributed in $[0, 1]$. All the data are shown in Figure 1, where the observed outcomes are shown in Figure 1b with darker color indicating a higher outcome. Readers may have some intuitive guess on our interested question: "Which subjects could have positive treatment effect if treated?", such as subjects on the top right corner. Yet we note that the underlying ground truth can be counter-intuitive while mostly correctly captured by our proposed algorithm (results in Section 1.4). Before showing the results, we formalize our question of interest in the next section.

1.2 Problem setup

Suppose we have n subjects in the dataset. Each subject i has potential control outcome Y_i^C, potential treated outcome Y_i^T, and the treatment indicator A_i for $i \in [n] \equiv \{1, 2, \ldots, n\}$. Our results allow the potential outcomes to either be viewed as random variables or fixed. The treatment effect of subject i is defined as $Y_i^T - Y_i^C$ and the observed outcome is $Y_i = Y_i^C(1 - A_i) + Y_i^T A_i$ under the standard causal assumption of consistency ($Y_i = Y_i^T$

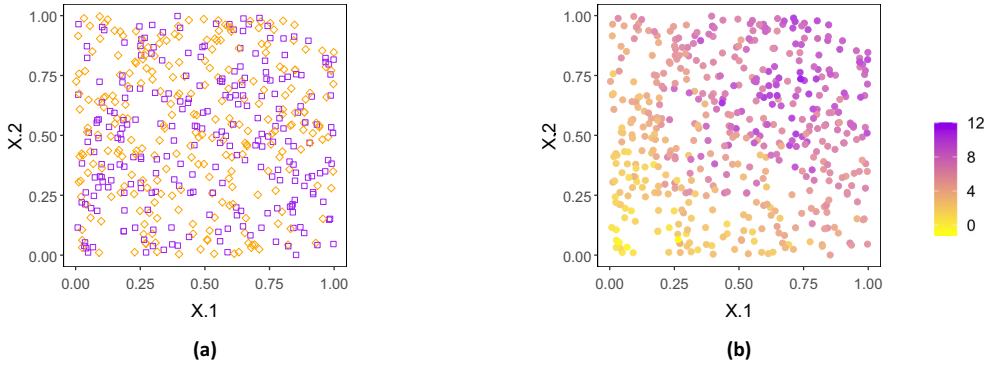

Figure 1: An illustrative example with 500 subjects, each has two recorded covariates. Every point is a subject. The treatment assignments are in the left plot (squares are treated, diamonds are not), and the observed outcomes are in the right plot. We hope to answer the question: which individuals have a positive treatment effect? (a) Subjects in the treated group and control group separated by two shapes and colors and (b) higher outcomes are indicated by darker color.

when $A_i = 1$ and $Y_i = Y_i^C$ when $A_i = 0$). Person i's covariate is denoted as X_i. We first focus on Bernoulli randomized experiments without interference:

(i) conditional on covariates, treatment assignments are independent coin flips:

$$\mathbb{P}[(A_1, \ldots, A_n) = (a_1, \ldots, a_n) | X_1, \ldots, X_n] = \prod_{i=1}^{n} \mathbb{P}(A_i = a_i) = (1/2)^n, \tag{1}$$

for any $(a_1, \ldots, a_n) \in \{0, 1\}^n$. The aforementioned setting is later extended to observational studies where the probabilities of receiving treatments can be heterogeneous and possibly unknown.

(ii) conditional on covariates, the outcome of one subject Y_i is independent of the assignment A_j of another subject, for any $i \neq j$:

$$Y_i \perp A_j | \{X_1, \ldots, X_n\} \quad \text{for } i \neq j, \tag{2}$$

which is implied by (1) when the potential outcomes are viewed as fixed values.

We do not assume the observed data (Y_i, A_i, X_i) are identically distributed. We consider heterogeneous effects in the sense that the distribution of $Y_i^T - Y_i^C$ varies, and aim at identifying those individuals with a positive treatment effect. (If the covariates X_i are not informative about the heterogeneity in $Y_i^T - Y_i^C$, our identification power could be low, and this is to be expected.) We choose to formalize and frame the problem in terms of multiple hypothesis testing, by first defining the null hypothesis for subject i as having zero treatment effect:

$$H_{0i}^{\text{zero}} : (Y_i^T | X_i) \overset{d}{=} (Y_i^C | X_i), \tag{3}$$

or equivalently, $H_{0i}^{\text{zero}} : (Y_i | A_i = 1, X_i) \overset{d}{=} (Y_i | A_i = 0, X_i)$.[1]

An extension is introduced in Appendix A.1, where we relax the null as those with a nonpositive effect, defined by stochastic dominance $(Y_i^T | X_i) \preccurlyeq (Y_i^C | X_i)$, meaning that $\mathbb{P}(Y_i^T \leq y | X_i) \leq \mathbb{P}(Y_i^C \leq y | X_i)$ or simply $Y_i^T \leq Y_i^C$ if the potential outcomes are fixed.

Our algorithms control the error of falsely identifying subjects whose null hypothesis is true (i.e., having zero effect), and aim at correctly identifying subjects with positive effects. Let $\succ$ denote stochastic dominance, as mentioned earlier. We say a subject has a *positive effect* if

$$(Y_i^T | X_i) \succ (Y_i^C | X_i). \tag{4}$$

When treating the potential outcomes and covariates as fixed, we simply write $Y_i^T > Y_i^C$.

The output of our proposed algorithms is a set of identified subjects, denoted as $\mathcal{R}$, with a guarantee that the expected proportion of falsely identified subjects is upper bounded. Specifically, denote the set of subjects that are true nulls as $\mathcal{H}_0 := \{i \in [n] : H_{0i}^{\text{zero}} \text{ is true}\}$. Then the number of false identifications is $|\mathcal{R} \cap \mathcal{H}_0|$. The expected proportion of false identifications is a standard error metric, and the false discovery rate (FDR):

$$\text{FDR} := \mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max\{|\mathcal{R}|, 1\}}\right]. \tag{5}$$

Given $\alpha \in (0, 1)$, we propose algorithms that guarantee FDR $\leq \alpha$ and have reasonably high *power*, which is defined as the expected proportion of correctly identified subjects:

$$\text{power} := \mathbb{E}\left[\frac{|\mathcal{R} \cap \text{Pos}|}{\max\{|\text{Pos}|, 1\}}\right],$$

where $\text{Pos} := \{i : (Y_i^T | X_i) \succ (Y_i^C | X_i)\}$ or $\text{Pos} := \{i : Y_i^T > Y_i^C\}$ are subjects with positive effects.

---

1 Alternatively, we can treat the potential outcomes and covariates as fixed, and frame the null hypothesis as $H_{0i}^{\text{zero}} : Y_i^T = Y_i^C$. A last, hybrid, version (e.g., Howard and Pimentel [4]) is to treat the two potential outcomes as random with joint distribution $(Y_i^T, Y_i^C) | X_i \sim P_i$, and the null posits $H_{0i}^{\text{zero}} : Y_i^T = Y_i^C$ almost surely-P_i, meaning that P_i is supported on $\{(x, y) : x = y\}$. All our theoretical results work with any interpretation, but we stick to (3) by default.

1.3 Related problem: error control in subgroup identification

We note that our problem setup is not exactly the same as most work in subgroup identification, such as Foster et al. [5], Zhao et al. [6], and Imai and Ratkovic [7]. The identified subgroups are usually defined by functions of covariates, rather than a subset of the investigated subjects as in our paper. While defining the subgroup by a function of covariates makes it easy to generalize the finding in the investigated sample to a larger population, it does not seem straightforward to nonasymptotically control the error of false identifications using the former definition, which is a major distinction between previous studies and our work. Most existing work does not have an error control guarantee (see an overview in Lipkovich et al. [1], Table XV), except a few discussing error control on the level of subgroups as opposed to the level of individuals in our article. The difference between FDR control at a subgroup level and at an individual level is detailed below.

1.3.1 Subgroup FDR control

Karmakar et al. [8], Gu and Shen [9], and Xie et al. [10] discuss FDR control at a subgroup level, where the latter two have little discussion on incorporating continuous covariates and require parametric assumptions on the outcomes. Thus, we follow the setup presented in Karmakar et al. [8] to compare the FDR control at a subgroup level (in their article) and individual level (in our paper). Let the subgroups be nonoverlapping sets $\{\mathcal{G}_1, \ldots, \mathcal{G}_G\}$. The null hypothesis for a subgroup $\mathcal{G}_g$ is defined as follows:

$$\mathcal{H}_{0g} : H_{0i}^{\text{zero}} \text{ is true for all } i \in \mathcal{G}_g,$$

or equivalently, $\mathcal{H}_{0g} : \mathcal{G}_g \subseteq \mathcal{H}_0$ (recall $\mathcal{H}_0$ is the set of subjects with zero effect). Let D_g be the 0/1-valued indicator function for whether $\mathcal{H}_{0g}$ is identified or not. The FDR at a subgroup level is defined as the expected proportion of falsely identified subgroups:

$$\text{FDR}^{\text{subgroup}} := \mathbb{E}\left[\frac{|\{g \in [G] : \mathcal{G}_g \subseteq \mathcal{H}_0, D_g = 1\}|}{\max\{|\{g \in [G] : D_g = 1|, 1\}}\right], \tag{6}$$

which collapses to the FDR at an individual level as defined in (5) when each subgroup has exactly one subject. Although our interactive procedure is designed for FDR control at an individual level, we propose extensions to FDR control at a subgroup level in Section 6. As a brief summary, Karmakar et al. [8] propose to control $\text{FDR}^{\text{subgroup}}$ by constructing a p-value for each subgroup and apply the classical BH method [11]. However, it is not trivially applicable to control FDR at an individual level, because their p-values would only take value 1/2 or 1 when each subgroup has exactly one subject, leading to zero identification power following either the classic BH procedure or the more recent AdaPT framework in Lei and Fithian [12]. In other cases where subgroups have more than one subject, the aforementioned error control does not imply whether subjects within a rejected subgroup are mostly non-nulls, or if many are nulls with zero effect. Such error control can be too weak to effectively tell apart most subgroups. Our article appears to be the first to propose methods for identifying subjects having positive effects with (finite sample) FDR control. Individual level inferences are more fine-grained, from which (a) researchers can proceed with a follow-up analysis on identified individuals whom they strongly believe benefit from the treatment; (b) the identified individuals have more trust in the treatment, and as an example, in industry, one can recommend a new product to identified customers with much higher confidence). Importantly, the identified individuals need not have been originally treated, as seen in our representative example (Figures 1 and 3(b)).

We end by noting that we do propose an extension to our method for subgroup identification as well, and compare it to the study by Karmakar et al. [8] in Section 6.

1.3.2 Other related error control at a subgroup level

Cai et al. [13] and Athey and Imbens [14] develop confidence intervals for the averaged treatment effect within subgroups, where the former assumes the size of each subgroup to be large, and the latter requires a separate sample for inference. These intervals can potentially be used to generate a p-value for each subgroup and control FDR at a subgroup level via standard multiple testing procedures, but no explicit discussion is provided. Lipkovich et al. [15], Lipkovich and Dmitrienko [16], Sivaganesan et al. [17], and Berger et al. [18] propose methods with control on a different error metric: the global type-I error, which is the probability of identifying any subgroup when no subject has nonzero treatment effect (i.e., H_{0i}^{zero} is true for all subjects). Our FDR control guarantee implies valid global type-I error, and FDR control is more informative on the correctness of the identified subgroups/subjects when there exist subjects having nonzero effects.

1.4 An overview of our procedure

As discussed, it appears to be new and practically interesting to provide FDR control guarantees at an individual level. Another merit of our proposed method is that it allows a human analyst and an algorithm to interact, to better accomplish the goal.

Interactive testing is a recent idea that emerged in response to the growing practical needs of allowing human interaction in the process of data analysis. In practice, analysts tend to try several methods or models on the same dataset until the results are satisfying, but this violates the validity of standard testing methods (e.g., invalid FDR control). In our context of identifying positive effects, the appealing advantages of an interactive test include that (a) an analyst is allowed to use (partial) data, together with prior knowledge, to design a strategy of selecting subjects potentially having positive effects, and (b) it is a multistep iterative procedure during which the analyst can monitor performance of the current strategy and make adjustments on the selection strategy at any step (at the cost of not altering earlier steps). Despite the flexibility of an analyst to design and alter the algorithm using (partial) data, our proposed procedure always maintains valid FDR control. We name our proposed algorithm I^3 (I-cube), for interactive identification of individual treatment effects.

The core idea that enables human interaction is to separate the information used for selecting subjects with positive effects and that for error control, via "masking and unmasking" (Figure 2). In short, masking means we hide the treatment assignment $\{A_i\}_{i=1}^n$ from the analyst. The algorithm alternates between two steps – selection and error control – until a simple stopping criterion introduced later is reached. Below is a intuitive description of the framework, and we give the full method in Section 2.

- **Selection**. Consider a set of candidate subjects to be identified as having a positive effect (whose null to be rejected), denoted as rejection set $\mathcal{R}_t$ for iteration t. We start with all the subjects included, $\mathcal{R}_0 = [n]$. At each iteration, the analyst excludes possible nulls (i.e., subjects that are unlikely to have positive effects) from the previous $\mathcal{R}_{t-1}$, using all the available information (outcomes Y_i and covariates X_i for all subjects $i \in [n]$, and progressively unmasked A_i from the step of error control, and possible prior information). Note that our method does not automatically use prior information and the revealed data. The analyst is free to use any black-box prediction algorithm that uses the available information and evaluates the subjects possibly using an estimated probability of having a positive treatment effect. This step is where a human is allowed to incorporate her subjective choices.
- **Error control (and unmasking)**. The algorithm uses the complete data $\{Y_i, A_i, X_i\}$ to estimate FDR of the current candidate rejection set $\widehat{\text{FDR}}(\mathcal{R}_t)$, as feedback to the analyst. If the estimated FDR is above the target level $\widehat{\text{FDR}}(\mathcal{R}_t) > \alpha$, the analyst goes back to the step of selection, along with additional information: the excluded subjects ($i \notin \mathcal{R}_t$) have their A_i unmasked (revealed), which could improve her understanding of the data and guide her choices in the next selection step.

The algorithms we propose in the main article build on and modify the aforementioned procedure to achieve reasonably high power and develop various extensions.

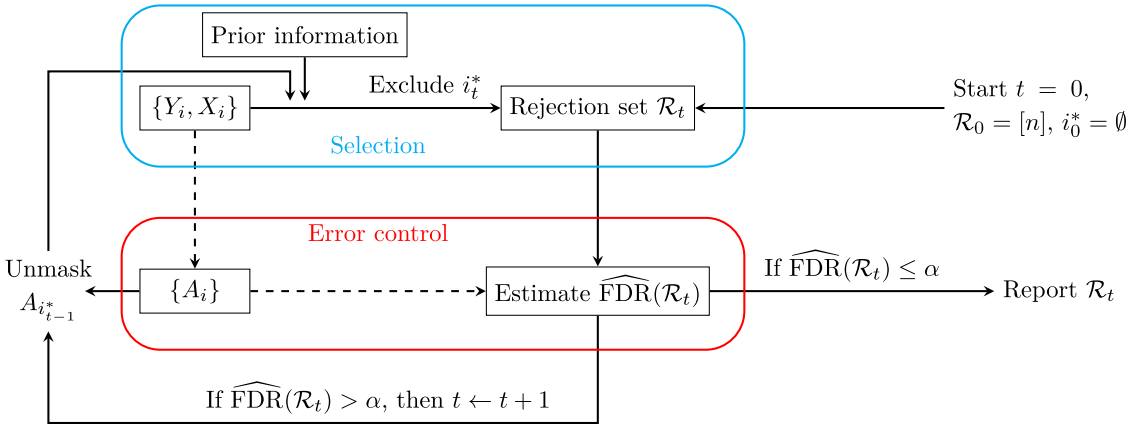

Figure 2: A schematic of the I^3 algorithm. All treatment assignments are initially kept hidden: only $(Y_i, X_i)_{i\in[n]}$ are revealed to the analyst, while all $\{A_i\}$ remain 'masked'. The initial candidate rejection set is $\mathcal{R}_0 = [n]$ (thus, no subject is excluded initially and $i_0^* = \varnothing$). The false discovery proportion $\widehat{\mathrm{FDR}}$ of the current candidate set $\mathcal{R}_t$ is estimated by the algorithm (dashed lines), and reported to the analyst. If $\widehat{\mathrm{FDR}}(\mathcal{R}_t) > \alpha$, the analyst chooses a subject i_t^* to remove it from the proposed rejection set $\mathcal{R}_t = \mathcal{R}_{t1}\backslash\{i_t^*\}$, whose assignment $A_{i_t^*}$ is then "unmasked" (revealed). Importantly, using any available prior information, covariates and working model, the analyst can choose subject i_t^* and shrink $\mathcal{R}_t$ in any manner. This process continues until $\widehat{\mathrm{FDR}}(\mathcal{R}_t) \le \alpha$ (or $\mathcal{R}_t = \varnothing$).

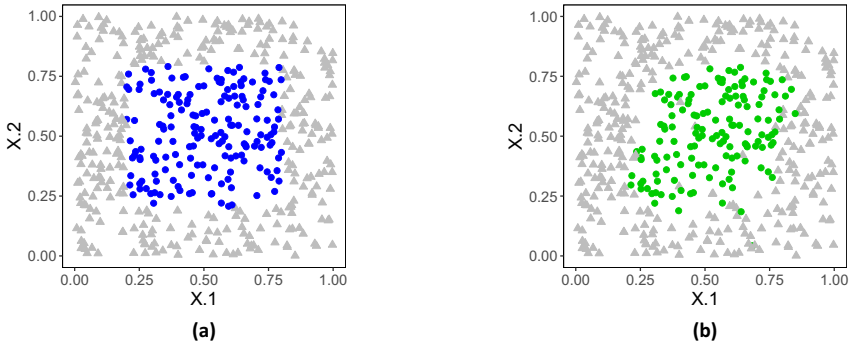

Figure 3: An illustrative example with 500 subjects, each has two recorded covariates. The Crossfit-I^3 identifies most subjects with positive effects, although about half of them did not receive treatment and their potential treated outcomes were not observed. (a) Blue round dots represent subjects with true positive effect (unknown ground truth), and (b) green round dots represent subjects identified by the Crossfit-I^3.

Recall our illustrative example in Section 1.1, the underlying groundtruth and the identifications made by the Crossfit-I^3 (our central algorithm) are shown in Figure 3. Although the observed outcomes tend to be higher for subjects in the top right corner, the subjects with true positive effects are in the center. Such discrepancy is because the *control* outcome varies with covariates (could often happen in practice) and is designed to be higher in the top right corner, such that their observed outcomes could be high regardless of whether the treatment has any effect. Nonetheless, our proposed algorithm can tell the difference of high observed outcomes caused by high control outcomes versus those caused by positive treatment effects, and correctly identify most true positive effects.

1.4.1 Related work in testing

Testing procedures that allow human interaction are first proposed by Lei and Fithian [12] and Lei et al. [19] for the problem of FDR control in multiple testing, followed by several works for other error metrics, such as Duan et al. [20,21]. These articles focus on generic multiple testing problems, which operate on the *p*-values and

ignore the process of generating p-values from data. In contrast, this article applies the idea of interactive testing to observed data and propose tests in the context of causal inference for the treatment effect. To our knowledge, we are not aware of p-values for individual identification in causal inference that are not binary and impose no model assumption. An example of binary p-value is $\mathbb{1}(\widehat{\Delta}_i \le 0)$, where $\widehat{\Delta}_i$ can be viewed as a treatment effect estimator and defined in (8), and the p-value satisfies the mirror-conservative property in Lei and Fithian [12]. However, the masking framework then results in trivial guesses because each masked p-value is $\{0, 1\}$ for all individuals. We view our contribution as (1) framing formally the question of selecting individuals with a positive effect as a multiple testing problem with FDR control; and (2) overcoming the challenge of designing p-value of individual zero effect hypothesis without any model assumption, by connecting the key property that enables masking framework in Lei and Fithian [12], independence between two functions of a single numeric p-value, and the key observation from our paper – independence between treatment assignment A_i and the vector of outcome and covariates $\{Y_i, X_i\}$. The interactive tests first stem from the knockoff method for regression by Barber and Candès [22], which relies on the symmetry of the carefully-designed test statistics. Here, we construct symmetric statistics in the framework of causal inference.

1.4.2 Paper outline

The rest of the article is organized as follows. In Section 2, we describe an interactive algorithm wrapped by a cross-fitting framework, which identifies subjects with positive effects with FDR control. We evaluate our proposed algorithm numerically in Section 3 and provide theoretical power analysis in simple settings in Section 4. Section 5 presents an extension to the setting of observational studies, and we implement the proposed algorithm to real datasets in Section 8. More extensions can be found in Appendix A. Section 9 concludes the article with a discussion on the potential of our proposed interactive procedures.

2 An interactive algorithm with FDR control

To enable us to effectively infer the treatment effect, we use the following *working model*:

$$Y_i^C = f(X_i) + U_i \text{ and } Y_i^T = \Delta(X_i) + f(X_i) + U_i, \tag{7}$$

where U_i is zero-mean noise (unexplained variance) that is independent of A_i. When working with such a model, we effectively want to identify subjects with a positive treatment effect $\Delta(X_i)$. Importantly, model (7) needs not be correctly specified or accurately reflect reality in order for the algorithms in this article to have a valid FDR control. However, the more ill-specified or inaccurate the model is, the more power may be hurt; related numerical experiments are included in Appendix D.

To identify subjects with positive effects, we first introduce an estimator of the treatment effect $\Delta(X_i)$ following the working model (7). Denote the expected outcome given the covariates as $m(X_i) \coloneqq \mathbb{E}(Y_i|X_i)$, and let $\widehat{m}(X_i)$ be an arbitrary estimator of $m(X_i)$ using the outcomes and covariates $\{Y_i, X_i\}_{i=1}^n$. Define the *residual* as $E_i \coloneqq Y_i - \widehat{m}(X_i)$, and an estimator of $\Delta(X_i)$ is

$$\widehat{\Delta}_i \coloneqq 4(A_i - 1/2)\cdot E_i, \tag{8}$$

which, under randomized experiments, is equivalent to the nonparametric estimator of the conditional treatment average effect $\mathbb{E}(Y_i^T|X_i) - \mathbb{E}(Y_i^C|X_i)$ in several recent papers [23,24] and can be traced back to the semiparametrics literature with Robinson [25]. A critical property of $\widehat{\Delta}_i$ that later leads to FDR control is that[2]

$$\mathbb{P}(\widehat{\Delta}_i > 0 | \{Y_j, X_j, E_j\}_{j=1}^n) \le 1/2, \tag{9}$$

2 Note that property (9) uses the fact that the outcome estimator $\widehat{m}(X_i)$ is independent of A_i, so it is important that the estimation of $\widehat{m}$ does not use the assignments $\{A_i\}_{i=1}^n$; however, it should not affect the estimation much because $m(X_i) \equiv \mathbb{E}(Y_i|X_i)$ is not a function of A_i.

under H_{0i}^{zero} in (3), because H_{0i}^{zero} implies $A_i \perp\!\!\!\perp \{Y_i, X_i\}$ and $\mathbb{P}(A_i - 1/2 > 0) = 1/2$. With the rigorous argument for validity in Appendix B.1, we briefly describe the reasoning here: the condition in (9) corresponds to the information used for selecting subjects (recall in Figure 2), because the treatment assignments A_i are hidden for candidates in $\mathcal{R}_t$ and assignments are mutually independent. The aforementioned property indicates that the estimated effect $\widehat{\Delta}_i$ is no more likely to be positive than negative if the selected subject has zero effect, regardless of how the analyst decides which subject to select. Therefore, the sign of $\widehat{\Delta}_i$ can be used to estimate the number of false identifications and achieve FDR control, which we elaborate next.

