Biometrika (2015), **102**, 4, *pp.* 767–782
Printed in Great Britain
doi: 10.1093/biomet/asv034
Advance Access publication 18 September 2015

Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods

By JESSE Y. HSU

Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
hsu9@mail.med.upenn.edu

JOSÉ R. ZUBIZARRETA

Division of Decision, Risk and Operations, Columbia University, New York, New York 10027, U.S.A.
zubizarreta@columbia.edu

DYLAN S. SMALL and PAUL R. ROSENBAUM

Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
dsmall@wharton.upenn.edu rosenbaum@wharton.upenn.edu

Summary

An effect modifier is a pretreatment covariate that affects the magnitude of the treatment effect or its stability. When there is effect modification, an overall test that ignores an effect modifier may be more sensitive to unmeasured bias than a test that combines results from subgroups defined by the effect modifier. If there is effect modification, one would like to identify specific subgroups for which there is evidence of effect that is insensitive to small or moderate biases. In this paper, we propose an exploratory method for discovering effect modification, and combine it with a confirmatory method of simultaneous inference that strongly controls the familywise error rate in a sensitivity analysis, despite the fact that the groups being compared are defined empirically. A new form of matching, strength-k matching, permits a search through more than k covariates for effect modifiers, in such a way that no pairs are lost, provided that at most k covariates are selected to group the pairs. In a strength-k match, each set of k covariates is exactly balanced, although a set of more than k covariates may exhibit imbalance. We apply the proposed method to study the effects of the earthquake that struck Chile in 2010.

Some key words: Closed testing; Design sensitivity; Optimal matching; Sensitivity analysis; Truncated product of p-values.

1. Introduction

1·1. *Subgroups suggested by the data*

In experiments and observational studies, effect modification refers to the possibility that the magnitude or stability of a treatment effect changes with the observed covariates. There is

effect modification with pairs matched to have the same value of observed covariates, x, if the treated-minus-control pair difference in outcomes Y varies with x in magnitude or stability. For instance, in an observational study of a treatment intended to reduce the incidence of malaria by controlling mosquitoes, Hsu et al. (2013) examined treatment-control pairs matched for age and gender. The outcome was the level of malaria parasites found in the blood, and the benefits of the treatment were found to be much greater for young children than for adults.

In observational studies, the magnitude and stability of a treatment effect strongly influence its sensitivity to biases from unmeasured covariates. Small and unstable treatment effects can often plausibly be explained away as being created by small biases in the assignment of individuals to treatment or control groups, whereas large and stable treatment effects can only be explained as noncausal if the unmeasured biases are large (Rosenbaum, 2004, 2005, 2010). If there is effect modification, then for some values of x the study may be sensitive to small biases while for other values of x it may be insensitive to moderately large biases.

Hsu et al. (2013) proposed a method for empirically identifying a few promising subgroups of pairs based on a multivariate x, and then testing the null hypothesis of no treatment effect by pooling evidence from the subgroups. They used the data twice: once to build promising groups, and again to test the null hypothesis H_0 of no treatment effect and examine the sensitivity of that test to unmeasured biases. They discovered a few promising subgroups by a tree-based regression of $|Y|$ on multivariate x, and showed that discovering the groups in this way invalidated neither randomization tests nor sensitivity analyses for these tests, essentially because the signs of the outcomes Y had not been used. They showed further that this method of testing H_0 is almost as effective as knowing a priori which subgroup would be insensitive to unmeasured biases; more precisely, they showed that the overall pooled test of H_0 has the largest design sensitivity of the several subgroup tests.

Hsu et al. (2013) tested the one null hypothesis H_0 of no treatment effect at all, but they did this by pooling evidence from subgroups. If H_0 is rejected, then it is natural to ask which subgroups were affected. Although this question sounds natural, it is quite unlike conventional problems of simultaneous inference for several subgroups specified a priori. With subgroups discovered by an analysis of the data, a different assignment of treatments would have produced different subgroups, hence different null hypotheses, and perhaps even a different number of null hypotheses. What does it even mean to speak of the probability of falsely rejecting a true null hypothesis H^* if H^* would not have been tested had randomization selected a different treatment assignment? Our main goals in the present paper are to make sense of the idea of testing hypotheses that vary from one dataset to the next, to propose a method of strong control for the error rate in such tests, to propose a new form of covariate balance that facilitates studying effect modification, and to integrate these ideas with sensitivity analyses in observational studies.

1·2. *Example: the Chilean earthquake*

On 27 February 2010, a powerful earthquake of magnitude 8·8 struck off the coast of Chile. What effects did the earthquake have on the labour market in Chile? In this paper, we take the outcome to be the change in individual work income from before to after the earthquake. Using the data described in Zubizarreta et al. (2013), we constructed 2106 matched pairs of individuals, one in a severely shaken region of Chile and the other remote from the earthquake. The matching controlled for covariates measured before the earthquake: sex, marital status, number of persons in household, self-reported health problems, self-reported health perception, quartile of work income, age, self-reported psychological problems, disability, health insurance status, years of education, employment status, per capita household income, poverty status, housing status, quality of housing structure, and overcrowding. Some innovations in matching methodology

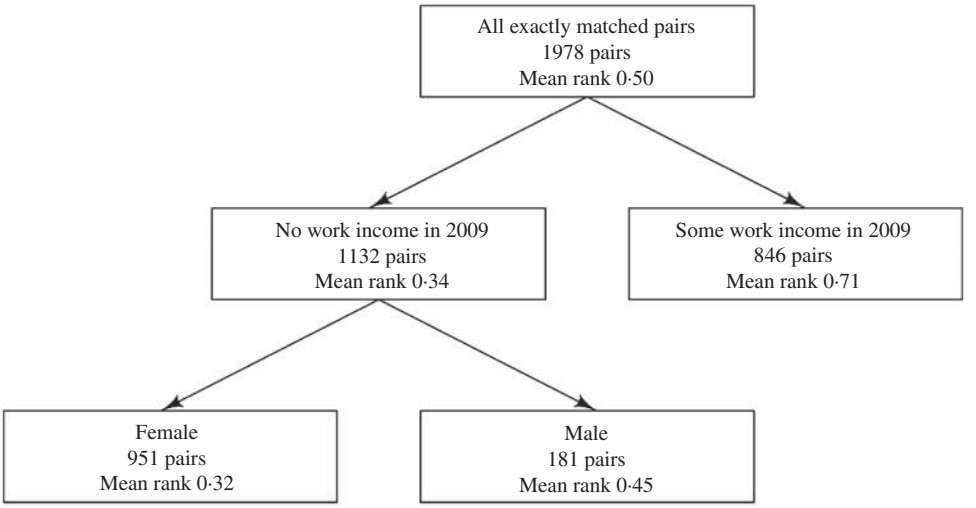

Fig. 1. Regression tree built from the ranks of the absolute differences $|Y_i|$ in changes in work income predicted from x_i. The ranks were divided by 1978, so that they fall in [0, 1]. The tree was constructed using the 1978 pairs that were exactly matched for the basic covariates, and it produced three groups: all people with positive work income, females with zero work income, and males with zero work income.

used here are described in § 6. Although matched individuals were similar in the ways just mentioned, people and jobs in regions remote from the earthquake may differ in other ways from those in the severely shaken centre of Chile.

We considered six covariates as possible effect modifiers: gender, male or female; health problems, yes or no; self-rated health, poor, fair, or good; quartile of individual work income in 2009; number of persons in household, 1, 2, 3, 4 or 5, or 6 or more; and marital status, married/cohabiting or other. Because most people, especially many women and elderly individuals, did not have individual work income in 2009, the quartiles of work income defined only three groups. The six basic covariates define $2 \times 2 \times 3 \times 3 \times 5 \times 2 = 360$ types of individuals. A total of $I = 2106$ matched pairs were formed, so many of the 360 types are represented by only a few pairs. Here, Y is the exposed-minus-control difference in the after-minus-before change in work income. The regression tree method of Breiman et al. (1984) was used to predict the rank of $|Y|$ from x, and reduced the 360 types of individuals to the three subgroups shown in Fig. 1, defined in terms of gender and work income prior to the earthquake: all people with positive work income, females with zero work income, and males with zero work income. The tree with three subgroups in Fig. 1 was constructed using only the 1978 pairs that are exactly matched for the basic covariates, but all 2106 pairs will be used in making inferences about treatment effects; see § 6 for details. Given that we discovered these three subgroups by using the current data, what can be said about which subgroups were affected by the earthquake?

