The Annals of Applied Statistics
2022, Vol. 16, No. 2, 1111–1129
https://doi.org/10.1214/21-AOAS1535

DETECTING HETEROGENEOUS TREATMENT EFFECTS WITH INSTRUMENTAL VARIABLES AND APPLICATION TO THE OREGON HEALTH INSURANCE EXPERIMENT

BY MICHAEL JOHNSON[1,a], JIONGYI CAO[2,c] AND HYUNSEUNG KANG[1,b]

[1]*Department of Statistics, University of Wisconsin-Madison,* [a]*mwjohnson8@wisc.edu,* [b]*hyunseung@stat.wisc.edu*
[2]*Department of Statistics, University of Chicago,* [c]*jiongyi@uchicago.edu*

There is an increasing interest in estimating heterogeneity in causal effects in randomized and observational studies. However, little research has been conducted to understand effect heterogeneity in an instrumental variables study. In this work we present a method to estimate heterogeneous causal effects using an instrumental variable with matching. The method has two parts. The first part uses subject-matter knowledge and interpretable machine-learning techniques, such as classification and regression trees, to discover potential effect modifiers. The second part uses closed testing to test for statistical significance of each effect modifier while strongly controlling the familywise error rate. We apply this method on the Oregon Health Insurance Experiment, estimating the effect of Medicaid on the number of days an individual's health does not impede their usual activities by using a randomized lottery as an instrument. Our method revealed Medicaid's effect was most impactful among older, English-speaking, non-Asian males and younger, English-speaking individuals with, at most, a high school diploma or General Educational Development.

1. Introduction.

1.1. *Motivation*: *Utilization of Medicaid in Oregon and the complier average causal effect*. In January of 2008, Oregon reopened its Medicaid-based health insurance plan for its eligible residents and, for a brief period, allowed a limited number of individuals to enroll in the program. Specifically, a household in Oregon was randomly selected by a lottery system run by the state, and any eligible individual in the household could choose to enroll in the new health insurance plan; households that were not selected by the lottery could not enroll whatsoever.

For policymakers, Oregon's randomized lottery system was a unique opportunity, specifically, a natural experiment to study Medicaid's causal effect on a variety of health and economic outcomes, as directly randomizing Medicaid (or withholding it) to individuals would be infeasible and unethical. In this natural experiment, commonly referred to as the Oregon Health Insurance Experiment (OHIE), Finkelstein et al. (2012) used the randomized lottery as an instrumental variable (see Section 2.2 for details) to study the complier average causal effect (CACE) or the effect of Medicaid among individuals who enrolled in Medicaid after winning the lottery (Angrist, Imbens and Rubin (1996)). The CACE reflects Medicaid's impact among a subgroup of individuals and differs from the average treatment effect for the entire population (ATE) or the intent-to-treat (ITT) effect of the lottery itself on the outcome. In this paper we focus on studying the CACE; see Imbens (2010), Swanson and Hernán (2013, 2014) for additional discussions on the CACE.

Received March 2020; revised August 2021.

Key words and phrases. Causal inference, complier average causal effect, heterogeneous treatment, instrumental variables, matching, machine learning, Oregon Health Insurance Experiment.

Often, in studying the CACE the population of compliers is assumed to be homogeneous, whereby two compliers are alike and have the same treatment effect. But no two individuals are the same, and it is plausible that some compliers may benefit more from the treatment than other compliers. For example, sick individuals who enroll in Medicaid after winning the lottery may benefit more from Medicaid than healthy individuals. Also, the perceived benefit of enrolling in Medicaid among sick vs. healthy individuals may create heterogeneity in the compliance rate, that is, the number of people who sign up when they win the lottery, with sick people, presumably, signing up more than healthy people. Alternatively, if people are equally likely to enroll in Medicaid when they win the lottery, those who are unemployed may benefit more from Medicaid in terms of reducing out-of-pocket healthcare spending and medical debt than those who are employed. The theme of this paper is to explore these issues, specifically, the heterogeneity of CACE and how to discover them in an honest manner by using well-known matching methods and recent tree-based methods in heterogeneous treatment effect estimation.

1.2. *Prior work and our contributions.* Traditional approaches to study heterogeneous effects required subgroups to be specified a priori rather than allowing for unknown subgroups to be discovered by the data (Rothwell (2005), Stallones (1987), Yusuf et al. (1991)). In recent years there have been many works in causal inference using tree-based methods to estimate effect heterogeneity or to identify data-driven subgroups when there is full compliance; see Athey and Imbens (2016), Athey, Tibshirani and Wager (2019), Chernozhukov et al. (2018), Hahn, Murray and Carvalho (2020), Hill (2011), Lee, Bargagli-Stoffi and Dominici (2021), Su et al. (2009), Wager and Athey (2018), Wang and Rudin (2021) and references therein. Notably, Wang and Rudin (2021), Lee, Small and Dominici (2021), and Lee, Bargagli-Stoffi and Dominici (2021) used data to suggest novel effect modifiers, aiding domain experts to identify new subgroups when there are too many possible subgroups to consider. The majority of the aforementioned work utilizes sample splitting or subsampling to obtain honest inference. Here, honest inference refers to a procedure that controls the type I error rate (or the familywise error rate) of testing a null hypothesis about a treatment effect at a desired level α; see Section 2.4 for additional discussions. However, Hsu, Small and Rosenbaum (2013) used pair matching and classification and regression trees (CART) (Breiman et al. (1984)) to conduct honest inference, all without sample splitting. A follow-up work by Hsu et al. (2015) formally showed that the procedure strongly controls the familywise error rate for testing heterogeneous treatment effects, again without sample splitting. Subsequent works by Lee et al. (2018), Lee, Small and Rosenbaum (2018), and Lee, Small and Dominici (2021) extended this idea to increase statistical power of detecting such effects.

There is also work on nonparametrically estimating treatment effects using instrumental variables (IV), mostly using likelihood, series, sieve, minimum distance, and/or moment-based methods; see Abadie (2003), Ai and Chen (2003), Athey, Tibshirani and Wager (2019), Blundell, Chen and Kristensen (2007), Blundell and Powell (2003), Chen and Pouzo (2012), Darolles et al. (2011), Hall and Horowitz (2005), Newey and Powell (2003), Su, Murtazashvili and Ullah (2013) and references therein. Recently, Bargagli-Stoffi and Gnecco (2018) and Bargagli-Stoffi, De-Witte and Gnecco (2019) explored effect heterogeneity in the CACE by using causal trees (Athey and Imbens (2015)) and Bayesian causal forests (Hahn, Murray and Carvalho (2020)), specifically by estimating heterogeneity in the ITT effect and dividing it by the compliance rate. However, to the best of our knowledge, none have used matching, a popular, intuitive, and easy-to-understand method in causal inference, as a device to nonparametrically estimate treatment heterogeneity in the CACE and to guarantee strong familywise type I error control. Works on using matching with an instrument by Baiocchi et al. (2010)

and Kang et al. (2013, 2016) only focused on the population CACE; they do not explore heterogeneity in the CACE. Also, aforementioned works by Hsu, Small and Rosenbaum (2013) and Hsu et al. (2015), using matching and CART, did not consider instruments.

The goal of this paper is to propose a matching-based method to study effect heterogeneity and to identify novel, data-driven subgroups in instrumental variables settings. Specifically, the target estimand of interest is what we call the *heterogeneous* complier average causal effect (H-CACE). A heterogeneous complier average causal effect (H-CACE) is the usual complier average causal effect, but for a subgroup of individuals defined by their preinstrument covariates. At a high level, H-CACE explores treatment heterogeneity in the complier population, where we suspect that not all compliers in the data react to the treatment in the same way. Some subgroup of compliers may respond to the treatment differently than another subgroup of compliers, who may not respond to the treatment at all; some may even be more likely to be compliers if they believe the treatment would benefit them, and they may actually benefit from the treatment. The usual CACE obscures the underlying heterogeneity among compliers by averaging across different types of compliers, whereas H-CACE attempts to expose it. Also, in the case where the four compliance types in Angrist, Imbens and Rubin (1996), specifically, compliers, never-takers, always-takers, and defiers have identical effects; the H-CACE can identify the heterogeneous treatment effect for the entire population using an instrument. Section 2.3 formalizes H-CACE and provides additional discussions.

Methodologically, to study H-CACE we combine existing ideas of heterogeneous treatment effect estimation in non-IV matching contexts by Hsu et al. (2015) and matching with IVs by Baiocchi et al. (2010) and Kang et al. (2016). Specifically, we first follow Baiocchi et al. (2010) and Kang et al. (2016) and conduct pair matching on a set of preinstrument covariates. Second, we follow Hsu et al. (2015), where we obscure the difference in the outcomes between treated and controls by using absolute differences, and use CART to discover novel subgroups of study units without contaminating downstream inference. Specifically, we use closed testing to test the H-CACE in different subgroups while strongly controlling for familywise error rate (Marcus, Peritz and Gabriel (1976)). Simulation studies are conducted to evaluate the performance of our proposed method under varying levels of compliance and effect heterogeneity. The simulation study also compares our method to the recent aforementioned method by Bargagli-Stoffi, De-Witte and Gnecco (2019). We then use our method to analyze heterogeneity in the effect of Medicaid on increasing the number of days a complying individual's health does not hamper their usual activities.