2.1 The I^3 algorithm

This section presents the I^3 algorithm and proves that it controls FDR. We introduce a modification based on cross-fitting that improves identification power in the next section.

The I^3 proceeds as progressively shrinking a candidate rejection set $\mathcal{R}_t$ at iteration t,

$$[n] = \mathcal{R}_0 \supseteq \mathcal{R}_1 \supseteq \ldots \supseteq \mathcal{R}_n = \varnothing,$$

where recall $[n]$ denotes the set of all subjects. We assume without loss of generality that one subject is excluded in each step. Denote the subject excluded at iteration t as i_t^*. The choice of i_t^* can use the information available to the analyst before iteration t, where we follow Lei and Fithian [12] to describe the information available to the analysis formally as a filtration (sequence of nested σ-fields):

$$\mathcal{F}_{t-1} = \sigma\left(\{Y_j, X_j\}_{j\in\mathcal{R}_{t-1}}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}}, \sum_{j\in\mathcal{R}_{t-1}} \mathbb{1}\{\widehat{\Delta}_j > 0\}\right), \tag{10}$$

or equivalently,

$$\mathcal{F}_{t-1} = \sigma\left(\{Y_j, X_j\}_{j\in[n]}, \{A_j\}_{j\notin\mathcal{R}_{t-1}}, \sum_{j\in\mathcal{R}_{t-1}} \mathbb{1}\{\widehat{\Delta}_j > 0\}\right), \tag{11}$$

where we unmask (reveal) the treatment assignments A_j for subjects excluded from $\mathcal{R}_{t-1}$, and the sum $\sum_{i\in\mathcal{R}_{t-1}}\mathbb{1}\{\widehat{\Delta}_i > 0\}$ is mainly used for FDR estimation as we describe later. The aforementioned available information include arbitrary functions of the revealed data, such as the residuals $\{E_j\}_{j=1}^n$ defined aforementioned equation (8). Similar to property (9), for each candidate subject $i \in \mathcal{R}_{t-1}$, we have

$$\mathbb{P}(\widehat{\Delta}_i > 0 | \{Y_j, X_j\}_{j\in\mathcal{R}_{t-1}}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}}) \le 1/2, \tag{12}$$

which ensures the FDR control as we explain next.

To control FDR, the number of false identifications is estimated by (12). The idea is to partition the candidate rejection set $\mathcal{R}_t$ into $\mathcal{R}_t^+$ and $\mathcal{R}_t^-$ by the sign of $\widehat{\Delta}_i$:

$$\mathcal{R}_t^- \coloneqq \{i \in \mathcal{R}_t : \widehat{\Delta}_i \le 0\}, \quad \mathcal{R}_t^+ \coloneqq \{i \in \mathcal{R}_t : \widehat{\Delta}_i > 0\}.$$

Notice that our proposed procedure only identifies the subjects whose estimated effect is positive, i.e., those in $\mathcal{R}_t^+$. Thus, the FDR is $\mathbb{E}\left[\frac{|\mathcal{R}_t^+ \cap \mathcal{H}_0|}{\max\{|\mathcal{R}_t^+|, 1\}}\right]$ by definition, where recall $\mathcal{H}_0$ is the set of true nulls. Intuitively, the number of false identifications $|\mathcal{R}_t^+ \cap \mathcal{H}_0|$ can be approximately upper bounded by $|\mathcal{R}_t^- \cap \mathcal{H}_0|$, since the number of positive signs should be no larger than the number of negative signs for the falsely identified nulls, according to property (9). Note that the set of true nulls $\mathcal{H}_0$ is unknown, so we use $|\mathcal{R}_t^-|$ to upper bound $|\mathcal{R}_t^- \cap \mathcal{H}_0|$, and propose an estimator of FDR for the candidate rejection set $\mathcal{R}_t$:

$$\widehat{\text{FDR}}(\mathcal{R}_t) = \frac{|\mathcal{R}_t^-| + 1}{\max\{|\mathcal{R}_t^+|, 1\}}. \tag{13}$$

Such FDR estimators using the count of positive or negative individuals stem from Barber and Candès [22], and have been extensively used in the literature of interactive testing such as Lei and Fithian [12], Lei et al. [19], and Duan et al. [20,21]. Overall, the I^3 shrinks $\mathcal{R}_t$ until time $\tau \coloneqq \inf\{t : \widehat{\mathrm{FDR}}(\mathcal{R}_t) \le \alpha\}$ and identifies only the subjects in $\mathcal{R}^+_\tau$, as summarized in Algorithm 1. We state the FDR control of I^3 in Theorem 1, and the proof can be found in Appendix B.1.

Theorem 1. *In a randomized experiment with assumptions* (1) *and* (2)*, and for any analyst that updates their working model(s) at any iteration t using the information in $\mathcal{F}_{t-1}$, the set R^+_τ rejected by the* I^3 *algorithm has FDR controlled at level α, meaning that*

$$\mathbb{E}\left[\frac{|\mathcal{R}^+_\tau \cap \mathcal{H}_0|}{\max\{|\mathcal{R}^+_\tau|, 1\}}\right] \le \alpha,$$

for the null hypothesis (3).

Consider a simple case where model (7) is accurate for every subject with a constant treatment effect $\Delta(X_i) = \delta > 0$. If δ is larger than the maximum noise, we have $\mathcal{R}^+_0 = [n]$, and the algorithm can stop at the very first step identifying all subjects. At the other extreme, if the effect δ is too small, the algorithm may also return an empty set, and this makes sense because while small *average* treatment effects can be learned using a large population, larger treatment effects are needed for *individual-level* identification.

Algorithm 1. The I^3 (interactive identification of individual treatment effect) procedure.

Initial state: Statistician (S) knows covariates and outcomes $\{X_i, Y_i\}_{i=1}^n$.
Computer (C) knows the treatment assignments $\{A_i\}_{i=1}^n$.
Target FDR level α is public knowledge.
Initial exchange: Both players initialize $\mathcal{R}_0 = [n]$ and set $t = 1$.
1. S builds a prediction model $\widehat{m}$ from X_i to Y_i.
2. S informs C about residuals $E_i \equiv Y_i - \widehat{m}(X_i)$.
3. C estimates the treatment effect as $\widehat{\Delta}_i \equiv 4(A_i - 1/2)E_i$.
4. C then divides $\mathcal{R}_t$ into $\mathcal{R}^-_t \coloneqq \{i \in \mathcal{R}_t : \widehat{\Delta}_i \le 0\}$ and $\mathcal{R}^+_t \coloneqq \{i \in \mathcal{R}_t : \widehat{\Delta}_i > 0\}$.
5. C reveals only $|\mathcal{R}^+_t|$ to S (who infers $|\mathcal{R}^-_t|$).
Repeated interaction: 6. S checks if $\widehat{\mathrm{FDR}}(\mathcal{R}_t) \equiv \frac{|\mathcal{R}^-_t| + 1}{\max\{|\mathcal{R}^+_t|, 1\}} \le \alpha$.
7. If yes, S sets $\tau = t$, reports $\mathcal{R}^+_\tau$ and exits.
8. Else, S picks any $i^*_t \in \mathcal{R}_{t-1}$ using everything S currently knows.
(S tries to pick an i^*_t that S thinks is null, i.e., S hopes that $\widehat{\Delta}_{i^*_t} \le 0$.)
9. C reveals $A_{i^*_t}$ to S, who also infers $\widehat{\Delta}_{i^*_t}$ and its sign.
10. S updates $\mathcal{R}_{t+1} = \mathcal{R}_t \backslash \{i^*_t\}$, and also $|\mathcal{R}^+_{t+1}|$ and $|\mathcal{R}^-_{t+1}|$.
11. Increment t and go back to Step 6.

We end the section with a remark. In step 8 of Algorithm 1, we hope to exclude subjects that are unlikely to have positive effects, based on the revealed data in $\mathcal{F}_{t-1}$. In other words, we should guess the sign of treatment effect $\widehat{\Delta}_i$, which depends on both the revealed data $\{Y_i, X_i\}$ and the hidden assignment A_i. However, notice that at the first iteration, we may learn/guess the opposite signs for all the subjects; when all assignments $\{A_i\}_{i=1}^n$ are hidden at $t = 1$, the likelihood of $\{A_i\}_{i=1}^n$ being the true values (leading to all correct signs for $\widehat{\Delta}_i$) is the same as the likelihood of all opposite values (leading to all opposite signs for $\widehat{\Delta}_i$), no matter what working model we use. Consequently, the subjects with large positive effects could be guessed as having large negative effects, causing them to be excluded from the rejection set. To improve power, we propose to wrap around the I^3 by a cross-fitting framework as described in the next section.

2.2 Improving stability and power with Crossfit-I^3

Cross-fitting refers to the idea of splitting the samples into two halves. We perform the I^3 on each half separately, so that for each half, the complete data (including the assignments) of the other half is revealed to the analyst to help infer the sign of treatment effect, addressing the issue of learning the opposite signs and improving the identification power.

Specifically, split the subjects randomly into two sets of equal size, denoted as $\mathcal{I}$ and $\mathcal{II}$, where $\mathcal{I} \cup \mathcal{II} = [n]$. The I^3 (Algorithm 1) is implemented on each set separately: at the start of I^3 on set $\mathcal{I}$, the analyst has access to the complete data for all subjects in set $\mathcal{II}$, and tries to identify subjects with positive effects in set $\mathcal{I}$ with FDR control at level $\alpha/2$; similar is the I^3 on set $\mathcal{II}$. Mathematically, let the candidate rejection set of implementing the I^3 on set $\mathcal{I}$ be $\mathcal{R}_t(\mathcal{I})$, where the initial set is $\mathcal{R}_0(\mathcal{I}) = \mathcal{I}$. The available information at iteration t is defined as follows:

$$\mathcal{F}_{t-1}(\mathcal{I}) = \sigma\left(\{Y_i, X_i\}_{i\in\mathcal{R}_{t-1}(\mathcal{I})}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}(\mathcal{I})}, \sum_{i\in\mathcal{R}_{t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i > 0\}\right), \tag{14}$$

which includes the complete data $\{Y_j, A_j, X_j\}$ for $j \in \mathcal{II}$ at any iteration $t \geq 0$[3]. Similarly, we define $\mathcal{R}_t(\mathcal{II})$ and $\mathcal{F}_{t-1}(\mathcal{II})$ for the I^3 implemented on set $\mathcal{II}$. The final rejection set is the union of rejections in $\mathcal{I}$ and $\mathcal{II}$ (Algorithm 2). We call this algorithm the Crossfit-I^3.

Algorithm 2 The Crossfit-I^3.

Input: Covariates, outcomes, treatment assignments $\{Y_i, A_i, X_i\}_{i=1}^n$, target level α;
Procedure:
1. Randomly split the sample into two subsets of equal size, denoted as $\mathcal{I}$ and $\mathcal{II}$;
2. Implement Algorithm 1 at level $\alpha/2$, where E initially knows $\{Y_k, X_k\}_{k=1}^n \cup \{A_j\}_{j\in\mathcal{II}}$ and sets $\mathcal{R}_0(\mathcal{I}) = \mathcal{I}$, getting a rejection set $\mathcal{R}_\tau^+(\mathcal{I}) \subseteq \mathcal{I}$;
3. Implement Algorithm 1 at level $\alpha/2$, where E initially knows $\{Y_k, X_k\}_{k=1}^n \cup \{A_j\}_{j\in\mathcal{I}}$ and sets $\mathcal{R}_0(\mathcal{II}) = \mathcal{II}$, getting a rejection set $\mathcal{R}_\tau^+(\mathcal{II}) \subseteq \mathcal{II}$;
4. Combine two rejection sets as the final rejection set, $\mathcal{R}_\tau^+ = \mathcal{R}_\tau^+(\mathcal{I}) \cup \mathcal{R}_\tau^+(\mathcal{II})$.

As long as the I^3 on two sets do not exchange information, Algorithm 2 has a valid FDR control (see the proof in Appendix B.2).

Theorem 2. *Under assumption* (1) *and* (2) *of randomized experiments, $\mathcal{R}_\tau^+$ rejected by the Crossfit-*I^3 *has FDR controlled at level α for the null hypothesis* (3).

In addition to addressing the issue of learning the opposite $\widehat{\Delta}_i$ in the original I^3, another benefit of using the crossing-fitting framework is that with the complete data revealed for at least half of the sample, the analyst does not have to deal with the problem of inferring missing data (the assignment A_i), which probably needs some parametric probabilistic modeling and the EM algorithm. Instead, because the assignments are revealed for subjects not in the candidate rejection set (at least half of the sample), their signs of $\widehat{\Delta}_j$ can be correctly calculated and used as "training data". The analyst can then employ a black-box prediction model, such as a random forest, to predict the signs of $\widehat{\Delta}_i$ for the subjects whose assignments are masked (hidden). As an example, we propose an automated strategy as follows to select a subject at step 8 in Algorithm 1.

3 For notational clarity, we use i to denote candidate subjects $i \in \mathcal{R}_t(\mathcal{I})$, and j for noncandidate subjects $j \notin \mathcal{R}_t(\mathcal{I})$, while k is used to index all subjects $k \in [n]$.

Algorithm 3 An automated heuristic to select i_t^* in the Crossfit-I^3.

Input: Current rejection set $\mathcal{R}_{t-1}(\mathcal{I})$, and available information for selection $\mathcal{F}_{t-1}(\mathcal{I})$;
Procedure:

1. Train a random forest classifier where the label is $\mathrm{sign}(\widehat{\Delta}_j)$ and the predictors are Y_j, X_j and the residuals E_j, using noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;
2. Estimate the probability of $\widehat{\Delta}_i$ being positive as $\widehat{p}(i,t)$ for subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$;
3. Select $i_t^* = \mathrm{argmin}\{\widehat{p}(i,t) : i \in \mathcal{R}_{t-1}(\mathcal{I})\}$.

We remark that in practice, the analyst can interactively change the prediction model, such as exploring parametric models to see which fits the data better. In principle, the analyst can perform any exploratory analysis on data in $\mathcal{F}_{t-1}(\mathcal{I})$ to decide a heuristic or score for selecting subject i_t^*; and the FDR control is valid as long as she does not use the assignments A_i for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$. For computation efficiency, we usually update the prediction models (or their parameters) once every 100 iterations (say).

To summarize, the Crossfit-I^3 described in Algorithm 2 involves two rounds of the I^3 (Algorithm 1), where step 8 of selecting a subject is allowed to involve human interaction; alternatively, step 8 can be an automated heuristic as presented in Algorithm 3.

Remark 1. We contrast our algorithms to identify individual effects with many existing algorithms targeting at alternative goals. As examples, two commonly discussed goals include testing whether *any* individual has any effect [4,26,27] and estimating the averaged treatment effect [28–30]. While the aforementioned two questions discuss causal inference at an integrated level for the investigated population, we study the problem of making claims for each individual subject, a harder problem by its nature. Therefore, it is expected that a larger effect size is required, compared to the previous two goals, to have reasonably high power for individual effect identification. In the following sections, we demonstrate through repeated numerical experiments and theoretical analysis that the Crossfit-I^3 has reasonably high power.

3 Numerical experiments

To assess our proposed procedure, we first describe a baseline method, which calculates a p-value for each subject under the assumption of linear models and applies the classical BH method [11]. We call this method the linear-BH procedure.

3.1 Two baselines: the BH procedure (assuming well-specified model) and selective SeqStep+

BH procedure under linear assumptions. For the treated group and control group, we first separately learn a linear model to predict Y_i using X_i, denoted as $\widehat{l}^T$ and $\widehat{l}^C$. By imputing the unobserved potential outcomes, we obtain estimators of the potential outcomes $\widetilde{Y}_i^T = Y_i \mathbb{1}\{A_i = 1\} + \widehat{l}^T(X_i)\mathbb{1}\{A_i = 0\}$ and $\widetilde{Y}_i^C = \widehat{l}^C(X_i)\mathbb{1}\{A_i = 1\} + Y_i \mathbb{1}\{A_i = 0\}$, and the treatment effect for subject i can be estimated as $\widehat{\Delta}_i^{\mathrm{BH}} := \widetilde{Y}_i^T - \widetilde{Y}_i^C$. If the potential outcomes are linear functions of covariates with standard Gaussian noises (which we refer to as the linear assumption), the estimated treatment effect asymptotically follows a Gaussian distribution. For each subject $i \in [n]$, we calculate a p-value for the zero-effect null (3) as follows:

$$P_i = 1 - \Phi\left(\widehat{\Delta}_i^{\mathrm{BH}} \Big/ \sqrt{\widehat{\mathrm{Var}}(\widehat{\Delta}_i^{\mathrm{BH}})}\right), \tag{15}$$

where the estimated variance is $\widehat{\mathrm{Var}}(\widehat{\Delta}_i^{\mathrm{BH}}) = \widehat{\mathrm{Var}}(\widetilde{Y}_i^T) + \widehat{\mathrm{Var}}(\widetilde{Y}_i^C)$, and Φ denotes the CDF of a standard Gaussian. To identify subjects having positive effects with FDR control, we apply the BH procedure to the above p-values. Note that, unlike our methods, the error control for BH would not hold when the linearity assumption is violated (see Appendix B.7 for the formal FDR control guarantee; see Section 7 for more numerical experiments when the linear assumption holds or does not).

Selective SeqStep+. Once we construct the estimated treatment effect $\widehat{\Delta}_i$ and call out the critical property (9) on the probability of the estimated effect sign, we can apply the Selective SeqStep+ by Barber and Candès [22]. Specifically, we set the p-value for each individual as $p_i = 1 - \frac{1}{2}\mathbb{1}(\widehat{\Delta}_i > 0)$, which equals 1/2 when estimated treatment effect $\widehat{\Delta}_i > 0$ and equals 1 when $\widehat{\Delta}_i <= 0$. The hypotheses can be ordered by $|E_i|$ decreasingly, and the constant c is chosen as 1/2 to maximize the power. We also note that the Selective SeqStep+ can be viewed as an automated version of I^3 where the Statistician (S) picks individuals by the ranking of $|E_i|$ in step 8 of Algorithm 1. We expect the power of our proposed Crossfit-I^3 to be higher than the Selective SeqStep+ (an automated version of I^3), because Crossfit-I^3 additionally uses information from the revealed treatment assignments when picking/ordering individuals, which are especially helpful to inform the direction of treatment effect.

3.2 Numerical experiments and power comparison

We run a simulation with 500 subjects ($n = 500$). Each subject is recorded with two binary attributes (e.g., female/male and senior/junior) and one continue attribute (e.g., body weight), denoted as a vector $X_i = (X_i(1), X_i(2), X_i(3)) \in \{0, 1\}^2 \times \mathbb{R}$. Among n subjects, the binary attributes are marginally balanced, and the subpopulation with $X_i(1) = 1$ and $X_i(2) = 1$ is of size 30. The continuous attribute is independent of the binary ones and follows the distribution of a standard Gaussian.

The outcomes are simulated as a function of the covariates X_i and the assignment A_i following the generating model (7). Recall that we previously used model (7) as a working model, which is not required to be correctly specified. Here, we generate data from such a model in simulation for a clear evaluation of the considered methods. We specify the noise U_i as a standard Gaussian, and the expected control outcome as $f(X_i) = 5(X_i(1) + X_i(2) + X_i(3))$, and the treatment effect as follows:

$$\Delta(X_i) = S_\Delta \cdot [5X_i^3(3)\mathbb{1}\{X_i(3) > 1\} - X_i(1)/2], \tag{16}$$

where $S_\Delta > 0$ encodes the scale of the treatment effect. In this setup, around 15% subjects have positive treatment effects with a large scale, and 43% subjects have a mild negative effect.[4]

For the Crossfit-I^3, we use random forests (with default parameters in R) to compute $\widehat{m}$, and use the automated selection strategy Algorithm 3 to select a subject at step 8 in Algorithm 1. The linear-BH procedure results in a substantially higher FDR than desired because the linear assumption does not hold in the underlying truth (A26) (Figure 4), whereas the Selective SeqStep+ and our proposed Crossfit-I^3 control FDR at the target level as expected. At the same time, the Crossfit-I^3 appears to have higher power than the Selective SeqStep+ and the linear-BH procedure to correctly identify subjects with true positive effects. More experiments that explore different treatment effect setups can be found in Section 7.

4 Asymptotic power analysis in simple settings

In addition to the numerical experiments, we provide a theoretical power analysis in some simple cases to understand the advantages and limitations of our proposed Crossfit-I^3.

First, consider the case without covariates. Our analysis is inspired by the work of Arias-Castro and Chen [31] and Rabinovich et al. [32], who study the power of methods with FDR control under a sparse Gaussian sequence model. Let there be n hypotheses, each associated with a test statistic V_i for $i = 1, \ldots, n$. They consider a class of methods called *threshold procedures* such that the final rejection set $\mathcal{R}$ is in the form

4 R code to fully reproduce all plots in the paper are available at https://github.com/duanby/I-cube.

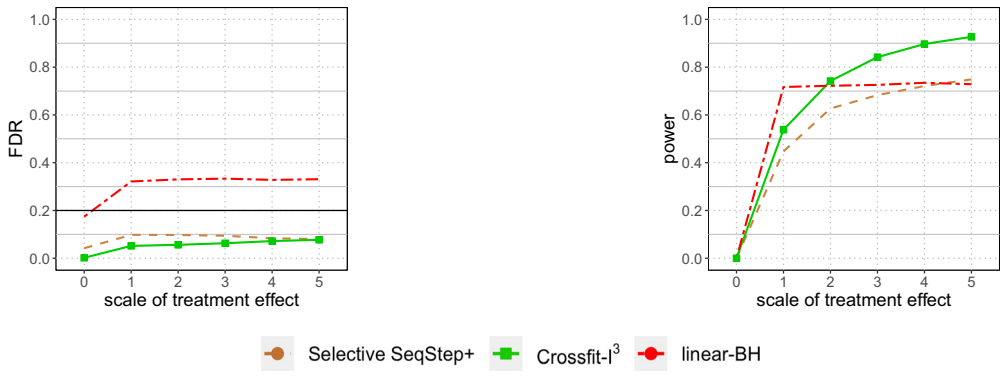

Figure 4: FDR (left) and power (right) of the Crossfit-I^3 compared with the linear-BH procedure and the Selective SeqStep+, with the treatment effect specified as model (A26) and the scale S_Δ varying in $\{0, 1, 2, 3, 4, 5\}$. The FDR control level is 0.2, marked by a horizontal line in error control plots. For all plots in this article, the FDR and power are averaged over 500 repetitions. The Crossfit-I^3 controls FDR while the linear-BH procedure does not because the treatment effect is nonlinear. Also, the Crossfit-I^3 can achieve higher power than both the Selective SeqStep+ and the linear-BH procedures.