2. Notation and review

2·1. *Notation*

There are I matched pairs, indexed by $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one in the treatment group with $Z_{ij} = 1$ and the other in the control group with $Z_{ij} = 0$, so that $Z_{i1} + Z_{i2} = 1$ for each i. Pairs are matched for an observed covariate x, so that $x_{i1} = x_{i2}$ for all i; but we assume that there is concern about an unmeasured covariate u not controlled by the matching,

where $0 \leqslant u \leqslant 1$, so that quite possibly $u_{i1} \neq u_{i2}$ for many or all i. Write $Z = (Z_{11}, \ldots, Z_{I2})^{\mathrm{T}}$ for the $2I$-dimensional vector containing the Z_{ij}, and write $\mathcal{Z}$ for the set containing the 2^I possible values z of Z, so $z \in \mathcal{Z}$ if $z = (z_{11}, \ldots, z_{I2})^{\mathrm{T}}$ with $z_{i1} + z_{i2} = 1$ and $z_{ij} = 0$ or $z_{ij} = 1$ for each (i, j). Conditioning on the event $Z \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. Write $|\mathcal{A}|$ for the number of elements of a finite set $\mathcal{A}$; for instance, $|\mathcal{Z}| = 2^I$.

Each subject has a potential response $r_{\mathrm{T}ij}$ if treated with $Z_{ij} = 1$, a potential response $r_{\mathrm{C}ij}$ if assigned to control with $Z_{ij} = 0$, and an observed response $R_{ij} = Z_{ij} r_{\mathrm{T}ij} + (1 - Z_{ij}) r_{\mathrm{C}ij}$ under the treatment actually received, whereas the effect of the treatment, namely $r_{\mathrm{T}ij} - r_{\mathrm{C}ij}$, is not observed for any subject; see Neyman (1923) and Rubin (1974). The sharp null hypothesis of no treatment effect (Fisher, 1935) is $H_0 : r_{\mathrm{T}ij} = r_{\mathrm{C}ij}$ for all i, j. Importantly, if H_0 is true, then $R_{ij} = r_{\mathrm{C}ij}$ does not change with the treatment assignment Z_{ij}, but if H_0 is false, then at least some of the R_{ij} do change with Z_{ij}. Write $\mathcal{F} = \{(r_{\mathrm{T}ij}, r_{\mathrm{C}ij}, x_{ij}, u_{ij}) : i = 1, \ldots, I;\ j = 1, 2\}$. The treated-minus-control pair difference in observed responses for pair i is $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, and it equals $(Z_{i1} - Z_{i2})(r_{\mathrm{C}i1} - r_{\mathrm{C}i2})$ if H_0 is true. Also, write $r_{\mathrm{C}} = (r_{\mathrm{C}11}, r_{\mathrm{C}12}, \ldots, r_{\mathrm{C}I2})^{\mathrm{T}}$ and $R = (R_{11}, \ldots, R_{I2})^{\mathrm{T}}$ for the vectors of dimension $2I$, and write $Y = (Y_1, \ldots, Y_I)^{\mathrm{T}}$ for the vector of dimension I.

2·2. *Randomization inference in experiments*

In a paired randomized experiment, subjects are paired on the basis of observed covariates, x_{ij}, and then a fair coin is flipped independently I times to determine the treatment assignments Z_{i1} and $Z_{i2} = 1 - Z_{i1}$; that is, $\mathrm{pr}(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = 1/2$ for each (i, j) and $\mathrm{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) = 2^{-I}$ for each $z \in \mathcal{Z}$. The null distribution of a test statistic $t(Z, R)$ under Fisher's H_0 is its permutation distribution, namely

$$\mathrm{pr}\{t(Z, R) \geqslant k \mid \mathcal{F}, \mathcal{Z}\} = \mathrm{pr}\{t(Z, r_{\mathrm{C}}) \geqslant k \mid \mathcal{F}, \mathcal{Z}\} = \frac{|\{z \in \mathcal{Z} : t(z, r_{\mathrm{C}}) \geqslant k\}|}{|\mathcal{Z}|}, \tag{1}$$

because $R = r_{\mathrm{C}}$ if H_0 is true, where r_{C} is fixed by conditioning on $\mathcal{F}$, and the distribution of Z is uniform on $\mathcal{Z}$ in a randomized experiment. For instance, if $t(Z, R)$ is Wilcoxon's signed rank statistic, then (1) would be its usual exact null distribution.

Similarly, Maritz (1979) tested H_0 using (1) and a suitably defined M-statistic, the quantity that is equated to zero in defining Huber's M-estimates, specifically, $t(Z, R) = \sum_{i=1}^{I} \psi(Y_i/a)$, where a is a quantile of the $|Y_i|$ and $\psi(\cdot)$ is a monotone increasing odd function, i.e., $\psi(d) = -\psi(-d)$, so $\psi(0) = 0$. We will use a version of Maritz's test statistic to analyse the earthquake data in § 4. Under H_0, the pair difference is $Y_i = (Z_{i1} - Z_{i2})(r_{\mathrm{C}i1} - r_{\mathrm{C}i2}) = \pm|r_{\mathrm{C}i1} - r_{\mathrm{C}i2}|$, so $|Y_i| = |r_{\mathrm{C}i1} - r_{\mathrm{C}i2}|$ is fixed by conditioning on $\mathcal{F}$ in (1), and hence a is also fixed; then, $t(Z, R) = \sum_{i=1}^{I} q_i \,\mathrm{sign}(Y_i)$, where $q_i = \psi(|r_{\mathrm{C}i1} - r_{\mathrm{C}i2}|/a)$ is fixed by conditioning on $\mathcal{F}$ and $\mathrm{sign}(Y_i) = 1$, 0 or -1 according to whether $Y_i > 0$, $Y_i = 0$ or $Y_i < 0$, respectively. As a consequence, under H_0, (1) is the distribution of the sum of I independent random variables taking the values $\pm\psi(|r_{\mathrm{C}i1} - r_{\mathrm{C}i2}|/a)$ with equal probabilities 1/2 if $|r_{\mathrm{C}i1} - r_{\mathrm{C}i2}| > 0$ or taking the value 0 with probability 1 if $|r_{\mathrm{C}i1} - r_{\mathrm{C}i2}| = 0$.

2·3. *Sensitivity analysis in observational studies*

A sensitivity analysis in an observational study supposes that, in the population prior to matching, individuals are independently assigned to the treatment or control groups with unknown probabilities, $\pi_{ij} = \mathrm{pr}(Z_{ij} = 1 \mid \mathcal{F})$, that may depend on both the observed covariates x_{ij} and the unobserved covariate u_{ij} as recorded in $\mathcal{F}$. The sensitivity analysis assumes that two subjects ij and $i'j'$ with the same observed covariates, $x_{ij} = x_{i'j'}$, may differ in their odds of treatment by

at most a factor of $\Gamma \geqslant 1$; that is, $\Gamma^{-1} \leqslant \pi_{ij}(1-\pi_{i'j'})/\{\pi_{i'j'}(1-\pi_{ij})\} \leqslant \Gamma$. It is easy to show that this is equivalent to assuming that $\log\{\pi_{ij}/(1-\pi_{ij})\} = \kappa(x_{ij}) + \gamma u_{ij}$ with $\gamma = \log(\Gamma)$ and $0 \leqslant u_{ij} \leqslant 1$ for some unknown function $\kappa(\cdot)$; see Rosenbaum (2002, § 4), where the proof consists in constructing u_{ij} from π_{ij} and conversely. Therefore, as u_{ij} is effectively a transformation of π_{ij}, u_{ij} may reflect the combined impact on treatment assignment of several unmeasured variables. The distribution of Z is then restricted to $\mathcal{Z}$ by conditioning on $Z \in \mathcal{Z}$. If pairs are matched for observed covariates x_{ij} so that $\kappa(x_{i1}) = \kappa(x_{i2})$, then $\text{pr}(Z_{i1} = 1 \mid \mathcal{F}, Z_{i1} + Z_{i2} = 1) = \exp(\gamma u_{i1})/\{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})\}$ and, for $z \in \mathcal{Z}$,

