

1  
2  
3  
4 **Strong control of the familywise error rate in observational**  
5 **studies that discover effect modification by exploratory**  
6 **methods**  
7

8  
9 BY JESSE Y. HSU

10 *Department of Biostatistics and Epidemiology, Perelman School of Medicine,*  
11 *University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.*

12 hsu9@mail.med.upenn.edu  
13

14 JOSÉ R. ZUBIZARRETA

15  
16 *Division of Decision, Risk and Operations, and Department of Statistics, Columbia University,*  
17 *New York, New York 10027, U.S.A.*

18 zubizarreta@columbia.edu  
19

20 DYLAN S. SMALL AND PAUL R. ROSENBAUM

21 *Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia,*  
22 *Pennsylvania 19104, U.S.A.*

23 dsmall@wharton.upenn.edu rosenbaum@wharton.upenn.edu  
24

25 SUMMARY  
26

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

40  
41 *Some key words:* Closed testing; Design sensitivity; Optimal matching; Sensitivity analysis; Truncated product of  
42  $p$ -values.  
43

44 1. INTRODUCTION

45 1.1. *Subgroups suggested by the data*

46  
47 In experiments and observational studies, effect modification refers to the possibility that  
48 the magnitude or stability of a treatment effect changes with the observed covariates. There is

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

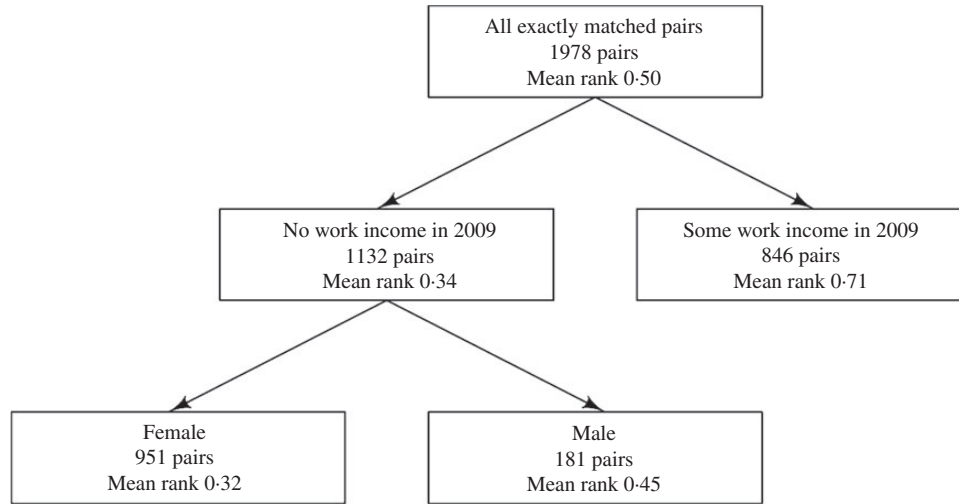
55 In observational studies, the magnitude and stability of a treatment effect strongly influence its  
56 sensitivity to biases from unmeasured covariates. Small and unstable treatment effects can often  
57 plausibly be explained away as being created by small biases in the assignment of individuals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

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

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

### 85 86 1.2. *Example: the Chilean earthquake*

87 On 27 February 2010, a powerful earthquake of magnitude 8.8 struck off the coast of Chile.  
88 What effects did the earthquake have on the labour market in Chile? In this paper, we take the  
89 outcome to be the change in individual work income from before the earthquake to after. Using  
90 the data described in [Zubizarreta et al. \(2013\)](#), we constructed 2106 matched pairs of individ-  
91 uals, one in a severely shaken region of Chile and the other remote from the earthquake. The  
92 matching controlled for covariates measured before the earthquake: sex, marital status, number  
93 of persons in household, self-reported health problems, self-reported health perception, quartile  
94 of work income, age, self-reported psychological problems, disability, health insurance status,  
95 years of education, employment status, per capita household income, poverty status, housing sta-  
96 tus, quality of housing structure, and overcrowding. Some innovations in matching methodology



112  
113  
114  
115  
116  
117

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.

118  
119  
120

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.

121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137

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?