$\mathcal{R} = \{i : V_i \geq \tau(V_1, \ldots, V_n)\}$, for some threshold $\tau(V_1, \ldots, V_n)$; they discuss two types of thresholds; see Appendix C for details of their results. An example of the threshold procedure is the BH procedure. Our proposed I^3 can also be simplified to a threshold procedure when using an automated selection strategy at step 8 of Algorithm 1: at each iteration, we exclude the subject with the smallest absolute value of the estimated treatment effect $|\widehat{\Delta}_i|$ (note that this strategy satisfies our requirement of not using assignments since $|\widehat{\Delta}_i| = |4(A_i - 1/2)(Y_i - \widehat{m}(X_i))| = 2|Y_i - \widehat{m}(X_i)|$). The resulting (simplified and automated) I^3 is a threshold procedure where $V_i = \widehat{\Delta}_i$. Note that our original interactive procedure is highly flexible, making the power analysis less obvious, so we discuss the power of Crossfit-I^3 with the aforementioned simplified selection strategy.

To contextualize our power analysis, we paraphrase one of the results in Arias-Castro and Chen [31], Rabinovich et al. [32]. Assume the test statistics $V_i \sim N(\mu_i, 1)$ are independent, with $\mu_i = 0$ under the null and $\mu_i = \mu > 0$ otherwise. Denote the number of non-nulls as n_1 and the *sparsity* of the non-nulls is parameterized by $\beta \in (0, 1)$ such that $n_1/n = n^{-\beta}$. Let the signal μ increase with n as $\mu = \sqrt{2r \log n}$, where the *signal strength* is encoded by $r \in (0, 1)$. Their power analysis is characterized by the signal r and sparsity β, which are also critical parameters to characterize the power in our context as we state later. These authors effectively prove that *for any fixed FDR level $\alpha \in (0, 1)$, no threshold procedure can have nontrivial power if $r < \beta$, but there exist threshold procedures with asymptotic power one if $r > \beta$.*

Our analysis differs from theirs in the non-null distribution of the test statistics. Given n subjects, suppose the potential outcomes for subject i are distributed as follows: $Y_i^C \sim N(0, 1)$ and $Y_i^T \sim N(\mu_i, 1)$, where the alternative mean is $\mu_i = 0$ if subject i is null, or $\mu_i = \mu > 0$ if i is non-null. Thus, the observed outcome of a null is $N(0, 1)$, and that of a non-null is a *mixture* of $N(\mu, 1)$ and $N(0, 1)$ (depending on the treatment assignment), instead of a shift of the null distribution as assumed in Arias-Castro and Chen [31], and the proof of the following result thus involves some modifications on their proofs (Appendix C.1).

Theorem 3. *Given a fixed FDR level $\alpha \in (0, 1)$ and let the number of subjects n go to infinity. When there is no covariate, the automated Crossfit-I^3 and the linear-BH procedure have the same power asymptotically: if $r < \beta$, their power goes to zero; if $r > \beta$, their power goes to 1/2. Further, among the treated subjects, their power goes to one.*

Remark 2. Power of both methods cannot converge to a value larger than 1/2 because without covariates, we cannot differentiate between the subjects with zero effect (whose outcome follows standard Gaussian regardless of treated or not) and the subjects with positive effects that are not treated (which also follows standard Gaussian). And the proportion of untreated subjects among those with positive effects is 1/2 because of the assumed randomization.

The aforementioned theorem discusses the case where there are no covariates to help guess which untreated subjects have positive effects. Next, we consider the case with an "ideal" covariate X_i: its value corresponds to whether a subject is a non-null (having positive effect) or not, $X_i = \mathbb{1}\{\mu_i > 0\}$. Here, we design the selection strategy (for step 8 of Algorithm 1) as a function of the covariates, because we hope that subjects with the similar covariates have similar treatment effects. Specifically, for the I^3 implemented on $\mathcal{I}$, we learn a prediction of $\widehat{\Delta}_j$ by X_j using noncandidate subjects $j \in \mathcal{II}$: $\text{Pred}(x) = \frac{1}{|\mathcal{II}|}\sum_{i\in\mathcal{II}}\widehat{\Delta}_j\mathbb{1}\{X_j = x\}$, where $x = \{0, 1\}$. Then for candidate subjects $i \in \mathcal{I}$, we exclude the ones whose $\text{Pred}(X_i)$ are lower. As we integrate information among subjects with the same covariate value, all non-null subjects (i.e., those with $X_i = 1$) would excluded after the nulls (with probability tending to one), regardless of whether they are treated; hence, we achieve power one.

Theorem 4. *Given a fixed FDR level $\alpha \in (0, 1)$ and let the number of subjects n go to infinity. With a covariate $X_i = \mathbb{1}\{\mu_i > 0\}$, the power of the automated Crossfit-*I^3 *converges to one for any fixed $r \in (0, 1)$ and $\beta \in (0, 1)$. In contrast, the power of the linear-BH procedure goes to zero if $r < \beta$. (When $r > \beta$, power of both methods converges to one.)*

Here is a short informal argument for why our power goes to one (see detailed proof in Appendix C.2). Since the nulls can be excluded before the non-nulls, we focus on the test statistics of the non-nulls. Let $\overset{d}{\to}$ denote convergence in distribution. The estimated effect $\widehat{\Delta}_i \overset{d}{\to} N(\mu, 1)$ for each non-null (since in the notation of Algorithm 1, $\widehat{m}(X_i = 1)$ converges to $\mu/2$ for the non-nulls, and thus, $E_i \overset{d}{\to} N(\mu/2, 1)$ for those with $A_i = 1$, and $E_i \overset{d}{\to} N(-\mu/2, 1)$ for those with $A_i = 0$.) Hence, at the time t_0 right after all the nulls are excluded (and all the non-nulls are in $\mathcal{R}_{t_0}$), the proportion of positive estimated effects $|\mathcal{R}^+_{t_0}|/|\mathcal{R}_{t_0}|$ converges to $\Phi(\mu)$, where Φ denotes the CDF of a standard Gaussian. We can stop before t_0 and identify subjects in $\mathcal{R}^+_{t_0}$ if $\widehat{\text{FDR}}(\mathcal{R}_{t_0})$, as a function of $|\mathcal{R}^+_{t_0}|/|\mathcal{R}_{t_0}|$, is less than α, which holds when $\Phi(\mu) > \frac{1}{1+\alpha}$. The power goes to one because μ grows to infinity for any fixed $r \in (0, 1)$, so that for large n, we stop before t_0 and the proportion of rejected non-nulls $|\mathcal{R}^+_{t_0}|/|\mathcal{R}_{t_0}|$ (which converges to $\Phi(\mu)$ as argued earlier) also goes to one. In short, the power guarantee does not depend on the sparsity β because of the designed selection strategy that incorporates covariates.

We note that our theoretical power analysis discusses two extreme cases, one with no covariate to assist the testing procedure (Theorem 3), and the other with a single "ideal" covariate that equals the indicator of non-nulls (Theorem 4). The numerical experiments in Section 3 consider more practical settings, where the analyst is provided with a mixture of covariates informative about the heterogeneous effect ($X_i(1)$ and $X_i(3)$ in our example) and some uninformative ones; still, the Crossfit-I^3 tends to have reasonably high power. So far, the article discusses the setup of a randomized experiment where each subject has 1/2 probability to be treated; in the following, we present the variant of Crossfit-I^3 for observational studies, where the probabilities of receiving treatment can depend on covariates and unknown.

5 Crossfit-I^3 in observational studies

For clear notation, we denote the true propensity score (probability of receiving treatment) for subject i as π_i, and the estimated one as $\widehat{\pi}_i$ (which we introduce soon). For the setup in observational studies, we introduce an alternative set of assumptions to replace assumption in (1):

(iii) conditional on covariates, treatment assignments are independent:

$$\mathbb{P}[(A_1, \ldots, A_n) = (a_1, \ldots, a_n)|X_1, \ldots, X_n] = \prod_{i=1}^{n}\mathbb{P}(A_i = a_i|X_1, \ldots, X_n) = \prod_{i=1}^{n}\pi_i, \tag{17}$$

for any $(a_1, \ldots, a_n) \in \{0, 1\}^n$ and $\{\pi_i\}_{i=1}^n$ can be unknown.

(iv) the propensity scores are bounded away from 0 and 1:

$$0 < \pi_{\min} \le \pi_i \le \pi_{\max} < 1 \text{ for all } i \in [n]. \tag{18}$$

Because the propensity scores π_i are unknown, we estimate them using the revealed data – specifically in the cross-fitting framework using the complete data for subjects in $\mathcal{II}$. To be exact, we modify the Crossfit-I^3 in algorithm 2 in two components:

- prior to step 2 of implementing the I^3, we estimate the bounds for the propensity scores as $\widehat{\pi_{\min}}(\mathcal{I})$ and $\widehat{\pi_{\min}}(\mathcal{I})$ by the complete data in $\mathcal{II}$. For example, we can estimate individual propensity scores by a logistic regression on covariates X_j using the complete data from noncandidate subjects $j \in \mathcal{II}$, and take their minimum and maximum as the estimations; and
- we modify the implementation of Algorithm 1 in the FDR estimator:

$$\widehat{\mathrm{FDR}}_t^{\hat{\pi}}(\mathcal{R}_t(\mathcal{I})) := \left\{\frac{1}{1 - \max\{1 - \widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}} - 1\right\}\frac{|\mathcal{R}_t^-(\mathcal{I})| + 1}{|\mathcal{R}_t^+(\mathcal{I})| \vee 1}; \tag{19}$$

and we similarly modify the procedure on set $\mathcal{II}$. We call the resulting algorithm Crossfit-I$^3_{\hat{\pi}}$, which have asymptotic FDR control when the propensity scores are well estimated (proof in Appendix B.5). Note that Crossfit-I$^3_{\hat{\pi}}$ does not involve a modification of the treatment effect estimator $\widehat{\Delta}_j$, which will not affect the FDR control as stated in Theorem 5. As a future direction, it would be an interesting extension to modify $\widehat{\Delta}_j$ and utilize the information of estimated propensity scores, and potentially improve the identification power.

Theorem 5. *Suppose there are n samples for identification. In the cross-fitting framework, let* $\widehat{\pi_{\min}}(\mathcal{I})$ *and* $\widehat{\pi_{\max}}(\mathcal{I})$ *be the estimated lower and upper bound of the propensity scores based on data information in* $\mathcal{F}_0(\mathcal{I})$. *Denote the estimation error as follows*:

$$\varepsilon_n^{\pi}(\mathcal{I}) := \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}[\max\{\widehat{\pi_{\min}}(\mathcal{I}) - \pi_{\min}, \pi_{\max} - \widehat{\pi_{\max}}(\mathcal{I}), 0\}],$$

and similarly define $\varepsilon_n^{\pi}(\mathcal{II})$. *Let* $\varepsilon_n^{\pi} = \varepsilon_n^{\pi}(\mathcal{I}) + \varepsilon_n^{\pi}(\mathcal{II})$, *and the FDR of Crossfit-*I$^3_{\hat{\pi}}$ *is upper bounded*:

$$\mathrm{FDR} \le \alpha\left\{1 + \varepsilon_n^{\pi}\left(\frac{4}{\max\{1 - \pi_{\min}, \pi_{\max}\}(1 - \max\{1 - \pi_{\min}, \pi_{\max}\})}\right)\right\}, \tag{20}$$

when $\max\{\varepsilon_n^q(\mathcal{I}), \varepsilon_n^q(\mathcal{II})\} \le \frac{1}{2}\max\{1 - \pi_{\min}, \pi_{\max}\}$, *under assumptions* (17), (18), *and* (2) *in observational studies, for the null hypothesis* (3).

Corollary 1. *Crossfit-*I$^3_{\hat{\pi}}$ *has asymptotic FDR control for the zero-effect null* (3) *when the estimation of propensity score bounds is consistent in the sense that* $\varepsilon_n^{\pi} \to 0$ *as sample size n goes to infinity*.

Remark 3. The Crossfit-I$^3_{\hat{\pi}}$ would have a larger FDR than the target level if the propensity score estimation is not consistent, and this inflation increases as the true bounds of propensity score get close to 0 and 1. Nonetheless, note that the inflation only depends on the minimum and maximum of the propensity scores (rather than for each individual), and only concerns the *one-sided* error for their estimation. Intuitively, we could have exact FDR control if the estimated minimum propensity score is lower than the true minimum and the estimated maximum larger than the true maximum – if the estimation is conservative to capture the extreme cases. Such desirable FDR control comes with a risk of having lower power because the conservative propensity score estimations would increase the FDR estimation in Crossfit-I$^3_{\hat{\pi}}$, and in turn, more subjects need to be excluded before claiming rejections.

Remark 4. Another variation is proposed in Appendix A.1, as what we call MaY-I$^3_{\hat{\pi}}$, which can have a stronger error control guarantee at the cost of reserving (masking) more information for FDR control and could potentially have lower identification power. Specifically, the variant can have doubly robust asymptotic FDR control: either when the propensity scores are well-estimated as earlier, or when the conditional

outcomes given covariates $m(X_i) \equiv \mathbb{E}(Y_i|X_i)$ are well-estimated by $\widehat{m}(X_i)$. In addition, another advantage of the MaY-I$^3_{\widehat{\pi}}$ is that it controls FDR for nonpositive effect, as we detail in Appendix A.1 and show in numerical experiments in the next section.

5.1 Numerical experiments

We follow the simulation setting in Section 3.2, except different propensity scores specified as a function of covariates. Let the treatment effect be

$$\Delta(X_i) = 15X_i^3(3)\mathbb{1}\{X_i(3) > 1\} - 3X_i(1)/2, \tag{21}$$

which is the treatment effect in (A26) with $S_\Delta = 3$. Consider the case where subjects with positive effects coincides with those having higher propensity scores:

$$\pi_i = \pi(X_i) = (1/2 + \delta)\mathbb{1}\{\Delta(X_i) > 0\} + 1/2\mathbb{1}\{\Delta(X_i) = 0\} + (1/2 - \delta)\mathbb{1}\{\Delta(X_i) < 0\}, \tag{22}$$

where $\delta \in (0, 0.5)$ denotes the deviation of the propensity score bounds to 1/2.

In the case of unknown propensity scores, several approaches are explored: estimating the propensity scores as in Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$; and falsely treating all propensity scores as 1/2 and implementing the original algorithms, referred to as Crossfit-I^3 and MaY-I^3 (detailed in Appendix A.1 for controlling error of nonpositive effects). They are compared with the oracle algorithms where the true propensity scores are plugged into Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$, denoted as Crossfit-I$^3_\pi$ and MaY-I$^3_\pi$. We are interested in the sensitivity of Crossfit-I^3 and MaY-I^3 because we might assume propensity scores to be 1/2 while they differ in practice.

Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$ with estimated propensity scores appear to control FDR at the target level for their corresponding null hypotheses, respectively (Figure 5). They have less power compared with the Crossfit-I$^3_\pi$ and MaY-I$^3_\pi$, which is expected since the latter two methods make use of the true propensity scores.

When all propensity scores are falsely treated as 1/2, we can implement Crossfit-I^3 and MaY-I^3 (Figure 6). In our experiments, the FDR for the zero-effect null seems to be controlled below the target level even when the true propensity scores are extreme (with $\pi_{\min} = 0.1$ and $\pi_{\max} = 0.9$ when $\delta = 0.4$). It coincides with our claim on doubly robust FDR control once noticing that MaY-I^3 is equivalent to MaY-I$^3_{\widehat{\pi}}$ when $\widehat{\pi}_{\min} = \widehat{\pi}_{\max} = 1/2$. In such a case, the propensity scores are poorly estimated $|\widehat{\pi}_{\min} - \pi_{\min}| = |\widehat{\pi}_{\max} - \pi_{\max}| = 0.4$, but FDR can be small when the expected outcome $\mathbb{E}(Y_i|\{X_i\}_{i=1}^n)$ is well-estimated by $\widehat{m}^{-I}(X_i)$. The FDR for the nonpositive-effect null can exceed the target level when the deviation δ is large and the propensity score estimation is poor,

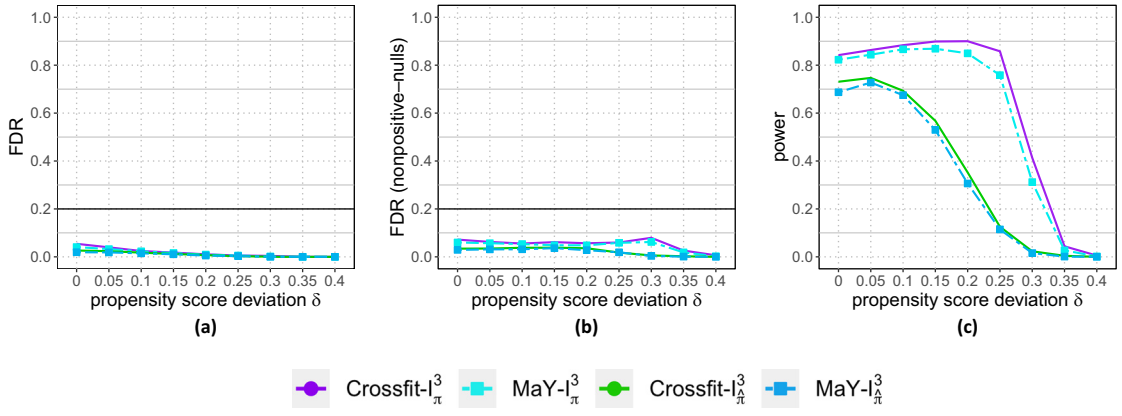

Figure 5: Performance of Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$, which estimate the propensity scores, compared with Crossfit-I$^3_\pi$ and MaY-I$^3_{\widehat{\pi}}$, which use the knowledge of the true propensity scores, when the treatment effect specified as model (21) and the propensity score deviates from 1/2 by δ where δ varies in {0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4}. Both Crossfit-I$^3_{\widehat{\pi}}$ and MaY-I$^3_{\widehat{\pi}}$ appears to control FDR, and have similar power. Their power are lower than Crossfit-I$^3_\pi$ and MaY-I$^3_{\widehat{\pi}}$ because the latter additionally use the true propensity scores. (a) FDR for the zero-effect null (3), (b) FDR for the nonpositive-effect null (A1), and (c) power of identifying subjects with positive effects.

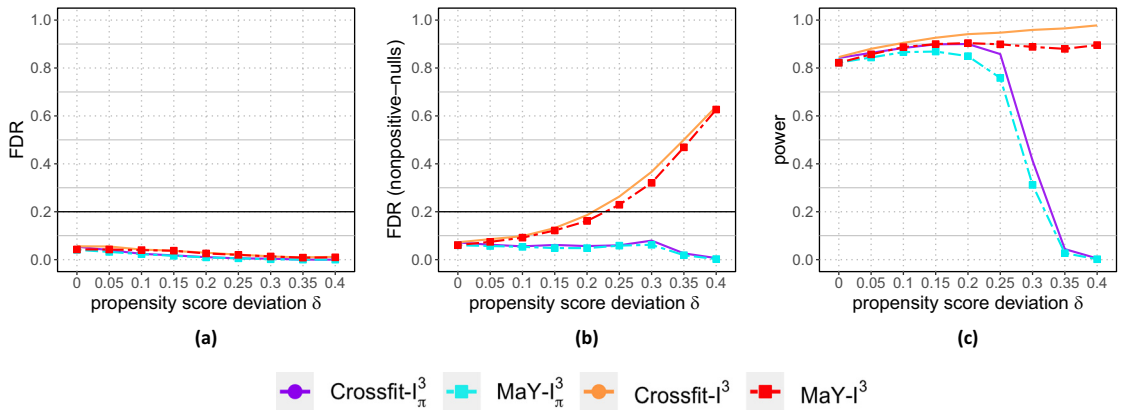

Figure 6: Performance of Crossfit-I^3 and MaY-I^3, which falsely treat all propensity scores as 1/2, compared with Crossfit-I$^3_\pi$ and MaY-I$^3_\pi$, which use the true propensity scores, when the treatment effect specified as model (21) and the propensity score deviates from 1/2 by δ where δ varies in $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. Power of the Crossfit-I^3 and MaY-I^3 increases because they do not suffer from conservative FDR estimator as δ increases. Although FDR for the nonpositive-effect null grows to exceed the target level when δ is larger than 0.2, FDR control for the zero-effect null seems to hold even when the true propensity scores are vastly different from 1/2. (a) FDR for the zero-effect null (3), (b) FDR for the nonpositive-effect null (A1), and (c) power of identifying subjects with positive effects.

corresponding to the case where $\pi_{\min} \le 0.25$ and $\pi_{\max} \ge 0.75$. The power of Crossfit-I^3 and MaY-I^3 does not follow the same trend as Crossfit-I$^3_\pi$ and MaY-I$^3_\pi$ when δ grows large, because their FDR estimator does not suffer from conservativeness introduced by extreme propensity scores.

6 FDR control at a subgroup level

Our proposed interactive methods control FDR on *individual* level, which means upper bounding the proportion of falsely identified subjects. In this section, we show that the idea of interactive testing can be extended to control FDR on *subgroup* level, where we aim at identifying multiple subgroups with positive effects and upper bounding the proportion of falsely identified subgroups. Recall that FDR control at a subgroup level is studied by Karmakar et al. [8] as we review in Section 1.3.

6.1 Problem setup

Let there be G nonoverlapping subgroups $\mathcal{G}_g$ for $g \in [G] \equiv \{1, \ldots, G\}$. The null hypothesis for each subgroup is defined as zero effect for all subjects within:

$$\mathcal{H}_{0g} : H_{0i}^{\text{zero}} \text{ is true for all } i \in \mathcal{G}_g, \tag{23}$$

or equivalently, $\mathcal{H}_{0g} : \mathcal{G}_g \subseteq \mathcal{H}_0$ (recall that $\mathcal{H}_0$ is the set of true null subjects). Let D_g be the decision function receiving the values 1 or 0 for whether $\mathcal{H}_{0g}$ is rejected or not rejected, respectively, and the FDR at a subgroup level is defined as follows:

$$\text{FDR}^{\text{subgroup}} := \mathbb{E}\left[\frac{|\{g \in [G] : \mathcal{G}_g \subseteq \mathcal{H}_0, D_g = 1\}|}{\max\{|\{g \in [G] : D_g = 1\}|, 1\}}\right].$$

Same as the algorithms at an individual level, the algorithms we propose at a subgroup level can be applied to samples that are paired or unpaired. For simple notation, we use $\{Y_i, A_i, X_i\}$ to denote the observed data for subject i when the samples are unpaired, and for pair i when the samples are paired (where $Y_i = \{Y_{i1}, Y_{i2}\}$ and similarly for A_i and X_i).

6.2 An interactive algorithm to identify subgroups

We first follow the same steps of Karmakar et al. [8] to define subgroups and generate the p-value for each subgroup. Specifically, the subgroups $\mathcal{G}_g$ for $g \in [G]$ is defined using the outcomes and covariates $\{Y_j, X_j\}_{j=1}^n$ (by an arbitrary algorithm or strategy, such as grouping subjects with the same covariates). For each subgroup $\mathcal{G}_g$, we can compute a p-value P_g by the classical Wilcoxon test (or using a permutation test, which obtains the null distribution by permuting the treatment assignment $\{A_i\}_{i=1}^n$).

