Electronic Journal of Statistics
Vol. 12 (2018) 3232–3253
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1476

False discovery rate control for effect modification in observational studies

Bikram Karmakar

Department of Statistics,
The Wharton School,
University of Pennsylvania, Pennsylvania, U.S.A.
e-mail: bikramk@wharton.upenn.edu

Ruth Heller

Department of Statistics and Operations Research,
Tel-Aviv University, Tel-Aviv, Israel.
e-mail: ruheller@post.tau.ac.il

and

Dylan S. Small

Department of Statistics,
The Wharton School,
University of Pennsylvania, Pennsylvania, U.S.A.
e-mail: dsmall@wharton.upenn.edu

Abstract: In an observational study, a difference between the treatment and control group's outcome might reflect the bias in treatment assignment rather than a true treatment effect. A sensitivity analysis determines the magnitude of this bias that would be needed to explain away as non-causal a significant treatment effect from a naive analysis that assumed no bias. Effect modification is the interaction between a treatment and a pretreatment covariate. In an observational study, there are often many possible effect modifiers and it is desirable to be able to look at the data to identify the effect modifiers that will be tested. For observational studies, we address simultaneously the problem of accounting for the multiplicity involved in choosing effect modifiers to test among many possible effect modifiers by looking at the data and conducting a proper sensitivity analysis. We develop an approach that provides finite sample false discovery rate control for a collection of adaptive hypotheses identified from the data on matched-pairs design. Along with simulation studies, an empirical study is presented on the effect of cigarette smoking on lead level in the blood using data from the U.S. National Health and Nutrition Examination Survey. Other applications of the suggested method are briefly discussed.

Keywords and phrases: Classification and regression trees, effect modification, sensitivity analysis, simultaneous testing, treatment effect.

Received May 2017.

Contents

1. Introduction

In a randomized study, we know the distribution of treatment assignment, but in an observational study, the distribution of treatment assignment is unknown. Consequently in an observational study, assuming that distribution of treatment assignment is random could introduce bias in inferences about the treatment effect. If there are no unmeasured confounders then it is possible to remove such bias by matching on observed covariates and conducting inference that assumes treatment is randomly assigned within matched sets (see e.g., Aakvik, 2001; Gemenisa and Rosemab, 2014; Pimentel, Yoon and Keele, 2016). On the other hand, if there do exist unmeasured confounders, then this analysis would be biased. In such a scenario a sensitivity analysis tries to answer the question of how much bias due to unmeasured confounders has to be present in the data to alter the inference based on the assumption of no unmeasured confounders (see e.g., Keele and Minozzi, 2013; Zubizarreta et al., 2012).

An effect modifier is defined as a pretreatment covariate for which the treatment effect differs according to the levels of the covariate. Effect modifiers are of inherent interest for personalizing treatment. In addition, although we can test for no treatment effect on any subject, i.e., Fisher's sharp null hypothesis (Fisher, 1935; Neyman, 1923), without having to consider effect modifiers, consideration of possible effect modification can increase power and reduce sensitivity to bias. Larger treatment effects are less sensitive to bias due to unmeasured confounders, than small treatment effects. Taking advantage of this fact Hsu, Small and Rosenbaum (2013) suggested forming a collection of subgroups of subjects in which subjects in a subgroup are expected to have the same level of treatment effect and then pool evidence from the subgroups to infer about Fisher's sharp null and effects within subgroups. There is though an operational difficulty in such analysis. Which variables are effect modifiers and at what levels of the variables the effect modification occurs are uncertain. It

is not always possible to have a priori knowledge of effect modifiers in a study. Many studies use the data to form the subgroups and then carry out analysis based on the learned groups. So, in such analysis the data is used twice, the first time to identify the subgroups and the second time to make inferences. To avoid always finding something if we look at enough data, it is important to control the error rate for having looked at the data to identify the subgroups of interest. Hsu et al. (2015) provided an algorithm that guarantees family wise error rate (FWER) control when the subgroups are built from the data.

The false discovery rate (FDR) is another choice for error control in multiple testing that is less conservative and has more power. Both FWER and FDR control the probability of falsely rejecting at least one null when all nulls are true. They differ when at least one null is false, for example, if we reject 20 nulls, 19 of which are false, FWER regards this as a failure while FDR regards this as a success for controlling the FDR at level 0.05 ($= 1/20$). Glickman, Rao and Schultz (2014) has strongly advocated FDR control in epidemiological studies on philosophical grounds. Many scientific communities have widely adopted FDR control as the norm for controlling for multiple testing. Our principal goal is to extend the work of Hsu et al. (2015) to propose a procedure that provides false discovery rate control. We show that in the same set-up of Hsu et al. (2015) we can use the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) on groups defined from the data, to provide FDR control on the group hypotheses.

The rest of the paper is organized as follows. In the following subsection we motivate our work by a real data example. Section 2.1 introduces required notation for technical purposes. Section 2.2 is dedicated to a short review of sensitivity analysis in an observational study. In Section 3 we derive the main technical results of our study. A detailed simulation study is presented in Section 4 and in Section 5 we revisit our motivating example.

1.1. Motivating example: lead level in the blood of smokers

Does smoking cause an increase in lead level in the blood? We consider data from the U.S. National Health and Nutrition Survey (NHANES) for the years 2009–2014. Hsu and Small (2013) studied similar data to elaborate on a different aspect of sensitivity analysis. We consider the data on the 9,103 adults 20 years or older who can be classified as smokers or non-smokers. A smoker is an individual who has reported smoking more than 100 cigarettes in his/her lifetime, has smoked every day for the last 30 days and has smoked one or more packs per days in the last 30 days. A non-smoker is someone who reported smoking less than 100 cigarettes in his/her lifetime and has not smoked any cigarettes in the last 30 days. There are 1,485 smokers and 7,618 non-smokers. Following previous observational studies of the effect of smoking (Rosenbaum, 2007a; Hsu and Small, 2013; Rosenbaum, 2017), we compare heavy smokers (as defined above) to non-smokers because making the two groups sharply differ in exposure dose increases the insensitivity of the study to unobserved biases when there is an exposure effect and no bias (i.e., it increases the design sensitivity, Rosenbaum, 2004).

We control for the following pretreatment covariates: age, gender, education (encoded in five indicator variables for categories of education level: less than 9th grade, between 9th and 11th grade, high school graduate/GED or equivalent, Some college or AA degree and College graduate or above), race (categorized as Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian and Other Race), income-to-poverty ratio, and an indicator variable for income information missing or not. We control for these pretreatment covariates by pair matching. In a pair matching each treated individual, i.e., a smoker, is matched to a control individual, i.e., a non-smoker, based on their observed covariates. In our study we use rank based Mahalanobis distance with a propensity score caliper (see Rosenbaum, 2010, Ch 8 for details) to form the matched pairs. This matching algorithm is implemented using the `pairmatch` function of the `optmatch` package in `R` (Hansen, 2007).

When considering change in lead level in the blood as an effect of smoking, genetic and environmental factors are potential confounding variables that we do not have any information on. A genetic factor can be associated with level of lead in the blood and might also affect the smoking habit of an individual. Also an individual in a certain industrial locality might be prone to a higher lead level and smoking behavior could be associated with locality. Consequently, in absence of information on locality and genetic aspects, a sensitivity analysis becomes important for properly evaluating the exposure effect.

We test Fisher's sharp null of no treatment effect with Huber-Maritz M-statistics using the `senmv` function (with parameters `inner = 0.1, trim = 1.5`) of the `sensitivitymv` package in `R` (Rosenbaum, 2015). An M-statistic, proposed by Maritz (1979), is the quantity equated to zero in defining Huber's M-estimates (Huber, 1981). We consider the one-sided alternative of smoking increasing lead level in the blood. The p-value, assuming there is no unmeasured confounding is less than 4.55×10^{-15}. Using the sensitivity analysis method of Rosenbaum (2007b), we can compute a minimum $\Gamma(\geq 1)$ at which the conclusion that smoking causes an increase in lead level is sensitive to bias from unmeasured confounding, where Γ is the maximum odds ratio for being a smoker vs. a non-smoker among two subjects matched on the measured confounders. If there are no unmeasured confounders, then $\Gamma = 1$; the more unmeasured confounding there is, the larger Γ is. In this study, we find that the effect is insensitive to hidden bias until $\Gamma = 2.6$, i.e., if the influence of unmeasured confounding is less than $\Gamma = 2.6$, we would still have strong evidence that smoking causes higher lead levels.