$$\text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) = \prod_{i=1}^{I} \frac{z_{i1}\exp(\gamma u_{i1}) + z_{i2}\exp(\gamma u_{i2})}{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})} = \frac{\exp(\gamma z^{\mathrm{T}} u)}{\sum_{b \in \mathcal{Z}} \exp(\gamma b^{\mathrm{T}} u)} \tag{2}$$

for some $u = (u_{11}, \ldots, u_{I2})^{\mathrm{T}} \in \mathcal{U} = [0,1]^{2I}$, where $\mathcal{U}$ is the $2I$-dimensional unit cube. When $\Gamma = 1$ so that $\gamma = 0$, expression (2) equals the randomization distribution, $\text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) = 2^{-I}$. Using (2), if γ and u were known, then under H_0 the distribution of the test statistic $T = t(Z, R) = t(Z, r_{\mathrm{C}})$ would be the sum of the probabilities in (2) over the z in $\{z \in \mathcal{Z} : t(z, r_{\mathrm{C}}) \geqslant k\}$. The sensitivity analysis asks: how large a departure Γ from randomization must be present to materially alter inferences based on the naive model which claims that adjustments for observed covariates x_{ij} suffice to remove all bias? Each value of $\Gamma \geqslant 1$ yields an interval of possible p-values or point estimates or endpoints of confidence intervals, and the question is: how large must Γ be if this interval is to be so long as to be uninformative, say permitting both acceptance and rejection of H_0?

The present paper considers analyses of subsets of the I pairs. Let $s \subseteq \{1, \ldots, I\}$ be a fixed nonempty subset of the I pairs, with $|s| \geqslant 1$. Much of our concern later on will be with sets of pairs selected on the basis of the data, but the complications introduced by a data-dependent set of pairs are deferred to § 2·4 and after. In the current paragraph, s is a set of pairs determined a priori, such as in a planned subgroup analysis for pairs of women. A vector Z appended with a subscript s, as in Z_s, means the vector of dimension $2 \times |s|$ containing those coordinates of Z that correspond to pairs $i \in s$. Notation such as R_s, $\mathcal{F}_s$ and $\mathcal{U}_s$ is interpreted similarly; moreover, H_{0s} denotes the hypothesis of no treatment effect for all pairs $i \in s$, that is, $H_{0s} : r_{\mathrm{T}ij} = r_{\mathrm{C}ij}$ for $i \in s$ and $j = 1, 2$.

If, as in § 2·2, the test statistic is of the form $T_s = t(Z_s, R_s) = \sum_{i \in s} q_{si}\,\text{sign}(Y_i)$ where $q_{si} \geqslant 0$ is a function of $\mathcal{F}_s$ under H_{0s}, then T_s is a function of aspects of just the pairs in s, and we can provide a sharp bound on the distribution of interest, $\text{pr}(T_s \geqslant k \mid \mathcal{F}, \mathcal{Z})$. Define $\bar{T}_{\Gamma s}$ to be the sum of $|s|$ independent random variables, the ith random variable being q_{si} with probability $\Gamma/(1+\Gamma)$ and $-q_{si}$ with probability $1/(1+\Gamma)$, provided $q_{si} > 0$; otherwise, the ith random variable is 0 with probability 1 if $q_{si} = 0$. Define $\tilde{T}_{\Gamma s}$ analogously but with $\Gamma/(1+\Gamma)$ and $1/(1+\Gamma)$ interchanged. Then it is not difficult to show that for each fixed $\Gamma = \exp(\gamma)$, as u_s ranges over $\mathcal{U}_s$, the unknown distribution $\text{pr}(T_s \geqslant k \mid \mathcal{F}, \mathcal{Z})$ of T_s under H_{0s} and (2) is sharply bounded by two known distributions,

$$\text{pr}(\tilde{T}_{\Gamma s} \geqslant k \mid \mathcal{F}, \mathcal{Z}) \leqslant \text{pr}(T_s \geqslant k \mid \mathcal{F}, \mathcal{Z}) \leqslant \text{pr}(\bar{T}_{\Gamma s} \geqslant k \mid \mathcal{F}, \mathcal{Z}); \tag{3}$$

see Rosenbaum (2002 § 4, 2007, 2015). When $0 = \gamma = \log(\Gamma)$, there is equality in (3), and both bounds in (3) equal the randomization distribution (1). The bounds in (3) are sharp, being attained for particular u_s in $\mathcal{U}_s$, so they cannot be improved except with additional information about the unobserved u_s. The bounds (3) yield bounds on p-values, point estimates and confidence intervals. An unobserved covariate u that produces a Δ-fold increase in the odds of a positive

outcome difference, $Y_i > 0$, and a Λ-fold increase in the odds of treatment, $Z_{i1} - Z_{i2} > 0$, is the same as one with $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$ (Rosenbaum & Silber, 2009a).

Other methods of sensitivity analysis in observational studies are discussed by, for instance, Cornfield et al. (1959), Egleston et al. (2009), Gastwirth (1992), Hosman et al. (2010), Imbens (2003), Lin et al. (1998), Liu et al. (2013), McCandless et al. (2007), Wang & Krieger (2006), Yanagawa (1984) and Yu & Gastwirth (2005).

2·4. *Effect modifiers when testing for no effect*

Hsu et al. (2013, § 4) tested Fisher's null hypothesis H_0 of no effect by first dividing the pairs $i \in \{1, \ldots, I\}$ into several groups based on $x_{i1} = x_{i2} = x_i$, say, looking for possible effect modification, i.e., larger or more stable treatment effects in some groups than in others. More precisely, $G \geqslant 1$ mutually exclusive and exhaustive groups $\mathcal{G} = \{s_1, \ldots, s_G\}$ of the pairs $i = 1, \ldots, I$ were formed, where each $s_g \subseteq \{1, \ldots, I\}$. These groups are formed by regressing a function of the absolute differences $|Y_i|$ on x_i in some fashion that yields non-overlapping groups, for instance by using a regression tree as in Fig. 1, in which $G = 3$ groups of pairs were formed. Under Fisher's H_0, the absolute difference in responses $|Y_i| = |r_{Ci1} - r_{Ci2}|$ is fixed by conditioning on $\mathcal{F}$, as discussed in § 2·2, so the grouping produced by the regression of $|Y_i|$ on x_i is also fixed. Under model (2), when H_0 is true, a test statistic $T_s = t(Z_s, R_s) = \sum_{i \in s} q_{si} \operatorname{sign}(Y_i)$ for $s \in \mathcal{G}$ has the usual bounds on its null distribution, namely (3), because these bounds refer to the conditional distribution given $(\mathcal{F}, \mathcal{Z})$ when H_0 is true. In particular, in a randomized experiment, model (2) holds with $0 = \gamma = \log(\Gamma)$, so that under H_0, the group-specific statistic T_s has its usual randomization distribution despite the data-dependent nature of the groups $\mathcal{G}$.

There is a reason to hope that a grouping based on the regression of $|Y_i|$ on x_i will construct useful groups. If H_0 is false with $Y_i = \rho(x_i) + \xi_i$, where $\rho(\cdot) \geqslant 0$ and ξ_i are independent and identically distributed with a continuous unimodal distribution symmetric about zero, then $|Y_i|$ is stochastically larger than $|Y_{i'}|$ if $\rho(x_i) > \rho(x_{i'})$ (Jogdeo, 1977, Theorem 2.2). Therefore, the regression of $|Y_i|$ on x_i may form groups with different typical effects under this simple model.