The interactive procedure we propose differs from the study by Karmakar et al. [8] by how we process the p-values of the subgroups. We adopt the work of Lei et al. [19] that proposes an interactive procedure with FDR control for generic multiple testing problems. The key property that allows human interaction while guaranteeing valid FDR control is similar to that in the I^3: the independence between the information used for selection and that used for FDR control. Here, with the p-values of subgroups, the two independent parts are as follows:

$$P_g^1 \coloneqq \min\{P_g, 1 - P_g\},$$

which is revealed to the analyst for selection and

$$P_g^2 \coloneqq 2 \cdot \mathbb{1}\left\{P_g < \frac{1}{2}\right\} - 1,$$

which is masked (hidden) for FDR control. Notice that for a null subgroup with a uniform p-value, (P_g^1, P_g^2) are independent, and we have that

$$\mathbb{P}(P_g^2 = 1 | P_g^1, [Y_i, X_i]_{i \in \mathcal{G}_g}) \leq 1/2, \tag{24}$$

because the p-values obtained by permutating assignments is uniform when conditional on the outcomes and covariates. We remark that the aforementioned property is similar to property (9) and (A5) that lead to valid FDR control at an individual level.

Algorithm 4 An interactive procedure for subgroup identification.

Initial state: Explorer (E) knows the covariates, outcomes $\{Y_i, X_i\}_{i=1}^n$.
Oracle (O) knows the treatment assignments $\{A_i\}_{i=1}^n$.
Target FDR level α is public knowledge.
Initial exchange: Set $t = 1$.

1. E defines subgroups $\{\mathcal{G}_g\}_{g=1}^G$ using $\{Y_i, X_i\}_{i=1}^n$.

2. Both players initialize $\mathcal{R}_0 = [G]$, and E informs O about the subgroup division.

3. O compute the p-value for each subgroup $\{P_g\}_{g=1}^G$, and decompose each p-value as $P_g^1 \coloneqq \min\{P_g, 1 - P_g\}$ and $P_g^2 \coloneqq 2 \cdot \mathbb{1}\{P_g < \frac{1}{2}\} - 1$.

4. O then divides $\mathcal{R}_t$ into $\mathcal{R}_t^- \coloneqq \{g \in \mathcal{R}_t : P_g^2 \leq 0\}$ and $\mathcal{R}_t^+ \coloneqq \{g \in \mathcal{R}_t : P_g^2 > 0\}$.

5. O reveals $\{P_g^1\}_{g=1}^G$, $|\mathcal{R}_t^-|$ and $|\mathcal{R}_t^+|$ to E.

Repeated interaction: 6. E checks if $\widehat{\mathrm{FDR}}^{\text{subgroup}}(\mathcal{R}_t) \equiv \frac{|\mathcal{R}_t^-| + 1}{\max\{|\mathcal{R}_t^+|, 1\}} \leq \alpha$.

7. If yes, E sets $\tau = t$, reports $\mathcal{R}_\tau^+$ and exits;

8. Else, E picks any $g_t^* \in \mathcal{R}_{t-1}$ using everything E currently knows.

(E tries to pick an g_t^* that they think is null; E hopes that $P_g^2 \leq 0$.)

9. O reveals $\{A_i\}_{i \in \mathcal{G}_g}$ to E, who also infers P_g^2.

10. E updates $\mathcal{R}_{t+1} = \mathcal{R}_t \backslash \{g_t^*\}$, and also $|\mathcal{R}_{t+1}^+|$ and $|\mathcal{R}_{t+1}^-|$;

11. Increment t and go back to Step 6.

Similar to the proposed methods at an individual level, the interactive procedure for subgroups progressively excludes subgroups and recursively estimates the FDR. Let the candidate rejection set $\mathcal{R}_t$ be a set of selected subgroups, starting from all subgroups included $\mathcal{R}_0 = [G]$. We interactively shrink $\mathcal{R}_t$ using the available information:

$$\mathcal{F}_{t-1}^{\text{subgroup}} = \sigma\left(\{P_g^1, [Y_i, X_i]_{i\in\mathcal{G}_g}\}_{g\in\mathcal{R}_{t-1}}, \{P_g, [Y_j, A_j, X_j]_{j\in\mathcal{G}_g}\}_{g\notin\mathcal{R}_{t-1}}, \sum_{g\in\mathcal{R}_{t-1}} P_g^2\right),$$

which masks (hides) the partial p-value P_g^2 and the treatment assignment A_i for candidate subgroups in $\mathcal{R}_{t-1}$; and the sum $\sum_{g\in\mathcal{R}_{t-1}} P_g^2$ is mainly provided for FDR estimation. Similar to our previously proposed interactive procedures, the FDR estimator is defined as follows:

$$\widehat{\text{FDR}}^{\text{subgroup}}(\mathcal{R}_t) = \frac{|\mathcal{R}_t^-| + 1}{\max\{|\mathcal{R}_t^+|, 1\}}, \tag{25}$$

with $\mathcal{R}_t^+ = \{g \in \mathcal{R}_t : P_g^2 = 1\}$ and $\mathcal{R}_t^- = \{g \in \mathcal{R}_t : P_g^2 = -1\}$. The algorithm shrinks $\mathcal{R}_t$ until time $\tau := \inf\{t : \widehat{\text{FDR}}^{\text{subgroup}}(\mathcal{R}_t) \le \alpha\}$, and identifies only the subgroups in $\mathcal{R}_\tau^+$, as summarized in Algorithm 4. Details of strategies to select subgroups based on the revealed p-value and covariates can be found in the study by Lei et al. [19]. As a comparison, Karmakar et al. [8] use the same set of p-values $\{P_g\}_{g\in[G]}$, and control FDR by the classical BH procedure.

6.3 Numerical experiments

We compare the performance of our proposed interactive procedure for subgroup identification with the method proposed by Karmakar et al. [8], following an experiment in their paper. Suppose each subject is recorded with two discrete covariates $X_i = \{X_i(1), X_i(2)\}$, where $X_i(1) \in \{1, \ldots, 40\}$ takes 40 levels with equal probability, and $X_i(2)$ is binary with equal probability (e.g., $X_i(1)$ could encode the city subject i lives in, and $X_i(2)$ the gender). The treatment effect $\Delta(X_i)$ is a constant δ if $X_i(1)$ is even, and we vary δ in six levels. We conduct the aforementioned experiment in two cases: unpaired samples (n = 2,000) with independent covariates and paired samples (n = 1,000) whose covariate values are the same for subjects within each pair.

Recall that the subgroups can be defined by covariates and outcomes. Here, since the covariates are discrete, we define subgroups by different values of $(X_i(1), X_i(2))$, resulting in 80 subgroups. The interactive procedure tends to have higher power than the BH procedure (Figure 7(a) and (b)) because it focuses on the subgroups that are more likely to be the non-nulls using the excluding process, and utilizes the covariates together with the p-values to guide the algorithm. Meanwhile, the BH procedure does not account for covariates once the p-values are calculated. Nonetheless, the interactive procedure can have lower power when the total number of subgroups that are truly non-null is small. We simulate the case where a subject has a positive effect δ if $X_i(1)$ a multiplier of 4 (i.e., $X_i(1)/2$ is even), so that there are 20 non-null subgroups in total (previously 40 non-nulls). The power of the interactive procedure is lower than the BH procedure (Figure 7(c)) because the FDR estimator in (25) can be conservative when $|\mathcal{R}^+|$ is small due to a small number of true non-nulls (e.g., with FDR control at α = 0.2, we need to shrink $\mathcal{R}_t$ until $|\mathcal{R}^-| < 3$ when $|\mathcal{R}^+|$ is around 20).

A side note is that we define the subgroups by distinct values of the covariates, whereas Karmakar et al. [8] suggest forming subgroups by regressing the outcomes on covariates using a tree algorithm. In their experiments and several numerical experiments we tried, we find that the number of subgroups defined by the tree algorithm is usually less than 10. However, we think the FDR control is less meaningful when the total number of subgroups is small. To justify our comment, note that an algorithm with valid FDR control at level α can make zero rejection with probability $1 - \alpha$ and reject all subgroups with probability α, which can happen when the total number of subgroups is small. In contrast, with a large number of subgroups, a reasonable algorithm is unlikely to jump between the extremes of making zero rejection and rejecting all n subgroups; and thus, controlling FDR indeed informs that the proportion of false identifications is low for the evaluated algorithm.

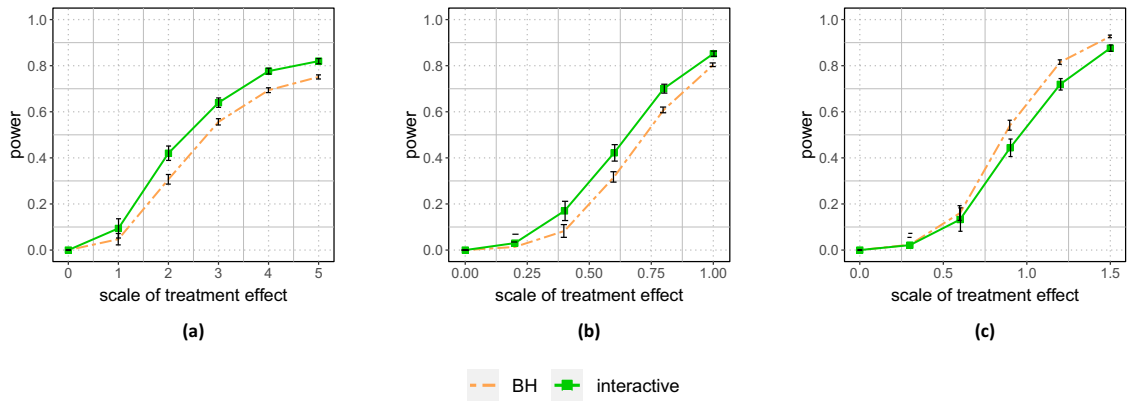

Figure 7: Performance of methods to identify subgroups with positive effects: the BH procedure and the interactive procedure (for 80 subgroups defined by the distinct values of covariates). We vary the scale of treatment effect under unpaired or paired samples. In both cases, the interactive procedure can have higher power than the BH procedure. When the number of non-null subgroups is too small (less than 20), the BH procedure can have higher power. The error bar marks two standard deviations from the center. (a) Unpaired samples, (b) paired samples, and (c) only a few subgroups are non-nulls.

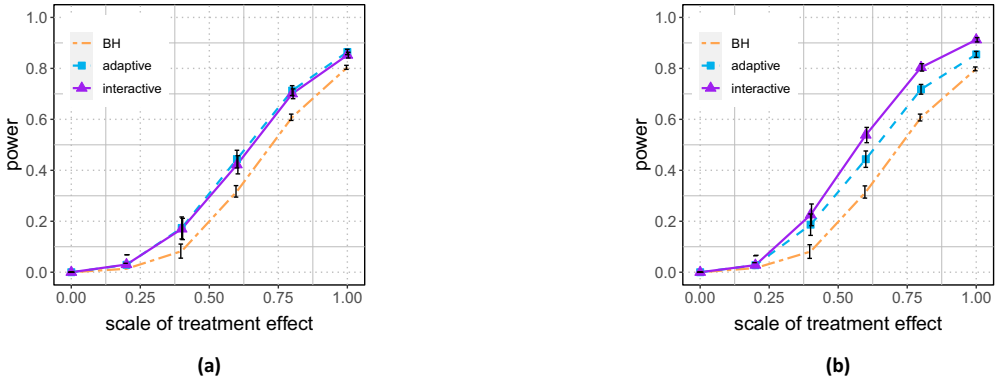

Figure 8: Power of two methods for subgroup identification: the BH procedure proposed by Karmakar et al. [8], the adaptive procedure, and the interactive procedure under different types of treatment effects (we define 80 subgroups by discrete values of the covariates). Our proposed interactive procedure tends to have higher power than the BH procedure because (1) it excludes possible nulls (shown by higher power of the adaptive procedure than the BH procedure in both plots); and (2) it additionally uses the covariates (shown when the treatment effect can be well learned as a function of covariates in the right plot). (a) Effect as discrete function of covariates and (b) effect as a simpler function of covariates.

6.4 Explanation of the higher power achieved by the interactive procedure

Although the interactive procedure and the BH procedure define the same set of subgroups and corresponding p-values, the interactive procedure has two properties that potentially improve the power from the BH procedure: (a) it excludes possible null subgroups so that it can be less sensitive to a large number of nulls, whereas the BH procedure considers all the subgroups at once; (b) the interactive procedure additionally uses the covariates. We can separately evaluate the effect of the aforementioned two properties by implementing two versions of Algorithm 4, which differ in the strategy to select subgroups in step ɑ. Specifically, the adaptive procedure selects the subgroup whose revealed (partial) p-value P_g^1 is the smallest (not using the covariates); and the interactive procedure selects the subgroup by an estimated probability of the P_g^2 to be positive (using the revealed P_g^1, the covariates, and the outcomes).

To see if both properties of Algorithm 4 contribute to the improvement of power from the BH procedure, we tested the methods under two simulation settings. Recall that the previous experiment defines a positive

treatment effect when the discrete covariate $X_i(1) \in \{1, \ldots, 40\}$ is even. Here, we add another case where the treatment effect is positive when $X_i(1) \leq 20$, so that the density of subgroups with positive effects is the same as previous, but the treatment effect is a simpler function of the covariates. Hence, in the latter case, we would expect the interactive procedure to learn this function of covariates rather accurately, and have higher power than the adaptive procedure which does not use the covariates; as confirmed in Figure 8(b). In the former simulation setting where the treatment effect is not a smooth function of the covariates and hard to be learned, the adaptive procedure and interactive procedure have similar power (Figure 8(a)). Still, they have higher power than the BH procedure because they exclude possible null subgroups.

7 Additional numerical experiments for variations in treatment effect

We have seen the numerical results of the proposed methods in previous sections where the treatment effect is defined in (A26) with sparse and strong positive effect, and dense and weak negative effect. This section presents three more examples of the treatment effect.

7.1 Linear effect

Let the treatment effect be

$$\Delta(X_i) = S_\Delta \cdot [2X_i(1)X_i(2) + 2X_i(3)], \tag{26}$$

where $S_\Delta > 0$. In this case, all subjects have treatment effects, and the scale correlates with the covariates (with interaction terms) linearly. Thus, the linear-BH procedure has valid error control as shown in Figure 9(a) (unlike other cases with nonlinear treatment effect).

7.2 Sparse and strong effect that is positive

Let the treatment effect be

$$\Delta(X_i) = S_\Delta \cdot [5X_i^3(3)\mathbb{1}\{X_i(3) > 1\}], \tag{27}$$

where $S_\Delta > 0$. Here, the subjects with $X_i(3) > 1$ have positive treatment effects. Although linear-BH procedure seems to have high power, its FDR is largely inflated since the assumption of linear correlation does not hold (see Figure 9(b)). In contrast, our proposed methods and Selective SeqStep+ have valid FDR control, and our proposed methods have higher power.

7.3 Sparse and strong effect in both directions

Let the treatment effect be

$$\Delta(X_i) = S_\Delta \cdot [5X_i^3(3)\mathbb{1}\{|X_i(3)| > 1\}], \tag{28}$$

where $S_\Delta > 0$. Here, the subjects with $X_i(3) > 1$ have positive treatment effects and those with $X_i(3) < -1$ have negative treatment effects; the scale and proportion of effects in both directions are the same. The power of Selective SeqStep+ is trivial because $|E_i|$ used in ordering cannot inform the direction of treatment effect. In contrast, the power of Crossfit-I^3 and MaY-I^3 are slightly lower than the previous setting since there is additionally negative effect in this example (Figure 9(c)).

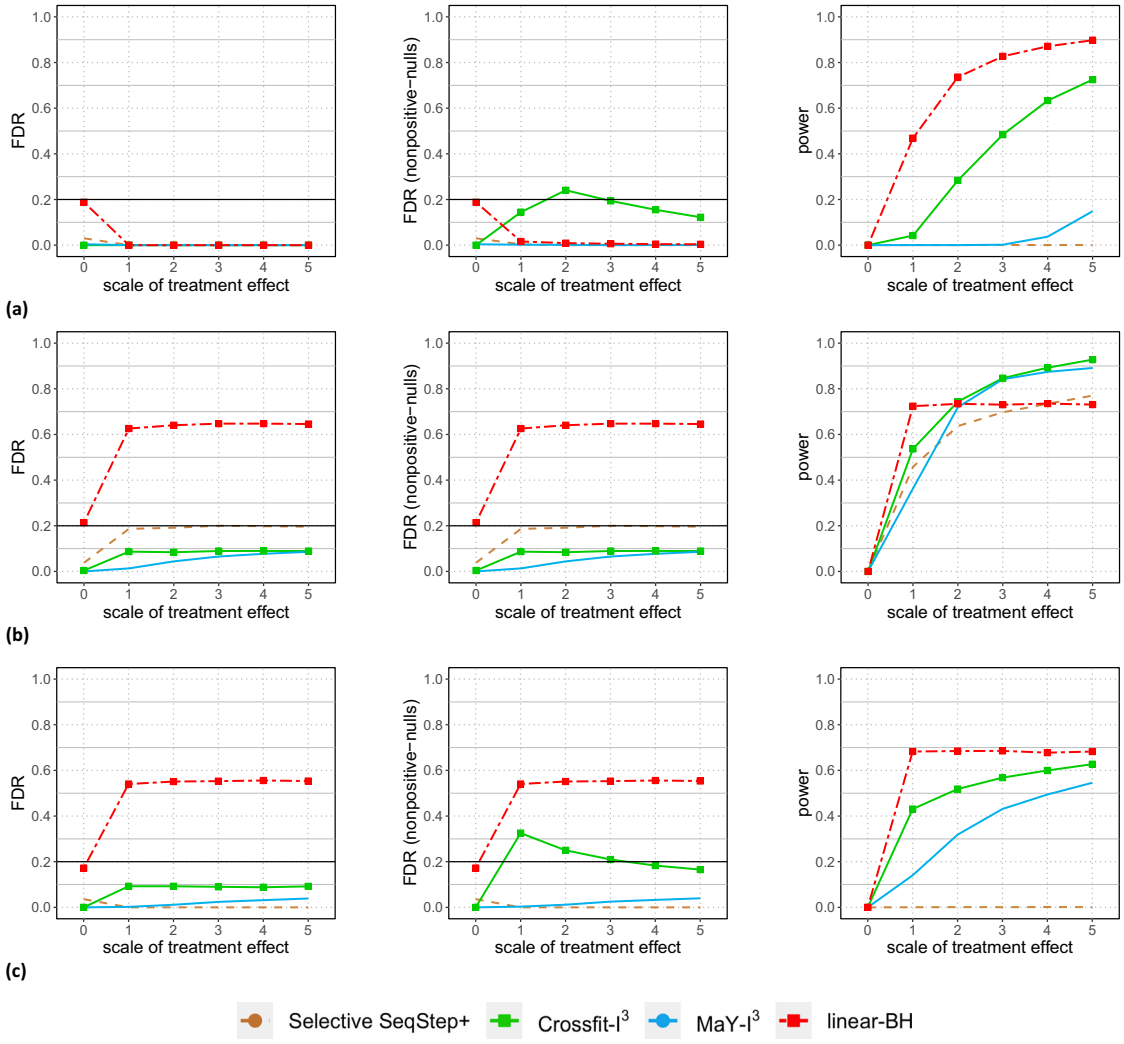

Figure 9: FDR for the zero-effect null (3) (the first column), and FDR for the nonpositive effect null (A1) (the second column), and power (the third column) of four methods: linear-BH procedure, Selective SeqStep+, Crossfit-I^3, MaY-I^3, under three types of treatment effect when varying the scale of treatment effect S_Δ in $\{0, 1, 2, 3, 4, 5\}$. When the linear assumption holds as in the first row, the linear-BH procedure has valid FDR control and high power, but its FDR is large when the treatment is a nonlinear function of the covariates as shown in the last two rows. In contrast, the Selective SeqStep+, Crossfit-I^3 and MaY-I^3 have valid FDR control for their target null hypotheses, respectively. Our proposed Crossfit-I^3 and MaY-I^3 have higher power than Selective SeqStep+, especially when positive effects have comparable scale as negative effects as shown in the first and the third rows, because the proposed methods can additionally use the revealed treatment assignments to inform the direction of treatment effects. (a) Dense two-sided effect (linear) as in model (26), (b) sparse and strong effect that is positive (nonlinear) in model (27), and (c) sparse and strong effect in both directions (nonlinear) in model (28).

8 A prototypical application to ACIC challenge dataset

We implement our proposed methods on datasets generated by Atlantic Causal Inference Conference (ACIC), which intend to evaluate methods for average treatment effect (ATE) estimation and uses real data covariates and modified outcomes to simulate cases with heterogeneous treatment effect, heterogeneous propensity scores, etc. We take an example dataset with 500 subjects, each of which is recorded with 22 continuous covariates. The proportion of treated subjects is 0.7, indicating that the propensity scores might not be 1/2 as in a standard randomized experiment. The actual ATE is 0.1, rather small compared to the outcomes range [14, 76], but the treatment effect could be positive and large for a subgroup of subjects and our proposed algorithms can be used to identify them.

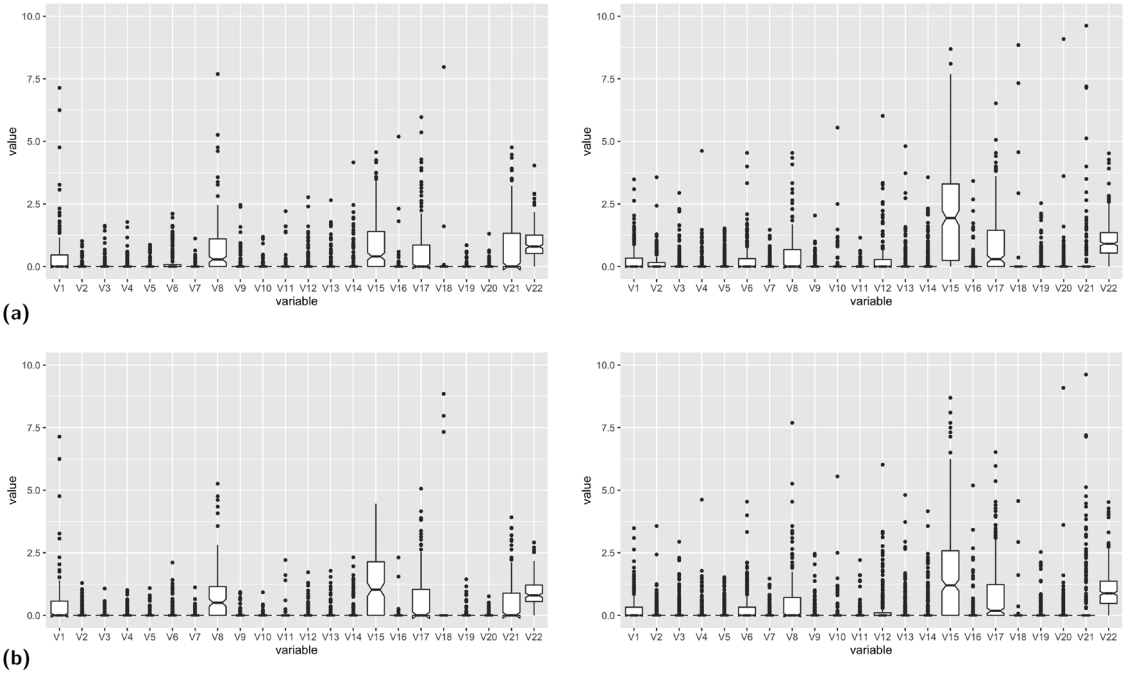

Figure 10: Characteristics of identified subjects: they tend to have larger value for variable 8, 21 and smaller value for variable 6, 15, 17, compared with not identified subjects. (a) Boxplot of covariates for subjects identified as nonzero effect (left) versus those not being identified (right) and (b) boxplot of covariates for subjects identified as positive effect (left) versus those not being identified (right).