Individuals with different measured covariate values may have different magnitudes of treatment effect, i.e., different magnitude of increase in lead level due to smoking. Thus these covariates may work as effect modifiers. Genetic and environmental factors, which are unmeasured, can also be correlated with the magnitude of treatment effect. It is not known a priori how these covariates form relevant subgroups that show similar level of treatment effect. We use our data to form such potential subgroups. Figure 1 shows five subgroups which are created based on a regression tree (Breiman et al., 1984) model of the rank of the absolute difference of lead level in the blood in smokers and non-smokers

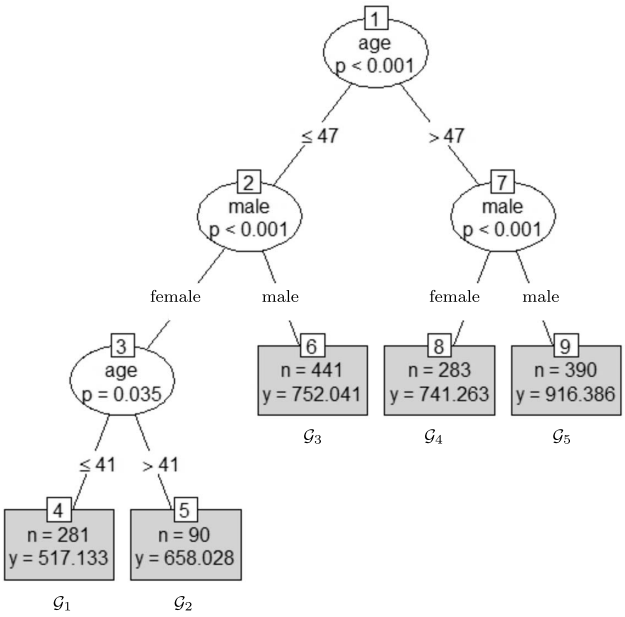

FIG 1. *Fitted regression tree of the rank of the absolute response difference $|Y_i|$ on the observed covariates. The nodes of the tree are the groups of individuals with n representing the number of such individuals and y the average rank of the absolute response differences of that group.*

within each pair on the observed matched covariates. This is implemented using the `ctree` function of the `R` package `party`. Covariates not matched exactly within pairs are averaged. We shall elaborate more on our choice of this method to build the subgroups in Section 3. We have five non overlapping subgroups, which are the leaf nodes of the tree in Figure 1. The key methodological question we will address is, how can we make valid inferences and perform valid sensitivity analyses that account for the fact that we have chosen the subgroups to examine based on the data? In Section 5 we shall report results of our analysis which compares our results in contrast to closed testing method of suggested by of (Marcus, Peritz and Gabriel, 1976) suggested by Hsu et al. This analysis will show evidence of smoking causing increase in lead level in the blood which is much less sensitive to bias compared to an analysis not incorporating effect modification.

2. Notation and reviews

2.1. Notation

In a matched pair study we denote I as the number of matched pairs. The matched pairs are formed based on observed covariates. For the ith matched pair there are two subjects denoted by $j = 1$ and 2. In shorthand we write ij

to denote jth subject for the ith matched pair. The treatment assignment for subject ij is denoted by Z_{ij} taking value 1 for treated or 0 for control. Let $\mathbf{x}_{ij}$ be the vector of observed covariates and u_{ij} be the summary of all unobserved covariates scaled between 0 and 1 for subject ij. By the architecture of the matching algorithm we have the constraints $Z_{i1} + Z_{i2} = 1$ and we seek to have $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_i$. It is possible though that $u_{i1} \neq u_{i2}$ since u_{i1} and u_{i2} are unobserved. In our motivating example of studying the effect of smoking on lead level in the blood, we have $I = 1,485$ pairs, i.e., $1,485 \times 2 = 2,970$ individuals.

For each subject ij, there is a pair of potential outcomes (r_{Tij}, r_{Cij}) corresponding to whether the subject is treated or not, i.e., $Z_{ij} = 1$ or 0 (Neyman, 1923; Rubin, 1974). To analyze the treatment effect we are interested in the difference $r_{Tij} - r_{Cij}$. For each individual we only observe one of these two responses based on the observed treatment assignment. We write the observed response for subject ij as $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$. We collect the characteristics of the subjects that are fixed regardless of treatment assignment and denote them by $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, 2, \cdots, I; j = 1, 2\}$. The only variable that is assumed to be random is the treatment assignment Z_{ij}, in other words the inference is conditional on $\mathcal{F}$. The difference in response between treated and control for the ith matched pair is given by $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$.

Consider a subset $\mathcal{G}$ of the indices $\{1, 2, \ldots, I\}$. We denote by $\mathbf{Z}_\mathcal{G}$ the vector of length $2 \times |\mathcal{G}|$ that collects the treatment assignments of matched pairs with indices in $\mathcal{G}$. In the same spirit we can introduce notation $\mathbf{r}_{T\mathcal{G}}, \mathbf{r}_{C\mathcal{G}}, \mathbf{x}_\mathcal{G}, \mathbf{u}_\mathcal{G}, \mathbf{R}_\mathcal{G}$ and $\mathbf{Y}_\mathcal{G}$. As a final piece of notation let $\mathcal{Z}_\mathcal{G}$ denote the collection of all possible treatment assignments. That is, $\mathcal{Z}_\mathcal{G}$ is the collection of $2^{|\mathcal{G}|}$ many $\mathbf{Z}_\mathcal{G}$ satisfying the constraint $Z_{i1} + Z_{i2} = 1$ for $i \in \mathcal{G}$. When $\mathcal{G}$ is the full set of indices we will simply drop the term $\mathcal{G}$ from the above set of notation.

Consider Fisher's sharp null hypothesis of no treatment effect restricted to subgroup $\mathcal{G}$. Let us denote this by $H_{0,\mathcal{G}}$. Suppose there are no unmeasured confounders, then we know that the treatment assignment in each pair is exactly randomized, i.e., $Pr(Z_{ij} = 1 \mid \mathcal{F}_\mathcal{G}, \mathcal{Z}_\mathcal{G}) = 1/2$ for each subject ij in the i-th pair ($i \in \mathcal{G}$). Since under the null hypothesis of no treatment effect among the subjects in $\mathcal{G}$, $\mathbf{r}_{C\mathcal{G}} = \mathbf{R}_\mathcal{G}$, we can calculate the null distribution of a test statistic $T(\mathbf{Z}_\mathcal{G}, \mathbf{R}_\mathcal{G})$ as

$$Pr(T(\mathbf{Z}_\mathcal{G}, \mathbf{R}_\mathcal{G}) \geq k \mid \mathcal{F}, \mathcal{Z}) = \frac{|\{\mathbf{z}_\mathcal{G} \in \mathcal{Z}_\mathcal{G} \mid T(\mathbf{z}_\mathcal{G}, \mathbf{R}_\mathcal{G}) \geq k\}|}{2^{|\mathcal{G}|}}.$$

There are various competing candidates for the choice of the test statistic $T(\cdot, \cdot)$. For our simulation we choose to use Wilcoxon's signed rank statistic due to its familiarity. Theory and computation of our work holds as is for other choices of test statistic suggested in the literature, e.g., Huber's M-statistics (Rosenbaum, 2007b) and U statistics (Rosenbaum, 2011).

2.2. Sensitivity to hidden bias

The sensitivity analysis approach we study here is based on two principal assumptions (see Rosenbaum, 1987, 2002, for more details). First, subjects are

assigned to treatment and control independently of each other. The propensity of treatment assignment is denoted by $\pi_{ij} = Pr(Z_{ij} = 1|\mathcal{F})$. When π_{ij}, even though unknown, is only a function of the observed covariates $\mathbf{x}_{ij}$, we can produce correct inference based on a paired randomized experiment as discussed at the end of the last section. When this may not be the case, we assume that two subjects with the same observed covariates may differ in their odds of treatment assignment by at most a factor of $\Gamma \geq 1$. That is, if for two subjects ij and $i'j'$ if $\mathbf{x}_{ij} = \mathbf{x}_{i'j'}$ then

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}/(1-\pi_{ij})}{\pi_{i'j'}/(1-\pi_{i'j'})} \leq \Gamma.$$

The factor Γ is termed as the level of hidden bias due to unmeasured confounders. If $\Gamma = 1$ then this model would correspond to paired randomized assignment. Let $\gamma = \log(\Gamma)$. Rosenbaum (2002) shows that this assumption is equivalent to writing the following treatment assignment distribution

$$Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = \prod_{i=1}^{I} \frac{z_{i1}exp(\gamma u_{i1}) + z_{i2}exp(\gamma u_{i2})}{exp(\gamma u_{i1}) + exp(\gamma u_{i2})}, \tag{2.1}$$

for some $\mathbf{u} \in [0,1]^{2I}$.