2. Method.

2.1. *Notation.* Let $i = 1, \dots, I$ index the I matched pairs and $j = 1, 2$ index the units within each matched pair i. Let Z_{ij} be a binary instrument for unit j in matched pair i where one unit in the pair receives the instrument value $Z_{ij} = 1$ and the other receives the value $Z_{ij} = 0$. In the OHIE data, $Z_{ij} = 1$ and $Z_{ij} = 0$ denotes an individual winning or losing the Medicaid lottery, respectively. Let $\mathbf{Z}$ be the vector of instruments, $\mathbf{Z} = (Z_{11}, Z_{12}, \dots, Z_{I1}, Z_{I2})$ and $\mathcal{Z}$ denote an event of instrument assignments for all units.

For unit j in matched pair i, let d_{1ij} and d_{0ij} denote the binary potential treatment/exposure, given the instrument value of $Z_{ij} = 1$ and $Z_{ij} = 0$, respectively. Further, define the potential response $r_{1ij}^{(d_{1ij})}$ for unit j in matched set i with exposure d_{1ij} receiving instrument value $Z_{ij} = 1$; we define $r_{0ij}^{(d_{0ij})}$ similarly but with instrument value $Z_{ij} = 0$. For the OHIE data, d_{1ij} denotes whether an individual enrolled in Medicaid and $r_{1ij}^{(d_{1ij})}$ denotes the potential outcome when the individual wins the lottery $Z_{ij} = 1$. For unit j in matched set i, the observed response is defined as $R_{ij} = r_{1ij}^{(d_{1ij})} Z_{ij} + r_{0ij}^{(d_{0ij})}(1 - Z_{ij})$ and the observed treatment is defined as $D_{ij} = d_{1ij} Z_{ij} + d_{0ij}(1 - Z_{ij})$. The notation assumes that the Stable Unit Treatment Value Assumption (SUTVA) holds (Rubin (1980)). Define $\mathcal{F} = \{(r_{1ij}^{(d_{1ij})}, r_{0ij}^{(d_{0ij})}, d_{1ij}, d_{0ij}, \mathbf{X}_{ij}, u_{ij}), i = 1, \dots, I, j = 1, 2\}$ to be the set of potential outcomes, treatments, and covariates, both observed, $\mathbf{X}_{ij}$, and unobserved, u_{ij}.

When partitioning the matched sets into subgroups for discovering effect heterogeneity, the following notation is used. We define a "set of sets" or grouping $\mathcal{G}$, which contains mutually exclusive and exhaustive subsets of the pairs $s_g \subseteq \{1, \dots, I\}$, so that $\mathcal{G} = \{s_1, \dots, s_G\}$. The subscript g in s_g is used to denote a unit partitioned into the gth subset s_g. To avoid overloading the notation, s and s_g will be used interchangeably when it is not necessary to specify a subgroup g. The set of potential outcomes, treatments, and covariates for subset s_g are defined as $\mathcal{F}_{s_g} = \{(r_{1sij}^{(d_{1sij})}, r_{0sij}^{(d_{0sij})}, d_{1sij}, d_{0sij}, \mathbf{X}_{sij}, u_{sij}) : s_g \subseteq \{1, \dots, I\}, i \in s_g, j = 1, 2\}$, where $\mathcal{F} = \bigcup_s \mathcal{F}_s$. For example, consider a grouping of two subgroups, $\mathcal{G} = \{s_1, s_2\}$, for $I = 10$ matched pairs. Suppose the first few pairs and the last pair make up the first subgroup and the rest are in the second subgroup, say $s_1 = \{1, 2, 3, 10\}$ and $s_2 = \{4, 5, 6, 7, 8, 9\}$. The set of potential responses, treatments, and covariates for the first group is then $\mathcal{F}_{s_1} = \{(r_{1s_1ij}^{(d_{1s_1ij})}, r_{0s_1ij}^{(d_{0s_1ij})}, d_{1s_1ij}, d_{0s_1ij}, \mathbf{X}_{s_1ij}, u_{s_1ij}) : s_1 = \{1, 2, 3, 10\}, i \in s_1, j = 1, 2\}$. The observed response, binary instrument, and exposure for a given unit in subset s_g is denoted as Z_{s_gij}, R_{s_gij}, and D_{s_gij}, respectively.

2.2. *Review*: *Matching*, *instrumental variables*, *and the CACE*. Matching is a popular nonparametric technique in observational studies to balance the distribution of the observed covariates between treated and control units by grouping units based on the similarity of their covariates; see Stuart (2010), Chapters 3 and 8 of Rosenbaum (2010, 2020) for overviews of matching. Pair matching is a specific type of matching where each treated unit is only matched to one control unit. In the context of instrumental variables and pair matching, the instrument serves as the treatment/control variable, and the matching algorithm creates I matched pairs, where the two units in a matched pair are similar in their observed covariates x_{ij}, but one receives the instrument value $Z_{ij} = 1$ and the other receives the instrument value $Z_{ij} = 0$.

Instrumental variables (IV) is a popular approach to analyze causal effects when unmeasured confounding is present and is based on using a variable called an instrument (Angrist, Imbens and Rubin (1996), Baiocchi, Cheng and Small (2014), Hernán and Robins (2006)). The instrument must satisfy three core assumptions: (A1) the instrument is related to the exposure or treatment, or $\sum_{i=1}^{I} \sum_{j=1}^{2} (d_{1ij} - d_{0ij}) \neq 0$ (commonly referred to as instrument relevance); (A2) the instrument is not related to the outcome in any way, except through the treatment, or $r_{0ij}^{(d)} = r_{1ij}^{(d)} \equiv r_{ij}^{(d)}$ for a fixed d (commonly referred to as the exclusion restriction), and (A3) the instrument is not related to any unmeasured confounders that affect the treatment and the outcome, or $P(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ within each pair i (commonly referred to as instrument ignorability or exchangeability). If these core assumptions are satisfied, it is possible to obtain bounds on the average treatment effect (Balke and Pearl (1997)). To point identify a treatment effect, one needs to make additional assumptions. Here, we assume (A4) monotonicity where the potential treatment is a monotonic function of the instrument values, or $d_{0ij} \leq d_{1ij}$. Assumption (A4) can be interpreted in terms of four subpopulations: compliers, always-takers, never-takers, and defiers (Angrist, Imbens and Rubin (1996)). Compliers are units which their treatment values follow their instrument values, or $d_{0ij} = 0, d_{1ij} = 1$. Always-takers always take the treatment regardless of their instrument values, or $d_{0ij} = d_{1ij} = 1$. Never-takers never take the treatment regardless of their instrument values, or $d_{0ij} = d_{1ij} = 0$. Defiers act against their instrument values, or $d_{0ij} = 1, d_{1ij} = 0$. Assumption (A4) then states that no defiers exist.

Let N_{CO} be the total number of compliers in the population. Under the IV assumptions (A1)–(A4), the CACE, formally defined as

$$\lambda = \frac{\sum_{i=1}^{I}(r_{1ij}^{(1)} - r_{0ij}^{(0)})I(d_{1ij}=1, d_{0ij}=0)}{\sum_{i=1}^{I}\sum_{j=1}^{2} d_{1ij} - d_{0ij}} = \frac{1}{N_{\mathrm{CO}}}\sum_{i=1}^{I}(r_{1ij}^{(1)} - r_{0ij}^{(0)})I(ij \text{ is a complier}),$$

can be identified from data by taking the ratio of the estimated ITT effect over the estimated compliance rate. In the context of matching and instrumental variables, Baiocchi et al. (2010) and Kang et al. (2016) proposed a test statistic to test the null $H_0 : \lambda = \lambda_0$ by using differences in the adjusted outcomes,

$$T(\lambda_0) = \frac{2}{I}\sum_{i=1}^{I}\sum_{j=1}^{2} Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - (1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij}), \tag{1}$$

along with an estimator for the variance of $T(\lambda_0)$,

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}\sum_{j=1}^{2}\bigl(Z_{ij}(R_{ij} - \lambda_0 D_{ij}) - (1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij}) - T(\lambda_0)\bigr)^2. \tag{2}$$

Under the null, Baiocchi et al. (2010) and Kang et al. (2016) showed that $\frac{T(\lambda_0)}{S(\lambda_0)}$ asymptotically follows a standard Normal distribution. For point estimation the same set of authors proposed a Hodges–Lehmann type estimator (Hodges and Lehmann (1963)) which involves solving λ in the equation $T(\lambda)/S(\lambda) = 0$. For a $1 - \alpha$ % confidence interval, the equation $T(\lambda)/S(\lambda) \leq z_{1-\alpha/2}$ is solved for λ, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard Normal distribution; see Kang et al. (2016) and Kang, Peck and Keele (2018) for details.