Hsu et al. (2013, § 4) test H_0 by computing G different p-values of the form (1) or p-value bounds of the form (3), using just the pairs $i \in s_g$ $(g = 1, \ldots, G)$, and combining these p-values using a generalization of Fisher's method for combining independent p-values, namely the truncated product of p-values of Zaykin et al. (2002); see Rosenbaum (2015) for the R package sensitivitymv. The truncated product uses as its test statistic the product of those p-values that are no larger than a prespecified cut-off $\tilde{\alpha}$ with $0 < \tilde{\alpha} \leqslant 1$, and for $\tilde{\alpha} = 1$ it is equivalent to Fisher's procedure; see Benjamini & Heller (2008) for simultaneous inference using Fisher's procedure. Hsu et al. (2013) showed that in the presence of even a small amount of effect modification, this procedure has higher power in a sensitivity analysis and larger design sensitivity than a test that ignores the groups.

In large samples, the power of a sensitivity analysis is determined by the design sensitivity (Rosenbaum, 2004), and a formula for the design sensitivity of the M-test of Maritz (1979) is given in Rosenbaum (2013, Corollary 1). Other things being equal, the design sensitivity is larger, and hence the sensitivity analysis has greater power in large samples, when the effect is larger, for example if the typical Y_i is larger, or when the dispersion of the Y_i is smaller for a given typical size (Rosenbaum, 2004, 2010 § 15, 2013). Combining separate p-values within groups $\mathcal{G}$ can increase the power of a sensitivity analysis when either the size or the dispersion of the Y_i vary from group to group (Hsu et al., 2013, § 3.3).

So far, the discussion has focused on testing the null hypothesis H_0 of no effect at all, and that hypothesis played a key role in permitting the groups $\mathcal{G}$ to be determined from the data by regressing a function of $|Y_i|$ on x_i. A more interesting question not addressed by Hsu et al. (2013)

is whether subhypotheses H_{0s} of no effect for pairs $i \in s$ with $s \in \mathcal{G}$ can be tested using (3) when H_0 may be false. If H_0 is false, then there is at least one pair i for which $r_{Ti1} \neq r_{Ci1}$ or $r_{Ti2} \neq r_{Ci2}$ or both, and in this case $\mathcal{G}$ is not a function of $(\mathcal{F}, \mathcal{Z})$ because $R \neq r_C$ in the sense that r_C is determined by $\mathcal{F}$ but R varies with Z. If we reject the null hypothesis H_0 of no effect on anyone, it is not clear from the argument of this section that we can say anything about just one of the groups, e.g., about H_{0s}. If H_0 is false in a randomized experiment, then the grouping $\mathcal{G}$ depends on Z: had randomization yielded a different treatment assignment Z, it might easily have yielded different groups $\mathcal{G}$, and the hypothesis H_{0s} is not even a hypothesis in any conventional sense, because the hypotheses change as the treatment assignments Z change. This issue is central to the current paper and is explored in § 3.

3. Familywise error rate with groups constructed from the data

3·1. *Consequences of data-dependent grouping for null distributions*

To address the issue raised at the end of § 2·4, the following conditions are assumed to hold.

Condition 1. The distribution of Z given $(\mathcal{F}, \mathcal{Z})$ is (2) for a specific $\gamma = \log(\Gamma) \geqslant 0$ and an unknown $u \in \mathcal{U}$.

Condition 2. Mutually exclusive and exhaustive groups $\mathcal{G} = \{s_1, \ldots, s_G\}$ are formed as a function of $|Y_i|$ and $x_i = x_{i1} = x_{i2}$ $(i = 1, \ldots, I)$.

Condition 2 says that Z was not explicitly used in constructing the groups but that R was used. Here, G and $\mathcal{G}$ are random quantities given $\mathcal{F}$ and $\mathcal{Z}$ because H_0 may be false and, if so, $R \neq r_C$ depends on Z and hence the groups in Condition 2 may also depend indirectly on Z through the dependence of R on Z. If the groups $\mathcal{G}$ are random quantities depending indirectly on Z, then taking the groups to be fixed, conditioning on $\mathcal{G}$, may alter the distribution of Z. Propositions 1 and 2 say that this genuine problem can be kept under control.

Let $h \subseteq \{1, \ldots, I\}$ be the union of all the groups s_g for which there is no treatment effect, i.e., the union of those s_g such that $r_{Ti1} = r_{Ci1}$ and $r_{Ti2} = r_{Ci2}$ for all $i \in s_g$; possibly, $h = \emptyset$. Obviously, the investigator does not know h.

As $\mathcal{G}$ is a random quantity, h is also a random quantity because it is a union of some of the s_g. Indeed, the set h is a function of $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$. Conditionally given $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$, the set h is fixed. Conditionally given $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$, if $h = \emptyset$ then there are affected pairs in every group $s \in \mathcal{G}$, so every H_{0s} is false, and false rejection of a true H_{0s} cannot occur. Conversely, conditionally given $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$, if $h \neq \emptyset$ then some group or groups $s \in \mathcal{G}$ contain no affected individuals, and false rejection of a true H_{0s} is possible. Proposition 1 and its corollary concern the distribution of the test statistic $T_h = t(Z_h, R_h)$ given $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ where the pairs $i \in h$ are all unaffected by the treatment but the grouping $\mathcal{G}$ itself, and hence also h, may have been affected by the treatment. Stated informally, Proposition 1 says that the data-dependent grouping $\mathcal{G}$ does not alter the null distribution of T_h even when H_0 is false so that $\mathcal{G}$ changes with Z: the null distribution of T_h is still bounded by (3) with $s = h$. To emphasize this point, T_h is computed from the union h of all groups s_g for which there is no treatment effect, and because the investigator does not know h, she cannot know when she has computed T_h. Proposition 1 is a step in the development of a multiple inference procedure that strongly controls false rejections, as discussed in § 3·2. Although the proof of Proposition 1 requires some attention to detail, the idea is not difficult: Z affects the groups $\mathcal{G}$ only indirectly by affecting R, but for $i \in h$ the treatment assignment Z_{ij} does not affect R, so the distribution of Z_{ij} for $i \in h$ is unchanged by conditioning on $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$.

Proposition 1. *Assume Conditions 1 and 2. If $h \neq \emptyset$, then the conditional distribution* $\text{pr}(T_h \geqslant k \mid \mathcal{F}, \mathcal{Z}, \mathcal{G})$ *of* $T_h = t(Z_h, R_h)$ *given* $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ *is sharply bounded by the bounds in* (3) *with* $s = h$.

Proof. Assume $h \neq \emptyset$. Let $\mathcal{N} \subseteq \{1, \ldots, I\}$ be the set of pairs with no treatment effect, so $r_{Ti1} = r_{Ci1}$ and $r_{Ti2} = r_{Ci2}$ if and only if $i \in \mathcal{N}$; let $\mathcal{E} \subseteq \{1, \ldots, I\}$ be the complementary set of affected pairs. Of course, $\mathcal{N} \supseteq h \neq \emptyset$, so $\mathcal{N} \neq \emptyset$. Let z be a possible value of Z_h, so z is a $2|h|$-dimensional vector $z = (z_{11}, z_{12}, \ldots, z_{\ell j}, \ldots, z_{|h|,1}, z_{|h|,2})^{\mathrm{T}}$ with $z_{\ell j} = 1$ or $z_{\ell j} = 0$ and $z_{\ell 1} + z_{\ell 2} = 1$ for each ℓ. Write $\mathcal{D}$ for the combination of the data $\{(r_{Ci1}, r_{Ci2}, x_i), i \in \mathcal{N}\}$ and the data $\{(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, x_i), i \in \mathcal{E}\}$. Because pairs $i \in \mathcal{N}$ are unaffected, with $R_{ij} = r_{Cij}$ for $i \in \mathcal{N}$ and $j = 1, 2$, the grouping $\mathcal{G} = \{s_1, \ldots, s_G\}$ is a function of $\mathcal{D}$. Because the grouping $\mathcal{G}$ is a function of $\mathcal{D}$, conditioning on $(\mathcal{G}, \mathcal{D}, \mathcal{F}, \mathcal{Z})$ is the same as conditioning on $(\mathcal{D}, \mathcal{F}, \mathcal{Z})$. For $i \in \mathcal{E}$, the information in $(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, x_i)$ that is not in $(\mathcal{F}, \mathcal{Z})$ is precisely $Z_{i1} = 1 - Z_{i2}$ for $i \in \mathcal{E}$; that is, one could construct $(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, x_i)$ from $(\mathcal{F}, \mathcal{Z})$ if one were told Z_{i1}. Putting this all together under (2), the $Z_{i1} = 1 - Z_{i2}$ for $i \in \mathcal{N}$ satisfy