For various approaches to sensitivity analysis in observational studies, see Cornfield et al. (2009), Gastwirth (1992), Handorf et al. (2013), Hosman, Hansen and Holland (2010), Imbens (2003), McCandless, Gustafson and Levy (2007), Liu, Kuramoto and Stuart (2013), Wang and Krieger (2006), Yanagawa (1984) and Yu and Gastwirth (2005).

Based on the above setting a sensitivity analysis provides bounds for the quantities which are unknown because of unobserved u's. For example, suppose $T_{\mathcal{G}}$ is the statistic to be used for testing the null $H_{0,\mathcal{G}}$. Because the π_{ij} are unknown we do not know the null distribution of $T_{\mathcal{G}}$. But under model assumption (2.1) we can produce two statistic $\underline{T}_{\Gamma,\mathcal{G}}$ and $\overline{T}_{\Gamma,\mathcal{G}}$ whose distribution under the null are known and they sharply bound $T_{\mathcal{G}}$ (using first order stochastic dominance) from below and above respectively. Thus we can produce lower and upper bounds on the p-value as $\underline{p}_{\Gamma}$ and $\bar{p}_{\Gamma}$. When $\Gamma = 1$, these two bounds are the same. At significance level α we would reject the null hypothesis if $\bar{p}_{\Gamma=1}$ is less than α. An inference is sensitive to hidden bias Γ at level α if $\bar{p}_{\Gamma} \geq \alpha$.

In the context of testing the null hypothesis for a subgroup $\mathcal{G}$, we can relax our model (2.1) to be valid for pairs in $\mathcal{G}$ only to write

$$Pr(\mathbf{Z}_{\mathcal{G}} = \mathbf{z}_{\mathcal{G}}|\mathcal{F}_{\mathcal{G}}, \mathcal{Z}_{\mathcal{G}}) = \prod_{i\in\mathcal{G}} \frac{z_{i1}exp(\gamma u_{i1}) + z_{i2}exp(\gamma u_{i2})}{exp(\gamma u_{i1}) + exp(\gamma u_{i2})}, \tag{2.2}$$

for some $\mathbf{u}_{\mathcal{G}} \in [0,1]^{2|\mathcal{G}|}$. The corresponding lower and upper bounds for the p-value are $\underline{p}_{\Gamma(\mathcal{G})}$ and $\bar{p}_{\Gamma(\mathcal{G})}$.

2.3. Effect modification

In the presence of effect modification, our strategy is to divide the set of individuals into subgroups. These subgroups can be constructed based on a priori information. But in most situations when there is not sufficient a priori information, we would like to be able to derive these groups from observed data. We denote the subgroups by $\{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G\}$. The subgroups are non-overlapping. An ideal grouping, that explores the effect modifiers perfectly, would aim for subgroups such that asymptotically, in the number of pairs, every matched pair in a single subgroup $\mathcal{G}_g$ has the same treatment effect and treatment effects of different groups differ. In finite samples there is a bias-variance trade-off in constructing the partition in terms of the size of the subgroups G. To be practically useful, we would like to have the grouping such that we have a moderate number of subgroups and two matched pairs which have similar treatment effects are in the same subgroup.

In our motivating example subjects were divided into $G = 5$ groups based on the data (cf. Figure 1). The procedure that identifies these 5 partitions of the data first calculates the absolute difference in lead level in the blood in each pair, ranks these absolute response differences and the ranks are used as outcome in a regression tree fitting on the observed covariates. The use of a regression tree model allows us to identify these groups in terms of the observed covariates.

$$\begin{aligned}
\mathcal{G}_1 &= \{\text{female of age less than 41 years}\},\\
\mathcal{G}_2 &= \{\text{female of age between 41 and 47 years}\},\\
\mathcal{G}_3 &= \{\text{male of age less than or equal to 47 years}\},\\
\mathcal{G}_4 &= \{\text{female of age more than 47 years}\},\\
\mathcal{G}_5 &= \{\text{male of age more than 47 years}\}.
\end{aligned}$$

Each subgroup $\mathcal{G}_g$ corresponds to a null hypothesis $H_{0,\mathcal{G}_g}$ of no exposure effect in all pairs in the group. Then we want to test the collection of G many hypotheses $\{H_{0,\mathcal{G}} \mid \mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G\}\}$. When the groups are derived from the data, the grouping is a random quantity and this collection of hypotheses is also random.

Hsu, Small and Rosenbaum (2013) tested the global (intersection) null hypothesis H_0 by first computing p-values for each of the hypotheses and then combining those G p-values. Their suggestion was to use the product of truncated p-values (Zaykin et al., 2002) $\bar{P}_{\Gamma\wedge} = \prod_{g=1}^{G} \left(\bar{p}_{\Gamma\mathcal{G}_g}\right)^{\chi(\bar{p}_{\Gamma\mathcal{G}_g} \le \tilde{\alpha})}$, where $\tilde{\alpha}$ is a parameter taking value in $(0, 1]$. Here and later in this paper we use $\chi(E)$ to denote the indicator function of an event E. For $\tilde{\alpha} = 1$ this would be equivalent to Fisher's method of combining p-values. When the groups are formed a priori, Hsu, Small and Rosenbaum provided upper bounds on the null distribution of $\bar{P}_{\Gamma\wedge}$ which can be used to carry out a sensitivity analysis.

As we advocated earlier, in presence of effect modifier we would like to provide inference on the collection of hypotheses considered. Hsu et al. (2015) suggested using a closed testing approach (Marcus, Peritz and Gabriel, 1976), that provides the FWER guarantee that the probability of at least one false rejection is at

most α. We will develop a method to control the FDR when choosing effect modifier hypotheses based on the data.

In our simulation in Section 4 and the analysis of data set studying smoking effect on the lead level in the blood in Section 5, we see considerable gain in considering effect modifiers in the sense that the evidence from such an analysis is much less sensitive to bias over an analysis that ignores effect modification. Through extensive simulation, Hsu et al. showed that, in the presence of effect modification, the closed testing procedure is less sensitive to unmeasured confounders than the global test of no effect. In our simulation and data analysis we observe that the proposed procedure is more robust to unmeasured confounding than the closed testing procedure.

2.4. False discovery rate

For simultaneous hypotheses, the false discovery rate introduced by Benjamini and Hochberg (1995) is defined as the expected value of the proportion of falsely rejected hypotheses out of all rejected hypotheses. Let D_g be the decision function receiving the values 1 or 0 for whether $H_{0,\mathcal{G}_g}$ is rejected or not rejected, respectively. Let $\mathcal{G}_0$ be the collection of pairs with no treatment effect. Then

$$FDR = E\left(\frac{\sum_{g=1}^{G}\chi(\mathcal{G}_g \subseteq \mathcal{G}_0, D_g = 1)}{max(\sum_{g=1}^{G}\chi(D_g = 1), 1)}\right).$$

3. Adaptive inference under effect modification

The selection of the subgroups from the data must only use information about a pair that would be unchanged if there is no treatment effect in the pair and the treatment assignment were reversed. In our motivating example we used a regression tree of the rank of the absolute difference of responses in a pair (Y_i) on the observed covariates $(\mathbf{x}_i)$. This approach is motivated by the fact that when the difference of responses Y_i is related to observed covariates via a non-negative function with additive noise, then the absolute response regressed on the covariates would group the similar effects (Jogdeo, 1977; Hsu et al., 2015). For our theoretical result we assume the algorithm of building the groups satisfies the condition below,

Condition A. The groups of pairs are mutually exclusive and are formed only as a function of $|Y_i|$ and the matched covariates $\mathbf{x}_i$.

The condition says that explicitly we are not allowed to use raw information of the treatment assignment. Let $\mathcal{G}_0$ be the collection of pairs with no treatment effect. Then the condition above allows us to use the information $\mathcal{I}$ that is the union of $\{(r_{Ci1}, r_{Ci2}, \mathbf{x}_i) \mid i \in \mathcal{G}_0\}$ and $\{(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, \mathbf{x}_i) \mid i \in \mathcal{G}_0^c\}$. To see this, note that for $i \in \mathcal{G}_0$ we have $|Y_i| = |r_{Ci1} - r_{Ci2}|$ and for $i \in \mathcal{G}_0^c$ we would have $|Y_i| = |R_{i1} - R_{i2}|$ with $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$. Our main

result is that the Benjamini-Hochberg (BH) procedure with level q applied on $\{\bar{p}_{\Gamma,\mathcal{G}} \mid \mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G\}\}$ provides the following guarantee

$$E\left(\frac{\sum_{g=1}^{G} \chi(\mathcal{G}_g \subseteq \mathcal{G}_0, D_g = 1)}{max(\sum_{g=1}^{G} \chi(D_g = 1), 1)} \,\middle|\, \mathcal{F}, \mathcal{Z}, \mathcal{I}\right) \leq q. \tag{3.1}$$

Following are the two lemmas that are key to establishing (3.1).