2.3. *Heterogeneous complier average causal effect* (*H-CACE*). We formally define the target estimand of interest in the paper, the heterogeneous treatment effect among compliers, or H-CACE. Formally, the H-CACE is defined as the CACE for a subgroup of compliers with a specific value of covariates,

$$\lambda(\mathbf{x}) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{1ij}^{(1)} - r_{0ij}^{(0)})I(d_{1ij}=1, d_{0ij}=0, \mathbf{X}_{ij}=\mathbf{x})}{\sum_{i=1}^{I}\sum_{j=1}^{2}(d_{1ij} - d_{0ij})I(\mathbf{X}_{ij}=\mathbf{x})}. \tag{3}$$

Because two units are assumed to have identical covariate values within each matched pair, $\lambda(\mathbf{x})$ can be rewritten as taking a subset of I matched pairs with identical covariates $\mathbf{x}$, say $s \subseteq \{1, \dots, I\}$,

$$\lambda_s = \frac{\sum_{i\in s}\sum_{j=1}^{2} r_{1sij}^{(d_{1sij})} - r_{0sij}^{(d_{0sij})}}{\sum_{i\in s}\sum_{j=1}^{2} d_{1sij} - d_{0sij}}.$$

Since each H-CACE λ_s has the same form as the original CACE, we can apply the test statistic in Section 2.2. Formally, consider the subset-specific hypothesis $H_{0s} : \lambda_s = \lambda_0$ against $H_{1s} : \lambda_s \neq \lambda_0$. We can use the test statistic (1) with variance (2) among the pairs specific to subset s.

Also, under assumptions (A1)–(A4), for a mutually exclusive and exhaustive grouping $\mathcal{G} = \{s_1, \dots, s_G\}$ of a set of pairs $s_g \subseteq \{1, \dots, I\}$ with at least one complier within each subgroup s_g, the original CACE is equal to a weighted version of H-CACE,

$$\lambda = \sum_{g=1}^{G} w_{s_g}\lambda_{s_g}, \quad w_{s_g} = \frac{\sum_{i\in s_g}\sum_{j=1}^{2} d_{1sij} - d_{0sij}}{N_{\mathrm{CO}}}.$$

An implication of this expression is that typical analysis of the CACE hides underlying effect heterogeneity. For example, suppose there are two subgroups defined by a binary covariate, say male or female, and consider two scenarios. In the first scenario, among compliers, 80% are male, and 20% are female. Also, the H-CACE of male is 1.25, and the H-CACE of female is 0. In the second scenario the male/female complier proportions remain the same, but the H-CACE of male is now 1.5, and the H-CACE of female is -1. In both scenarios the CACE is 1. But in the second scenario, females have a negative treatment effect. By only studying the CACE, as is typical in practice, variations in the treatment effects, defined by H-CACEs, would have been masked. The next section presents a way to unwrap the CACE and discover novel H-CACEs.

2.4. *Discovering and testing novel H-CACE.* A naive approach to finding and testing novel H-CACE would be to exhaustively test every H-CACE for every subset of matched pairs and gradually aggregate them, based on their covariate similarities with appropriate statistical tests. However, this procedure will not only lead to false discoveries, but it will also be grossly underpowered.

Instead, based on the work by Hsu et al. (2015), we propose to use exploratory machine learning methods, such as CART, to discover and aggregate matched pairs into subgroups with similar treatment effects, formulating grouping $\mathcal{G}$. We will then use closed testing to test effect heterogeneity, defined by these groups, while strongly controlling the familywise error rate; see Algorithm 1 for details.

We explain in some detail the key steps in Algorithm 1. First, the specification of the null value λ_0 is for testing the sharp null of the form $H_0 : r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})} = \lambda_0(d_{1ij} - d_{0ij})$; this sharp null implies the "weak" or composite null $H_0 : \lambda = \lambda_0$ (Baiocchi et al. (2010)). Setting $\lambda_0 = 0$ would test whether the H-CACE is zero or not and is the typical choice in most applications, unless other null values are of scientific interest. Second, under the sharp null the absolute value of the difference in adjusted outcomes between pairs, $|Y_i| = |(Z_{i1} - Z_{i2})(R_{i1} - \lambda_0 D_{i1} - (R_{i2} - \lambda_0 D_{i2}))|$, obscures the instrument assignment vector making $|Y_i|$ a function of $\mathcal{F}$ only, a fixed (and unknown) quantity. In contrast, Y_i is a function of both $\mathcal{F}$ and $\mathbf{Z}$. Consequently, conditional on $\mathcal{F}$, building a CART tree based on $|Y_i|$ as the response and $\mathbf{X}_i$ as the explanatory variables, does not affect the distribution of $\mathbf{Z}$. The distribution of $\mathbf{Z}$ within each pair remains 1/2, as stated in assumption (A3), and is a key ingredient to achieve familywise error rate control for downstream inference; see our discussion on honest inference below.

Third, Algorithm 1 applies closed testing, a multiple inference procedure by Marcus, Peritz and Gabriel (1976), to test for multiple hypotheses about H-CACEs generated by CART's grouping $\mathcal{G} = \{s_1, \dots, s_G\}$. Broadly speaking, closed testing will test sharp null hypotheses, defined by every parent and child node of the estimated tree from CART, and reject/accept these hypotheses while controlling for multiple testing issues; see Section 4.4 and Figure 5 for visualizations. A bit more formally, closed testing will test the global sharp null hypothesis $H_0 : r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})} = \lambda_0(d_{1ij} - d_{0ij})$ and subsequent subset-specific hypotheses $H_{0\mathcal{L}} : r_{1s_gij}^{(d_{1s_gij})} - r_{0s_gij}^{(d_{0s_gij})} = \lambda_0(d_{1s_gij} - d_{0s_gij})$ for all $g \in \mathcal{L}$, where $\mathcal{L}$ is a subset of the G groups formed by CART. We note that the difference between the global null and the subset-specific nulls is only in the pairs under consideration; all the nulls use the test statistics introduced in Section 2.2. Also, the subset-specific hypotheses imply $H_{0\mathcal{L}} : \lambda_s = \lambda_0$ for $s = \bigcup_{g \in \mathcal{L}} s_g$. Closed testing would only reject the subset-specific hypotheses $H_{0\mathcal{L}}$ if all of the p-values from superset hypotheses $H_{0\mathcal{L}'}$, $\mathcal{L} \subseteq \mathcal{L}'$, are less than α.

As mentioned earlier, the key step of using $|Y_i|$ in CART allows for both discovery and downstream honest testing of H-CACEs via closed testing; again, honesty refers to control

```
Given : Observed outcome $R$, binary instrument $Z$, exposure $D$, covariates $X$, null
        value $\lambda_0$ for testing, and desired familywise error rate $\alpha$
1 Pair match on observed covariates.
2 Calculate absolute value of pairwise differences for each matched pair
      $$|Y_i| = \big|(Z_{i1} - Z_{i2})\big(R_{i1} - \lambda_0 D_{i1} - (R_{i2} - \lambda_0 D_{i2})\big)\big|$$
3 Construct mutually exclusive and exhaustive grouping using CART. Here, CART takes
  $|Y_i|$ as the outcome and $\mathbf{X}_i$ from each matched pair as the predictors. CART outputs a
  partition of covariates, which we use to define $\mathcal{G} = \{s_1, \dots, s_G\}$ and, consequently,
  H-CACEs.
4 Run closed testing (Marcus, Peritz and Gabriel (1976)) to test statistical significance of
  H-CACEs for every subset $\mathcal{L} \subseteq \{1, \dots, G\}$ of $G$ groups where each subset defines the
  null hypothesis of the form $H_{0\mathcal{L}} : r_{1ij}^{(d_{1ij})} - r_{0ij}^{(d_{0ij})} = \lambda_0(d_{1ij} - d_{0ij})$ for all $g \in \mathcal{L}$.
  Formally, run
      for $\mathcal{L} \subseteq \{1, \dots, G\}$ do
          if $H_{0\mathcal{L}}$ has not been accepted then
              Calculate $T_s(\lambda_0)$ and $S_s(\lambda_0)$ for $s = \bigcup_{g \in \mathcal{L}} s_g$
              if $|\frac{T_s(\lambda_0)}{S_s(\lambda_0)}| \leq z_{1-\alpha/2}$ then
                  Accept the null hypothesis $H_{0\mathcal{K}} : \lambda_{\mathcal{K}} = \lambda_0$ for all $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \dots, \mathcal{G}\}$
              end
              else
                  Reject $H_{0\mathcal{L}}$
              end
          end
      end

Output: Estimated and inferential quantities for H-CACEs (e.g., effect size, confidence
        interval, $p$-value) and novel H-CACEs from closed testing.
```

Algorithm 1: Proposed method to discover and test effect heterogeneity in IV with matching

of the familywise error rate at level α when testing multiple hypotheses about H-CACEs that were discovered by data. Because $|Y_i|$ is not a function of $\mathbf{Z}$, the original distribution of $\mathbf{Z}$ is preserved, and we can use the standard randomization inference null distribution to honestly test each H-CACE discovered by CART. In fact, as noted in Hsu et al. (2015), this honesty property is preserved for any supervised machine-learning algorithm that forms groups based on $\mathbf{X}$ and $|Y|$ as well as subsequent visual heuristics to check the algorithms' performance. Also, in recent work on estimating heterogeneous causal effects (Athey, Tibshirani and Wager (2019), Chernozhukov et al. (2018), Park and Kang (2020)), the notion of "honest" inference is often tied to sample splitting, where one subsample is used to discover different subgroups or to estimate nuisance parameters and the other subgroup is used to test the causal effect. Our approach does not have to use sample splitting to obtain honest inference, and Proposition 1 shows this principle formally; Web Appendix A (Johnson, Cao and Kang (2022)) shows this principle numerically.