## 138 139

### 2. NOTATION AND REVIEW

#### 140 141

#### 2.1. Notation

142  
143  
144

There are  $I$  matched pairs, indexed by  $i = 1, \dots, I$ , 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,

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

150 Each subject has a potential response  $r_{Tij}$  if treated with  $Z_{ij} = 1$ , a potential response  $r_{Cij}$  if  
 151 assigned to control with  $Z_{ij} = 0$ , and an observed response  $R_{ij} = Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij}$  under  
 152 the treatment actually received, whereas the effect of the treatment, namely  $r_{Tij} - r_{Cij}$ , is not  
 153 observed for any subject; see Neyman (1923) and Rubin (1974). The sharp null hypothesis of no  
 154 treatment effect (Fisher, 1935) is  $H_0 : r_{Tij} = r_{Cij}$  for all  $i, j$ . Importantly, if  $H_0$  is true, then  $R_{ij} =$   
 155  $r_{Cij}$  does not change with the treatment assignment  $Z_{ij}$ ; but if  $H_0$  is false, then at least some of the  
 156  $R_{ij}$  do change with  $Z_{ij}$ . Write  $\mathcal{F} = \{(r_{Tij}, r_{Cij}, x_{ij}, u_{ij}) : i = 1, \dots, I; j = 1, 2\}$ . The treated-  
 157 minus-control pair difference in observed responses for pair  $i$  is  $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ ,  
 158 and it equals  $(Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$  if  $H_0$  is true. Also, write  $r_C = (r_{C11}, r_{C12}, \dots, r_{CI2})^\top$  and  
 159  $R = (R_{11}, \dots, R_{I2})^\top$  for the vectors of dimension  $2I$ , and write  $Y = (Y_1, \dots, Y_I)^\top$  for the vector  
 160 of dimension  $I$ .

## 161 162 2.2. Randomization inference in experiments

163 In a paired randomized experiment, subjects are paired on the basis of observed covariates,  $x_{ij}$ ,  
 164 and then a fair coin is flipped independently  $I$  times to determine the treatment assignments  $Z_{i1}$   
 165 and  $Z_{i2} = 1 - Z_{i1}$ ; that is,  $\text{pr}(Z_{ij} = 1 | \mathcal{F}, \mathcal{Z}) = 1/2$  for each  $(i, j)$  and  $\text{pr}(Z = z | \mathcal{F}, \mathcal{Z}) = 2^{-I}$   
 166 for each  $z \in \mathcal{Z}$ . The null distribution of a test statistic  $t(Z, R)$  under Fisher's  $H_0$  is its permutation  
 167 distribution, namely  
 168

$$169 \text{pr}\{t(Z, R) \geq k | \mathcal{F}, \mathcal{Z}\} = \text{pr}\{t(Z, r_C) \geq k | \mathcal{F}, \mathcal{Z}\} = \frac{|\{z \in \mathcal{Z} : t(z, r_C) \geq k\}|}{|\mathcal{Z}|}, \quad (1)$$

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

174 Similarly, Maritz (1979) tested  $H_0$  using (1) and a suitably defined  $M$ -statistic, the quantity  
 175 that is equated to zero in defining Huber's  $M$ -estimates, specifically,  $t(Z, R) = \sum_{i=1}^I \psi(Y_i/a)$   
 176 where  $a$  is a quantile of the  $|Y_i|$  and  $\psi(\cdot)$  is a monotone increasing odd function (i.e.,  $\psi(d) =$   
 177  $-\psi(-d)$ , so  $\psi(0) = 0$ ). We will use a version of Maritz's test statistic to analyse the earthquake  
 178 data in § 4. Under  $H_0$ , the pair difference is  $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) = \pm|r_{Ci1} - r_{Ci2}|$ , so  
 179  $|Y_i| = |r_{Ci1} - r_{Ci2}|$  is fixed by conditioning on  $\mathcal{F}$  in (1), and hence  $a$  is also fixed; then,  $t(Z, R) =$   
 180  $\sum_{i=1}^I q_i \text{sign}(Y_i)$ , where  $q_i = \psi(|r_{Ci1} - r_{Ci2}|/a)$  is fixed by conditioning on  $\mathcal{F}$  and  $\text{sign}(Y_i) = 1,$   
 181  $0$  or  $-1$  according to whether  $Y_i > 0$ ,  $Y_i = 0$  or  $Y_i < 0$ , respectively. As a consequence, under  $H_0$ ,  
 182 (1) is the distribution of the sum of  $I$  independent random variables taking the values  $\pm\psi(|r_{Ci1} -$   
 183  $r_{Ci2}|/a)$  with equal probabilities  $1/2$  if  $|r_{Ci1} - r_{Ci2}| > 0$  or taking the value  $0$  with probability  $1$   
 184 if  $|r_{Ci1} - r_{Ci2}| = 0$ .