Lemma 1. *Suppose the subgroups are built based on Condition A. Conditional on* $(\mathcal{F}, \mathcal{Z}, \mathcal{I})$ *the treatment assignments on groups* $\{\mathbf{Z}_{\mathcal{G}_g} \mid g = 1, 2, \ldots, G\}$ *are independently distributed.*

Proof. Recall, the assumption of our sensitivity model that the subjects are assigned to treatment and control independently of each other. Consequently, after conditioning on $\mathcal{I}$ the treatment assignments $\mathbf{Z}_{\mathcal{G}_0}$ would be independent and the others are fixed. □

Lemma 2. *Suppose the subgroups are built based on Condition A. Conditional on* $(\mathcal{F}, \mathcal{Z}, \mathcal{I})$ *for any* $\mathcal{G}_g$ *with no treatment effect the sensitivity model of* (2.2) *holds with* $\mathcal{G} = \mathcal{G}_g$.

Proof. If $\mathcal{G}_g$ has no treatment effect then $\mathcal{G}_g \subseteq \mathcal{G}_0$. Thus for each $i \in \mathcal{G}_g$ the information about i in $\mathcal{I}$ is already contained in $(\mathcal{F}_{\mathcal{G}_g}, \mathcal{Z}_{\mathcal{G}_g})$. Therefore finding the propensity score of treatment assignment conditioning on $(\mathcal{F}, \mathcal{Z}, \mathcal{I})$ is the same as conditioning on $(\mathcal{F}_{\mathcal{G}_g}, \mathcal{Z}_{\mathcal{G}_g})$. Thus we get

$$\begin{aligned} Pr(\mathbf{Z}_{\mathcal{G}_g} = \mathbf{z}_{\mathcal{G}_g} | \mathcal{F}, \mathcal{Z}, \mathcal{I}) &= Pr(\mathbf{Z}_{\mathcal{G}_g} = \mathbf{z}_{\mathcal{G}_g} | \mathcal{F}_{\mathcal{G}_g}, \mathcal{Z}_{\mathcal{G}_g}) \\ &= \prod_{i \in \mathcal{G}_g} \frac{z_{i1} exp(\gamma u_{i1}) + z_{i2} exp(\gamma u_{i2})}{exp(\gamma u_{i1}) + exp(\gamma u_{i2})}. \end{aligned}$$

□

As a consequence of Lemma 2, following the argument of Rosenbaum (2002) we can bound the p-values for testing the random null $H_{0,\mathcal{G}_g}$ by $\underline{p}_{\Gamma\mathcal{G}_g}$ and $\bar{p}_{\Gamma\mathcal{G}_g}$ from below and above respectively.

Theorem 1. *If the subgroups are built based on Condition A and the bias is no more than* Γ*, then the BH procedure applied to the collection* $\{\bar{p}_{\Gamma,\mathcal{G}} \mid \mathcal{G} \in \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G\}\}$ *at level* q *satisfies* (3.1).

The proof of the theorem is given in the Appendix.

Condition A guarantees that the grouping is determined by the information set $\mathcal{I}$. Different grouping strategies (that satisfy Condition A) may result in different groups, and hence have different power to reject non-null groups. The BH procedure on the p-value upper bounds guarantees that the FDR is controlled given the grouping, for any configuration of null and non-null groups. Moreover, the FDR control is guaranteed for any treatment assignment distribution of the nonnull pairs. Since the grouping is unchanged for any possible treatment assignments for the nonnull pairs when the absolute difference in response is used for group construction, by marginalizing over the information set $\mathcal{I}$, the FDR is unchanged. This is formalized in the following corollary.

Corollary 1.1. *Under the structure of Theorem 1*

$$E\left(\frac{\sum_{g=1}^{G}\chi(\mathcal{G}_g \subseteq \mathcal{G}_0, D_g = 1)}{\max(\sum_{g=1}^{G}\chi(D_g = 1), 1)} \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \leq q.$$

Proof. Given the knowledge of $\mathcal{F}$ and $\mathcal{Z}$, for a treatment assignment $\mathbf{Z}$, the information set $\mathcal{I}$ is determined from the treatment assignments on the non-null pairs, i.e., $\mathbf{Z}_{\mathcal{G}_0^c}$. Let $\mathcal{I}(\mathbf{z}_{\mathcal{G}_0^c})$ be the information set of the grouping strategy when the treatment assignment on the pairs in $\mathcal{G}_0^c$ is $\mathbf{z}_{\mathcal{G}_0^c}$. Then

$$\begin{aligned}
&E\left(\frac{\sum_{g=1}^{G}\chi(\mathcal{G}_g \subseteq \mathcal{G}_0, D_g = 1)}{\max(\sum_{g=1}^{G}\chi(D_g = 1), 1)} \,\middle|\, \mathcal{F}, \mathcal{Z}\right) \\
&= \sum_{\mathbf{z}_{\mathcal{G}_0^c} \in \mathcal{Z}_{\mathcal{G}_0^c}} E\left(\frac{\sum_{g=1}^{G}\chi(\mathcal{G}_g \subseteq \mathcal{G}_0, D_g = 1)}{\max(\sum_{g=1}^{G}\chi(D_g = 1), 1)} \,\middle|\, \mathcal{F}, \mathcal{Z}, \mathbf{z}_{\mathcal{G}_0^c}\right) Pr(\mathbf{Z}_{\mathcal{G}_0^c} = \mathbf{z}_{\mathcal{G}_0^c} \mid \mathcal{F}, \mathcal{Z}) \\
&= \sum_{\mathbf{z}_{\mathcal{G}_0^c} \in \mathcal{Z}_{\mathcal{G}_0^c}} E\left(\frac{\sum_{g=1}^{G}\chi(\mathcal{G}_g \subseteq \mathcal{G}_0, D_g = 1)}{\max(\sum_{g=1}^{G}\chi(D_g = 1), 1)} \,\middle|\, \mathcal{F}, \mathcal{Z}, \mathcal{I}(\mathbf{z}_{\mathcal{G}_0^c})\right) Pr(\mathbf{Z}_{\mathcal{G}_0^c} = \mathbf{z}_{\mathcal{G}_0^c} \mid \mathcal{F}, \mathcal{Z}).
\end{aligned}$$

The rest of the proof now follows from Theorem 1 which proves (3.1) for any information set $\mathcal{I} \equiv \mathcal{I}(\mathbf{z}_{\mathcal{G}_0^c})$. □

A motivation for considering effect modification in the sensitivity analysis of treatment effects is that a larger treatment effect tends to be less sensitive to hidden bias than a smaller treatment effect. The next theorem proves that our procedure provides FDR control for such a sensitivity analysis.

Theorem 2. *Suppose the subgroups are built based on stated Condition A. For hidden bias level* $\Gamma_1, \Gamma_2, \ldots, \Gamma_G$*, if the bias is at most* Γ_g *in* $g \in \mathcal{G}$ *then the BH procedure applied at level* q *on the collection* $\{\bar{p}_{\Gamma_g, \mathcal{G}_g} \mid g = 1, 2, \ldots, G\}$ *then FDR is controlled at level* q*.*

The proof of Theorem 2 is very similar to that of Theorem 1, and therefore omitted.

A simple three step procedure for our sensitivity analysis is as follows.

Procedure 3.1. The sensitivity analysis procedure for FDR control:

- *Create groups of pairs* $\mathcal{G}_1, \ldots, \mathcal{G}_G$ *using an algorithm that satisfies Condition A.*
- *Fix* $\Gamma_g : g = 1, 2, \ldots, G$ *and compute* $\bar{p}_{\Gamma_g, \mathcal{G}_g} : g = 1, 2, \ldots, G$*.*
- *Apply the BH procedure at the desired level* q*.*

For the second step of choosing the Γ_g values, subject matter information about what values of Γ_g are of concern and information from $(\mathcal{F}, \mathcal{Z}, \mathcal{I})$ can be used.