PROPOSITION 1 (Familywise error rate control of Algorithm 1). *Under the sharp null hypotheses $H_{0\mathcal{L}}$ in Algorithm 1, the conditional probability, given $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$, that the algorithm makes at least one false rejection of the set of hypotheses is, at most, α.*

We now discuss some important limitations of Proposition 1 and the proposed algorithm. First, our algorithm's guarantee on controlling the familywise error rate is only for testing sharp nulls. As noted in Section 2, page 289 of Rosenbaum (2002a), testing for sharp nulls does not necessarily imply that the true data generating process always follows the sharp null, and, as such, the proposition makes no claims about how the true data generating process actually looks like. Having said that, the limitation of testing a sharp null vs. a weak null has been discussed extensively; see Sections 3 and 4 of Rosenbaum (2002b), Ding (2017), Fogarty (2018, 2020). But a recent work by Fogarty et al. (2021) has shown that testing the sharp null, based on our test, is an asymptotically valid test for the weak null; see Remark 1 of their Proposition 1. This suggests that the guarantees from Proposition 1 will likely hold, even if we are testing weaker nulls with our algorithm. Second, a price we pay for using $|Y_i|$ to achieve honest inference is that we collapse the sign of the effect, and, therefore, CART treats subgroups with positive or negative effects equally. This is potentially problematic in settings where two different covariate values lead to identical effects (in magnitude) but different in signs; see Hsu, Small and Rosenbaum (2013) for additional discussions and Web Appendix D (Johnson, Cao and Kang (2022)) for a numerical illustration. For our Medicaid example, if there is a partition of the covariates that leads to two identical H-CACEs in magnitude, but different in signs, our algorithm may not be able to detect the two subgroups. But, since using Medicaid is unlikely to be harmful, we do not believe this will be a significant concern in our example, especially compared to the alternatives of not obtaining honest inference. Third, Proposition 1 does not describe the algorithm's statistical power to detect effect heterogeneity. The next section uses a simulation study to address power and other factors influencing discovery of H-CACEs.

3. Simulations.

We conduct a simulation study to measure the performance of the proposed algorithm in two ways: (1) statistical power to test H-CACEs and (2) recovering effect modifiers. Throughout the simulation study we vary the the compliance rate because prior works have shown that performance of IV methods depends heavily on the compliance rate or, more generally, on the instrument's association to the treatment (i.e., instrument strength). In particular, problems can arise when the compliance rate is low; see Staiger and Stock (1997), Stock, Wright and Yogo (2002), and references therein for more details.

Following Hsu et al. (2015), each simulation setting fixes the potential outcomes $r_{0ij}^{(d_{0ij})}$ and $r_{1ij}^{(d_{1ij})}$, potential treatments d_{0ij} and d_{1ij}, and covariates $\mathbf{X}_{ij}$ of each unit j within each of the $I = 2000$ pairs. There are six preinstrument covariates, each generated from independent Bernoulli trials with 0.5 probability of success. At most, two covariates, x_1 and x_2, modify the treatment effect. That is, H-CACEs, defined by $\lambda(x_1, \dots, x_6)$ in equation (3), depend on, at most, two covariates, x_1 and x_2. Also, because both x_1 and x_2 are binary, there are, at most, four different H-CACEs, defined by different combinations of binary variables λ_{00}, λ_{01}, λ_{10}, and λ_{11}; for notational simplicity we use $\lambda_{x_1 x_2}$ to represent equation (3). Similar to the design of the OHIE, the data is generated under the assumption of one-sided compliance. This means that, for every unit, the potential treatment having not received the instrument is 0, $d_{0ij} = 0$. The potential treatment, having received the instrument, d_{1ij}, is then a Bernoulli trial with success rate π; π is also the compliance rate. In Web Appendix C (Johnson, Cao and Kang (2022)) we consider the setting in which the compliance rate may depend on x_1 and x_2, say via $\pi_{x_1 x_2}$. Finally, the potential outcomes, having not received the instrument $r_{0ij}^{(d_{0ij})}$, are from a standard normal distribution $r_{0ij}^{(d_{0ij})} \sim N(0,1)$, and the potential outcomes, having received the instrument $r_{1ij}^{(d_{1ij})}$, are a function of the H-CACE $r_{1ij}^{(d_{1ij})} = r_{0ij}^{(d_{0ij})} + d_{1ij}\lambda_{x_1x_2}$. Once all the potential treatment and outcomes are generated, the observed treatment and outcome are determined, based on the value of the instrument and SUTVA. Finally, the regression tree in Algorithm 1 is estimated in R using the package *rpart*, version 4.1-15 (Therneau, Atkinson and Ripley (2015)). Unless specified otherwise, we use a complexity parameter of 0.005 (half of the default setting) and use defaults for the rest of *rpart*'s parameters. Our proposed method is referred to as "H-CACE" in the results below.

For comparison, we also apply a recent method by Bargagli-Stoffi, De-Witte and Gnecco (2019) to discover and test H-CACEs. Briefly, their method, which we refer to as "BCF-IV" in the results below, utilizes modern tree-based methods (Athey and Imbens (2015), Hahn, Murray and Carvalho (2020)) to estimate heterogeneous intent-to-treat (ITT) effects and suggests different subpopulations of interest; we remark that, unlike our proposal, their method does not use matching and uses the original, untransformed R_{ij} inside the tree fitting step. Then, for each subpopulation the method estimates and tests its H-CACE, using the two-stage least square estimator. We use the *bcf_iv* function available on the authors' Github repository and use the default parameters of *rpart* and *bcf* (Hahn, Murray and Carvalho (2020)).

3.1. *Statistical power.* To measure a method's statistical power when the subgroup-specific null hypotheses are not specified a priori, we divide the number of false null hypotheses rejected by the total number of false null hypotheses suggested by the method. We refer to this rate as the true discovery rate; note that, if the number of false nulls being suggested is fixed, the true discovery rate is one minus the proportion of false null hypotheses retained.

We compute the true discovery rate at varying levels of instrument strength and four heterogeneous treatment settings: (a) No Heterogeneity, (b) Slight Heterogeneity, (c) Strong Heterogeneity, and (d) Complex Heterogeneity. In setting (a) there are no effect modifiers, resulting in one subgroup with equal treatment effects, $\lambda_{00} = \lambda_{01} = \lambda_{10} = \lambda_{11} = 0.5$. In setting (b) there is one effect modifier x_1, resulting in two subgroups with similar but different treatment effects, $\lambda_{00} = \lambda_{01} = 0.7$ and $\lambda_{10} = \lambda_{11} = 0.3$. In setting (c) there is one effect modifier x_1, resulting in two subgroups with dissimilar treatment effects, $\lambda_{00} = \lambda_{01} = 0.9$ and $\lambda_{10} = \lambda_{11} = 0.1$. And in setting (d) there are two effect modifiers x_1 and x_2, resulting in three subgroups, one with a strong effect, two with no effects, and the last group with the average effect, $\lambda_{00} = 1.5$, $\lambda_{01} = \lambda_{10} = 0$ and $\lambda_{11} = 0.5$. In all four settings the overall complier average causal effect is $\lambda = 0.5$.

We repeat the simulation 1000 times for each treatment heterogeneity and instrument strength combination. We remark that the null hypothesis is that of no treatment effect (i.e., $\lambda_0 = 0$), and only the hypotheses consisting of pairs with $\lambda_{x_1x_2} = 0$ are true null hypotheses.

Figure 1 shows the true discovery rate under four treatment heterogeneity settings. We see that as the compliance rate (i.e., instrument strength) increases, the true discovery rate of our method grows across all settings. In particular, our approach has the best power in the region where the compliance rate is low, roughly under 40%. Even when the compliance rate is high, we see that BCF-IV generally has lower power than our method across different heterogeneity settings, especially in the No Heterogeneity, Slight Heterogeneity, and Strong Heterogeneity settings. In the Complex Heterogeneity setting we see the true discovery rate is rather similar between the two methods. This is because this setting has the largest discrepancies in H-CACEs between subgroups, and, thus, it is easy for CART to correctly split on the covariates x_1 and x_2. Further, with the large magnitudes of H-CACEs in this setting, the null hypothesis tests are more easily rejected in favor of the alternative.