Four of our proposed methods are implemented with FDR control at level $\alpha = 0.2$: the Crossfit-I^3 and MaY-I^3 which assume the propensity scores to be 1/2 for all subjects, and Crossfit-I$^3_{\hat{\pi}}$ and MaY-I$^3_{\hat{\pi}}$ which estimate the propensity scores. The numbers of identifications by Crossfit-I^3, MaY-I^3, Crossfit-I$^3_{\hat{\pi}}$ and MaY-I$^3_{\hat{\pi}}$ are 446, 429, 238, 162. Among them, 234 subjects are commonly identified by Crossfit-I^3 and Crossfit-I$^3_{\hat{\pi}}$, which control the expected proportion of falsely identifying subjects with zero effect (approximately if the propensity scores are not 1/2); and 158 subjects are commonly identified by MaY-I^3 and MaY-I$^3_{\hat{\pi}}$, which control the expected proportion of falsely identifying subjects with nonpositive effect (approximately if the propensity scores are not 1/2). Compared with the rest subjects, the ones identified as having positive effect tend to have larger values for covariate 8, 21 and smaller values for covariate 6, 15, 17 (Figure 10).

9 Summary

We discuss the problem of identifying subjects with positive effects. Most existing methods identify *subgroups* with positive treatment effects, and they cannot upper bound the proportion of falsely identified *subjects* within an identified subgroup. In contrast, we propose Crossfit-I^3 with finite-sample FDR control (i.e., the expected proportion of subjects with zero effect is no larger than α among the identified subjects). One advantage of the Crossfit-I^3 is allowing human interaction – an analyst (or an algorithm) can incorporate various types of prior knowledge and covariates using any working model; she can also adjust the model at any step, potentially improving the identification power. Despite this flexibility, the Crossfit-I^3 achieves valid FDR control. Notably, because Crossfit-I^3 incorporates covariates, it can identify subjects with positive effects, including those not treated.

Our proposed interactive procedure can extended to various settings:

- FDR control of nonpositive effects in randomized experiments (Appendix A.1);
- FDR control of zero/nonpositive effects in observational studies (Section 5 and Appendix A.1.4);
- paired samples (Appendix A.2);
- FDR control at a subgroup level (Section 6).

The error control for our interactive procedures is based on the independence properties between the data used for FDR control and the revealed data for interaction, such as property (9) for the zero-effect null and (A5) for the nonpositive-effect null.

The idea of interactive testing can be generalized to many other problems as long as we can construct two parts of data that are (conditionally) independent. As an example, for alternative definitions of the zero-effect null, such as those involving conditional expectations or conditional quantiles, one can explore specific functions of $\{Y_i, X_i\}$ that are independent of treatment assignment under certain model assumptions, and then plug them into the Crossfit-I^3 framework. Importantly, our interactive procedures using the idea of "masking and unmasking" should be contrasted with data-splitting approaches. We remark that no test, interactive or otherwise, can be run twice from scratch (with a tweak made the second time to boost power) after the entire data has been examined; this amounts to *p*-hacking. We view our interactive tests as one step towards enabling experts (scientists and statisticians) to work together with statistical models and machine learning algorithms to discover scientific insights with rigorous guarantees.

Acknowledgements: We thank Edward Kennedy and Bikram Karmakar for their comments on an early draft of the paper. AR acknowledges support from NSF CAREER 1916320.

Funding information: AR acknowledges support from NSF CAREER 1916320.

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results and approved the final version of the manuscript. All authors developed methodology and experiment design. BD developed the model code and performed the simulations. All authors prepared the manuscript.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets generated during and/or analysed during the current study are available in https://github.com/duanby/I-cube.

References

[1] Lipkovich I, Dmitrienko A, B D'Agostino R. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med. 2017;36(1):136–96.

[2] Powers S, Qian J, Jung K, Schuler A, Shah NH, Hastie T, et al. Some methods for heterogeneous treatment effect estimation in high dimensions. Stat Med. 2018;37(11):1767–87.

[3] Loh WY, Cao L, Zhou P. Subgroup identification for precision medicine: A comparative review of 13 methods. Wiley Interdiscipl Rev Data Mining Knowledge Discovery. 2019;9(5):e1326.

[4] Howard SR, Pimentel SD. The uniform general signed rank test and its design sensitivity. Biometrika. 2021;108:381–96.

[5] Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Stat Med. 2011;30(24):2867–80.

[6] Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Amer Stat Assoc. 2012;107(499):1106–18.

[7] Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat. 2013;7(1):443–70.

[8] Karmakar B, Heller R, Small DS. False discovery rate control for effect modification in observational studies. Electron J Stat. 2018;12(2):3232–53.

[9] Gu J, Shen S. Oracle and adaptive false discovery rate controlling methods for one-sided testing: theory and application in treatment effect evaluation. Econometrics J. 2018;21(1):11–35.

[10] Xie Y, Chen N, Shi X. False discovery rate controlled heterogeneous treatment effect detection for online controlled experiments. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2018. p. 876–85.

[11] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological). 1995;57(1):289–300.

[12] Lei L, Fithian W. AdaPT: an interactive procedure for multiple testing with side information. J R Stat Soc Ser B (Statistical Methodology). 2018;80(4):649–79.

[13] Cai T, Tian L, Wong PH, Wei L. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics. 2011;12(2):270–82.

[14] Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Nat Acad Sci. 2016;113(27):7353–60.

[15] Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med. 2011;30(21):2601–21.

[16] Lipkovich I, Dmitrienko A. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. J Biopharm Stat. 2014;24(1):130–53.

[17] Sivaganesan S, Laud PW, Müller P. A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. Stat Med. 2011;30(4):312–23.

[18] Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification. J Biopharm Stat. 2014;24(1):110–29.

[19] Lei L, Ramdas A, Fithian W. A general interactive framework for false discovery rate control under structural constraints. Biometrika. 2021;108(2):253–67.

[20] Duan B, Ramdas A, Balakrishnan S, Wasserman L. Interactive martingale tests for the global null. Electr J Stat. 2020;14(2):4489–551.

[21] Duan B, Ramdas A, Wasserman L. Familywise error rate control by interactive unmasking. In: International Conference on Machine Learning. PMLR; 2020. p. 2720–9.

[22] Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. Ann Stat. 2015;43(5):2055–85.

[23] Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika. 2021;108(2):299–319.

[24] Kennedy EH. Towards optimal doubly robust estimation of heterogeneous causal effects. Electr J Stat. 2023;17(2):3008–49.

[25] Robinson PM. Root-N-consistent semiparametric regression. Econometr J Econometr Soc.1988;56(4):931–54.

[26] Rosenbaum PR. Covariance adjustment in randomized experiments and observational studies. Stat Sci. 2002;17(3):286–327.

[27] Rosenblum M, Van Der Laan MJ. Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. Biometrics. 2009;65(3):937–45.

[28] Lin W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedmanas critique. Ann Appl Stat. 2013;7(1):295–318.

[29] Fogarty CB. Regression-assisted inference for the average treatment effect in paired experiments. Biometrika. 2018;105:994–1000.

[30] Guo K, Basse G. The generalized Oaxaca-Blinder estimator. J Amer Stat Assoc. 2023;118(541):524–36.

[31] Arias-Castro E, Chen S. Distribution-free multiple testing. Electr J Stat. 2017;11(1):1983–2001.

[32] Rabinovich M, Ramdas A, Jordan MI, Wainwright MJ. Optimal rates and trade-offs in multiple testing. Stat Sinica. 2020;30:741–62.

[33] Li A, Barber RF. Accumulation tests for FDR control in ordered hypothesis testing. J Amer Stat Assoc. 2017;112(518):837–49.

[34] Fan J, Hall P, Yao Q. To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? J Amer Stat Assoc. 2007;102(480):1282–8.

Appendix

A Extensions of the I^3

A.1 FDR control of nonpositive effects

The Crossfit-I^3 controls the false identifications of subjects with zero treatment effect, as defined in the null hypothesis (3) and its corresponding footnote. In this section, we develop a modification to additionally control the error of falsely identifying subjects with nonpositive treatment effects, by defining a different null hypothesis.

A.1.1 Problem setup

We define the null hypothesis for subject i as nonpositive effect:

$$H_{0i}^{\text{nonpositive}} : (Y_i^T|X_i) \preccurlyeq (Y_i^C|X_i), \tag{A1}$$

or equivalently, $H_{0i}^{\text{nonpositive}} : (Y_i|A_i = 1, X_i) \preccurlyeq (Y_i|A_i = 0, X_i)$. As mentioned earlier, our algorithm applies to two alternative definitions of the null hypothesis. In the context of treating the potential outcomes and covariates as fixed, the null hypothesis is

$$H_{0i}^{\text{nonpositive}} : Y_i^T \le Y_i^C, \tag{A2}$$

and in the hybrid version where the potential outcomes are random with joint distribution $(Y_i^T, Y_i^C)|X_i \sim P_i$, the null posits

$$H_{0i}^{\text{nonpositive}} : Y_i^T \le Y_i^C \text{ almost surely-}P_i. \tag{A3}$$

Note that the nonpositive-effect null is less strict than the zero-effect null. Thus, an algorithm with FDR control for $H_{0i}^{\text{nonpositive}}$ must have valid FDR control for H_{0i}^{zero}, but the reverse needs not be true. Indeed, we observe in numerical experiments (Figure A2(b)) that the Crossfit-I^3 does not control FDR for the nonpositive-effect null. This section presents a variant of Crossfit-I^3 that controls false identifications of nonpositive effects, possibly more practical when interpreting the identified subjects. For example, when controlling FDR for the nonpositive-effect null at level $\alpha = 0.2$, we are able to claim that the expected proportion of identified subjects with positive effects is no less than 80%.

A.1.2 An interactive procedure in randomized experiments

Recall that the FDR control of the Crossfit-I^3 is based on property (9) that when the null hypothesis is true for subject i, we have $\mathbb{P}(\widehat{\Delta}_i|\{Y_j, X_j, E_j\}_{j=1}^n) \le 1/2$, but this statement no longer holds when the null hypothesis is defined as $H_{0i}^{\text{nonpositive}}$ in (A1). Fortunately, this issue can be fixed by making the condition in (9) coarser and removing the outcomes:

$$\mathbb{P}(\widehat{\Delta}_i|\{X_j\}_{j=1}^n) \le 1/2,$$

which is reflected in the interactive procedure as reducing the available information for selecting subject i_t^* (at step 8 of Algorithm 1) – we additionally mask (hide) the outcome Y_i of the candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$ when implementing the I^3 on set $\mathcal{I}$. We call the resulting interactive algorithm MaY-I^3, as it masks the outcomes.

Specifically, the MaY-I^3 modifies Crossfit-I^3 where we define the available information to select subjects when implementing Algorithm 1 on set $\mathcal{I}$ as follows:

$$\mathcal{F}_{t-1}^{-Y}(\mathcal{I}) = \sigma\left(\{X_i\}_{i\in\mathcal{R}_{t-1}(\mathcal{I})}, \{Y_j, A_j, X_j\}_{j\notin\mathcal{R}_{t-1}(\mathcal{I})}, \sum_{i\in\mathcal{R}_{t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i > 0\}\right). \tag{A4}$$

To calculate $\widehat{\Delta}_i$ at $t = 0$ when Y_i for all $i \in \mathcal{I}$ are masked, let $\widehat{m}^{-\mathcal{I}}(X_i)$ be an estimator of $\mathbb{E}(Y_i|X_i)$ that is learned using data from noncandidate subjects $\{Y_j, X_j\}_{j\notin\mathcal{I}}$, and let the residuals be $E_i^{-\mathcal{I}} := Y_i - \widehat{m}^{-\mathcal{I}}(X_i)$. Define $\Delta_i^{-\mathcal{I}} := 4(A_i - 1/2)\cdot E_i^{-\mathcal{I}}$, and similar to property (9) for the zero-effect null, we have

$$\mathbb{P}(\widehat{\Delta}_i^{-\mathcal{I}} > 0 | \{X_j\}_{j\in\mathcal{I}} \cup \{Y_j, X_j, E_j^{-\mathcal{I}}\}_{j\notin\mathcal{I}}) \le 1/2, \tag{A5}$$

under $H_{0i}^{\text{nonpositive}}$, leading to valid FDR control for nonpositive effects. Overall, the MaY-I^3 follows Algorithm 2, except the estimated treatment effect $\widehat{\Delta}_i$ replaced by $\widehat{\Delta}_i^{-\mathcal{I}}$, and the available information for selection $\mathcal{F}_{t-1}(\mathcal{I})$ replaced by $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$. See Appendix B.3 for the proof of FDR control.

Theorem A1. *Under assumptions* (1) *and* (2) *of randomized experiments, the MaY-*I^3 *has a valid FDR control at level α for the nonpositive-effect null hypothesis under any of definitions* (A1), (A2), *or* (A3). *For the last definition, FDR control also holds conditional on the covariates and potential outcomes.*

Similar to Algorithm 3 for the Crossfit-I^3, we can design an automated algorithm for the MaY-I^3 to select a subject in step 8 of Algorithm 1, but the available information $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$ no longer includes the outcomes of candidate subjects. One naive strategy is to follow Algorithm 3 in the main article, which is designed for the Crossfit-I^3, with the outcomes removed from the predictors; however, it appears to result in less accurate prediction of the effect signs, and in turn rather low power (numerial results are in the next paragraph). Here, we take a different approach by predicting the treatment effect instead of their signs, because the treatment effect might be better predicted as a function of the covariates (without outcomes) than a binary sign, especially when the treatment effect is indeed a smooth and simple function of the covariates. Specifically, we first estimate the treatment effect for the noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$ using a well-studied doubly robust estimator ([24] and the references therein):

$$\Delta_j^{\text{DR}} = 4(A_j - 1/2)\cdot(Y_j - \widehat{\mu}_A(X_j)) + \widehat{\mu}_1(X_j) - \widehat{\mu}_0(X_j), \tag{A6}$$

where $(\widehat{\mu}_0, \widehat{\mu}_1)$ are random forests trained to predict the outcomes for the control and treated group, respectively. By using the provided covariates X_i, we can predict Δ_i^{DR} for the candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$. The subject with the smallest prediction of Δ_i^{DR} is then excluded. This automated strategy is described in Algorithm 5.

Algorithm 5. An automated heuristic to select i_t^* in the MaY-I^3.

Input: Current rejection set $\mathcal{R}_{t-1}(\mathcal{I})$, and available information for selection $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$;
Procedure:

1. Estimate the treatment effect for noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$ as Δ_j^{DR} in (A6);
2. Train a random forest where the label is the estimated effect Δ_j^{DR} and the predictors are the covariates X_j, using noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;
3. Predict Δ_i^{DR} for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$ via the above random forest, denoted as $\widehat{\Delta}_i^{\text{DR}}$;
4. Select i_t^* as $\text{argmin}\{\widehat{\Delta}_i^{\text{DR}} : i \in \mathcal{R}_{t-1}(\mathcal{I})\}$.

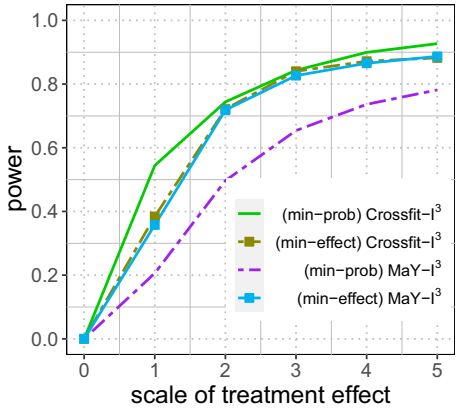

Figure A1: Power of the Crossfit-I^3 and MaY-I^3 with two strategies to select subjects: the min-prob strategy and the min-effect strategy, under the treatment effect defined in (A26) of the main paper with the scale S_Δ varies in $\{0, 1, 2, 3, 4, 5\}$. The Crossfit-I^3 tends to have higher power when using the min-prob strategy, and the MaY-I^3 tends to have higher power when using the min-effect strategy.

To summarize, we have presented two types of strategy for selecting subjects: the Crossfit-I^3 chooses the one with the smallest predicted probability of a positive $\widehat{\Delta}_i$ (Algorithm 3 in the main paper), which we denote here as the *min-prob strategy*; and the MaY-I^3 chooses the one with the smallest prediction of estimated effect Δ_j^{DR} (Algorithm 5), which we denote here as the *min-effect strategy*. Note that the proposed interactive algorithm can use arbitrary strategy as long as the available information for selection is restricted. That is, the Crossfit-I^3 can use the same min-effect strategy, and the MaY-I^3 can use the min-prob strategy (after removing the outcomes from the predictors, which we elaborate in the next paragraph). However, we observe in numerical experiments that both interactive procedures have higher power when using their original strategies, respectively (Figure A1).

A.1.3 Numerical experiments

Before details of the experiment results, we first describe the min-prob strategy for the MaY-I^3, where the available information $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$ does not include the outcomes for candidate subjects. Similar to the min-prob strategy in Algorithm 3 of the main paper, we hope to use the outcome Y_i and residual $E_i = Y_i - \widehat{m}(X_i)$ as predictors, and predict the sign of treatment effect for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$, but Y_i and E_i for the

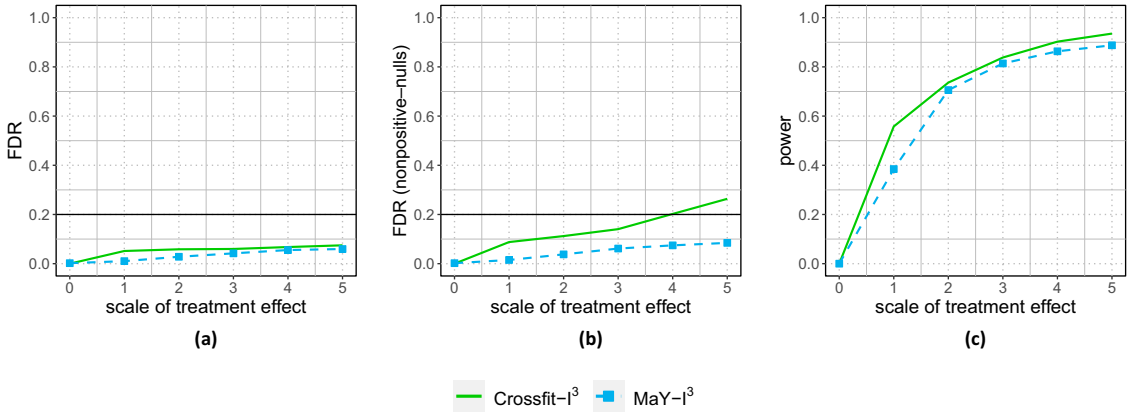

Figure A2: Performance of two interactive methods, Crossfit-I^3 and MaY-I^3, with the treatment effect specified as model (A26) and the scale S_Δ varying in $\{0, 1, 2, 3, 4, 5\}$. The MaY-I^3 controls FDR for a more relaxed null (nonpositive effects) than the Crossfit-I^3, while the Crossfit-I^3 has slightly higher power than the MaY-I^3. (a) FDR for the zero-effect null (3), (b) FDR for the nonpositive-effect null (A1), and (c) power of identifying subjects with positive effects.

candidate subjects are not available in $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$. Thus, we propose algorithm 6, where we first estimate Y_i and E_i using the covariates (see steps 1 and 2); and steps 3–5 are similar to Algorithm 3, which obtain the probability of having a positive treatment effect.

Algorithm 6. The min-prob strategy to select i_t^* in the MaY-I^3.

Input: Current rejection set $\mathcal{R}_{t-1}(\mathcal{I})$, and available information for selection $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$;
Procedure:

1. Predict the outcome Y_k of each subject $k \in [n]$ by covariates, denoted as $\widehat{Y}^{-\mathcal{I}}(X_k)$, where $\widehat{Y}^{-\mathcal{I}}$ is learned using noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;

2. Predict the residual $E_k = Y_k - \widehat{m}(X_K)$ of each subject $k \in [n]$ by covariates, denoted as $\widehat{E}^{-\mathcal{I}}(X_k)$, where $\widehat{E}^{-\mathcal{I}}$ is learned using noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;

3. Train a random forest classifier where the label is $\text{sign}(\widehat{\Delta}_j)$ and the predictors are $(\widehat{Y}^{-\mathcal{I}}(X_j), X_j, \widehat{E}^{-\mathcal{I}}(X_j))$, using noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$;

4. Predict the probability of $\widehat{\Delta}_i$ being positive as $\widehat{p}(i,t)$ for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$;

5. Select $i_t^* = \text{argmin}\{\widehat{p}(i,t) : i \in \mathcal{R}_{t-1}(\mathcal{I})\}$.

The Crossfit-I^3 has higher power when using the min-prob strategy than the min-effect strategy because the former additionally uses the outcome as a predictor. For the MaY-I^3, the min-effect strategy leads to higher power because the estimated treatment effect Δ_j^{DR} in (A6) can provide reliable evidence of which subjects have a positive effect. If using the min-prob strategy, it could be harder to learn an accurate prediction by Algorithm 6 where two of the predictors $\widehat{Y}^{-\mathcal{I}}(X_j)$ and $\widehat{E}^{-\mathcal{I}}(X_j)$ are obtained by estimation, increasing the complexity in modeling. Therefore, we present the Crossfit-I^3 and MaY-I^3 with the min-prob and min-effect strategies, respectively, as preferred in numerical experiments. Nonetheless, we remark that our proposed interactive frameworks for the Crossfit-I^3 and MaY-I^3 allow arbitrary strategies to select subjects, and an analyst can design her own strategy based on her domain knowledge.

We compare the Crossfit-I^3 and MaY-I^3 using the same experiment as Section 3.2. In terms of the error control, both the Crossfit-I^3 and MaY-I^3 control FDR for the zero-effect null at the target level (Figure A2(a)). When the null is defined as having a nonpositive effect, the Crossfit-I^3 can violate the error control (Figure A2(b)), whereas the MaY-I^3 preserves valid FDR control. In terms of the power, the Crossfit-I^3 has slightly higher power since the analyst can select subjects using information defined by $\mathcal{F}_{t-1}(\mathcal{I})$ in (10), which is richer compared to $\mathcal{F}_{t-1}^{-Y}(\mathcal{I})$ in (A4) for the MaY-I^3.

To summarize, the error control of the MaY-I^3 is more strict than the Crossfit-I^3, controlling false identifications of both zero effects and negative effects, while its power is slightly lower. We recommend the Crossfit-I^3 if one only concerns the error of falsely identifying subjects with zero effect. Alternatively, we recommend the MaY-I^3 when it is desired to control the error of falsely identifying subjects with nonpositive effects.