$$\begin{aligned}\text{pr}(Z_{i1} = 1 \mid \mathcal{F}, \mathcal{Z}, \mathcal{G}, \mathcal{D}) &= \text{pr}(Z_{i1} = 1 \mid \mathcal{F}, \mathcal{Z}, \mathcal{D}) \\ &= \text{pr}(Z_{i1} = 1 \mid \mathcal{F}, \mathcal{Z}) = \frac{\exp(\gamma u_{i1})}{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})}\end{aligned}$$

because (i) $\mathcal{G}$ is a function of $\mathcal{D}$, and (ii) the Z_{i1} for $i \in \mathcal{N}$ are conditionally independent of the $Z_{i'1}$ for $i' \in \mathcal{E}$ and, apart from $Z_{i'j}$ for $i' \in \mathcal{E}$, the rest of $(R_{i'1}, R_{i'2}, Z_{i'1}, Z_{i'2}, x_{i'})$ for $i' \in \mathcal{E}$ is already fixed by conditioning on $(\mathcal{F}, \mathcal{Z})$. Using (2) again together with the fact that h is fixed by conditioning on $(\mathcal{G}, \mathcal{F}, \mathcal{Z})$ yields

$$\text{pr}(Z_h = z \mid \mathcal{F}, \mathcal{Z}, \mathcal{G}, \mathcal{D}) = \prod_{\ell \in h} \frac{z_{\ell 1} \exp(\gamma u_{\ell 1}) + z_{\ell 2} \exp(\gamma u_{\ell 2})}{\exp(\gamma u_{\ell 1}) + \exp(\gamma u_{\ell 2})}. \tag{4}$$

The right-hand side of (4) depends on $(\mathcal{G}, \mathcal{F}, \mathcal{Z})$ because h depends on $(\mathcal{G}, \mathcal{F}, \mathcal{Z})$, but it does not depend on $\mathcal{D}$ given $(\mathcal{G}, \mathcal{F}, \mathcal{Z})$; therefore (4) equals $\text{pr}(Z_h = z \mid \mathcal{F}, \mathcal{Z}, \mathcal{G})$. It follows that the distribution of Z_h, namely $\text{pr}(Z_h = z \mid \mathcal{F}, \mathcal{Z}, \mathcal{G})$, and hence also the distribution $\text{pr}(T_h \geqslant k \mid \mathcal{F}, \mathcal{Z}, \mathcal{G})$, is identical to the distribution that produced the bounds in (3) with $s = h$, proving the result. □

Corollary 1. *Assume Condition 2. In a randomized experiment, if $h \neq \emptyset$, then the conditional distribution of* $T_h = t(Z_h, R_h)$ *given* $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ *is its randomization distribution, namely* (3) *with* $\gamma = 0$.

3·2. *Closed testing with groups built from the data*

Let $\mathcal{K} \subseteq \{1, \ldots, G\}$ be a nonempty subset of the groups. If the groups $\mathcal{G}$ were fixed a priori, then the hypothesis $H_{\mathcal{K}}$ could be defined to say that there is no treatment effect in the pairs $i \in \bigcup_{g \in \mathcal{K}} s_g$; that is, $H_{\mathcal{K}}$ asserts that $r_{Tij} = r_{Cij}$ for $j = 1, 2$ for all $i \in s_g$ for all $g \in \mathcal{K}$. A test of the a priori hypothesis $H_{\mathcal{K}}$ with a priori groups $\mathcal{G}$ could be based on T_s in § 2·3 with $s = \bigcup_{g \in \mathcal{K}} s_g$ and, in particular, for each fixed $\Gamma = \exp(\gamma) \geqslant 1$ a level-α test could be constructed using the upper bound in (3), and this would be a conventional randomization test if $\Gamma = 1$. With a priori groups $\mathcal{G}$, the closed testing procedure of Marcus et al. (1976) would reject $H_{\mathcal{K}}$ at level α if and only if it had rejected at level α all $H_{\mathcal{L}}$ with $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \ldots, G\}$, and it would strongly control the familywise error rate, i.e., it would falsely reject at least one true $H_{\mathcal{K}}$ with probability at most α no matter which hypotheses $H_{\mathcal{M}}$ are true for $\mathcal{M} \subseteq \{1, \ldots, G\}$. See Hochberg & Tamhane (1987, Ch. 1) for discussion of the familywise error rate, and see Rosenbaum & Silber (2009b)

and Rosenbaum (2015) for discussion in the context of a sensitivity analysis. Weak control of the familywise error rate is no longer regarded as adequate, so we do not discuss it further; it says that the chance of falsely rejecting $H_{\mathcal{K}}$ is at most α if H_0 is true, but if H_0 is false then there are no promises about false rejection of $H_{\mathcal{K}}$. Does a similar result hold when the groups $\mathcal{G}$ are built using the data subject to Condition 2?

Proposition 2 says that we may apply closed testing using groups constructed from the data at hand, yet strongly control the familywise error rate in a sensitivity analysis. Setting $\Gamma = 1$ yields the corollary to Proposition 2.

Algorithm 1. Construct groups $\mathcal{G} = \{s_1, \ldots, s_G\}$ by a method that satisfies Condition 2. Fix $\Gamma \geqslant 1$, and for each $\mathcal{L} \subseteq \{1, \ldots, G\}$ determine the value $k_{\Gamma,\mathcal{L}}$ from the upper bound in (3) with $s = \bigcup_{g\in\mathcal{L}} s_g$ as the smallest value such that $\mathrm{pr}(\bar{T}_{\Gamma s} \geqslant k_{\Gamma,\mathcal{L}} \mid \mathcal{F}, \mathcal{Z}) \leqslant \alpha$ for a fixed α with $0 < \alpha < 1$. Reject the null hypothesis $H_{\mathcal{K}}$ if $\bar{T}_{\Gamma s} \geqslant k_{\Gamma,\mathcal{L}}$ with $s = \bigcup_{g\in\mathcal{L}} s_g$ for all $\mathcal{L}$ such that $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \ldots, G\}$.

Proposition 2. *Assume that Condition 1 holds with the specified* Γ. *If the bias is no larger than* Γ, *then the conditional probability given* $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ *that Algorithm 1 makes at least one false rejection is at most* α.

Proof. The proof adapts the reasoning in Marcus et al. (1976). If $h = \emptyset$, then all the $H_{\mathcal{K}}$ are false and there is nothing to prove, so assume $h \neq \emptyset$. By the definition of h, hypothesis $H_{\mathcal{K}}$ is true if $h \supseteq s = \bigcup_{g\in\mathcal{K}} s_g$, and otherwise $H_{\mathcal{K}}$ is false. If $H_{\mathcal{K}}$ is false, there is no risk of falsely rejecting it, so for the remainder of the proof assume that $H_{\mathcal{K}}$ is true. To reject $H_{\mathcal{K}}$, Algorithm 1 must reject $H_{\mathcal{T}}$ with $h = \bigcup_{g\in\mathcal{T}} s_g$ and $\mathcal{K} \subseteq \mathcal{T}$, where $H_{\mathcal{T}}$ is true by the definition of h. Rejecting $H_{\mathcal{T}}$ requires that $T_h \geqslant k_{\Gamma,\mathcal{T}}$. Under the stated conditions, Proposition 1 tells us that the distribution of T_h is not distorted by conditioning on the grouping $\mathcal{G}$, so $\mathrm{pr}(T_h \geqslant k_{\Gamma,\mathcal{T}} \mid \mathcal{F}, \mathcal{Z}, \mathcal{G}) \leqslant \mathrm{pr}(\bar{T}_{\Gamma s} \geqslant k_{\Gamma,\mathcal{T}} \mid \mathcal{F}, \mathcal{Z}, \mathcal{G}) \leqslant \alpha$. □

Corollary 2. *In a randomized paired experiment, the conditional probability given* $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ *that Algorithm 1 makes at least one false rejection is at most* α.