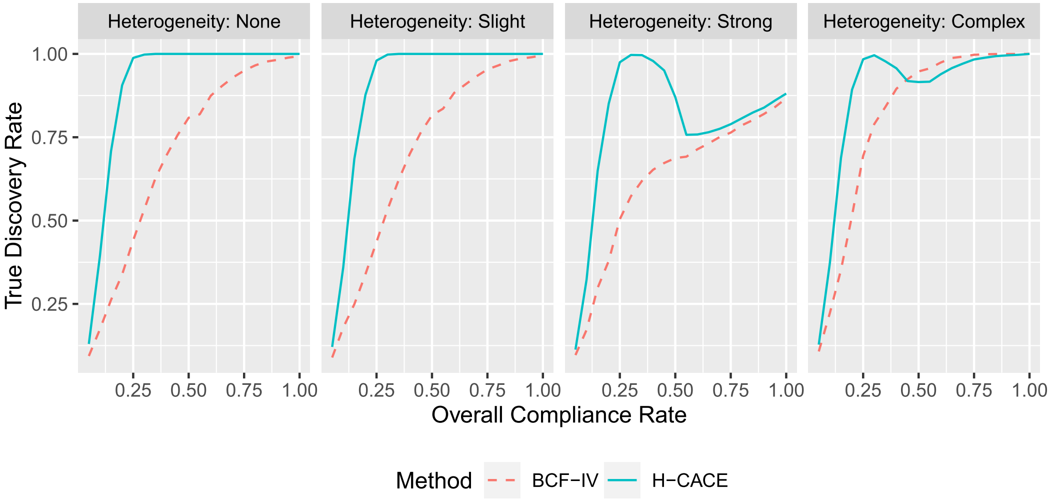


FIG. 1. *True discovery rate as a function of the compliance rate and heterogeneity settings. The dashed and solid lines denote the BCF-IV procedure and our proposed algorithm, respectively.*

We also take a moment to explain a counter-intuitive dip in our method's true discovery rate under the Strong and Complex Heterogeneity settings in Figure 1. Briefly, this drop in the true discovery rate is due to the formation of leaves with smaller treatment effects. As the compliance rate becomes large, these small effects begin to be get suggested by CART. But the power to reject the null in favor of these small effects are small, and the overall true discovery rate dips briefly. However, as the compliance rate reaches one, we see the true discovery rate of our method begin to climb again; Web Appendix E (Johnson, Cao and Kang (2022)) contains additional details surrounding this phenomena.

Another interpretation of this dip reflects a limitation of the true discovery rate as a metric for statistical power in certain settings of multiple testing with hypotheses adaptively generated by tree-based algorithms. Specifically, because the true discovery rate only considers hypotheses generated by the tree, if a tree were to not split (or rarely split) and the overall effect is strong, there would only be one hypothesis in the denominator of the true discovery rate, and the lone hypothesis would likely be rejected, leading to a true discovery rate of one. The Strong and Complex Heterogeneity settings under low compliance rates, especially before the dip, is a reflection of this phenomena where our tree fails to split, resulting in one single hypothesis that is eventually rejected; see Web Appendix E (Johnson, Cao and Kang (2022)) for additional discussions. Nevertheless, as Figure 1 shows, our method consistently shows a higher true discovery rate, compared to BCF-IV, across many settings, and our method is more likely to discover true nonzero effects than BCF-IV.

3.2. *False positive rate and F-score.* We also assess our algorithm's ability to predict effect modifiers from $\mathbf{X}_{ij}$. Specifically, we say that a method predicts a variable to be an effect modifier when the tree splits on the variable and rejects one of the hypotheses of the split's children. In contrast, if either: (a) the tree splits on a variable, but none of the hypotheses defined by the split is rejected, or (b) the tree does not split on the variable, the variable is not predicted to be an effect modifier. For example, for a given tree that splits only on covariate x_1, if at least one of the subgroup-specific null hypotheses is rejected, x_1 is predicted to be an effect modifier. Instead, if none of the subgroup-specific null hypotheses are rejected, then x_1 as well as other variables not selected by the tree are not predicted to be effect modifiers. We then use the F-score and the false positive rate (FPR) common in the classification literature to

TABLE 1
Binary classification table for effect modifiers

	True condition	
Method's prediction	Variable is an effect modifier	Variable is not an effect modifier
Predicted as effect modifier	True Positive (TP)	False Positive (FP)
Not predicted as effect modifier	False Negative (FN)	True Negative (TN)

measure a method's ability to correctly predict effect modifiers. The F-score is the harmonic mean of recall and precision, or, alternatively,

$$F = \frac{\mathrm{TP}}{\mathrm{TP} + 0.5(\mathrm{FP} + \mathrm{FN})},$$

where TP stands for true positives, FP stands for false positives and FN stands for false negatives; see Table 1 for details. The F-score ranges from zero to one with a value closer to one implying greater accuracy. The FPR is defined as $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ and ranges from zero to one, with a value close to zero being preferred.

We use the same four heterogeneity settings of: (a) No Heterogeneity, (b) Slight Heterogeneity, (c) Strong Heterogeneity, and (d) Complex Heterogeneity. Figure 2 shows the results of the F-score and FPR from our proposed algorithm and BCF-IV. Across the four settings our proposal has a false positive rate of nearly zero, never falsely declaring a variable to be an effect modifier. In contrast, BCF-IV has a larger false positive rate, declaring variables to be effect modifiers when they do not actually modify the compliers' effect. For example, in setting (a), without any effect modifiers, BCF-IV has a false positive rate hovering above 50%, whereas our method has a false positive rate of 0%. In other words, BCF-IV falsely declared at least one of the six covariates as an effect modifier roughly 50% of the time, whereas our method never declared any of the six covariates as effect modifiers.

However, our algorithm's F-score is generally smaller than that from BCF-IV, unless the compliance rate is high and the effect heterogeneity is strong. In particular, when the compli-

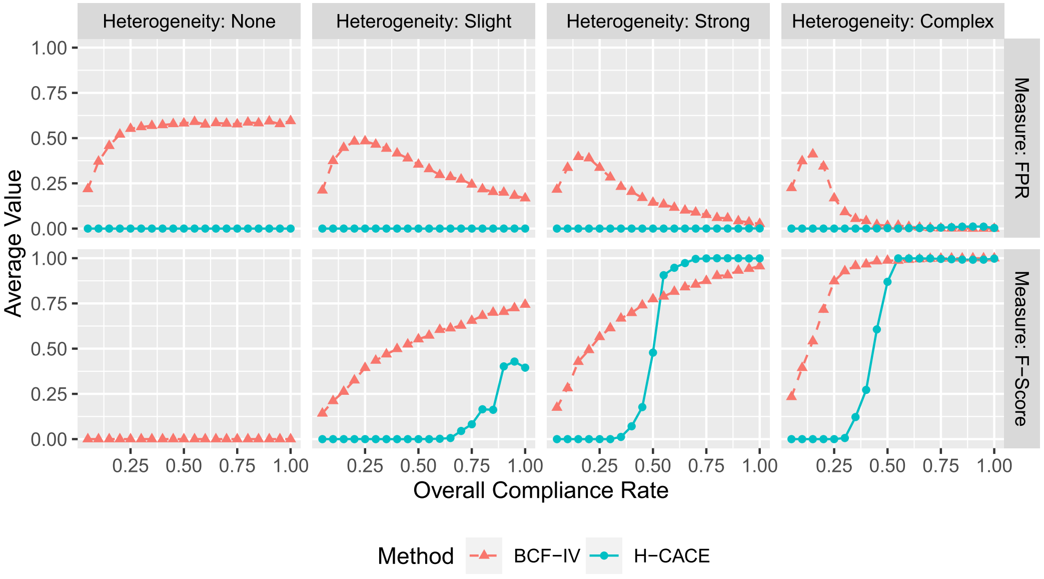


FIG. 2. *F-score and false positive rate as a function of compliance rate and heterogeneity settings. The solid lines with circles denote our proposed algorithm, and the dashed lines with triangles denote BCF-IV.*

ance rate is roughly under 50% or if two subgroups have similar effect sizes, our method cannot predict the effect modifiers as well as BCF-IV. But, when the compliance rate is above 50% and the effect heterogeneity is strong, our algorithm has a similar F-score as BCF-IV. Overall, the low F-score is a price that our algorithm pays for making sure that the FPR is small. In contrast, BCF-IV has a higher F-score but pays a price with a high FPR.

In the Supplementary Material Web Appendix C, D, and F (Johnson, Cao and Kang (2022)), we conduct additional simulation studies where we: (i) vary the compliance rate by covariates, (ii) allow H-CACEs to be equal in magnitude but opposite in direction to measure the effect of using $|Y_i|$ in our algorithm, and (iii) demonstrate the two methods in a simulation that closely resembles the data from the OHIE, where there are more than two effect modifiers. To summarize the results, for (i) and (iii) the story is very similar to what's presented here, where our method has high true discovery rate, low FPR, and F-score, compared to those from BCF-IV. For (ii), as expected, we find that our method has a low true discovery rate, FPR, and F-score. But, as soon as the magnitudes of the H-CACEs are dissimilar, our method returns to the case presented here.