A.1.4 Extensions to observational studies

The extension from randomized experiments to observational studies for the nonpositive-effect nulls is similar to that for the zero-efffect nulls from Crossfit-I^3 to Crossfit-I$^3_{\widehat{\pi}}$ – we estimate the propensity scores π_i using the revealed data (detailed steps described in Section 5). We call the resulting algorithm MaY-I$^3_{\widehat{\pi}}$.

FDR control for MaY-I$^3_{\widehat{\pi}}$ holds even when the bounds of true propensity scores reach 0 or 1:
(iv) the propensity scores are bounded between 0 and 1:

$$0 \le \pi_{\min} \le \pi_i \le \pi_{\max} \le 1 \text{ for all } i \in [n], \tag{A7}$$

which is less stringent than assumption (18) for Crossfit-$\text{I}^3_{\widehat{\pi}}$. The error control of the MaY-$\text{I}^3_{\widehat{\pi}}$ is doubly robust: when the propensity score estimation is poor, FDR would still be close to target level when the expected outcomes are well-estimated. To characterize the error of expected outcome estimation, we define a "centered" CDF Φ as follows:

$$\Phi_i(\varepsilon) \coloneqq \mathbb{P}(Y_i - \mathbb{E}(Y_i|X_i) \le \varepsilon | \{X_i\}_{i=1}^n)$$

with "upper" and "lower" bounds defined as $\Phi_{\max}(\varepsilon) \coloneqq \max_{i\in[n]}\Phi_i(\varepsilon)$ and $\Phi_{\min}(\varepsilon) \coloneqq \min_{i\in[n]}\Phi_i(\varepsilon)$. If the estimation error ε is small and the outcome distribution is symmetric and continuous, the centered CDF is close to 1/2, leading to less FDR inflation as we describe later. We define several estimation errors when performing the I^3 on set $\mathcal{I}$ as follows.

- Let the error of propensity score estimation be $\varepsilon_n^{\pi}(\mathcal{I}) \coloneqq \max_{i\in\mathcal{I}}|\pi_i - \widehat{\pi}_i(\mathcal{I})|$, where $\widehat{\pi}_i(\mathcal{I})$ is the estimated propensity score using data information in $\mathcal{F}_0^{-Y}(\mathcal{I})$.
- Let the error of expected outcome estimation be

$$\varepsilon_n^{Y}(\mathcal{I}) \coloneqq \max_{i\in\mathcal{I}}\{|\mathbb{E}(Y_i|X_i) - \widehat{m}^{-\mathcal{I}}(X_i)|\},$$

  where $\widehat{m}^{-\mathcal{I}}(X_i)$ is the estimated expected outcome learned using data information in $\mathcal{F}_0^{-Y}(\mathcal{I})$, which includes the complete data of noncandidate subjects $j \in [n]\backslash\mathcal{I}$. Recall that the sign of $\Delta_{\widehat{m}}(X_i)$ in MaY-I^3 depends on the sign of $Y_i - \widehat{m}^{-\mathcal{I}}(X_i)$, whose probability of being positive deviates from 1/2 at most by $\max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}$.
- For a candidate subject $i \in \mathcal{I}$ such that the zero-effect null hypothesis (3) is true, denote the probability of $\Delta_{\widehat{m}}(X_i)$ being positive as follows:

$$q_i(\mathcal{I}) \coloneqq \mathbb{P}((A_i - 1/2)\cdot(Y_i - \widehat{m}(X_i)) > 0 | \mathcal{F}_0^{-Y}(\mathcal{I})), \tag{A8}$$

  which is upper bounded by

$$q_i(\mathcal{I}) \le \min\{\max\{\pi_{\max}, 1 - \pi_{\min}\}, \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\} \coloneqq q_{\max}(\mathcal{I}) \tag{A9}$$

  which is close to 1/2 (the ideal case) when *either* the true propensity score is close to 1/2 *or* the error of the expected outcome estimation $\varepsilon_n^Y(\mathcal{I})$ is small.
- The estimation error of $q_i(\mathcal{I})$ is upper bounded as follows:

$$\varepsilon_n^q(\mathcal{I}) \coloneqq \varepsilon_n^{\pi}(\mathcal{I}) - \max\{0, \max\{\pi_{\max}, 1 - \pi_{\min}\} - \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\}. \tag{A10}$$

Similar error terms can be derived for the procedure on set $\mathcal{II}$.

Theorem A2. *The FDR control of MaY-*$\text{I}^3_{\widehat{\pi}}$ *is upper bounded*:

$$\mathbb{E}[\text{FDP}^{\widehat{\pi}}_{\hat{t}}] \le \alpha\left\{1 + \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\varepsilon_n^q(\mathcal{I})\left(\frac{4}{q_{\max}(\mathcal{I})(1 - q_{\max}(\mathcal{I}))}\right)\right] + \mathbb{E}_{\mathcal{F}_0(\mathcal{II})}\left[\varepsilon_n^q(\mathcal{II})\left(\frac{4}{q_{\max}(\mathcal{II})(1 - q_{\max}(\mathcal{II}))}\right)\right]\right\},$$

when $\varepsilon_n^q(\mathcal{I}) \le q_{\max}(\mathcal{I})/2$ *and* $\varepsilon_n^q(\mathcal{II}) \le q_{\max}(\mathcal{II})/2$, *in an observational setting satisfying assumptions* (17), (A7), *and* (2) *for the zero-effect null* (3).

Corollary A1. (Doubly-robust FDR control) *As sample size n goes to infinity, the MaY-*$\text{I}^3_{\widehat{\pi}}$ *has asymptotic FDR control for the zero-effect null* (3) *when either*

- (*a*) *the propensity score estimation is consistent in the sense that* $\mathbb{E}_{\mathcal{F}_0(\mathcal{I})}[\varepsilon_n^{\pi}(\mathcal{I})] \to 0$; *and* (*b*) $\mathbb{E}_{\mathcal{F}_0(\mathcal{II})}[\varepsilon_n^{\pi}(\mathcal{II})] \to 0$, *and the true propensity scores are bounded away from 0 and 1; or*
- (*a*) *the expected outcome estimation is consistent in the sense that* $\varepsilon_n^Y(\mathcal{I}) \to 0$ *almost surely over the conditional distribution given* $\mathcal{F}_0(\mathcal{I})$; *and* (*b*) *same for* $\varepsilon_n^Y(\mathcal{II})$; *and* (*c*) *the difference between bounds on true propensity scores and 1/2 is larger than its estimation error*: $\max\{\pi_{\max}, 1 - \pi_{\min}\} - 1/2 \ge \varepsilon_n^{\pi}(\mathcal{I}) \vee \varepsilon_n^{\pi}(\mathcal{II})$ *almost surely; and (d) the outcome distribution is symmetric.*

Note that the aforementioned theorem states the FDR guarantee for the zero-effect null (3), and the error control for the nonpositive-effect null (A1) is discussed in Appendix B.6, whose condition to ensure asymptotic error control is similar, but practically it could be hard to have consistent estimation for the expected outcomes. Also, the aforementioned theorem provides an upper bound of the FDR in terms of the maximum estimation error over all subjects, while in practice, we expect the FDR to be close to the target level when the estimation error is small for most subjects.

A.2 Paired samples

A.2.1 Problem setup

Our discussion has focused on the case where samples are not paired, and the proposed algorithms can be extended to the paired-sample setting. Suppose there are n pairs of subjects. Let outcomes of subjects in the ith pair be Y_{ij}, treatment assignments be indicators A_{ij}, covariates be X_{ij} for $j = 1, 2$ and $i \in [n]$. We deal with randomized experiments without interference, and assume that

(i) conditional on covariates, the treatment assignments are independent coin flips:

$$\mathbb{P}[(A_{11}, \ldots, A_{n1}) = (a_1, \ldots, a_n) | X_1, \ldots, X_n] = \prod_{i=1}^{n} \mathbb{P}(A_i = a_i) = (1/2)^n, \quad \text{and} \quad A_{i1} + A_{i2} = 1 \text{ for all } i \in [n].$$

(ii) conditional on covariates, the outcome of one subject $Y_{i_1 j_1}$ is independent of the treatment assignment of another subject $A_{i_2 j_2}$ conditional on $A_{i_1 j_1}$, for any $(i_1, j_1) \neq (i_2, j_2)$.

As before, we can develop interactive algorithms for two types of error control (only the definitions when treating the potential outcomes as random variables are presented, but the FDR control still applies to all versions of the null):

$$H_{0i}^{(\text{zero, paired})} : (Y_{ij}^T | X_{ij}) \overset{d}{=} (Y_{ij}^C | X_{ij}) \text{ for both } j = 1, 2; \tag{A11}$$

$$H_{0i}^{(\text{nonpositive, paired})} : (Y_{ij}^T | X_{ij}) \preccurlyeq (Y_{ij}^C | X_{ij}) \text{ for both } j = 1, 2. \tag{A12}$$

Next, we present the extensions of Crossfit-I^3 for FDR control of zero effect and MaY-I^3 for FDR control of nonpositive effect for the paired-sample setting, by a few modifications in the effect estimator.

A.2.2 Interactive algorithms for paired samples

With the pairing information, the treatment effect can be estimated without involving $\widehat{m}$ as in (8):

$$\widehat{\Delta}_i^{\text{paired}} := (A_{i1} - A_{i2})(Y_{i1} - Y_{i2}), \tag{A13}$$

as used by Rosenbaum [26] and Howard and Pimentel [4], among others. The aforementioned estimation satisfies the critical property to guarantee FDR control: for a null pair i of two subjects with zero effects in (A11), we have

$$\mathbb{P}(\widehat{\Delta}_i^{\text{paired}} > 0 | \{Y_{j1}, Y_{j2}, X_{j1}, X_{j2}\}_{j=1}^n) \leq 1/2. \tag{A14}$$

Thus, the Crossfit-I^3 (Algorithm 2) with $\widehat{\Delta}_i$ replaced by $\widehat{\Delta}_i^{\text{paired}}$ has valid FDR control for the zero-effect null (A11), where the analyst excludes pairs using the available information, including $\{Y_{i1}, Y_{i2}, X_{i1}, X_{i2}\}$ for candidate subjects $i \in \mathcal{R}_{t-1}(\mathcal{I})$, and $\{Y_{j1}, Y_{j2}, A_{j1}, A_{j2}, X_{j1}, X_{j2}\}$ for noncandidate subjects $j \notin \mathcal{R}_{t-1}(\mathcal{I})$, and the sum $\sum_{i \in \mathcal{R}_{>t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i^{\text{paired}} > 0\}$ for FDR estimation. An automated strategy exclude pair i_t^* (at step 8 of Algorithm 1) under paired samples is the same as Algorithm 3, except $\widehat{\Delta}_i$ being replaced by $\widehat{\Delta}_i^{\text{paired}}$.

Recall the nonpositive-effect null under paired samples:

$$H_{0i}^{(\text{nonpositive, paired})} : (Y_{ij}|A_{ij} = 1, X_{ij}) \preccurlyeq (Y_{ij}|A_{ij} = 0, X_{ij}) \text{ for both } j = 1, 2,$$

and we observe that

$$\mathbb{P}(\widehat{\Delta}_i^{\text{paired}} > 0 | \{X_{j1}, X_{j2}\}_{j=1}^n) \leq 1/2, \tag{A15}$$

where $\widehat{\Delta}_i^{\text{paired}}$ is defined in (A13) of the main paper. Thus, the MaY-I^3 with $\widehat{\Delta}_i$ replaced by $\widehat{\Delta}_i^{\text{paired}}$ has valid FDR control for the nonpositive-effect null, where the analyst progressively excludes pairs using the available information:

$$\mathcal{F}_{t-1}^{-Y,\text{paired}} = \sigma\left(\{X_{i1}, X_{i2}\}_{i\in\mathcal{R}_{t-1}}, \{Y_{j1}, Y_{j2}, A_{j1}, A_{j2}, X_{j1}, X_{j2}\}_{j\notin\mathcal{R}_{t-1}}, \sum_{i\in\mathcal{R}_{t-1}(\mathcal{I})} \mathbb{1}\{\widehat{\Delta}_i^{\text{paired}} > 0\}\right).$$

We can also implement an automated version of the MaY-I^3 where the selection of the excluded subject follows a similar procedure as Algorithm 5. The difference is that in step 1, we estimate the treatment effect for noncandidate subjects $j \notin \mathcal{R}(\mathcal{I})$ directly as $\widehat{\Delta}_i^{\text{paired}} \equiv (A_{i1} - A_{i2})(Y_{i1} - Y_{i2})$ instead of Δ_i^{DR} to avoid estimating outcomes in $(\widehat{\mu}_0, \widehat{\mu}_1)$.

A.2.3 Numerical experiments

We compare the power of the interactive procedures with and without the pairing information, using the same experiments as previous. When the subjects within each pair have the same covariate values, the power under paired samples is higher than treating them as unpaired (Figure A3(a)), because the noisy variation in the observed outcomes that results from the potential control outcomes can be removed by taking the difference in outcomes within each pair.

The advantage of procedures under paired samples becomes less evident when the subjects within a pair do not match exactly. We simulate unmatched pairs by introducing a parameter ε such that for each pair i, the covariates of the two subjects within satisfy: $\mathbb{P}(X_{i1}(1) \neq X_{i2}(1)) = \varepsilon$, $\mathbb{P}(X_{i1}(2) \neq X_{i2}(2)) = \varepsilon$, $X_{i1}(3) = X_{i2}(3) + U(0, 2\varepsilon)$, where $U(0, 2\varepsilon)$ is uniformly distributed between 0 and 2ε, and a larger ε leads to a larger degree of mismatch. As ε increases, the power of procedures using the pairing information decreases (Figure A3(b)), because the estimated treatment effect $\widehat{\Delta}_i^{\text{paired}}$ becomes less accurate for the mismatching setting.

To further investigate the power decrease, we extend the definition of mismatch for $\varepsilon \in (0, 1)$ to a larger ε: $\mathbb{P}(X_{i1}(1) \neq X_{i2}(1)) = \min\{\varepsilon, 1\}$ and $\mathbb{P}(X_{i1}(2) \neq X_{i2}(2)) = \min\{\varepsilon, 1\}$ and $X_{i1}(3) = X_{i2}(3) + U(0, 2\varepsilon)$, where $U(0, 2\varepsilon)$ is uniformly distributed between 0 and 2ε, and a larger ε leads to a larger degree of mismatch. As ε increases,

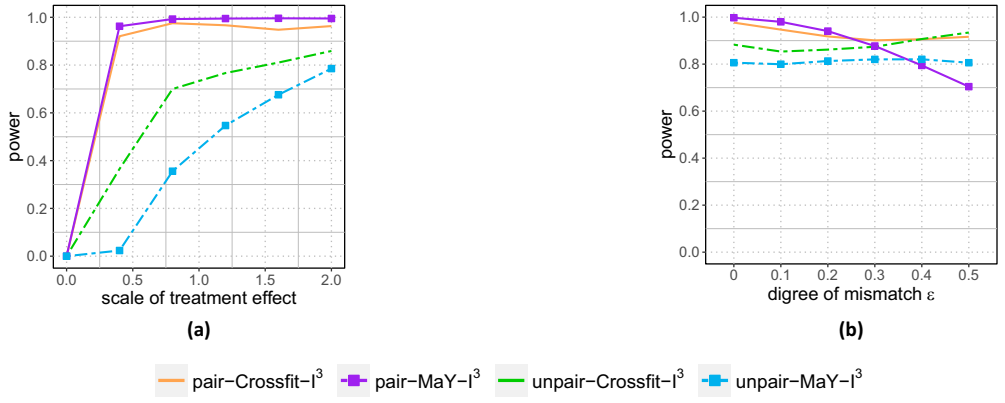

Figure A3: Power under paired samples with treatment effects specified by model (A26) when our proposed algorithms (Crossfit-I^3 and MaY-I^3) utilize the pairing information, which is higher than treating all subjects as unpaired. The advantage is less evident when the subjects within each pair are not exactly matched to have the same covariate values. (a) Exact pairs and (b) subjects within the pair do not match exactly.

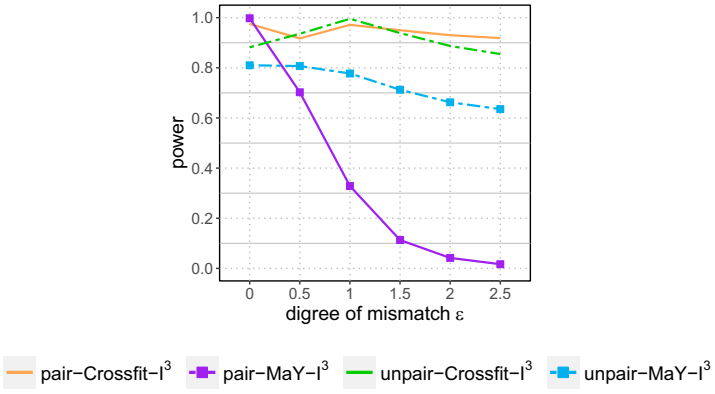

Figure A4: Power of identifying subjects with positive effects of the proposed algorithms (Crossfit-I^3 and MaY-I^3) with or without pairing information, when the scale of treatment effect is fixed at 2 and the degree of mismatch ε varies. The power of algorithms without pairing information first increase and then decrease as ε becomes larger.

the power under the unpaired samples first increases (Figure A4). It is because the treatment effect is positive when $X_i(3) > 1$, which only takes 15% proportion if without mismatching; thus, the pattern of treatment effect is not easy to learn. In contrast, when there is a positive shift on $X_i(3)$ as designed in the aforementioned mismatching setting, more subjects have positive effects so that the algorithm can more easily learn the effect pattern and hence increase the power. The power can slightly decrease when the degree of mismatch is too large ($\varepsilon > 1$), because there are fewer subjects without treatment effect, also affecting the estimation of treatment effect.

B Proof of error controls

The proofs are based on an optional stopping argument, as a variant of the ones presented in Lei and Fithian [12], Lei et al. [19], Li and Barber [33], and Barber and Candès [22].

Lemma A1. (Lemma 2 of Lei and Fithian [12]) *Suppose that, conditionally on the σ-field $\mathcal{G}_{-1}$, $b_1, \ldots, b_n$ are independent Bernoulli random variables with*

$$\mathbb{P}(b_i = 1 | \mathcal{G}_{-1}) = \rho_i \geq \rho > 0, \quad \textit{almost surely}.$$

Let $(\mathcal{G}_t)_{t=0}^{\infty}$ be a filtration with $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots$ and suppose that $[n] \supseteq C_0 \supseteq C_1 \supseteq \ldots$, with each subset C_{t+1} measurable with respect to $\mathcal{G}_t$. If we have

$$\mathcal{G}_t = \sigma\left(\mathcal{G}_{-1}, C_t, (b_i)_{i \notin C_t}, \sum_{i \in C_t} b_i\right), \tag{A16}$$

and τ is an almost-surely finite stopping time with respect to the filtration $(\mathcal{G}_t)_{t \geq 0}$, then

$$\mathbb{E}\left[\frac{1 + |C_\tau|}{1 + \sum_{i \in C_\tau} b_i} \middle| \mathcal{G}_{-1}\right] \leq \rho^{-1}.$$

B.1 Proof of Theorem 1

Proof. We show that the I^3 controls FDR by Lemma A1, where

$$b_i \coloneqq \mathbb{1}\{(A_i - 1/2) \cdot E_i \leq 0\} \quad \text{and} \quad \mathcal{G}_{-1} \coloneqq \sigma(\{Y_j, X_j\}_{j=1}^n) \quad \text{and} \quad C_t \coloneqq \mathcal{R}_t \cap \mathcal{H}_0,$$

for $t = 0, 1, \ldots$. The assumptions in Lemma A1 are satisfied: (a) $\mathbb{P}(b_i = 1 | \mathcal{G}_{-1}) \geq 1/2$ for subjects with zero effect $i \in \mathcal{H}_0$:

$$\begin{aligned}\mathbb{P}((A_i - 1/2)\cdot E_i \le 0|\mathcal{G}_{-1}) &= \mathbb{P}(A_i = 1)\mathbb{1}(E_i \le 0|\mathcal{G}_{-1}) + \mathbb{P}(A_i = 0)\mathbb{1}(E_i \ge 0|\mathcal{G}_{-1}),\\ &\qquad \text{because } A_i \text{ is independent of } \mathcal{G}_{-1}\\ &= 1/2[\mathbb{1}(E_i \le 0|\mathcal{G}_{-1}) + \mathbb{1}(E_i \ge 0|\mathcal{G}_{-1})] \ge 1/2;\end{aligned}$$

and (b) the filtration in our algorithm satisfies $\mathcal{F}_t \subseteq \mathcal{G}_t$, so the time of stopping the algorithm $\hat{t} \coloneqq \min\{t : \widehat{\mathrm{FDR}}(\mathcal{R}_t) \le \alpha\}$ is a stopping time with respect to $\mathcal{G}_t$; and (c) C_{t+1} is measurable with respect to $\mathcal{G}_t$. Thus, by Lemma A1, expectation, we have

$$\mathbb{E}\left[\frac{1 + |\mathcal{R}^+_{\hat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\hat{t}} \cap \mathcal{H}_0|} \;\middle|\; \mathcal{G}_{-1}\right] \le 2,$$

By definition, the FDR conditional on $\mathcal{G}_{-1}$ at the stopping time $\hat{t}$ is

$$\begin{aligned}\mathbb{E}\left[\frac{|\mathcal{R}^+_{\hat{t}} \cap \mathcal{H}_0|}{\max\{|\mathcal{R}^+_{\hat{t}}|, 1\}}\middle|\mathcal{G}_{-1}\right] &= \mathbb{E}\left[\frac{1 + |\mathcal{R}^-_{\hat{t}} \cap \mathcal{H}_0|}{\max\{|\mathcal{R}^+_{\hat{t}}|, 1\}} \cdot \frac{|\mathcal{R}^+_{\hat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\hat{t}} \cap \mathcal{H}_0|}\middle|\mathcal{G}_{-1}\right]\\ &\le \mathbb{E}\left[\widehat{\mathrm{FDR}}(\mathcal{R}_{\hat{t}})\cdot \frac{|\mathcal{R}^+_{\hat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\hat{t}} \cap \mathcal{H}_0|}\middle|\mathcal{G}_{-1}\right]\\ &\le \alpha\mathbb{E}\left[\frac{|\mathcal{R}^+_{\hat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\hat{t}} \cap \mathcal{H}_0|}\middle|\mathcal{G}_{-1}\right]\\ &= \alpha\mathbb{E}\left[\frac{1 + |\mathcal{R}^+_{\hat{t}} \cap \mathcal{H}_0|}{1 + |\mathcal{R}^-_{\hat{t}} \cap \mathcal{H}_0|} - 1\middle|\mathcal{G}_{-1}\right] \le \alpha,\end{aligned}$$

and the proof completes by applying the tower property of conditional expectation.