To summarize, Proposition 2 and Corollary 2 would be relatively straightforward applications of closed testing if the groups $\mathcal{G}$ had been specified a priori; see Hsu et al. (2013, § 3.4). However, in § 1·2, $I = 2106$ pairs of exposed and control individuals were collapsed to $G = 3$ groups $\mathcal{G} = \{s_1, s_2, s_3\}$ using a regression of $|Y_i|$ on $x_i = x_{i1} = x_{i2}$, so these groups are not specified a priori. Conditioning on $\mathcal{G}$ to fix the groups, and hence also fix the null hypotheses, distorts the distributions of some of the Z_{ij} when some of the H_{0s} are true and others are false. Proposition 1 says that under Condition 2, the distortion of the distribution of Z_{ij} is confined to groups s such that H_{0s} is false, and then Proposition 2 and Corollary 2 tell us that closed testing strongly controls the familywise error rate among groups selected on the basis of the data.

4. Change in work income following the Chilean earthquake

Table 1 compares two sensitivity analyses for the example described in § 1·2 and Fig. 1. The values in Table 1 are 100 times the upper bounds on one-sided p-values for various tests, and Algorithm 1 is applied to some of these. The combined sensitivity analysis in Table 1 uses all $I = 2106$ pair differences Y_i in an M-test of the type suggested by Maritz (1979), which is similar to a lightly trimmed mean, with Huber's ψ-function $\psi(d) = \max\{-1, \min(1, d)\}$ applied to Y_i/a where a is the upper 1% quantile of $|Y_i|$. The second sensitivity analysis uses the same statistic

Table 1. *Comparison of two sensitivity analyses for the change in work income following the Chilean earthquake, where the second sensitivity analysis pools subgroup analyses and yields simultaneous inferences about subgroups. The tabulated values are* 100 *times the upper bounds on one-sided p-values for a given value of* Γ*; in each column, the least sensitive p-value bound significant at the* 0·05 *level in closed testing is starred*

	Overall tests		Two groups			Individual groups		
Γ	Combined	zf × zm × p	zf × zm	zf × p	zm × p	zf	zm	p
1	0	0	0	0	0	0	0	0·9*
1·1	0·1	0	0	0·6	0·4	0·2	0·1	7·9
1·2	2·3*	0	0	1·8	0·7	0·8	0·2	28·8
1·25	6·7	0·1	0	3·1*	1·0	1·5*	0·4	43·7
1·3	15·1	0·2	0·1	5·1	1·3	2·6	0·6	58·9
1·35	28·0	0·4	0·2	8·0	1·8	4·1	0·8	72·4
1·4	44·0	3·7	2·4	100	2·4	6·2	1·1	82·9
1·45	60·6	4·8*	3·1*	100	3·1*	8·9	1·5*	90·3
1·5	74·9	6·1	4·0	100	4·0	12·3	2·0	94·8
1·6	92·5	9·6	6·5	100	6·5	21·0	3·3	98·8

Combined, a single M-test using all $I = 2106$ pairs with no attempt to discover effect modification; p, subgroup in Fig. 1 consisting of all people with positive work income; zm, subgroup of males with zero work income; zf, subgroup of females with zero work income; zf × zm × p, pooling of all three subgroup p-values; zf × zm, pooling of the p-values for the two groups with zero work income; zf × p, pooling of the p-values for the zero-income-female group and the positive-income group; zm × p, pooling of the p-values for the zero-income-male group and the positive-income group.

but computes three p-value bounds, one for each subgroup in Fig. 1, denoted by p, zm and zf in the table, and then pools the p-value bounds using the truncated product of p-values due to Zaykin et al. (2002) with $\tilde{\alpha} = 0{\cdot}1$.

The second sensitivity analysis in Table 1 tests for no effect at all in any of the subgroups, i.e., H_0, as suggested in Hsu et al. (2013), and if H_0 is rejected it then tests hypotheses about subgroups using Algorithm 1. The three tests within groups use the same test in each of the groups defined by the regression tree, namely zf for zero-work-income females, zm for zero-work-income males, and p for positive-work-income individuals. These individual p-values are combined using the truncated product of p-values truncated at 0·1; for example, zf × zm combines the two p-values for the pairs with zero work income before the earthquake. Closed testing starts with zf × zm × p, continuing to subhypotheses only if certain rejections take place. When testing the null hypothesis H_0 of no effect at all, the combined test is sensitive at $\Gamma = 1{\cdot}3$, while the truncated product zf × zm × p is insensitive at $\Gamma = 1{\cdot}45$. Although the null hypothesis of no effect is rejected for all groups at $\Gamma = 1$, at $\Gamma = 1{\cdot}45$ the null hypothesis of no effect is rejected only for men with no work income prior to the earthquake.

In Table 1, the truncated product test of no effect at all in any of the subgroups is less sensitive to bias than the combined test, the former being insensitive to $\Gamma = 1{\cdot}45$ and the latter being sensitive to $\Gamma = 1{\cdot}25$. There is a substantial difference between $\Gamma = 1{\cdot}45$ and $\Gamma = 1{\cdot}25$, as can be seen using the device in Rosenbaum & Silber (2009a); specifically, an unobserved covariate that doubled the odds of exposure to the treatment, i.e., doubled the odds of $Z_{i1} - Z_{i2} = 1$, and doubled the odds of a positive pair difference in outcomes, i.e., doubled the odds of $Y_i > 0$, corresponds to $\Gamma = 1{\cdot}25$, whereas an unobserved covariate that doubled the odds of exposure to the treatment and tripled the odds of a positive pair difference in outcomes corresponds to $\Gamma = 1{\cdot}4$. Proposition 2 permits more to be said. In the absence of bias in exposure to the earthquake, $\Gamma = 1$, the hypothesis of no effect on change in work income is rejected in all three subgroups.

For people with work income prior to the earthquake, this rejection is sensitive to a small bias of $\Gamma = 1{\cdot}1$; for women without work income it is insensitive to a bias of $\Gamma = 1{\cdot}25$; and for men without work income it is insensitive to a bias of $\Gamma = 1{\cdot}45$. Although the p-value bound is 0·026 for women without work income at $\Gamma = 1{\cdot}3$, Algorithm 1 never performs this test, as it is stopped by the p-value bound of 0·051 for the conjunction of the two groups of women without income and people with income. The strongest evidence of an effect of the earthquake on work income is among men without work income prior to the earthquake: those exposed to the earthquake were less likely to find jobs and have work income after the earthquake than similar men located far from the earthquake.

A novel aspect of Table 1 is that the three groups $\mathcal{G} = \{s_1, s_2, s_3\}$ were constructed using the data at hand, yet Proposition 2 implies that the familywise error rate has been controlled with data-dependent groups and multiple tests in a sensitivity analysis that allows for a bias of $\Gamma = 1{\cdot}45$. In this specific sense, we can discover groups exhibiting effect modification using the data, yet act as if those groups were specified a priori in closed testing of subgroup hypotheses, while strongly controlling the familywise error rate.

5. Simulation

5·1. *Structure of the simulation*

This simulation checks the claims of Propositions 1 and 2, investigates the ability of a regression tree (Breiman et al., 1984) of $|Y_i|$ on x_i to identify relevant subgroups, and examines various concepts of power. We would like to be able to report insensitivity to bias when the association between treatment Z_{ij} and response R_{ij} is produced by an actual treatment effect, and not by bias in assigning treatments. Therefore, the power of a sensitivity analysis is evaluated when, unknown to the investigator, the treatment is effective and there is no unmeasured bias. In this situation, the power of a level-α sensitivity analysis performed with a specific value of $\Gamma \geqslant 1$ is the probability that the upper bound on the p-value will be less than or equal to α when computed with this Γ; see Rosenbaum (2004, 2010 Part III, 2013).