3.3. *Takeaways from the simulation study.* Overall, the simulation study shows that our algorithm has large statistical power and low false positive rates across all settings. In contrast, the BCF-IV algorithm has low power and produces large FPRs, especially when no effect heterogeneity exists in the data; in other words, BCF-IV often falsely declares a variable to be an effect modifier. But our algorithm generally has a low F-score, compared to that from BCF-IV, except in regimes where the effect heterogeneity is strong and the compliance rate is high.

We remark that the simulations studies do not encapsulate every type of effect heterogeneity, and it is possible that our method may suffer in certain settings. In particular, as discussed above, because our method tends to be conservative in predicting effect modifiers in order to guard against discovering spurious heterogeneity, we suspect that if there are many effect modifiers compared to spurious effect modifiers, our method may not be able to detect all of the effect modifiers. This suspected degradation in performance was not observed when we had five effect modifiers among 15 potential effect modifiers in our simulation study that mimicked the data from the OHIE. But further investigation is warranted, especially if the number of effect modifiers and/or the number of covariates is high dimensional.

We also remark that the simulation results in Sections 3.1 and 3.2 do not necessarily contradict each other. Roughly speaking, the result in Section 3.1 concerns the ability for algorithms to have high *statistical power*, whereas the result in Section 3.2 concerns the ability for algorithms to *predict* variables. An algorithm like BCF-IV could liberally predict many effect modifiers, generally leading to a high F-score, but a high FPR. Also, the power to test the nulls, suggested by the predicted effect modifiers, could be low since the selected variables will define many (likely small) subgroups. In contrast, an algorithm like ours could conservatively predict effect modifiers, leading to a small F-score but a low FPR. Also, the power to test the nulls suggested by the predicted variables could be high since most of the selected variables will be effect modifiers. In short, our method is somewhat cautious but certain, whereas BCF-IV is optimistic but somewhat error-prone.

4. Analysis of the Oregon health insurance experiment.

4.1. *Data description.* We use our method to analyze the heterogeneous effects of Medicaid on the number of days an individual's physical or mental health prevented their usual activities in the past month. In brief, the OHIE collected administrative data on hospital discharges, credit reports and mortality, survey data on health care utilization, financial strain,

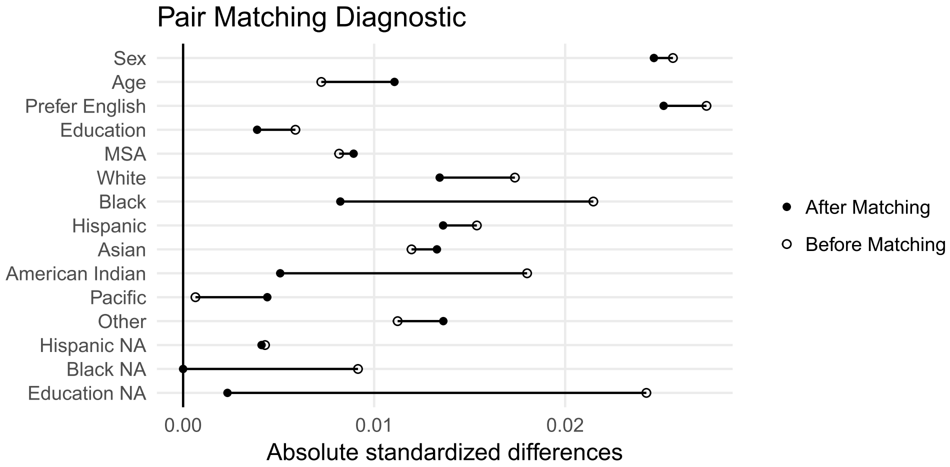


FIG. 3. *Covariate balance as measured by difference in means of the covariates between the treated and control groups, before and after matching.*

overall health, and prerandomization demographic data. There were 11,808 lottery winners and 11,933 lottery losers in the publicly available survey data for a total sample size of 23,741 individuals; see Finkelstein et al. (2012) for details.

We matched on the following demographic, prerandomization variables recorded by Finkelstein et al. (2012): sex, age, whether they preferred English materials when signing up for the lottery, whether they lived in a metropolitan statistical area (MSA), their education level (less than high school, high school diploma or General Educational Development (GED), vocational or two-year degree, four-year college degree or more), and self-identified race (as the individual reported in the survey). Since some of the covariates had missing data, namely, selfidentifying as Hispanic or Black and their level of education, we also matched on indicators of their missingness; see Section 9.4 of Rosenbaum (2010) for details. We used the R package *bigmatch*, version 0.6.1 (Yu (2019)) with an optimal caliper and a robust rank-based Mahalanobis distance to generate our optimal pair match. Figure 3 shows covariate balance before and after matching.

For the majority of covariates, the matching algorithm did little to change the absolute standard differences between lottery winners and losers. This is not surprising, given that the lottery was randomized. However, the indicator for missingness in education, self-identified American Indian, and Black were made to be more similar after matching. An absolute standardized difference of 0.25 is deemed acceptable (Rubin (2001), Stuart (2010)), which our covariates satisfied after matching.

4.2. *Instrument validity.* Before we present the results of our analysis using the proposed method, we discuss the plausibility of the lottery as an instrument. The lottery is randomized which ensures that the instrument is unrelated to unmeasured confounders and satisfying (A3). Winning the lottery, on average, increased enrollment of Medicaid by 30% (Finkelstein et al. (2012)), satisfying (A1). Assumption (A4), in the context of the OHIE, states that there are no individuals who defy the lottery assignment to take (or not take) Medicaid if they lost (or won) the lottery. This is guaranteed by the design of the lottery, since an individual who lost the lottery cannot have access to Medicaid. However, we remark that Finkelstein et al. (2012) measured the treatment as whether or not an individual has ever had Medicaid during the study, and a few individuals were already enrolled in Medicaid before the lottery winners were announced. Finally, assumption (A2) is the only assumption that could potentially be violated since individuals were not blind to their lottery results. Theoretically, this allowed