4. Simulation

Here we examine the performance of the suggested procedure 3.1 on data built groups in various simulation settings. On $I = 1,000$ pairs, the treatment assignment is assumed to be randomized so that there is no hidden bias in the study. First, we consider a single observed covariate (x) that takes five possible values 1 through 5 independently for each pair with equal probabilities. The response difference for the pairs is simulated from $N(\mu_x, 1)$. We denote the vector of treatment effects on covariate x as $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_5)$. When $\mu_j, j = 1, 2, \ldots, 5$ are not all the same, there is effect modification. Based on this simulated data we fit a regression tree on the rank of absolute response on the covariate x to build the groups. A second binary covariate with no effect modification is also used in the regression fit. Thus the groups tested (and the number of groups) are possibly different on each simulation. Table 1 reports the summary of the simulation study based on 2,000 simulations. Eight separate scenarios, starting with no effect (A) to different levels of effect modification (B–D) and then various scenarios of strong effect modification (E-1–E-4) are investigated.

In Table 1 the false discovery rate is controlled at the desired level $q = 0.05$ by both the closed testing and the BH procedure. To compare the power, we compare the closed testing procedure (Marcus, Peritz and Gabriel, 1976) to the BH procedure in terms of the false non-discovery rate (FNR) (Genovese and Wasserman, 2002) defined as the expected value of the ratio of false hypotheses not rejected out of all the hypotheses not rejected. In terms of estimated FNR, the BH procedure is never worse than closed testing and sometimes considerably better, e.g., in scenario E-4 the BH procedure has FNR 0.1% compared to 68% for $\Gamma = 3$ closed testing procedure or in scenario E-1 where FNR for the BH and the closed testing are 0% and 46%, respectively for $\Gamma = 1$.

We also consider two other type of errors, the proportion of pairs rejected having no effect and proportion of pairs not rejected but having treatment effect. While FDR and FNR are group level measures, these two measures compare the methods on the pair level. In terms of both these errors BH is as good as and often better than closed testing especially in scenarios of strong effect modification (E-1, E-3). For example in scenario E-4 at $\Gamma = 3$ on average, 77% of the pairs having treatment effect are not rejected in closed testing as compared to 0% (rounded to one significant digit) for the BH procedure.

The design above does not introduce any hidden bias to our simulation study, thus we have true $\Gamma = 1$ along with a treatment effect. The justification for considering this favorable situation for the power computation for sensitivity analysis is explained in (Hansen, Rosenbaum and Small, 2014, Section 3) and we review it here. Even though in practice we cannot know if we are in the favorable situation, by computing the power in this situation we are assessing the ability of our analyses to discriminate between two situations where we know unambiguously the desired result of the sensitivity analyses. In one situation, with moderate bias and no treatment effect, we expect that any associations between treatment and outcome can be explained by magnitude of bias at most Γ and by construction there can be at most a risk of at most α to report

TABLE 1
Average power, FDR and FNR in a setting with effect modification in five levels of a covariate.

Scenario $\boldsymbol{\mu}$ (effect modification)	# Nodes	Γ	# Rejected Nodes		FDR (%)		FNR (%)		% of pairs rejected with no effect		% of pairs not rejected with effect	
			Closed Testing	BH	Closed Testing	BH	Closed Testing	BH	Closed Testing	BH	Closed Testing	BH
(0,0,0,0,0)	1.05	1	0.05	0.05	4.2	5.2	0	0	5	5	n/a	n/a
A: Null Case	(1,1)	1.1	**0.005**	**0.006**	0.5	0.6	0	0	0.5	0.6	n/a	n/a
		1.3	0	0	0	0	0	0	0	0	n/a	n/a
(0.5, 0.5, 0.5,	1.05	1	1.05	1.05	0	0	0	0	n/a	n/a	0	0
0.5,0.5)	(1,1)	2	1.035	1.035	0	0	1.4	1.4	n/a	n/a	0.6	0.6
B: Constant		3	**0.22**	**0.22**	0	0	79	79	n/a	n/a	78	78
effect		5	0	0	0	0	100	100	n/a	n/a	100	100
(0.6, 0.6, 0.6,	1.17	1	1.17	1.17	0	0	0	0	n/a	n/a	0	0
0.4, 0.4)	(1,2)	2	1.06	1.06	0	0	8.6	8.6	n/a	n/a	4.6	4.7
C: Mild		3	**0.1**	**0.15**	0	0	91	91	n/a	n/a	90	88
modification		5	0	0	0	0	100	100	n/a	n/a	100	100
(1.5, 0, 0,	2.52	1	1.78	1.8	0.6	1.2	0	0	0.6	1	0	0
1.5, 0)	(1,4)	3	**0.07**	1.13	0	0	78	61	0	0	95	67
D: Complex		15	0	**0.07**	0	0	81	81	0	0	100	97
modification		20	0	0	0	0	78	78	0	0	100	100
(2, 0, 0, 0, 0)	2.05	1	1.05	1.05	2.4	2.6	0	0	3.8	3.9	0	0
E-1: Strong	(2,2)	2	**0.06**	1	0	0	46	0	0	0	93	0
effect		20	0	**0.25**	0	0	49.3	36.6	0	0	100	73
modification		25	0	0	0	0	49	49	0	0	100	100
(1.5,0.5,0.1,0,0)	2.17	1	1.9	1.9	0.1	0.1	17.5	17.3	0.1	0.1	13	12
E-2: Strong	(2,3)	2	**0.4**	1.06	0	0	96.9	96.1	0	0	89	76
effect		10	0	**0.74**	0	0	98	97	0	0	98	83
modification		15	0	0	0	0	100	100	0	0	100	100
(0, 1.2, 0,	1.31	1	1.12	1.12	0.5	0.5	0.3	0	0.2	0.2	0.3	0.1
0.8, 0)	(1,3)	2	**0.38**	0.47	0	0	59.6	57.5	0	0	64	61
E-3: Strong		3	0	0.1	0	0	93	91	0	0	99	97
effect		5	0	**0.09**	0	0	93	92	0	0	100	98
modification		10	0	0	0	0	93	93	0	0	100	100
(1, 2, 0, 0, 3)	4.05	1	3.04	3.04	1.2	1.2	0	0	1.8	1.9	0	0
E-4: Strong	(4,4)	3	**0.66**	3	0	0	68	0.1	0	0	77	0
effect		20	0	**0.6**	0	0	74	67	0	0	100	78
modification		25	0	0	0	0	75	75	0	0	100	100

Average of the number of terminal nodes (numbers in the parenthesis give the interquartile range), number of rejections, FDR, FNR, and power of closed testing and the BH procedure. There are two covariates, only one of these covariates is associated with treatment effect. This covariate is distributed as multinomial on 5 levels. Summary is based on 2,000 simulations of random treatment assignment and for the covariate value of x the response difference is simulated from $N(\boldsymbol{\mu}(x), 1)$. $\boldsymbol{\mu}$ is the vector of length 5 of μ_x values.

otherwise. In the second situation, when there is no bias and there is a treatment effect, then we hope to reject the null hypothesis. On the other hand, if we were considering a situation where there was large bias in treatment assignment and a small treatment effect, so that rejection of the null is nearly assured for all small or moderate Γ then we would not have been pleased to reject the null for small or moderate Γ because we know we would also have rejected the null in this situation had it been true.

As seen in Table 1, when there is no or very mild effect modification, e.g., in the first three scenarios A, B and C, two methods have similar level of sensitivity to hidden bias. Here we say that the test is insensitive to level Γ (≥ 1) if it rejects any of the terminal node hypotheses at that level of Γ. But when there is strong

TABLE 2
Average power, FDR and FNR in a setting with effect modification at 10 levels of a covariate.

Scenario $\boldsymbol{\mu}$ (effect modification)	# Nodes	Γ	# Rejected Nodes		FDR (%)		FNR (%)		% of pairs rejected with no effect		% of pairs not rejected with effect	
			Closed Testing	BH	Closed Testing	BH	Closed Testing	BH	Closed Testing	BH	Closed Testing	BH
$\mu(x) = 0$	3.6	1	0.02	0.04	2	5	0	0	1	1.3	n/a	n/a
No effect	(1,7.05)	1.1	0	0.01	0	1	0	0	0	0.001	n/a	n/a
		1.2	0	0	0	0	0	0	0	0	n/a	n/a
		3	0	0	0	0	0	0	0	0	n/a	n/a
$\mu(x) = 0$ if	3.7	1	2.94	3.22	1.1	1.6	23	16	0.3	0.4	8	5
x odd	(1,8)	2	0	0.14	0	0	85	85	0	0	97	97
$\mu(x) = 0.5$ if		3	0	0.005	0	0	87	87	0	0	100	100
x even		5	0	0	0	0	86	86	0	0	100	100
$\mu(x) = 0$ if	5.6	1	3.84	4.02	0.4	0.7	2	0	0.1	0.2	1.3	0
x odd	(2,13)	2	1.38	2.92	0	0	37	18	0	0	34	15
$\mu(x) = 1$ if		10	0	0	0	0	73	73	0	0	100	100
even												
$\mu(x) = 0$ if	8.5	1	5.6	6.2	0	1.4	7	0.6	0	0	6	0
x odd	(2,16.6)	5	0.34	3.74	0	0	79	64	0	0	94	70
$\mu(x) = x\,(mod\,6)$		10	0	0.01	0	0	72	72	0	0	100	100
if x even		15	0	0	0	0	72	72	0	0	100	100
$\mu(x) = 0$ if	8.8	1	4.84	4.96	0	1.7	0.2	0	0.1	0.6	0	0
x odd	(2,16.6)	5	0.68	4.34	0	0	45	11	0	0	62	11
$\mu(x) = 3$ if		10	0	2.52	0	0	65	53	0	0	99	79
x even		15	0	0.45	0	0	65	64	0	0	100	89
		18	0	0	0	0	65	65	0	0	100	100