The simulation has two versions, a limited version presented in the main article and a more extensive version reported in the Supplementary Material. A description of the limited version follows. In parallel with the example in § 1·2, the simulation considers six covariates x as potential effect modifiers, and there are $I = 2000$ pair differences Y_i. Each of these covariates is binary, and they are six independent Bernoulli trials with probability of success 1/2. Of these, at most two of the covariates interact with the treatment to modify the effect on the pair differences Y_i, but it is left to the regression tree to discover which covariates matter. For the two levels of the two active covariates, $a = 0, 1$ and $b = 0, 1$, the pair differences are distributed as $Y_i \sim N(\mu_{ab}, 1)$. In the null case, $\mu_{ab} = 0$ for all a and b. In all other cases, the expected effect or the average of the four μ_{ab} is 1/2. In the case of a constant effect without effect modification, $\mu_{ab} = 1/2$ for all a and b, and in this case it is a mistake to split the pairs into groups. How frequently does this mistake occur and how harmful is it when it does occur? In the case of slight effect modification, $\mu_{00} = \mu_{01} = 0{\cdot}6$ and $\mu_{10} = \mu_{11} = 0{\cdot}4$. Are there benefits to grouping when the effect modification is so slight? In the case of complex effect modification, $\mu_{00} = 1{\cdot}5$, $\mu_{01} = \mu_{10} = 0$ and $\mu_{11} = 0{\cdot}5$, so a tree-based procedure must split on both covariates to succeed in separating all pairs with different expectations. In this case, the null hypothesis of no effect is false, but there are subgroups with no effect, so falsely rejecting a true subgroup hypothesis is now possible. While $\sigma^2_{ab} = 1$ for all a and b in the simulation presented here in the main article, the extensive simulation in the Supplementary Material varies both μ_{ab} and σ^2_{ab} in a variety of other patterns.

As in Hsu et al. (2013), the tree is built from the regression tree fit of the ranks of 2000 absolute pair differences $|Y_i|$ on the x_i. The R package rpart (R Development Core Team, 2015) was used with complexity parameter set to 0·005. Each sampling situation was replicated 5000 times, so an estimated power or error rate has standard error of at most $\sqrt{(0{\cdot}5 \times 0{\cdot}5/5000)} = 0{\cdot}0071$.

5·2. *Evaluating the groups*

How can we judge whether groups $\mathcal{G}$ constructed by the regression tree are in fact good groups? In each sampling situation, let $\mu_i = E(Y_i)$ and $\sigma_i^2 = \text{var}(Y_i)$, and of course in a simulation we know μ_i and σ_i^2. In all simulated cases, μ_i and σ_i^2 vary with at most two of the binary covariates, so there are at most four values of each, and we take $\sigma_i^2 = 1$ except in the Supplementary Material. Write $\bar{\mu}_g = |s_g|^{-1} \sum_{i \in s_g} \mu_i$ for the average expectation in group g. We say that a tree is perfect if $\mu_i = \bar{\mu}_g$ for every $i \in s_g$ for every g, that is, if the groups always separate pair differences with different expectations. We quantify departures from perfection by

$$\iota_{\mathcal{G}} = \frac{\sum_{g=1}^{G} \sum_{i \in s_g} \{(\mu_i - \bar{\mu}_g)^2 + \sigma_i^2\}}{\sum_{g=1}^{G} \sum_{i \in s_g} \sigma_i^2},$$

which is the fractional increase in the mean squared error from grouping by $\mathcal{G}$ rather than by a perfect grouping. A perfect tree has $\iota_{\mathcal{G}} = 1$. For comparison, we also compute $\iota_{\mathcal{A}}$ where $\mathcal{A}$ is a single group of all the pairs, $\mathcal{A} = \{s_1\}$ with $s_1 = \{1, \ldots, I\}$.

In Table 2 we report the mean of $\iota_{\mathcal{G}}$ and $\iota_{\mathcal{A}}$ for four sampling situations, each replicated 5000 times. The grouping by the regression tree shows only small departures from perfection in all four situations, with mean squared error $\iota_{\mathcal{G}}$ ranging from 1 to 1·03. Not grouping had mean squared error $\iota_{\mathcal{A}}$ ranging from 1 to 1·38. Not grouping in the complex effect modification setting had a much higher mean squared error than grouping by the regression tree: an $\iota_{\mathcal{A}}$ of 1·38 compared with an $\iota_{\mathcal{G}}$ of 1·03. Table 2 also records the number of trees out of 5000 that had a single leaf, so that the regression tree produced just one group consisting of all 2000 pairs. In the case of no effect or a constant effect, more than 4939 of the 5000 trees had a single leaf; that is, the regression tree rarely created groups when there was no reason to create groups.

5·3. *Level of the tests*

Propositions 1 and 2 make assertions about the level of certain tests or testing procedures. Specifically, Proposition 1 says that whenever a group of pairs is entirely unaffected, a test with nominal level α will falsely reject with probability at most α, despite the fact that the groups were constructed using the data. Proposition 2 says that when closed testing is applied with component testing having nominal level α, the familywise error rate is strongly controlled at α: the chance of falsely rejecting at least one true hypothesis is at most α. Do the simulation results agree with these assertions?

In Table 2, the column headed False rejections, All is the proportion of null leaves in which the hypothesis of no effect was falsely rejected, and for $\Gamma = 1$ it is consistently near 0·05. Table 2 also records the number of null leaves. A false rejection cannot occur when every individual is affected by the treatment, and in these cases the false rejections section of the table is blank. When there is no unmeasured bias, as in all the simulated examples, but the sensitivity analysis entertains the possibility of such bias, i.e., $\Gamma > 1$, the chance of false rejection is much less than 0·05. The column headed False rejections, Family is the proportion of applications of closed testing that resulted in at least one false rejection, that is, a null leaf declared to have been affected.

Table 2. *Summary of the results of evaluating the groups, level of the tests, and power of the tests for the null hypothesis of no treatment effect with various* Γ *when matched pair differences have normal errors and constant variance; sensitivity analysis values in the last five columns are reported out of* 1000

Scenario $(\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11})$	Perfect MSE $(\iota_{\mathcal{G}}, \iota_{\mathcal{A}})$	# 1-leaf trees	Leaves Null (Total)	Γ	False rejections All	False rejections Family	Power to reject H_0 One	Power to reject H_0 Trunc	Reject false H_0
Null case, no effect									
(0, 0, 0, 0)	1	4939	1·01	1	52	52	51	52	
	(1, 1)		(1·01)	1·01	35	34	34	34	
				1·1	0	0	0	0	
				1·2	0	0	0	0	
				1·3	0	0	0	0	
Constant effect without effect modification									
(0·5, 0·5, 0·5, 0·5)	1	4945	0	1			1000	1000	1000
	(1, 1)		(1·01)	2·8			807	805	803
				3			378	378	377
				3·2			77	78	77
				3·4			7	8	7
Slight effect modification									
(0·6, 0·6, 0·4, 0·4)	0·18	4063	0	1			1000	1000	1000
	(1·01, 1·01)		(1·19)	2·8			796	803	711
				3			322	438	347
				3·2			59	211	128
				3·4			5	131	67
Complex effect modification									
(1·5, 0, 0, 0·5)	0·18	0	1·18	1	48	48	1000	1000	1000
	(1·03, 1·38)		(3·19)	2·3	0	0	822	1000	574
				2·5	0	0	284	1000	553
				15	0	0	0	999	499
				30	0	0	0	64	32

Perfect MSE, fractional increase in mean squared error compared with a perfect grouping; # 1-leaf trees, the number of single-leaf trees among 5000 replicates; Leaves, Null (Total), the averaged null (total) leaves over 5000 replicates; False rejections, All, the proportion of null leaves in which the hypothesis of no effect was falsely rejected; False rejections, Family, the proportion of applications of closed testing that issued at least one false rejection; Power to reject H_0, the power of two sensitivity analyses when testing H_0 of no effect at all; One, the combined test; Trunc, the truncated product; Reject false H_0, the proportion of pairs in a group for which the hypothesis of no effect is rejected by closed testing using the truncated product, averaging over affected pairs and then 5000 replicates.