## 185 186 187 2.3. Sensitivity analysis in observational studies

188 A sensitivity analysis in an observational study supposes that, in the population prior to match-  
 189 ing, individuals are independently assigned to the treatment or control groups with unknown  
 190 probabilities,  $\pi_{ij} = \text{pr}(Z_{ij} = 1 | \mathcal{F})$ , that may depend on both the observed covariates  $x_{ij}$  and the  
 191 unobserved covariate  $u_{ij}$  as recorded in  $\mathcal{F}$ . The sensitivity analysis assumes that two subjects  $i$   
 192 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  $\Gamma \geq 1$ ; that is,  $\Gamma^{-1} \leq \pi_{ij}(1 - \pi_{i'j'}) / \{\pi_{i'j'}(1 - \pi_{ij})\} \leq \Gamma$ . 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  $0 \leq u_{ij} \leq 1$  for some unknown function  $\kappa(\cdot)$ ; see Rosenbaum (2002, § 4), where the proof consists in constructing  $u_{ij}$  from  $\pi_{ij}$  and conversely. Therefore, as  $u_{ij}$  is effectively a transformation of  $\pi_{ij}$ ,  $u_{ij}$  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  $z \in \mathcal{Z}$ ,

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

for some  $u = (u_{11}, \dots, u_{I2})^T \in \mathcal{U} = [0, 1]^{2I}$ , 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  $\gamma$  and  $u$  were known, then under  $H_0$  the distribution of the test statistic  $T = t(Z, R) = t(Z, r_C)$  would be the sum of the probabilities in (2) over the  $z$  in  $\{z \in \mathcal{Z} : t(z, r_C) \geq k\}$ . The sensitivity analysis asks: how large a departure  $\Gamma$  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  $\Gamma \geq 1$  yields an interval of possible  $p$ -values or point estimates or endpoints of confidence intervals, and the question is: how large must  $\Gamma$  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  $s \subseteq \{1, \dots, I\}$  be a fixed nonempty subset of the  $I$  pairs, with  $|s| \geq 1$ . 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,  $H_{0s} : r_{Tij} = r_{Cij}$  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  $q_{si} \geq 0$  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,  $\text{pr}(T_s \geq k \mid \mathcal{F}, \mathcal{Z})$ . Define  $\tilde{T}_{\Gamma s}$  to be the sum of  $|s|$  independent random variables, the  $i$ th random variable being  $q_{si}$  with probability  $\Gamma/(1 + \Gamma)$  and  $-q_{si}$  with probability  $1/(1 + \Gamma)$ , provided  $q_{si} > 0$ ; otherwise, the  $i$ th 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  $\text{pr}(T_s \geq k \mid \mathcal{F}, \mathcal{Z})$  of  $T_s$  under  $H_{0s}$  and (2) is sharply bounded by two known distributions,

$$\text{pr}(\tilde{T}_{\Gamma s} \geq k \mid \mathcal{F}, \mathcal{Z}) \leq \text{pr}(T_s \geq k \mid \mathcal{F}, \mathcal{Z}) \leq \text{pr}(\tilde{T}_{\Gamma^{-1} s} \geq k \mid \mathcal{F}, \mathcal{Z}); \quad (3)$$

see Rosenbaum (2002, § 4; 2007). 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  $\Delta$ -fold increase in the odds of a positive outcome

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

243 Other methods of sensitivity analysis in observational studies are discussed by, for instance,  
 244 Cornfield et al. (1959), Eggleston et al. (2009), Gastwirth (1992), Hosman et al. (2010), Imbens  
 245 (2003), Lin et al. (1998), Liu et al. (2013), McCandless et al. (2007), Wang & Krieger (2006),  
 246 Yanagawa (1984) and Yu & Gastwirth (2005).