Average of the number of terminal nodes (numbers in the parenthesis give the interquartile range), number of rejections, FDR, FNR, and power of closed testing and the BH procedure. There are two covariates, only one of these covariates is associated with treatment effect. This covariate is distributed as multinomial on 10 levels. Summary is based on 2,000 simulations of random treatment assignment and for the covariate value of x the response difference is simulated from $N(\boldsymbol{\mu}(x), 1)$. $\boldsymbol{\mu}$ is the vector of length 10 of μ_x values.

effect modification of the covariate the BH procedure is insensitive to a much higher level of hidden bias compared to closed testing. For example, in scenario E-4 the average number of rejections by closed testing and BH, at $\Gamma = 3$, are 0.66 and 3, respectively; at $\Gamma = 20$ they are 0 and 0.66, respectively.

FDR control is particularly useful when we expect large number of groups. We carry out two sets of simulation studies to examine false discovery rate control with 10 groups and 20 groups. The simulation design is as before with $I = 1,000$ pairs and equal randomized treatment assignment. There are two covariates, one distributed as multinomial with 10 and 20 labels for the two studies respectively and another one is Bernoulli. The difference in response for paired individuals is modeled as normally distributed with unit variance and mean equal to $\mu(x)$, where x is the label of the multinomial covariate. Table 2 and Table 3 report results of simulation studies based on this design. The conclusions from Table 2 and Table 3 are consistent with those of the previous

TABLE 3
Average power, FDR and FNR in a setting with effect modification at 20 levels of a covariate.

Scenario	# Nodes	Γ	# Rejected Nodes	FDR (%)	FNR (%)	% of pairs rejected with no effect	% of pairs not rejected with effect
$\mu(x) \equiv 0$	6	1	0.05	4.9	0	1.6	n/a
No effect	(1, 18)	1.1	0.01	1	0	0.003	n/a
		1.2	0.002	0.002	0	0	n/a
		1.3	0	0	0	0	n/a
$\mu(x) = 0$ if	6	1	5.526	1.14	1.8	0.2	0.3
x odd	(1,18)	3	0.03	0	77	0	99
$\mu(x) = 0.5$		5	0	0	78.4	0	100
if x even							
$\mu(x) = 0$ if	8	1	6.15	1.6	0	0.3	0
x odd	(1, 30)	3	4.4	0	44	0	82
$\mu(x) = 1$ if		5	2.9	0	54.2	0	91
x even		12	0	0	66.2	0	100
$\mu(x) = 0$ if	17	1	10	2.04	0	0.6	0
x odd	(3,37)	10	8.4	0	26.2	0	68
$\mu(x) = 3$ if		30	0.7	0	55.7	0	95
x even		40	0	0	58	0	100
$\mu(x) = 0$ if	24	1	13	1.76	0	0.8	0
x odd	(2, 38)	5	11.5	0	17.7	0	27
$\mu(x) = x(mod\ 6)$		20	8.8	0	43	0	67
if x even		40	0	0	66	0	100

Average of the number of terminal nodes (numbers in the parenthesis give the interquartile range), number of rejections, FDR, FNR, and power of the BH procedure. There are two covariates, only one of these covariates is associated with treatment effect. This covariate is distributed as multinomial on 20 levels. Summary is based on 2,000 simulations of random treatment assignment and for the covariate value of x the response difference is simulated from $N(\boldsymbol{\mu}(x), 1)$. $\boldsymbol{\mu}$ is the vector of length 20 of μ_x values.

simulation. FDR is controlled throughout and when there are no effects (first row of Table 2 and Table 3) FDR reaches its nominal level by the BH procedure while closed testing is conservative. The power gain from the BH procedure over closed testing can be seen in estimated FNR values in Table 2. As an example, in the last scenario for $\Gamma = 5$ the FNR level is at 45% for closed testing compared to 11% for the BH procedure and in the same scenario on average closed testing fails to reject 62% of the pairs with treatment effect as compared to 11% for the BH procedure. Finally, the BH procedure is much less sensitive to hidden bias than closed testing. For example, in the last scenario of Table 2 closed testing is insensitive until hidden bias level of $\Gamma = 5$ whereas the BH procedure is insensitive until $\Gamma = 15$. The setting of Table 3 does not contain the closed testing results as closed testing procedure for 20 groups does 2^{20} comparisons which requires unmanageable amount of computing resources. The number of rejected nodes, FNR and % of pairs rejected with effect in Table 3 shows the considerable power of the proposed procedure.

We summarize our simulation results here. We have considered various scenarios of effect modifications along with different number of subgroups in our simulations. While both closed testing and our procedure controls for FDR at the nominal level, our sensitivity analysis BH procedure is much less sensitive to hidden bias compared to closed testing. In terms of power, the BH procedure always had higher simulated power compared to closed testing in all the measures we have considered in our comparisons. With large number of subgroups closed testing becomes computationally infeasible as the complexity is exponential in the number of subgroups, but the BH procedure has computational complexity which is at most quadratic in the number of subgroups.

5. Results for study of the effect of smoking on lead in the blood

We go back to our motivating example and analyze the data on 1,485 matched pairs of individuals for effect of smoking on lead level in the blood. As described in Section 1.1, we used a data based method consistent with Condition A to derive five mutually exclusive and exhaustive subgroups of the pairs as described in Section 2.3. These subgroups contain 281, 90, 441, 283 and 390 pairs of subjects respectively. Figure 1 shows average of the ranks of the absolute response differences in the terminal nodes.

We aim to assess whether smoking increases the level of lead in the blood. We use a Huber-Maritz M-statistics with specifications as given in Section 1.1. In the notation of Section 2.1, for an odd function $\psi(\cdot)$ with $\psi(0) = 0$, a Huber-Maritz M-statistics for group $\mathcal{G}$ is of the form $\sum_{i\in\mathcal{G}} \psi(y_i/s_{\mathcal{G}})$, where y_i is the difference in response between treatment and control for matched pair i and $s_{\mathcal{G}}$ is the median of the absolute difference in response for the matched pairs in $\mathcal{G}$. In our analysis we consider $\psi(y) = \text{sign}(y)\,(\min(|y|, 1.5) - 0.1)_+$, where $(a)_+ := \max(a, 0)$. For more discussion on M-statistics and related sensitivity analysis procedures see Rosenbaum (2007b, 2014).

We first consider the global Fisher's null hypothesis of no treatment effect for any subject. In Section 1.1 we noted that the test that combines all the pairs without consideration of effect modification allows for the rejection decision to be sensitive to hidden bias of $\Gamma = 2.6$. Using the group information Table 4 in the second to last column reports truncated product, with $\tilde{\alpha} = 0.20$, p-values $\bar{p}_{\Gamma\wedge}$ for the global null for different levels of hidden bias. These are computed using the `truncatedP` function of the `sensitivitymv` package in `R`. The decision to reject the global null based on truncated product p-values is sensitive to hidden bias of 2.8.

Simes test (Simes, 1986) can also be used to test the global null hypothesis of no treatment effect. It rejects the global null if and only if the BH procedure rejects at least one of the group-hypotheses. To simplify the notation suppose the group numbers are ordered by their p-values, $\bar{p}_{\Gamma_g,\mathcal{G}_g} \le \bar{p}_{\Gamma_{g'},\mathcal{G}_{g'}}$ for $1 \le g < g' \le G$. Simes' p-value is $\min_{g\ge 1} G \times \bar{p}_{\Gamma_g,\mathcal{G}_g}/g$. Simes' p-value is a valid p-value if it combines individual p-values that are independent or have PRDS dependence (Benjamini and Yekutieli, 2001). To see this note that

TABLE 4
Maximum of p-values for varied hidden bias levels.