Notice that when the potential outcomes are treated as fixed, the same proof applies to the null defined as $Y_j^T = Y_j^C$, because the independence between A_i and $\mathcal{G}_{-1}$ still holds for the nulls. In the hybird version of the null $H_{0i}^{\text{zero}} : Y_i^T = Y_i^C$ almost surely-P_i, the aforementioned proof applies with $\mathcal{G}_{-1} \coloneqq \sigma(\{Y_j, Y_j^T, Y_j^C, X_j\}_{j=1}^n)$. Thus, FDR is controlled at level α conditional on the potential outcomes and covariates $\{Y_j^T, Y_j^C, X_j\}_{j=1}^n$. □

B.2 Proof of Theorem 2

Proof. Let the set of false rejections in $\mathcal{R}(\mathcal{I})$ be $\mathcal{V}(\mathcal{I})$. We conclude that the FDR of the I^3 implemented on set $\mathcal{I}$ is controlled at level $\alpha/2$:

$$\mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I})|}{\max\{|\mathcal{R}(\mathcal{I})|, 1\}}\middle|\mathcal{G}_{-1}\right] \le \alpha/2,$$

following the error control of the I^3 in Section B.1, where the initial candidate rejection set is $R_0 = \mathcal{I}$, and thus, $C_0 = \mathcal{I} \cap \mathcal{H}_0$. Similarly, the FDR of the I^3 implemented on set $\mathcal{II}$ is also controlled at level $\alpha/2$. Therefore, the FDR of the combined set $\mathcal{R}(\mathcal{I}) \cup \mathcal{R}(\mathcal{II})$ is controlled at level α as claimed:

$$\begin{aligned}&\mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I}) \cup \mathcal{V}(\mathcal{II})|}{|\mathcal{R}(\mathcal{I}) \cup \max\{|\mathcal{R}(\mathcal{II})|, 1\}}\middle|\mathcal{G}_{-1}\right]\\ &\le \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I})|}{|\mathcal{R}(\mathcal{I}) \cup \max\{|\mathcal{R}(\mathcal{II})|, 1\}}\middle|\mathcal{G}_{-1}\right] + \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{II})|}{|\mathcal{R}(\mathcal{I}) \cup \max\{|\mathcal{R}(\mathcal{II})|, 1\}}\middle|\mathcal{G}_{-1}\right]\\ &\le \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{I})|}{\max\{|\mathcal{R}(\mathcal{I})|,\ 1\}}\middle|\mathcal{G}_{-1}\right] + \mathbb{E}\left[\frac{|\mathcal{V}(\mathcal{II})|}{|\max\{|\mathcal{R}(\mathcal{II})|,\ 1\}}\middle|\mathcal{G}_{-1}\right] \le \alpha,\end{aligned}$$

the proof completes for the null (3) in the main paper after applying the tower property of conditional expectation. The FDR control also applies to the other two definitions of the null with fixed or hybrid version of the outcomes, following the same arguments as the end of Section B.1. □

B.3 Proof of Theorem A1

Proof. We prove that the FDR control holds for the I^3 implemented on $\mathcal{I}$, and the same conclusion applies to $\mathcal{II}$, so the overall FDR control is guaranteed following the Proof of Theorem 2 in Section B.2.

We first present the proof when the potential outcomes are viewed as random variables. Define $\mathcal{G}_{-1} \coloneqq \sigma(\{X_i\}_{i=1}^n, \{Y_i, A_i\}_{i\notin\mathcal{I}})$, and $\mathcal{G}'_t = \sigma(\mathcal{G}_{-1}, C_t, (Y_i, A_i)_{i\notin C_t}, \sum_{i\in C_t} b_i)$, which contains more information than $\mathcal{G}_t$ as defined in (A16). We claim that Lemma A1 holds when we replace $\mathcal{G}_t$ by $\mathcal{G}'_t$, because the distribution of b_i conditional on $\mathcal{G}_t$ is the same as on $\mathcal{G}'_t$ for any $t = 0, \ldots, n$. Similar to the proof of Theorem 1 in Section B.1, we check that the assumption in Lemma A1 are satisfied: (a) the filtration in our algorithm satisfies $\mathcal{F}_t \subseteq \mathcal{G}'_t$, so the time of stopping the algorithm $\hat{t} \coloneqq \min\{t : \widehat{\mathrm{FDR}}(\mathcal{R}_t) \le \alpha\}$ is a stopping time with respect to $\mathcal{G}'_t$; and (b) C_{t+1} is measurable with respect to $\mathcal{G}'_t$; and (c) for subjects with nonpositive effect $i \in \mathcal{H}_0^{\text{nonpositive}}$:

$$\mathbb{P}((A_i - 1/2)\cdot E_i^{-\mathcal{I}} \le 0 | \mathcal{G}_{-1}) \ge 1/2. \tag{A17}$$

To see that the last assumption holds, notice that

$$\begin{aligned}
&\mathbb{P}((A_i - 1/2)\cdot(Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) \le 0 | \mathcal{G}_{-1}) \\
&\quad = \mathbb{P}(Y_i^C \ge \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1})\mathbb{P}(A_i = 0) + \mathbb{P}(Y_i^T \le \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1})\mathbb{P}(A_i = 1); \text{ and} \\
&\mathbb{P}((A_i - 1/2)\cdot(Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) > 0 | \mathcal{G}_{-1}) \\
&\quad = \mathbb{P}(Y_i^C < \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1})\mathbb{P}(A_i = 0) + \mathbb{P}(Y_i^C > \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1})\mathbb{P}(A_i = 1).
\end{aligned}$$

For any potential outcomes of the nulls such that $(Y_i^T | X_i) \preccurlyeq (Y_i^C | X_i)$, it holds that

$$\mathbb{P}(Y_i^C \ge D | X_i) \ge \mathbb{P}(Y_i^T > D | X_i), \text{ and } \mathbb{P}(Y_i^T \le D | X_i) \ge \mathbb{P}(Y_i^C < D | X_i),$$

for any constant D, so

$$\begin{aligned}
&\mathbb{P}(Y_i^C \ge \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1}) \ge \mathbb{P}(Y_i^T > \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1}), \text{ and} \\
&\mathbb{P}(Y_i^T \le \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1}) \ge \mathbb{P}(Y_i^C < \widehat{m}^{-\mathcal{I}}(X_i) | \mathcal{G}_{-1}),
\end{aligned}$$

because $m^{-\mathcal{I}}(X_i)$ is fixed given $\mathcal{G}_{-1}$. Because $\mathbb{P}(A_i = 1)$ is 1/2 for all subjects, we have

$$\mathbb{P}((A_i - 1/2)\cdot(Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) \le 0 | \mathcal{G}_{-1}) \ge \mathbb{P}((A_i - 1/2)\cdot(Y_i - \widehat{m}^{-\mathcal{I}}(X_i)) > 0 | \mathcal{G}_{-1}),$$

which proves Claim (A17) and in turn the FDR control of the MaY-I^3.

When the potential outcomes are treated as fixed, the aforementioned proof applies to the null defined as $Y_i^T \le Y_i^C$ in (A2) of the main paper, in which case $\mathbb{P}(Y_i^C \ge D | \mathcal{G}_{-1})$ is zero or one, and the above arguments still hold. For the hybird version of the null (A3) in the main paper, the aforementioned proof applies with $\mathcal{G}_{-1} \coloneqq \sigma(\{Y_i^T, Y_i^C, X_i\}_{i=1}^n, \{Y_i, A_i\}_{i\notin\mathcal{I}})$. Thus, FDR is controlled at level α conditional on the potential outcomes and covariates $\{Y_j^T, Y_j^C, X_j\}_{j=1}^n$. □

B.4 Preliminaries to proof of error controls under observational studies

Lemma A2. *Let q_i be the conditional probability of a positive estimated sign*:

$$q_i \coloneqq \mathbb{P}[(A_i - 1/2)\cdot(Y_i - \widehat{m}(X_i)) > 0 | \mathcal{G}_{-1}],$$

where $\widehat{m}$ is an summary statistic (mean or mean) of $Y_i|X_i$, learned using $\mathcal{G}_{-1}$. Denote the maximum as

$q_{\max}(\mathcal{I}) \equiv \max_{i\in\mathcal{I}} q_i$, and let its estimation of using information in $\mathcal{G}_{-1}$ be $\hat{q}_{\max}$, and the (one-sided) estimation error be $\varepsilon_n^q(\mathcal{I}) = \max\{q_{\max}(\mathcal{I}) - \hat{q}_{\max}(\mathcal{I}), 0\}$. Define the FDR estimator as follows:

$$\widehat{\mathrm{FDR}}_{\hat{t}}(\mathcal{I}) \equiv \left(\frac{1}{1 - \widehat{q_{\max}}(\mathcal{I})} - 1\right)\frac{|\mathcal{R}_{\hat{t}}^-| + 1}{|\mathcal{R}_{\hat{t}}^+| \vee 1},$$

then the FDR of I^3 *run by Analyst I at level $\alpha/2$ is bounded*:

$$\mathbb{E}[\mathrm{FDP}_{\hat{t}}(\mathcal{I})|\mathcal{G}_{-1}] \le \alpha/2\left\{1 + \varepsilon_n^q(\mathcal{I})\cdot\frac{4}{q_{\max}(\mathcal{I})(1 - q_{\max}(\mathcal{I}))}\right\},$$

when $\varepsilon_n^q(\mathcal{I}) \le q_{\max}(\mathcal{I})/2$.

Proof. By Lemma A1 where

$$b_i \coloneqq \mathbb{1}\{(A_i - 1/2)\cdot E_i \le 0\} \quad \text{and} \quad C_t \coloneqq \mathcal{R}_t \cap \mathcal{H}_0,$$

and the tower property of conditional expectation, we have

$$\mathbb{E}\left[\frac{|\mathcal{R}_{\hat{t}}^+ \cap \mathcal{H}_0|}{1 + |\mathcal{R}_{\hat{t}}^-|}\middle|\mathcal{G}_{-1}\right] \le \left(\frac{1}{1 - q_{\max}(\mathcal{I})} - 1\right),$$

where the stopping time is denoted as $\hat{t}$. The FDR at $\hat{t}$ is upper bounded:

$$\begin{aligned}
\mathbb{E}\left[\frac{|\mathcal{R}_{\hat{t}}^+ \cap \mathcal{H}_0|}{|\mathcal{R}_{\hat{t}}^+| \vee 1}\middle|\mathcal{G}_{-1}\right] &= \mathbb{E}\left[\frac{1 + |\mathcal{R}_{\hat{t}}^-|}{|\mathcal{R}_{\hat{t}}^+| \vee 1}\cdot\frac{|\mathcal{R}_{\hat{t}}^+ \cap \mathcal{H}_0|}{1 + |\mathcal{R}_{\hat{t}}^-|}\middle|\mathcal{G}_{-1}\right] \\
&\le \alpha/2\left(\frac{1}{1 - \widehat{q_{\max}}(\mathcal{I})} - 1\right)^{-1}\mathbb{E}\left[\frac{|\mathcal{R}_{\hat{t}}^+ \cap \mathcal{H}_0|}{1 + |\mathcal{R}_{\hat{t}}^-|}\middle|\mathcal{G}_{-1}\right] \\
&\le \alpha/2\left(\frac{1}{1 - \widehat{q_{\max}}(\mathcal{I})} - 1\right)^{-1}\left(\frac{1}{1 - q_{\max}(\mathcal{I})} - 1\right).
\end{aligned}$$

By Taylor expansion on $f(x) = \left(\frac{1}{1-x} - 1\right)^{-1}$ around $x_0 = q_{\max}(\mathcal{I})$, we have

$$f(x_0 - \varepsilon) \le f(x_0) + \frac{\varepsilon}{(x_0 - \varepsilon)^2} \le f(x_0) + \frac{4\varepsilon}{x_0^2}$$

when $0 \le \varepsilon \le \frac{x_0}{2}$. Thus, FDR is close to the target level when $q_{\max}(\mathcal{I}) - \widehat{q_{\max}}(\mathcal{I})$ is small:

$$\begin{aligned}
\mathbb{E}[\mathrm{FDP}_{\hat{t}}^{\hat{\pi}}(\mathcal{I})|\mathcal{G}_{-1}] &\le \alpha/2\left\{\left(\frac{1}{1 - q_{\max}(\mathcal{I})} - 1\right)^{-1} + 4\varepsilon_n^q(\mathcal{I})\left(\frac{1}{q_{\max}(\mathcal{I})}\right)^2\right\}\left(\frac{1}{1 - q_{\max}(\mathcal{I})} - 1\right) \\
&= \alpha/2\left\{1 + \varepsilon_n^q(\mathcal{I})\frac{4}{q_{\max}(\mathcal{I})(1 - q_{\max}(\mathcal{I}))}\right\},
\end{aligned}$$

when $\varepsilon_n^q(\mathcal{I}) \le q_{\max}(\mathcal{I})/2$. □

B.5 Proof of Theorem 5

Proof. By Lemma A2, the probability of a positive sign of the estimated treatment effect q_i is

$$\begin{aligned}
q_i &\coloneqq \mathbb{P}((A_i - 1/2)\cdot E_i > 0|\mathcal{G}_{-1}) \\
&= \pi_i\mathbb{P}(E_i > 0|\mathcal{G}_{-1}) + (1 - \pi_i)\mathbb{P}(E_i < 0|\mathcal{G}_{-1}) \le \max\{1 - \pi_{\min}, \pi_{\max}\}.
\end{aligned}$$

Thus, $q_{\max}(\mathcal{I}) = \max\{1 - \pi_{\min}, \pi_{\max}\}$ and $\hat{q}_{\max}(\mathcal{I}) = \max\{1 - \widehat{\pi_{\min}}(\mathcal{I}), \widehat{\pi_{\max}}(\mathcal{I})\}$, and $\varepsilon_n^q(\mathcal{I}) = q_{\max}(\mathcal{I}) - \hat{q}_{\max}(\mathcal{I})$, we have

$$\mathbb{E}[\mathrm{FDP}_{\hat{t}}^{\hat{\pi}}(\mathcal{I})] \le \alpha/2\left\{1 + \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\varepsilon_n^q(\mathcal{I}) \cdot \frac{4}{q_{\max}(\mathcal{I})(1 - q_{\max}(\mathcal{I}))}\right]\right\},$$

when $\varepsilon_n^q(\mathcal{I}) \le \frac{1}{2}\max\{1 - \pi_{\min}, \pi_{\max}\}$. □

B.6 Proof of Theorem A2

Proof. Denote the individual propensity score as $\mathbb{P}(A_i = 1|X_i) = \pi_i$. For the nulls, the probability of a positive sign of the estimated treatment effect as q_i:

$$\begin{aligned} q_i(\mathcal{I}) &\coloneqq \mathbb{P}((A_i - 1/2)\cdot(Y_i - \widehat{m}(X_i)) > 0|\mathcal{F}_0^{-Y}(\mathcal{I})) \\ &= \pi_i\mathbb{P}(Y_i - \widehat{m}(x_i) > 0|\mathcal{F}_0^{-Y}(\mathcal{I})) + (1 - \pi_i)\mathbb{P}(Y_i - \widehat{m}(x_i) < 0|\mathcal{F}_0^{-Y}(\mathcal{I})) \\ &\le \min\{\max\{\pi_i, 1 - \pi_i\}, \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\}, \end{aligned}$$

where $\mathbb{P}(Y_i - \widehat{m}(x_i) > 0|\mathcal{F}_0^{-Y}(\mathcal{I}))$ can be separated from $\mathbb{P}(A_i - 1/2 > 0|\mathcal{F}_0^{-Y}(\mathcal{I}))$ because they are independent for zero-effect nulls. Thus, let an upper bound be

$$q_{\max}(\mathcal{I}) = \min\{\max\{\pi_{\max}, 1 - \pi_{\min}\}, \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\}.$$

Let the estimator of q_i be

$$\hat{q}_i \coloneqq \max\{\hat{\pi}_i, 1 - \hat{\pi}_i\}.$$

To describe the resulting estimation error of q_i, we define a difference $d_i(\mathcal{I}) \coloneqq \max\{\pi_i, 1-\pi_i\} - \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}$, which takes large value if the propensity score deviates from 1/2 (smaller value if the outcome probability deviates from 1/2). The true q_i is upper bounded by estimated $\hat{q}_i$ plus some estimation error that depends on $d_i(\mathcal{I})$:

$$\begin{aligned} q_i - \hat{q}_i &\le \varepsilon_i^{\pi}(\mathcal{I}) && \text{if } d_i(\mathcal{I}) \le 0; \\ q_i - \hat{q}_i &\le \varepsilon_i^{\pi}(\mathcal{I}) - d_i(\mathcal{I}) && \text{if } d_i(\mathcal{I}) > 0, \end{aligned}$$

where $\varepsilon_i^{\pi}(\mathcal{I}) = \pi_i - \hat{\pi}_i(\mathcal{I})$; it can be written in one line as

$$q_i - \hat{q}_i \le \varepsilon_i^{\pi}(\mathcal{I}) - \max\{0, \max\{\pi_i, 1 - \pi_i\} - \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\}.$$

Thus, the estimation error for $q_{\max}(\mathcal{I})$ is upper bounded as follows:

$$\max_{i\in\mathcal{I}}\{\varepsilon_i^{\pi}(\mathcal{I}) - \max\{0, \max\{\pi_i, 1 - \pi_i\} - \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\}\} \tag{A18}$$

$$\le \varepsilon_n^{\pi}(\mathcal{I}) - \max\{0, \max\{\pi_{\max}, 1 - \pi_{\min}\} - \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\} \coloneqq \varepsilon_n^q(\mathcal{I}). \tag{A19}$$

By Lemma A2, we have

$$\mathbb{E}[\mathrm{FDP}_{\hat{t}}^{\hat{\pi}}(\mathcal{I})] \le \alpha/2\left\{1 + \mathbb{E}_{\mathcal{F}_0(\mathcal{I})}\left[\varepsilon_n^q(\mathcal{I})\left(\frac{4}{q_{\max}(\mathcal{I})(1 - q_{\max}(\mathcal{I}))}\right)\right]\right\},$$

when $\varepsilon_n^q(\mathcal{I}) \le q_{\max}(\mathcal{I})/2$. □

Remark A1. If we consider nulls as nonpositive effects, we have

$$\begin{aligned} q_i(\mathcal{I}) &\coloneqq \mathbb{P}((A_i - 1/2)\cdot(Y_i - \widehat{m}(X_i)) > 0 | \mathcal{F}_0^{-Y}(\mathcal{I})) \\ &= \pi_i \mathbb{P}(Y_i^T - \widehat{m}(x_i) > 0 | \mathcal{F}_0^{-Y}(\mathcal{I})) + (1-\pi_i)\mathbb{P}(Y_i^C - \widehat{m}(x_i) < 0 | \mathcal{F}_0^{-Y}(\mathcal{I})) \\ &\le \pi_i \mathbb{P}(Y_i^C - \widehat{m}(x_i) > 0 | \mathcal{F}_0^{-Y}(\mathcal{I})) + (1-\pi_i)\mathbb{P}(Y_i^C - \widehat{m}(x_i) < 0 | \mathcal{F}_0^{-Y}(\mathcal{I})) \\ &\le \min\{\max\{\pi_i, 1-\pi_i\}, \max\{\Phi_{\max}[\varepsilon_n^Y(\mathcal{I})], 1 - \Phi_{\min}[-\varepsilon_n^Y(\mathcal{I})]\}\}, \end{aligned}$$

where $\Phi_{\max}(c) \coloneqq \max_{i\in\mathcal{I}} \mathbb{P}(Y_i^C - \mathbb{E}(Y_i^C|X_i) \le c | X_i)$, $\Phi_{\min}(c) \coloneqq \min_{i\in\mathcal{I}} \mathbb{P}(Y_i^C - \mathbb{E}(Y_i^C|X_i) \le c | X_i)$ and $\varepsilon_n^Y(\mathcal{I}) = \max_{i\in\mathcal{I}} |\widehat{m}(X_i) - \mathbb{E}(Y_i^C|X_i)|$, and then, we can make the same claim as above. However, it is harder to have robust FDR control when the propensity scores are poorly estimated, because $\widehat{m}(x_i)$ is an estimator for $\mathbb{E}(Y_i|X_i)$, which can be very different from the expected *control* outcome for a subject with negative effect. (We can design an algorithm where $\widehat{m}(x_i)$ is an estimation of the expected control outcome, but when the estimation is well, it would have zero power to detect positive effect if the subject is not treated.)

B.7 Error control guarantee for the linear-BH procedure

Theorem A3. *Suppose the outcomes follow a linear model*: $Y_i = l^{\Delta}(X_i)A_i + l^f(X_i) + U_i$, *where* l *denotes a linear function, and* U_i *is standard Gaussian noise. The linear-BH procedure controls FDR of the nonpositive-effect null in* (A1) *of the main article asymptotically as the sample size* n *goes to infinity*.

Note that the error control would not hold when the linear assumption is violated. For example, if the expected treatment effect $\mathbb{E}(Y_i^T - Y_i^C|X_i)$ is some nonlinear function of the covariates, the estimated treatment effect $\widehat{\Delta}_i^{\mathrm{BH}}$ would not be consistent; in turn, for the null subjects with zero effect, the p-values would not be valid (i.e., not stochastically equal or larger than uniform). Hence, the linear-BH procedure would not guarantee the desired FDR control, as we show in the numerical experiments in Section 3 of the main article.