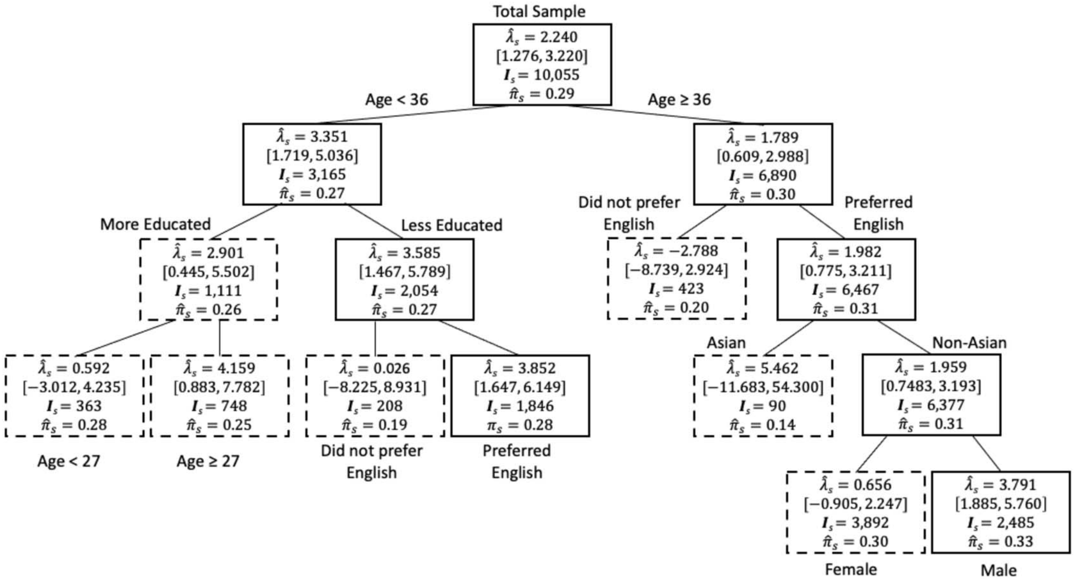


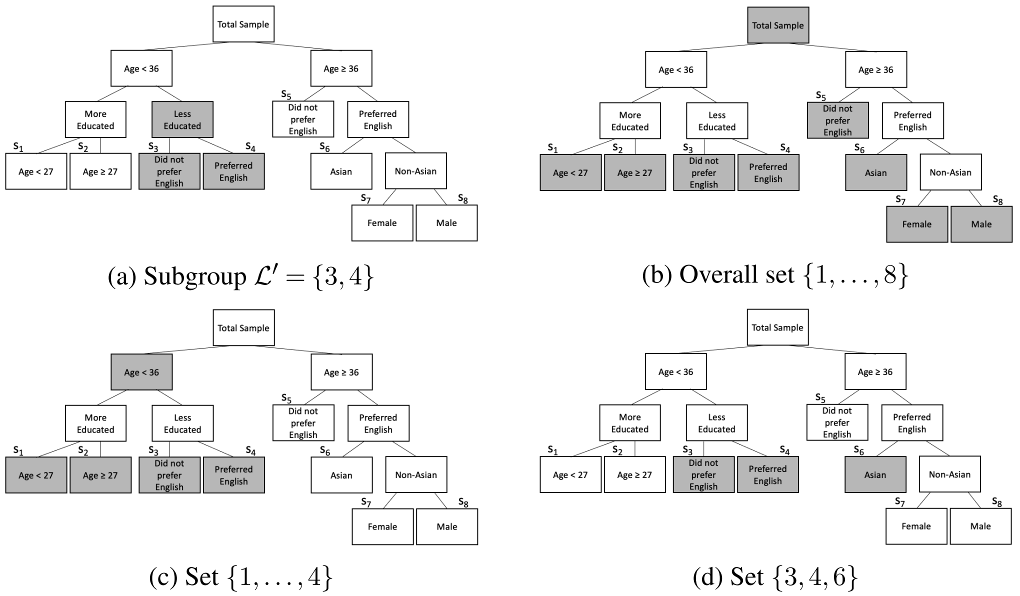
FIG. 4. *Results of our proposed method on the effect of enrolling in Medicaid on the number of days physical or mental health did not prevent usual activities. Here, less educated refers to pairs with, at most, a high school diploma or GED and more educated refers to pairs with a higher education. Also, positive effects are beneficial to individuals. Solid lined boxes denote hypothesis tests that were rejected, and dashed lined boxes denote hypotheses that were retained by closed testing. Within each box the subgroup-specific estimated H-CACE* $\hat{\lambda}_s$, *its* 95% *confidence interval, sample size of pairs* I_s, *and the estimated compliance rate* $\hat{\pi}_s$ *are provided.*

lottery losers to seek other health insurance or lottery winners to make less healthy decisions since they're now able to be insured. These changes in an individual's behavior could affect his/her outcome regardless of his/her treatment and thus may violate (A2).

4.3. *Analysis and results.* We run Algorithm 1 and present the results in Figure 4. We remark that we used *rpart* in R with a complexity parameter of 0 and maximum depth of 4. The depth of the tree was chosen by forming trees of larger depth and then pruning back until a more interpretable tree was obtained. For each node of the CART, we tested whether or not there is an effect of enrolling in Medicaid $H_{0s} : \lambda_s = 0$. In Figure 4 a solid lined box denotes a null hypothesis that was rejected, and a dashed lined box denotes a null hypothesis that was retained, both by the closed testing procedure. Each node contains its estimated H-CACE $\hat{\lambda}_s$, 95% confidence interval, the number of pairs I_s, and the estimated compliance rate $\hat{\pi}_s$. Here, a positive H-CACE implies a decrease in the number of days where the individual's physical and mental health prevented them from their usual activities, and a negative value implies an increase; in short, positive effects are beneficial to individuals. Also, some nodes imply a significant effect of Medicaid at level 0.05 but are enclosed in a dashed lined box. This is due to the closed testing procedure; an intersection of hypotheses containing the node in question was not rejected, and so any hypotheses in this intersection could not be rejected.

From Figure 4 we can see evidence of heterogeneous treatment effects among the complier population. Specifically, Medicaid had a strong effect: (1) among complying non-Asian men over the age of 36, who prefer English, as well as (2) complying individuals younger than 36, who prefer English, and do not have more than a high school diploma or GED. Interestingly, among non-Asians over the age of 36 and who prefer English, females did not benefit from Medicaid as much as males, even though the female subgroup was larger than the male subgroup and the compliance rates between the two subgroups were similar.

More generally, while there is some variation in the compliance rates between groups, most of them are minor and hover between 25% to 30%. The minor variation suggests that, while



(a) Subgroup $\mathcal{L}' = \{3, 4\}$

(b) Overall set $\{1, \dots, 8\}$

(c) Set $\{1, \dots, 4\}$

(d) Set $\{3, 4, 6\}$

FIG. 5. *Illustration of closed testing to test the null hypothesis H_{0s_4} for all $j = 1, 2$ and $i \in s_4$. Each subplot highlights subsets required to be tested and rejected as part of closed testing.*

some subgroups are more likely to be compliers than others, most of the effect heterogeneity is likely driven by the variation in how the treatment differentially changes the response across subgroups; a bit more formally, most of the effect heterogeneity is likely arising from the numerator of the H-CACE rather than the denominator of the H-CACE.

4.4. *An example of closed testing.* To better illustrate the closed testing portion of Algorithm 1, we walk through an example of the testing procedure, based on the OHIE. As seen in Figure 4, CART produced a tree with $G = 8$ leaves. Now, consider testing whether there is evidence of a heterogeneous effect of Medicaid for young individuals who prefer English and have at most a high school diploma or GED, that is, node s_4 in Figure 5 and $\mathcal{L} = \{4\}$ using Algorithm 1's notation. The null hypothesis of interest would be H_{0s_4}, for all $j = 1, 2$ and $i \in s_4$. We then test and reject all of the hypothesis tests containing group s_4. For example, we need to test the null hypothesis concerning the ancestor of s_4, say the subgroup of individuals who are younger than 36 and have, at most, a high school diploma or GED, denoted as $\mathcal{L}' = \{3, 4\}$; see part (a) of Figure 5. Additionally, we need to test and reject all of the supersets containing $\mathcal{L}'$ which include but are not limited to the overall set $\{1, \dots, 8\}$, $\{1, \dots, 4\}$, and $\{3, 4, 6\}$. If every superset hypothesis and H_{0s_4} are rejected at level α, we can declare the effect in node s_4 to be significant, and by Proposition 1 the familywise error rate is controlled at α. Repeating this process for every node in the tree will give the results in Figure 4.

5. Discussion. In this paper we propose a method, based on matching, to detect effect heterogeneity using an instrument. Under the usual IV assumptions our method discovers and tests heterogeneity in the complier average treatment effect by combining matching, CART, and closed testing, all without the need to do sample splitting. The latter is achieved by taking the absolute value of the adjusted pairwise differences to conceal the instrument assignment, and this allows our proposed method to control the familywise error rate. We also conducted a simulation study to examine the performance of our method and compared it to a recent method referred to as BCF-IV. Our method was then used to study the effect of

Medicaid on the number of days an individual's physical or mental health did not prevent their usual activities where we used the lottery selection as an instrument. We found that Medicaid benefited complying, older, non-Asian men who selected English materials at lottery sign-up and for complying, younger, less educated individuals who selected English materials at lottery sign-up.

We conclude by making some recommendations about how to properly use our algorithm in practice, especially in light of existing approaches. First, as explained in the Introduction, when there is noncompliance, exploring heterogeneity in the ITT alone with existing methods may provide an incomplete picture of the nature of the treatment effect. Relatedly, in settings where unmeasured confounding is unavoidable, our method based on an instrument is a promising way to discover and test effect heterogeneity.

Second, as alluded to in Section 3.3, the simulation results suggest that our algorithm tends to be conservative in discovering novel effect modifiers, reporting effect modifiers only if there is strong evidence for heterogeneity, and minimizing prediction of spurious effect modifiers. In other words, investigators can be reasonably confident that effect heterogeneity exists among the variables declared by our algorithm as "real" effect modifiers. But those variables that are not predicted by our algorithm may also be true effect modifiers, and in such cases, investigators may need additional samples to detect them using our method. In contrast, BCF-IV tends to be anticonservative, reporting more effect modifiers, some of which may be spurious effect modifiers. While this may be advantageous in situations where there is slight effect heterogeneity or where exploration for effect heterogeneity is encouraged, investigators may not feel as confident about whether the detected effect heterogeneity truly exists.

Third, how our method performs in settings with potentially high dimensional effect modifiers is not fully understood. In particular, while our method performed well when the structure of effect heterogeneity grew more complex or when the number of effect modifiers were five out of 15 potential effect modifiers, the simulations did not consider the setting of moderate to high dimensional effect modifiers, and future research is warranted.

Fourth, most recent approaches on effect heterogeneity, notably Chernozhukov et al. (2018), utilize sample splitting to achieve honest inference (i.e., type I error rate control), whereas our method uses absolute value of matched pairs to achieve it; note that both methods theoretically allow for a large class of machine learning methods to detect heterogeneous treatment effects, even though ours focused on CART for its simplicity and interpretability. While our method uses the full sample for both discovery and honest testing, compared to those based on sample splitting, one of the caveats of our method is that our method may not be able to detect subgroups with identical effect sizes but in opposite signs. Overall, every algorithm for effect heterogeneity carries some trade-offs, and we urge investigators to understand their strengths and limitations to solidify and strengthen causal conclusions about effect heterogeneity in IV studies.

Acknowledgments. We thank the Associate Editor and two anonymous referees for helpful comments.

Funding. The first author was supported by NIH Trainee Grant T32 LM012413. The third author was supported in part by NSF Grant DMS-1811414.