Here too, the 0·05 familywise level appears to have been preserved, consistent with the claim of Proposition 2.

In brief, building the groups by the regression tree method of regressing $|Y_i|$ on x_i does not appear to have increased the probability of falsely rejecting a true null hypothesis, which is consistent with Propositions 1 and 2.

5·4. *Power of the tests*

The two columns in Table 2 headed Power to reject H_0 give the power of two sensitivity analyses when testing for no effect at all, H_0. Here, the column labelled 'One' is for the combined test in Table 1, which performs a single test using all $I = 2000$ pairs. In contrast, the column labelled Trunc performs a separate test for each subgroup and combines their p-values using the truncated product of Zaykin et al. (2002) truncated at 0·05. The Supplementary Material considers other

methods for combining p-values. Consistent with the asymptotic results of Hsu et al. (2013) about design sensitivity and the limiting power of a sensitivity analysis, the combined method has worse power than the truncated product method except when the effect is constant.

The final column of Table 2 requires some explanation. There are $I = 2000$ pairs in each simulated sample, but only some of these are affected by the treatment. For a pair that is affected, we assign the score 1 if that pair is in a group for which the hypothesis of no effect is rejected by closed testing using the truncated product, and we assign the score 0 otherwise. The final column, Reject false H_0, is the average over 5000 replicates of the proportion of 1 scores among the affected pairs. In general, comparing the last column with the Power to reject H_0, One column, it is seen that the truncated product will often identify specific affected groups by closed testing at values of Γ for which the combined test has virtually no power to detect anything.

In brief, a single test for all pairs is inferior in terms of power in all simulated cases of effect modification, and it has only slightly higher power than the other methods when the effect is constant. Closed testing using the truncated product will often identify affected groups when a single test would accept the null hypothesis of no effect at all.

6. Strength-k matching: near-exact matching

In the earthquake data in § 1·2, there were $V = 6$ candidate covariates defining $2 \times 2 \times 3 \times 3 \times 5 \times 2 = 360$ types of individuals, but the final branchings in Fig. 1 had only three groups formed from just two covariates. It is often difficult to match exactly for many covariates x_{ij}, but not so difficult to match for two covariates and simply balance the rest. However, before we examined the matched pairs, we did not know which two covariates would be suggested as possible effect modifiers in Fig. 1, so we did not know which two covariates should be exactly matched and which other covariates could merely be balanced. In this section we propose a new and strong form of matching, strength-k matching, and show how it can aid in the study of effect modification.

Strength-k balance means that every subset of k covariates is exactly balanced. Here, balance refers to the distributions of covariates in exposed and control groups, not to who is paired with whom. In § 1·2, each of the $C(6, 3) = 20$ subsets of $k = 3$ of the $V = 6$ covariates is exactly balanced. For example, one of the 20 subsets had $3 \times 3 \times 5 = 45$ categories, and the exposed and control groups had exactly the same frequencies in each of these 45 categories. The term strength-k matching is intended to suggest a limited analogy with the orthogonal arrays used to construct fractional factorial designs; see Hedayat et al. (1999). Matching V covariates with strength k is fairly straightforward to implement with the R package mipmatch (Zubizarreta, 2012): essentially, one requests balance for $C(V, k) = 20$ covariates formed as direct products of k of the V covariates, ignoring the origin of these 20 covariates as built from six covariates.

By definition, strength-k matching means that, in addition to the constraint of strength-k balance, all $V = 6$ covariates should be exactly matched in the maximum number of pairs. This turned out to be $1978/2106 = 94\%$ of the pairs. The match used several additional but standard matching techniques that are described in the Supplementary Material.

The tree in Fig. 1 was built using the 1978 exactly matched pairs. Once Fig. 1 had selected $2 \leqslant k$ covariates as candidates for effect modification, the $128 = 2106 - 1978$ inexactly matched pairs were rematched to be exact for the two covariates and to be exact for as many as possible of the $V = 6$ covariates. Clearly, in a strength-k match of V covariates, one can always rematch the inexact pairs to be exact for k or fewer covariates. Also, this rematching does not alter the fact that all I rematched pairs constitute a strength-k match of V covariates, because that property is unaffected by who is matched to whom. So we built the groups using the 1978 exactly matched

pairs, discovered that the groups were a function of two covariates, and then re-paired the inexactly matched pairs to be exactly matched for the two covariates, knowing that all V covariates would exhibit a high degree of balance. In this way, all $I = 2106$ pairs were used and all were exactly matched for the three groups in Fig. 1.

Acknowledgement

We thank the referees, associate editor, and editor for their insightful comments. This work was supported in part by the U.S. National Science Foundation and the Alfred P. Sloan Foundation. Zubizarreta is also affiliated with the Department of Statistics at Columbia University.

Supplementary material

Supplementary material available at *Biometrika* online includes the details of implementing the strength-k match in § 6 using the R package mipmatch (Zubizarreta, 2012), together with a more extensive version of the simulation in § 5.

References

Benjamini, Y. & Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64**, 1215–22.
Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. New York: Chapman and Hall/CRC.
Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. & Wynder, E. (1959). Smoking and lung cancer. *J. Nat. Cancer Inst.* **22**, 173–203.
Egleston, B. L., Scharfstein, D. O. & MacKenzie, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics* **65**, 497–504.
Fisher, R. A. (1935). *Design of Experiments*. Edinburgh: Oliver & Boyd.
Gastwirth, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33**, 19–34.
Hedayat, A. S., Sloane, N. J. A. & Stufken, J. (1999). *Orthogonal Arrays*. New York: Springer.
Hochberg, Y. & Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.
Hosman, C. A., Hansen, B. B. & Holland, P. W. H. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Statist.* **4**, 849–70.
Hsu, J. Y., Small, D. S. & Rosenbaum, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Am. Statist. Assoc.* **108**, 135–48.
Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93**, 126–32.
Jogdeo, K. (1977). Association and probability inequalities. *Ann. Statist.* **5**, 495–504.
Lin, D. Y., Psaty, B. M. & Kronmal, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54**, 948–63.
Liu, W., Kuramoto, S. J. & Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev. Sci.* **14**, 570–80.
Marcus, R., Eric, P. & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–60.
Maritz, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66**, 163–6.
McCandless, L. C., Gustafson, P. & Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statist. Med.* **26**, 2331–47.
Neyman, J. (1923). On the application of probability theory to agricultural experiments (1990 translation). *Statist. Sci.* **5**, 463–80.
R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Rosenbaum, P. R. (2002). *Observational Studies*, 2nd ed. New York: Springer.
Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91**, 153–64.
Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Am. Statistician* **59**, 147–52.
Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests and confidence intervals in matched observational studies. *Biometrics* **63**, 456–64.

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. New York: Springer.
ROSENBAUM, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* **69**, 118–27.
ROSENBAUM, P. R. (2015). Two R packages for sensitivity analysis in observational studies. *Observ. Stud.* **1**, 1–17.
ROSENBAUM, P. R. & SILBER, J. H. (2009a). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J. Am. Statist. Assoc.* **104**, 501–11.
ROSENBAUM, P. R. & SILBER, J. H. (2009b). Amplification of sensitivity analysis in observational studies. *J. Am. Statist. Assoc.* **104**, 1398–405.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.* **66**, 688–701.
WANG, L. & KRIEGER, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statist. Med.* **25**, 2257–71.
YANAGAWA, T. (1984). Case-control studies: Assessing the effect of a confounding factor. *Biometrika* **71**, 191–4.
YU, B. B. & GASTWIRTH, J. L. (2005). Sensitivity analysis for trend tests: Application to the risk of radiation exposure. *Biostatistics* **6**, 201–9.
ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. & WEIR, B. S. (2002). Truncated product method of combining *P*-values. *Genet. Epidemiol.* **22**, 170–85.
ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Am. Statist. Assoc.* **107**, 1360–71.
ZUBIZARRETA, J. R., CERDA, M. & ROSENBAUM, P. R. (2013). Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design. *Epidemiology* **24**, 79–87.

[*Received May* 2014. *Revised April* 2015]