Proof. For simplicity, we treat all the covariates as fixed values and denote them as the covariance matrix $\mathbb{X}_a = (X_i : A_i = a)^T$ for $a \in \{T, C\}$, where we temporarily use $A_i = T$ to denote the case of being treated $A_i = 1$. Under the linear assumption, the estimated outcome $\widehat{l}^a$ asymptotically follows a Gaussian distribution, whose expected value is $l^{\Delta}(X_i)\mathbb{1}\{a = T\} + l^f(X_i)$. Its variance can be estimated as follows:

$$\widehat{\mathrm{Var}}(\widehat{l}^a(X_i)) = \hat{\sigma}_a^2 (X_i^T (\mathbb{X}_a^T \mathbb{X}_a)^{-1} X_i^T),$$

where the variance from noise is estimated as follows:

$$\hat{\sigma}_a^2 = \sum_{A_i = a} (Y_i - \widehat{l}^a(X_i))^2 \Bigg/ \left(\sum_i \mathbb{1}\{A_i = a\} - d - 1 \right),$$

and d is the number of covariates. Note that the observed outcome also follows a Gaussian distribution $N(l^{\Delta}(X_i)\mathbb{1}\{a = T\} + l^f(X_i), \sigma_a^2)$. Note that in each estimated effect $\widehat{\Delta}_i^{\mathrm{BH}}$, the observed outcome Y_i^a is independent of the estimated potential outcome $Y_i^{\bar{a}}$, where $\bar{a}$ is the complement of a: $\bar{a} \cup a = \{T, C\}$. Thus, the estimated effect asymptotically follow a Gaussian distribution whose expected value is $l^{\Delta}(X_i)$ (nonpositive under the null) and the variance is $\mathrm{Var}(\widehat{\Delta}_i^{\mathrm{BH}}) = \mathrm{Var}(\widetilde{Y}_i^T) + \mathrm{Var}(\widetilde{Y}_i^C)$, where an estimation is $\widehat{\mathrm{Var}}(\widetilde{Y}_i^a) = \hat{\sigma}_a^2 \mathbb{1}\{A_i = a\} + \widehat{\mathrm{Var}}(\widehat{l}^a(X_i))\mathbb{1}\{A_i = \bar{a}\}$. Therefore, the resulting p-value P_i as defined in (15) of the main paper is asymptotically valid (uniform or stochastically larger) if subject i is a null, and hence, the BH procedure leads to asymptotic FDR control [34]. □

C Proof of power analysis

Our proof of the power analysis mainly uses the results presented by Arias-Castro and Chen [31], who consider the setup with n hypotheses, each associated with a test statistic V_i for $i \in [n]$. Assume the test statistics are independent with the survival function $\mathbb{P}(V_i \geq x) = \Psi_i(x)$, which equals $\Psi(x - \mu_i)$ where $\mu_i = 0$ under the null and $\mu_i > 0$ otherwise. They focus on a class of distribution called asymptotically generalized Gaussian (AGG), whose survival function satisfies:

$$\lim_{x\to\infty} x^{-\gamma} \log \Psi(x) = -1/\gamma, \tag{A20}$$

with a constant $\gamma > 0$. For example, a normal distribution is AGG with $\gamma = 2$. They discuss a class of multiple testing methods called *threshold procedure*: the final rejection set $\mathcal{R}$ is in the form

$$\mathcal{R} = \{i : V_i \geq \tau(V_1, \ldots, V_n)\}, \tag{A21}$$

for some threshold $\tau(V_1, \ldots, V_n)$, and separately study two types of thresholds: the BH procedure [11] with threshold:

$$\tau_{\mathrm{BH}} = V_{(\iota_{\mathrm{BH}})}, \quad \iota_{\mathrm{BH}} := \max\{i : V_{(i)} \geq \Psi^{-1}(i\alpha/n)\}, \tag{A22}$$

where $V_{(1)} \geq \ldots \geq V_{(n)}$ are ordered statistics; and the Barber-Candès (BC) procedure [22] with a threshold on the absolute value of V_i:

$$\tau_{\mathrm{BC}} = \inf\{v \in |\mathbf{V}| : \widehat{\mathrm{FDP}}(v) \leq \alpha\}, \tag{A23}$$

where $|\mathbf{V}| := \{|V_i| : i \in [n]\}$ is the set of sample absolute values, and

$$\widehat{\mathrm{FDP}}(v) := \frac{|\{i : V_i \leq -v\}| + 1}{\max\{|\{i : V_i \geq v\}|, 1\}},$$

and the final rejection set is those with positive V_i and value larger than τ_{BC}. The stopping rule for the BC procedure is similar to our proposed algorithms, as detailed next.

Recall in Section 4 of the main article, we consider a simplified automated version of the I^3 that exclude the subject with the smallest absolute value of the estimated treatment effect $|\widehat{\Delta}_i|$. Thus, the automated I^3 is a BC procedure where the test statistic of interest is $V_i = \widehat{\Delta}_i = 4(A_i - 1/2)(Y_i - \widehat{m}_n(X_i))$. Following the aforementioned notations and let Φ be the CDF for standard Gaussian, we denote the survival function for the nulls as follows:

$$\Psi_n^{\text{null}}(x) = \frac{1}{2}(1 - \Phi(x + \widehat{m}_n)) + \frac{1}{2}(1 - \Phi(x - \widehat{m}_n)),$$

which is a mixture of two Gaussians, with $\widehat{m}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, and $\widehat{m}_n \overset{a.s.}{\to} 0$ by the strong law of large numbers. For the non-nulls, the survival function is

$$\Psi_n^{\text{non-null}}(x) = \frac{1}{2}(1 - \Phi(x + \widehat{m}_n - \mu)) + \frac{1}{2}(1 - \Phi(x - \widehat{m}_n)).$$

Note that our setting is slightly different from the discussion presented by Arias-Castro and Chen [31] because the non-nulls differ from the nulls by a shift on one of the Gaussian component (rather than a shift in the overall survival function Ψ_n^{null}). Similar to the characterization by AGG in (A20), both survival functions Ψ_n^{null} and $\Psi_n^{\text{non-null}}$ asymptotically satisfy a tail property that for any $x_n \to \infty$ as $n \to \infty$:

$$\lim_{n\to\infty} x_n^{-\gamma} \log \Psi_n(x_n) = -1/\gamma, \tag{A24}$$

with probability one and $\gamma = 2$, which we later refer to as asymptotic AGG. Conclusions in our article basically follows the proofs presented by Arias-Castro and Chen [31] with the test statistics V_i specified as the estimated treatment effect $\widehat{\Delta}_i$.

C.1 Proof of Theorem 3

We first present the proof for the power of the automated Crossfit-I^3, and the power of the linear-BH is proven similarly as shown later.

Proof. Zero power when $r < \beta$.

The argument of zero power indeed applies to any threshold procedure as defined in (A21): $\mathcal{R} = \{i : \widehat{\Delta}_i \geq d\}$, for some $d \in \mathbb{R}$. Following the proof of Theorem 1 in Arias-Castro and Chen Arias-Castro and Chen [31], we argue that the FDR control cannot be satisfied for any $\alpha \in (0, 1)$ unless the threshold d is large enough such that $d > \mu + \delta_n$ with $\delta_n = \log\log n$; but in this case, most non-nulls cannot be included in the rejection set, and thus the power goes to zero.

First, we claim that when $d \leq \mu + \delta_n$, the false discovery proportion (FDP) goes to one in probability. By the proof of Theorem 1 in Arias-Castro and Chen [31], we have that FDP goes to one in probability if $\frac{(n-n_1)\Psi_n^{\text{null}}(\mu+\delta_n)}{n_1} \to \infty$ with probability one, where n_1 is the number of non-nulls. Their proof also verifies that $\frac{(n-n_1)\Psi_n^{\text{null}}(\mu+\delta_n)}{n_1} \to \infty$ because Ψ_n^{null} satisfies property (A24) with probability one.

Next, we show that when $d > \mu + \delta_n$, the power goes to zero. Notice that power can be equivalently defined as $\mathbb{E}(1-\text{FNR})$, where FNR (false negative rate) is defined as the proportion of non-nulls not identified. Again by the proof of Theorem 1 in Arias-Castro and Chen [31], we have that the FNR converge to one in probability if $\Psi_n^{\text{non-null}}(\mu + \delta_n)$ goes to zero with probability one, which is true because $\delta_n \to \infty$.

Combining the aforementioned two arguments, we conclude that for any threshold procedure whose rejection set is in the form of (A21), the power goes to zero for any FDR control $\alpha \in (0, 1)$ when $r < \beta$.

Note that the aforementioned proof assumes that the test statistics $V_i = \widehat{\Delta}_i$ are mutually independent. For simplicity, we design the Crossfit-I^3 where $\widehat{m}_n$ for the I^3 implemented on $\mathcal{I}$ is computed using data in $\mathcal{II}$, to ensure the above mutual independence. Thus, the aforementioned proof applies to the I^3 implemented on each half, $\mathcal{I}$ and $\mathcal{II}$. The overall power behaves the same asymptotically since $\mathcal{I}$ and $\mathcal{II}$ result from a random split of all subjects $[n]$. For all cases hereafter, we prove the power claim for the I^3 implemented on $\mathcal{I}$ conditional on data in $\mathcal{II}$, and the same claim holds for the overall power as reasoned above.

Half power when $r > \beta$. We first prove the limit inferior of the power is at least 1/2, and then the limit superior is at most 1/2, mainly using the proof of Theorem 3 in Arias-Castro and Chen [31].

They consider a sequence of thresholds $d_n^* = (\gamma r^* \log n)^{1/\gamma}$ for some $r^* \in (\beta, r \wedge 1)$. We first claim that the FDR estimator at d_n^* is less than any $\alpha \in (0, 1)$ for large n, or mathematically $\widehat{\text{FDR}}(d_n^*) \leq \alpha$. It can be verified by the proof of Theorem 3 in Arias-Castro and Chen [31] where the survival function of $\widehat{\Delta}_i$ is $G(d_n^*) = (1-\varepsilon)\Psi_n^{\text{null}}(d_n^*) + \varepsilon\Psi_n^{\text{non-null}}(d_n^*)$ with $\varepsilon = n^{-\beta}$, and the fact that $\Psi_n^{\text{non-null}}(d_n^*) \to 1/2$ and $\Psi_n^{\text{null}}(d_n^*) \to n^{-r^*}$ (by property (A24)) with probability one. It follows that the true stopping threshold τ_n satisfies $\tau_n \leq d_n^*$. Also, by Lemma 1 in Arias-Castro and Chen [31], we have that the proportion of correctly identified non-nulls at threshold d_n^* is $\frac{1}{n_1}\sum_{i \notin \mathcal{H}_0} \mathbb{1}\{\widehat{\Delta}_i \geq d\} = \Psi_n^{\text{non-null}}(d_n^*) + o_{\mathbb{P}}(1)$, where $\Psi_n^{\text{non-null}}(d)$ decreases in d and converges to 1/2 when $d = d_n^*$. Recall that the true stopping threshold is no larger than d_n^*, so the limit inferior of the power is at least 1/2.

The power converges to 1/2 once we show that the limit superior of the power is at most 1/2. Consider a positive constant $d^0 \in (0, \infty)$, and we claim that the actual stopping threshold $\tau_n \geq d^0$ for large n because the FDR estimator goes to one, following similar arguments in the proof of Theorem 3 in Arias-Castro and Chen [31]. Specifically,

$$\widehat{\text{FDR}}(d^0) \equiv \frac{|\{i \in [n] : \widehat{\Delta}_i \leq -d^0\}| + 1}{\max\{|\{i \in [n] : \widehat{\Delta}_i \geq d^0\}|, 1\}} = \frac{1 + n(1 - \widehat{G}_n(-d^0))}{\max\{n\widehat{G}_n(d^0), 1\}},$$

where $\widehat{G}_n(d^0) = \frac{1}{n}\sum_{i\in[n]} \mathbb{1}(\widehat{\Delta}_i \geq d^0)$ denotes the empirical survival function. Use the fact that $G_n(d^0) = (1-\varepsilon)\Psi_n^{\text{null}}(d^0) + \varepsilon\Psi_n^{\text{non-null}}(d^0) \to 1 - \Phi(d^0)$ and $G_n(-d^0) \to \Phi(d^0)$ almost surely, we observe that $\widehat{\text{FDR}}(d^0) \to 1$ with probability one. Also, the proportion of correctly identified non-nulls at threshold d^0 is

$\Psi_n^{\text{non-null}}(d^0) + o_{\mathbb{P}}(1)$ (recall in the previous paragraph), where $\Psi_n^{\text{non-null}}(d^0) \to 1/2 + 1/2(1 - \Phi(d^0))$. Thus, the power for large n is smaller than $1/2 + 1/2(1 - \Phi(d^0))$ for all $d^* \in (0, \infty)$; in other words, the limit superior of the power is smaller than $\inf_{d \in (0,\infty)} 1/2 + 1/2(1 - \Phi(d^*)) = 1/2$.

With the limit inferior and superior of the power bounded by 1/2, we conclude that the power converges to 1/2. In fact, the aforementioned proof implies that the power of identifying non-null subjects that are treated is one (notice that $\Psi_n^{\text{non-null, treated}}(d_n^*) = 1 - \Phi(d_n^* + \widehat{m}_n - \mu) \to 1$ with probability one, so the limit inferior of the power for treated non-nulls is at least 1). □

Proof for the linear-BH procedure. The power of the linear-BH procedure when there is no covariates can be proved following similar steps as above, and using intermediate results of Theorem 2 in Arias-Castro and Chen [31]. In their notation, the linear-BH procedure uses $V_i = \widehat{\Delta}_i^{\text{BH}}$ as the test statistics, and we separately discuss power among the treated group and the control group in ensure the independence among V_i. The survival functions for the nulls and non-nulls in the treated group are

$$\Psi_n^{\text{null}}(x) = 1 - \Phi\left(\frac{x}{\sqrt{1 + 1/n^C}}\right) \text{ and } \Psi_n^{\text{non-null}}(x) = 1 - \Phi\left(\frac{x - \mu}{\sqrt{1 + 1/n^C}}\right),$$

where n^C is the number of untreated subjects. For the control group, the survival functions are

$$\Psi_n^{\text{null}}(x) = 1 - \Phi\left(\frac{x + \widehat{Y}^T}{\sqrt{1 + 1/n^T}}\right) \text{ and } \Psi_n^{\text{non-null}}(x) = 1 - \Phi\left(\frac{x + \widehat{Y}^T}{\sqrt{1 + 1/n^T}}\right),$$

where $\widehat{Y}^T = \sum_{A_i=1} Y_i \overset{a.s.}{\to} 0$, and n^T is the number of treated subjects. Since the aforementioned distributions converge to a Gaussian, these survival functions satisfy the AGG property asymptotically as defined in (A24).

Proof. First, we claim that the power goes to zero when $r < \beta$, following the proof in Section C.1 for any threshold procedure (separately for the treated group conditional on control group). Then, we prove that power converges to 1/2 when $r > \beta$: the power among untreated subjects is asymptotically zero because the survival functions for the nulls and non-nulls are the same; the power among treated subjects is asymptotically one following the proof of Theorem 2 in Arias-Castro and Chen [31] as detailed next.

Again consider the sequence of thresholds $d_n^* = (\gamma r^* \log n)^{1/\gamma}$ for some $r^* \in (\beta, r \wedge 1)$. We first claim that the FDR estimator at d_n^* is less than any $\alpha \in (0, 1)$ for large n, or mathematically $\widehat{\text{FDR}}(d_n^*) \le \alpha$. It can be verified by the proof of Theorem 2 in Arias-Castro and Chen [31], where $G(d_n^*) = (1 - \varepsilon)\Psi_n^{\text{null}}(d_n^*) + \varepsilon\Psi_n^{\text{non-null}}(d_n^*)$ with $\varepsilon = n^{-\beta}$, and the fact that $\Psi_n^{\text{non-null}}(d_n^*) \to 1$ for the treated group and $\Psi_n^{\text{null}}(d_n^*) \to n^{-r^*}$ (by property (A24)) with probability one. It follows that the true stopping threshold τ_n satisfies $\tau_n \le d_n^*$. Also, by Lemma 1 in Arias-Castro and Chen [31], we have that the proportion of correctly identified non-nulls at threshold d_n^* is $\frac{1}{n_1}\sum_{i \notin \mathcal{H}_0} \mathbb{1}\{\widehat{\Delta}_i^{\text{BH}} \ge d\} = \Psi_n^{\text{non-null}}(d_n^*) + o_{\mathbb{P}}(1)$, where $\Psi_n^{\text{non-null}}(d)$ for the treated group decreases in d and converges to 1 when $d = d_n^*$. Recall that the true stopping threshold is no larger than d_n^*, so the limit inferior of the power among the treated subjects is at least 1. Therefore, the overall power converges to 1/2. □

C.2 Proof of Theorem 4

We first consider the power when all subjects are non-nulls ($\beta = 0$).

Lemma A3. *Given any fixed FDR control level $\alpha \in (0, 1)$ and let the number of subjects n goes to infinity. When all subjects are non-nulls, the stopping time $\tau = 0$ with probability tending to one if $\mu > \Phi^{-1}(\frac{1}{1+\alpha})$, and in this case the power converges to $\Phi(\mu)$.*

For example, when $\alpha = 0.2$, the asymptotic power of the automated I^3 is larger than 0.8 if $\mu \geq 1$.

Proof. The stopping time $\tau = 0$ if and only if the FDR control is satisfied when all the subjects are included: $\widehat{\text{FDR}}_n(\mathcal{R}_0) = \frac{|\mathbb{R}_0^-| + 1}{\max\{|\mathbb{R}_0^+|, 1\}} \leq \alpha$, or equivalently, $\frac{|\mathbb{R}_0^+|}{n} \geq \frac{1+\frac{1}{n}}{1+\alpha}$. Notice that the proportion of positive $\widehat{\Delta}_i$ converges to a constant: $\frac{|\mathbb{R}_0^+|}{n} \xrightarrow{a.s.} \Phi(\mu)$, because $\widehat{\Delta}_i$ of each non-null follows a Gaussian distribution with mean μ and variance less than 2. Thus, if $\Phi(\mu) > \frac{1}{1+\alpha}$, for any $\varepsilon \in (0, 1)$, there exists N such that for all $n \geq N$, we have that (a) $|\frac{|\mathbb{R}_0^+|}{n} - \Phi(\mu)| < \varepsilon$ with probability at least $1 - \varepsilon$; and (b) $\widehat{\text{FDR}}_n(\mathcal{R}_0) = \frac{|\mathbb{R}_0^-| + 1}{\max\{|\mathbb{R}_0^+|, 1\}} \leq \alpha$ (hence $\tau = 0$) with probability at least $1 - \varepsilon$. (Notice that the threshold N can be chosen as not depending on μ, which is useful in the next proof.) In such a case, the power is no less than $(1 - \varepsilon)(\Phi(\mu) - \varepsilon)$ when $n \geq N$; and the power is no larger than $\Phi(\mu) - \varepsilon$; so the power converges to $\Phi(\mu)$. The proof completes once notice that the condition $\Phi(\mu) > \frac{1}{1+\alpha}$ is equivalent to $\mu > \Phi^{-1}(\frac{1}{1+\alpha})$. □

Proof of Theorem 4. Power of the Crossfit-I^3. Recall that the I^3 implemented on $\mathcal{I}$ exclude subjects based on the averaged estimated effect on $\mathcal{II}$: $\text{Pred}(x) = \overline{\widehat{\Delta}_i(X_i = x)}$, which converges to μ almost surely when $x = 1$ (the non-nulls), and 0 almost surely when $x = 0$ (the nulls). Thus, no non-nulls in $\mathcal{I}$ would be excluded before excluding all the nulls in $\mathcal{I}$ (with probability going to one) for any fixed $\mu > 0$. Combined with Lemma A3, we have that if $\mu > \Phi^{-1}(\frac{1}{1+\alpha})$, for any $\varepsilon \in (0, 1)$, there exists $N(\varepsilon, \alpha)$ such that for all $n \geq N$, the power of the I^3 is higher than $\Phi(\mu) - \varepsilon$. Also, the limit of power increases to one for any $r > 0$ (where the signal μ increases): there exists $N'(\varepsilon)$ such that for all $n \geq N'(\varepsilon)$, $\Phi(\mu) \geq 1 - \varepsilon$. Therefore, for any $\varepsilon \in (0, \frac{1}{1+\alpha})$, we have that for all $n \geq \max\{N'(\varepsilon), N(\varepsilon, \alpha)\}$, the power of I^3 implemented on $\mathcal{I}$ is no less than $1 - 2\varepsilon$; thus completes the proof.

Power of the linear-BH procedure. As mentioned earlier, we separately argue that the power for the treated group and the control group converges to zero when $r < \beta$, and converges to one when $r > \beta$. For a subject in the treated group with $X_i = x$, where $x \in \{0, 1\}$, the estimated effect is a Gaussian $\widehat{\Delta}_i^{\text{BH}} \sim N(0, 1 + \frac{1}{\sum_i \mathbb{1}(X_i = x, A_i = 0)})$ for the nulls and $\widehat{\Delta}_i^{\text{BH}} \sim N(\mu, 1 + \frac{1}{\sum_i \mathbb{1}(X_i = x, A_i = 0)})$ for the non-nulls. For a subject in the control group with $X_i = x$ where $x \in \{0, 1\}$, the estimated effect is a Gaussian $\widehat{\Delta}_i^{\text{BH}} \sim N(0, 1 + \frac{1}{\sum_i \mathbb{1}(X_i = x, A_i = 1)})$ for the nulls and $\widehat{\Delta}_i^{\text{BH}} \sim N(\mu, 1 + \frac{1}{\sum_i \mathbb{1}(X_i = x, A_i = 1)})$ for the non-nulls.

The power of the linear-BH procedure directly results from Theorem 2 in Arias-Castro and Chen [31] because in both the treated and control group, (a) the linear-BH procedure is the BH procedure where the random variable of interest is $V_i = \widehat{\Delta}_i^{\text{BH}}$; and (b) $\widehat{\Delta}_i^{\text{BH}}$ of non-nulls and nulls differ by a shift μ; and (c) the survival function of $\widehat{\Delta}_i^{\text{BH}}$ is asymptotically AGG (recall definition in (A24)) since it converges to a Gaussian distribution). □

D Experiments when data do not follow working model (7)

In previous discussions, we follow the working model (7), where the treated outcome is the sum of treatment effect $\Delta(X)$ and the control outcome, in both the modeling of Crossfit-I^3 algorithm and the simulated data for evaluating performance. Here, to understand the performance of Crossfit-I^3 when the data do not follow the working model (7), we change the simulated data while keeping the same modeling in Crossfit-I^3.

We specify the potential outcomes as follows:

$$Y_i^C = f(X_i) + U_i \text{ and } Y_i^T = L \cdot D(X_i) + (5 - L)D(X_i)f(X_i) + f(X_i) + U_i, \tag{A25}$$

where

$$D(X_i) = 5X_i^3(3)\mathbb{1}\{X_i(3) > 1\} - X_i(1)/2, \tag{A26}$$

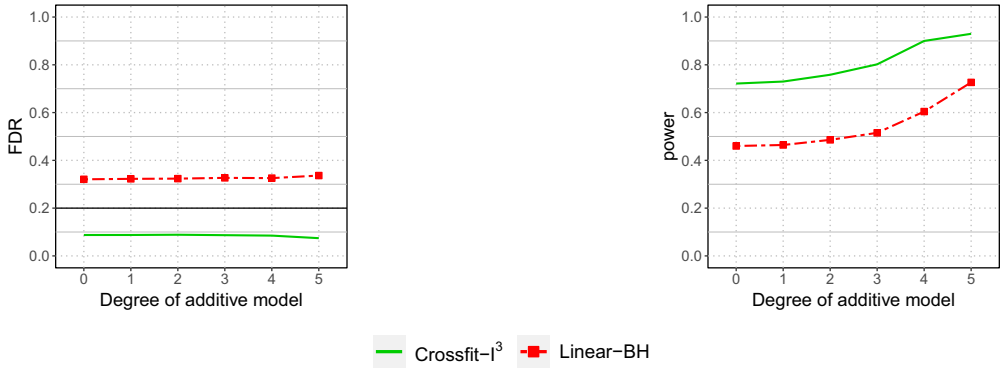

Figure A5: FDR (left) and power (right) of the Crossfit-I^3 compared with the linear-BH procedure, with the potential outcomes specified as model (A25) and the scale L varying in $\{0, 1, 2, 3, 4, 5\}$. The Crossfit-I^3 controls FDR and can achieve high power even when the data does not follow the additive model (7). The power is higher as data agrees more with the additive model, because Crossfit-I^3 can model the treatment effect more accurately.

and $L \in (0, 5)$ encodes the degree of agreement with the additive model (7) – $L = 5$ corresponds to the same simulation in Section 3.2 when $S_\Delta = 5$, and the data follow the additive model. Compared with the additive model (7) in the main article, the treatment effect becomes a function of the potential control outcome $L \cdot D(X_i) + (5 - L)D(X_i)f(X_i)$.

As expected, Crossfit-I^3 has valid FDR control, and the power becomes higher when the simulated data are more aligned with the working model (7). Also, the power of Crossfit-I^3 is consistently higher than the linear-BH procedure when varying L, and the power does not decay much when working model (7) does not reflect the groundtruth data, i.e., $L = 0$ (Figure A5).