SUPPLEMENTARY MATERIAL

Supplement to "Detecting heterogeneous treatment effects with instrumental variables and application to the Oregon health insurance experiment" (DOI: 10.1214/21-AOAS1535SUPPA; .pdf). The supplementary materials contain additional simulation studies and the proof of Proposition 1. Web Appendix A numerically demonstrates the honest simultaneous discovery and inference of Algorithm 1. Web Appendix B details the proof of Proposition 1. Web Appendix C is a simulation study demonstrating the performance of our algorithm when heterogeneity exists in both the treatment effect and compliance rate. Web Appendix D is a simulation study demonstrating the performance of the algorithm when effects are equal in magnitude but opposite in direction. Web Appendix E provides additional details surrounding the counter-intuitive dip in the true discovery rate observed in the simulations and drawbacks of the true discovery rate as a measure of statistical power. Web Appendix F is a simulation study based off the OHIE to demonstrate the performance of the algorithm under varying treatment effect magnitudes.

R code (DOI: 10.1214/21-AOAS1535SUPPB; .zip). R code used to conduct simulation studies and carry out analyses on the publicly available OHIE data.

REFERENCES

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. MR1960380 https://doi.org/10.1016/S0304-4076(02)00201-4

AI, C. and CHEN, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica* **71** 1795–1843. MR2015420 https://doi.org/10.1111/1468-0262.00470

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.

ATHEY, S. and IMBENS, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. Available at arXiv:1504.01132v1 [stat.ML].

ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582 https://doi.org/10.1002/sim.6128

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550 https://doi.org/10.1198/jasa.2010.ap09490

BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.

BARGAGLI-STOFFI, F. J., DE-WITTE, K. and GNECCO, G. (2019). Heterogeneous causal effects with imperfect compliance: A novel Bayesian machine learning approach. Available at arXiv:1905.12707 [stat.ME].

BARGAGLI-STOFFI, F. J. and GNECCO, G. (2018). Estimating heterogeneous causal effects in the presence of irregular assignment mechanisms. In 2018 *IEEE 5th International Conference on Data Science and Advanced Analytics* (*DSAA*) 1–10. IEEE, New York.

BLUNDELL, R., CHEN, X. and KRISTENSEN, D. (2007). Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica* **75** 1613–1669. MR2351452 https://doi.org/10.1111/j.1468-0262.2007.00808.x

BLUNDELL, R. and POWELL, J. L. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econom. Soc. Monogr.* **36** 312–357.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHEN, X. and POUZO, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* **80** 277–321. MR2920758 https://doi.org/10.3982/ECTA7888

CHERNOZHUKOV, V., DEMIRER, M., DUFLO, E. and FERNANDEZ-VAL, I. (2018). Generic machine learning inference on heterogenous treatment effects in randomized experiments. *National Bureau of Economic Research*.

DAROLLES, S., FAN, Y., FLORENS, J. P. and RENAULT, E. (2011). Nonparametric instrumental regression. *Econometrica* **79** 1541–1565. MR2883763 https://doi.org/10.3982/ECTA6539

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345. MR3695995 https://doi.org/10.1214/16-STS571

FINKELSTEIN, A., TAUBMAN, S., WRIGHT, B., BERNSTEIN, M., GRUBER, J., NEWHOUSE, J. P., ALLEN, H., BAICKER, K. and GROUP, O. H. S. (2012). The Oregon health insurance experiment: Evidence from the first year. *Q. J. Econ.* **127** 1057–1106.

FOGARTY, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika* **105** 994–1000. MR3877880 https://doi.org/10.1093/biomet/asy034

FOGARTY, C. B. (2020). Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *J. Amer. Statist. Assoc.* **115** 1518–1530. MR4143482 https://doi.org/10.1080/01621459.2019.1632072

FOGARTY, C. B., LEE, K., KELZ, R. R. and KEELE, L. J. (2021). Biased encouragements and heterogeneous effects in an instrumental variable study of emergency general surgical outcomes. *J. Amer. Statist. Assoc.*

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. MR4154846 https://doi.org/10.1214/19-BA1195

HALL, P. and HOROWITZ, J. L. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Ann. Statist.* **33** 2904–2929. MR2253107 https://doi.org/10.1214/009053605000000714

HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17** 360–372.

HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.* **34** 598–611. MR0152070 https://doi.org/10.1214/aoms/1177704172

HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **108** 135–148. MR3174608 https://doi.org/10.1080/01621459.2012.742018

HSU, J. Y., ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2015). Strong control of the family-wise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552 https://doi.org/10.1093/biomet/asv034

IMBENS, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* **48** 399–423.

JOHNSON, M., CAO, J. and KANG, H. (2022). Supplement to "Detecting heterogeneous treatment effects with instrumental variables and application to the Oregon health insurance experiment." https://doi.org/10.1214/21-AOAS1535SUPPA, https://doi.org/10.1214/21-AOAS1535SUPPB

KANG, H., PECK, L. and KEELE, L. (2018). Inference for instrumental variables: A randomization inference approach. *J. Roy. Statist. Soc. Ser. A* **181** 1231–1254. MR3876390 https://doi.org/10.1111/rssa.12353

KANG, H., KREUELS, B., ADJEI, O., KRUMKAMP, R., MAY, J. and SMALL, D. S. (2013). The causal effect of malaria on stunting: A Mendelian randomization and matching approach. *Int. J. Epidemiol.* **42** 1390–1398.

KANG, H., KREUELS, B., MAY, J. and SMALL, D. S. (2016). Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *Ann. Appl. Stat.* **10** 335–364. MR3480499 https://doi.org/10.1214/15-AOAS894

LEE, K., BARGAGLI-STOFFI, F. J. and DOMINICI, F. (2021). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. Preprint. Available at arXiv:2009.09036.

LEE, K., SMALL, D. S. and DOMINICI, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *J. Amer. Statist. Assoc.* **116** 569–580. MR4270004 https://doi.org/10.1080/01621459.2020.1870476

LEE, K., SMALL, D. S. and ROSENBAUM, P. R. (2018). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* **74** 1161–1170. MR3908134

LEE, K., SMALL, D. S., HSU, J. Y., SILBER, J. H. and ROSENBAUM, P. R. (2018). Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *J. Roy. Statist. Soc. Ser. A* **181** 535–546. MR3749529 https://doi.org/10.1111/rssa.12298

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056 https://doi.org/10.1093/biomet/63.3.655

NEWEY, W. K. and POWELL, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica* **71** 1565–1578. MR2000257 https://doi.org/10.1111/1468-0262.00459

PARK, C. and KANG, H. (2020). A groupwise approach for inferring heterogeneous treatment effects in causal inference. Preprint. Available at arXiv:1908.04427v2.

ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487 https://doi.org/10.1214/ss/1042727942

ROSENBAUM, P. R. (2002b). [Covariance adjustment in randomized experiments and observational studies]: Rejoinder. *Statist. Sci.* **17** 321–327. With comments and a rejoinder by the author. MR1962487 https://doi.org/10.1214/ss/1042727942

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

ROSENBAUM, P. R. (2020). Modern algorithms for matching in observational studies. *Annu. Rev. Stat. Appl.* **7** 143–176. MR4104189 https://doi.org/10.1146/annurev-statistics-031219-041058

ROTHWELL, P. M. (2005). Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *Lancet* **365** 176–186.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–188.

STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622 https://doi.org/10.2307/2171753

STALLONES, R. A. (1987). The use and abuse of subgroup analysis in epidemiological research. *Prev. Med.* **16** 183–194.

STOCK, J. H., WRIGHT, J. H. and YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist.* **20** 518–529. MR1973801 https://doi.org/10.1198/073500102288618658

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313

SU, L., MURTAZASHVILI, I. and ULLAH, A. (2013). Local linear GMM estimation of functional coefficient IV models with an application to estimating the rate of return to schooling. *J. Bus. Econom. Statist.* **31** 184–207. MR3055331 https://doi.org/10.1080/07350015.2012.754314

SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M. and LI, B. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.* **10** 141–158.

SWANSON, S. A. and HERNÁN, M. A. (2013). Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology* **24** 370–374.

SWANSON, S. A. and HERNÁN, M. A. (2014). Think globally, act globally: An epidemiologist's perspective on instrumental variable estimation [discussion of MR3264545]. *Statist. Sci.* **29** 371–374. MR3264549 https://doi.org/10.1214/14-STS491

THERNEAU, T., ATKINSON, B. and RIPLEY, B. (2015). Package 'rpart'. R package version 4.1-15. Available at https://cran.r-project.org/package=rpart.

WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

WANG, T. and RUDIN, C. (2021). Causal rule sets for identifying subgroups with enhanced treatment effect. Preprint. Available at arXiv:1710.05426.

YU, R. (2019). bigmatch: Making optimal matching size-scalable using optimal calipers. R package version 0.6.1. Available at https://CRAN.R-project.org/package=bigmatch.

YUSUF, S., WITTES, J., PROBSTFIELD, J. and TYROLER, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* **266** 93–98.