Γ	$\bar{p}_{\Gamma,\mathcal{G}_1}$	$\bar{p}_{\Gamma,\mathcal{G}_2}$	$\bar{p}_{\Gamma,\mathcal{G}_3}$	$\bar{p}_{\Gamma,\mathcal{G}_4}$	$\bar{p}_{\Gamma,\mathcal{G}_5}$	$\bar{p}_{\Gamma\wedge}$ ($\tilde{\alpha}=0.2$)	Simes' p-value
1	$1.2\cdot10^{-13}$	$5.5\cdot10^{-10}$	0	$6.3\cdot10^{-11}$	$2.2\cdot10^{-16}$	0	0
1.5	$3.0\cdot10^{-6}$	$2.2\cdot10^{-6}$	$4.4\cdot10^{-8}$	$1.1\cdot10^{-4}$	$5.3\cdot10^{-7}$	0	$2.2\cdot10^{-7}$
2	0.004	$1.4\cdot10^{-4}$	0.001	**0.032**	0.004	$1.7\cdot10^{-9}$	0.001
2.2	**0.02**	$4.1\cdot10^{-4}$	**0.01**	0.105	**0.024**	$1.4\cdot10^{-6}$	0.002
2.5	0.105	0.002	0.092	0.325	0.151	0.002	0.008
2.6	0.157	0.002	0.153	0.417	0.228	**0.012**	0.011
2.8	0.293	0.004	0.324	0.6	0.421	0.136	0.021
3	0.455	0.008	0.53	0.753	0.622	0.176	**0.038**
3.5	0.802	**0.023**	0.896	0.951	0.926	0.287	0.114
4	0.954	0.051	0.989	0.994	0.993	0.368	0.256

At bias level Γ, $\bar{p}_{\Gamma,\mathcal{G}_g}$ is the maximum p-value for group $\mathcal{G}_g$. $\bar{p}_{\Gamma\wedge}$ is the maximum p-value for the global test using the truncated product method of combining maximum p-values of the individual groups and Simes' p-value is the p-value for the global hypothesis using Simes' method. Values less than $9\cdot10^{-17}$ are rounded down to 0. In each column, largest p-value less than 0.05 is in bold.

$\{\bar{p}_{\Gamma_g,\mathcal{G}_g} : g = 1,2,\ldots,G\}$ conditional on $(\mathcal{F},\mathcal{Z},\mathcal{I})$ is an independent collection by Lemma 1 and by Lemma 2 each of $\bar{p}_{\Gamma_g,\mathcal{G}_g}$ is stochastically larger than the uniform distribution on $[0,1]$. Then conditional on $(\mathcal{F},\mathcal{Z},\mathcal{I})$, the validity of Simes' p-value follows from Simes (1986).

The last column of Table 4 reports the Simes' p-values. When $\Gamma_g = 2$ for all groups Simes' p-value is 1.0×10^{-3}. Using Simes' p-values the decision to reject the global hypothesis is sensitive for hidden bias levels of $\Gamma_g = 3.5$ for each group. Thus, consideration of effect modification has provided evidence that is much less sensitive to bias than not considering effect modification, $\Gamma = 3.5$ compared to $\Gamma = 2.6$ (from Section 1.1).

Now we consider testing the null hypothesis of no treatment effect for each group. Table 4 shows the maximum possible values of the p-values at different levels of hidden bias. Under the assumption of no hidden bias for each of the groups we have p-values 1.2×10^{-13}, 5.5×10^{-10}, $< 9.99\times10^{-16}$ (reported as 0), 6.3×10^{-11} and 2.2×10^{-16} supporting the hypothesis that smoking causes increase in lead level in the blood. These inferences individually are sensitive at different levels of hidden biases $2.5, 4, 2.5, 2.2$ and 2.5 for the five groups. This sensitivity analysis is not multiplicity corrected.

Table 5 present results of two multiplicity corrected analyses – the BH procedure that controls for the FDR and the closed testing procedure that controls for the FWER (Hsu et al., 2015). The BH procedure is less sensitive to bias than the closed testing procedure. For example, if the hidden bias of treatment assignment throughout the data is at most $\Gamma = 2.8$, then the BH procedure still finds effect of smoking causing increase in lead level in the blood for female between age 41 and 47 years ($\mathcal{G}_2$) while the closed testing procedure fails to reject the null hypotheses for all the groups. In closed testing, the smallest Γ for which all five of the hypotheses are sensitive to bias is $\Gamma = 2.6$ compared to $\Gamma = 3.5$ for the BH procedure.

TABLE 5

Adjusted p-values for closed testing procedure and adjusted p-values for the BH procedure for different level of hidden biases. Adjusted p-values are compared to the nominal level 0.05. *In each column for each comparison, largest p-value less than 0.05 is in bold.*

Γ	$\bar{p}_{\Gamma,\mathcal{G}_1}$	$\bar{p}_{\Gamma,\mathcal{G}_2}$	$\bar{p}_{\Gamma,\mathcal{G}_3}$	$\bar{p}_{\Gamma,\mathcal{G}_4}$	$\bar{p}_{\Gamma,\mathcal{G}_5}$
BH adjusted p-values					
1	$1.93 \cdot 10^{-13}$	$5.48 \cdot 10^{-10}$	0	$7.88 \cdot 10^{-11}$	$5.55 \cdot 10^{-16}$
1.5	$3.72 \cdot 10^{-6}$	$3.67 \cdot 10^{-6}$	$2.22 \cdot 10^{-7}$	$1.09 \cdot 10^{-4}$	$1.32 \cdot 10^{-6}$
2	0.005	0.001	0.003	**0.032**	0.005
2.2	**0.03**	0.002	**0.026**	0.105	**0.03**
2.5	0.175	0.008	0.175	0.325	0.188
2.6	0.262	0.011	0.262	0.417	0.285
2.8	0.527	0.021	0.527	0.6	0.527
3	0.753	**0.038**	0.753	0.753	0.753
3.5	0.951	0.114	0.951	0.951	0.951
4	0.994	0.256	0.994	0.994	0.994
Adjusted p-values for closed testing procedure					
1	$1.16 \cdot 10^{-13}$	$5.48 \cdot 10^{-10}$	0	$6.30 \cdot 10^{-11}$	$2.22 \cdot 10^{-16}$
1.5	$2.97 \cdot 10^{-6}$	$2.20 \cdot 10^{-6}$	$4.44 \cdot 10^{-8}$	$1.09 \cdot 10^{-4}$	$5.28 \cdot 10^{-7}$
2	0.004	0.001	0.001	**0.032**	0.004
2.2	**0.028**	0.005	**0.01**	0.105	**0.024**
2.5	0.224	**0.048**	0.094	0.325	0.178
2.6	0.343	0.105	0.177	0.417	0.295
2.8	0.623	0.359	0.501	0.623	0.623
3	0.867	0.68	0.832	0.867	0.867
3.5	0.999	0.996	0.999	0.999	0.999
4	1	1	1	1	1

6. Discussion

In this paper we investigated a sensitivity analysis with false discovery rate control for testing the hypothesis of no treatment effect under possible effect modification. Effect modification is the correlation of magnitude of treatment effect with pre-treatment covariates. Consideration of effect modification in practice leads us to make stronger conclusions about treatment effect. In the absence of prior knowledge about what covariates to consider as potential effect modifiers, we learn about what potential effect modifiers to test from the data. Our main theoretical result says that under appropriate restriction on the grouping method we can guarantee FDR control. In our simulation studies and data analysis, we have used the CART algorithm to construct groups from matched treatment-control pairs from the data. The interpretability of the regression tree makes it a promising choice. There have been new suggestions for interpretable classifiers (see e.g., Letham et al. (2015)). Such algorithms can also be used in practice. In the presence of effect modification one can consider different levels of hidden biases for different groups. In our simulations the FDR controlling method shows more power compared to closed testing. This is to be expected, since typically FDR controlling procedures have greater power than FWER controlling procedures. When there are many different levels of effect modification

leading to a large number of groups, FDR controlling procedures can have a large power advantage compared to FWER controlling procedures.

The method discussed in this article can potentially be applied to various situations involving historical controls. For example, the method can assist in planning a phase III clinical trial, based on single-arm phase II trial results. After a preliminary safety assessment of a new treatment in phase II, data is collected to understand the effectiveness of the treatment before conducting a full scale Phase III trial. Often phase II trials are single-arm studies with no control group and patients receiving the treatment are compared to historical controls. A subgroup analysis with sensitivity analysis in phase II can be a step towards planning a phase III trial. The sensitivity analysis discussed in this article would point to signals for the groups of patients for which treatment has possibly different levels of effects or no beneficial effect at all. These signals can be used to introduce blocking variables or develop enrichment strategies in designing a Phase III trial (Freidlin and Korn, 2014). In another application, the method can enhance various comparison studies to historical data. For example, Sammarco et al. (2016) used historical control data from NHANES database to assess the harmful effect of exposure to crude oil on petroleum hydrocarbon concentrations in the blood. The treatment group consisted of people who came in contact with crude oil in 2010 BP/Deepwater Horizon oil spill. The poor power of the study, particularly for long-chain hydrocarbons, can be improved by considering a subgroup based sensitivity analysis as person-to-person variability in the effects of exposure to crude oil may be high and may be affected by age, background and smoking pattern.

Appendix: Proof of Theorem 1

Proof. Let $C_r^{(-\mathcal{G}_g)}$ be the event that r groups are rejected along with $\mathcal{G}_g$. The proof follows that of Benjamini and Yekutieli (2001). The FDR is

$$E\left(\frac{\sum_{g=1}^{G}\chi(\mathcal{G}_g\subseteq\mathcal{G}_0,D_g=1)}{max(\sum_{g=1}^{G}\chi(D_g=1),1)}\Big|\mathcal{F},\mathcal{Z},\mathcal{I}\right)$$
$$=\sum_{g=1}^{G}\sum_{r=1}^{G}\frac{1}{r}\chi(\mathcal{G}_g\subseteq\mathcal{G}_0)Pr\Big(D_g=1,\sum_{k=1}^{G}\chi(D_k=1)=r\Big|\mathcal{F},\mathcal{Z},\mathcal{I}\Big).$$

We can write the right hand side of the identity above as,

$$=\sum_{g=1}^{G}\chi(\mathcal{G}_g\subseteq\mathcal{G}_0)\sum_{r=1}^{G}\frac{1}{r}Pr\Big(\bar{p}_{\Gamma,\mathcal{G}_g}\leq rq/G;C_r^{(-\mathcal{G}_g)}\Big|\mathcal{F},\mathcal{Z},\mathcal{I}\Big)$$
$$=\sum_{g=1}^{G}\chi(\mathcal{G}_g\subseteq\mathcal{G}_0)\sum_{r=1}^{G}\frac{1}{r}Pr\Big(\bar{p}_{\Gamma,\mathcal{G}_g}\leq\frac{rq}{G}\Big|\mathcal{F},\mathcal{Z},\mathcal{I}\Big)Pr\Big(C_r^{(-\mathcal{G}_g)}\Big|\mathcal{F},\mathcal{Z},\mathcal{I}\Big)$$
$$\leq q/G\sum_{g=1}^{G}\chi(\mathcal{G}_g\subseteq\mathcal{G}_0)\sum_{r=1}^{G}Pr\Big(C_r^{(-\mathcal{G}_g)}\Big|\mathcal{F},\mathcal{Z},\mathcal{I}\Big)$$

$$= q/G \sum_{g=1}^{G} \chi(\mathcal{G}_g \subseteq \mathcal{G}_0) \times 1 \leq q.$$

The second identity follows from Lemma 1. The first inequality follows from Lemma 2 and the fact that $\bar{p}_{\Gamma\mathcal{G}_g}$ being a upper bound on the true p-value is stochastically larger than uniform distribution on [0,1], i.e., $Pr(\bar{p}_{\Gamma\mathcal{G}_g} \leq a) \leq Pr(p_{\Gamma\mathcal{G}_g} \leq a) \leq a$. The final identity follows since $C_r^{(-\mathcal{G}_g)}$ for $r = 1, 2, \ldots, G$ are disjoint event, thus we have $\sum_{r=1}^{G} Pr(C_r^{(-\mathcal{G}_g)}|\mathcal{F}, \mathcal{Z}, \mathcal{I}) = Pr(\cup_{r=1}^{G} C_r^{(-\mathcal{G}_g)}|\mathcal{F}, \mathcal{Z}, \mathcal{I}) = 1$. Thus the theorem is proved. □

References

Aakvik, A. (2001). Bounding a Matching Estimator: The Case of a Norwegian Training Program. *Oxford Bulletin of Economics and Statistics* **63** 115–143.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 289–300. MR1325392

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

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. MR726392 (86b:62101)

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (2009). Smoking and lung cancer: recent evidence and a discussion of some questions. *International Journal of Epidemiology* **38** 1175–1191.

Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd, Edinburgh.

Freidlin, B. and Korn, E. L. (2014). Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nature Reviews Clinical Oncology* **11** 81–90.

Gastwirth, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in title vii cases to omitted variables. *Jurimetrics* **33** 19–34.

Gemenisa, K. and Rosemab, M. (2014). Voting Advice Applications and electoral turnout. *Electoral Studies* **36** 281–289.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 499–517. MR1924303

Glickman, M. E., Rao, S. R. and Schultz, M. R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology* **67** 850–857.

Handorf, E. A., Bekelman, J. E., Heitjan, D. F. and Mitra, N. (2013). Evaluating costs with unmeasured confounding: A sensitivity analysis for the treatment effect. *Ann. Appl. Stat.* **7** 2062–2080.

Hansen, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News* **7** 18–24.

Hansen, B. B., Rosenbaum, P. R. and Small, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association* **109** 133–144. MR3180552

Hosman, C. A., Hansen, B. B. and Holland, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. MR2758424

Hsu, J. Y. and Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69** 803–811. MR3146776

Hsu, J. Y., Small, D. S. and Rosenbaum, P. R. (2013). Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association* **108** 135–148. MR3174608

Hsu, J. Y., Zubizarreta, J. R., Small, D. S. and Rosenbaum, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552

Huber, P. (1981). *Robust Statistics*. New York: Wiley. MR0606374

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review* **93** 126–132.

Jogdeo, K. (1977). Association and Probability Inequalities. *Ann. Statist.* **5** 495–504. MR0448703

Keele, L. and Minozzi, W. (2013). How Much Is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data. *Political Analysis* **21** 193–216.

Letham, B., Rudin, C., McCormick, T. H. and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.* **9** 1350–1371. MR3418726

Liu, W., Kuramoto, J. and Stuart, E. (2013). Sensitivity Analysis for Unobserved Confounding in Nonexperimental Prevention Research. *Prevention Science* **14** 570–580.

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

Maritz, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166. MR0529161

McCandless, L. C., Gustafson, P. and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine* **26** 2331–2347. MR2368419

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. Translated and edited by D. M. Dabrowska and T. P. Speed. *Statist. Sci.* **5** 465–472. MR1092986

Pimentel, S. D., Yoon, F. and Keele, L. (2016). Variable ratio matching with fine balance in a study of the Peer Health Exchange. *Statistics in medicine* **34** 4070–4082. MR3431322

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915

ROSENBAUM, P. R. (2002). Observational Studies. In *Observational Studies. Springer Series in Statistics* Springer, New York. MR1899138

ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. MR2050466

ROSENBAUM, P. R. (2007a). Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics* **63** 1164–1171. MR2414594

ROSENBAUM, P. R. (2007b). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* **63** 456–464. MR2370804

ROSENBAUM, P. R. (2010). Observational Studies. In *Design of Observational Studies. Springer Series in Statistics* Springer, New York. MR2561612

ROSENBAUM, P. R. (2011). A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics* **67** 1017–1027. MR2829236

ROSENBAUM, P. R. (2014). Weighted m-statistics with superior design sensitivity in matched observational studies with multiple controls. *Journal of the American Statistical Association* **109** 1145–1158. MR3265687

ROSENBAUM, P. R. (2015). Two r packages for sensitivity analysis in observational studies. *Observational Studies* **1** 1–17. MR1353914

ROSENBAUM, P. R. (2017). The general structure of evidence factors in observational studies. *Statistical Science* **32** 514–530. MR3730520

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

SAMMARCO, P., KOLIAN, S., WARBY, R., BOULDIN, J., SUBRA, W. and PORTER, S. (2016). Concentrations in human blood of petroleum hydrocarbons associated with the BP/Deepwater Horizon oil spill, Gulf of Mexico. *Archives of Toxicology* **90** 829–837.

SIMES, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. MR0897872

WANG, L. and KRIEGER, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statistics in Medicine* **25** 2257–2271. MR2240099

YANAGAWA, T. (1984). Case-control studies: assessing the effect of a confounding factor. *Biometrika* **71** 191–194. MR0738341

YU, B. and GASTWIRTH, J. L. (2005). Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics* **6** 201–209.

ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. and WEIR, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology* **22** 170–185.

ZUBIZARRETA, J. R., NEUMAN, M., SILBER, J. H. and ROSENBAUM, P. R. (2012). Contrasting Evidence Within and Between Institutions That Provide Treatment in an Observational Study of Alternate Forms of Anesthesia. *Journal of the American Statistical Association* **107** 901–915. MR3010879