#### 247 2.4. Effect modifiers when testing for no effect

248  
 249 Hsu et al. (2013, § 4) tested Fisher's null hypothesis  $H_0$  of no effect by first dividing the pairs  
 250  $i \in \{1, \dots, I\}$  into several groups based on, say,  $x_{i1} = x_{i2} = x_i$ , looking for possible effect modi-  
 251 fication, i.e., larger or more stable treatment effects in some groups than in others. More precisely,  
 252  $G \geq 1$  mutually exclusive and exhaustive groups  $\mathcal{G} = \{s_1, \dots, s_G\}$  of the pairs  $i = 1, \dots, I$  were  
 253 formed, where each  $s_g \subseteq \{1, \dots, I\}$ . These groups are formed by regressing a function of the  
 254 absolute differences  $|Y_i|$  on  $x_i$  in some fashion that yields non-overlapping groups, for instance  
 255 by using a regression tree as in Fig. 1, in which  $G = 3$  groups of pairs were formed. Under Fisher's  
 256  $H_0$ , the absolute difference in responses  $|Y_i| = |r_{Ci1} - r_{Ci2}|$  is fixed by conditioning on  $\mathcal{F}$ ,  
 257 as discussed in § 2.2, so the grouping produced by the regression of  $|Y_i|$  on  $x_i$  is also fixed.  
 258 Under model (2), when  $H_0$  is true, a test statistic  $T_s = t(Z_s, R_s) = \sum_{i \in s} q_{si} \text{sign}(Y_i)$  for  $s \in \mathcal{G}$   
 259 has the usual bounds on its null distribution, namely (3), because these bounds refer to the con-  
 260 ditional distribution given  $(\mathcal{F}, \mathcal{Z})$  when  $H_0$  is true. In particular, in a randomized experiment,  
 261 model (2) holds with  $0 = \gamma = \log(\Gamma)$ , so that under  $H_0$ , the group-specific statistic  $T_s$  has its  
 262 usual randomization distribution despite the data-dependent nature of the groups  $\mathcal{G}$ .

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

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

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

285 So far, the discussion has focused on testing the null hypothesis  $H_0$  of no effect at all, and  
 286 that hypothesis played a key role in permitting the groups  $\mathcal{G}$  to be determined from the data by  
 287 regressing a function of  $|Y_i|$  on  $x_i$ . A more interesting question not addressed by Hsu et al. (2013)  
 288 is whether subhypotheses  $H_{0s}$  of no effect for pairs  $i \in s$  with  $s \in \mathcal{G}$  can be tested using (3) when

289  $H_0$  may be false. If  $H_0$  is false, then there is at least one pair  $i$  for which  $r_{Ti1} \neq r_{Ci1}$  or  $r_{Ti2} \neq r_{Ci2}$   
 290 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  
 291 determined by  $\mathcal{F}$  but  $R$  varies with  $Z$ . If we reject the null hypothesis  $H_0$  of no effect on anyone,  
 292 it is not clear from the argument of this section that we can say anything about just one of the  
 293 groups, e.g., about  $H_{0s}$ . If  $H_0$  is false in a randomized experiment, then the grouping  $\mathcal{G}$  depends  
 294 on  $Z$ : had randomization yielded a different treatment assignment  $Z$ , it might easily have yielded  
 295 different groups  $\mathcal{G}$ , and the hypothesis  $H_{0s}$  is not even a hypothesis in any conventional sense,  
 296 because the hypotheses change as the treatment assignments  $Z$  change. This issue is central to  
 297 the current paper and is explored in § 3.

### 300 3. FAMILYWISE ERROR RATE WITH GROUPS CONSTRUCTED FROM THE DATA

#### 301 3.1. Consequences of data-dependent grouping for null distributions

302 To address the issue raised at the end of § 2.4, the following conditions are assumed to hold.

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

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

307  
 308 Condition 2 says that  $Z$  was not explicitly used in constructing the groups but that  $R$  was used.  
 309 Here,  $G$  and  $\mathcal{G}$  are random quantities given  $\mathcal{F}$  and  $\mathcal{Z}$  because  $H_0$  may be false and, if so,  $R \neq r_C$   
 310 depends on  $Z$  and hence the groups in Condition 2 may also depend indirectly on  $Z$  through the  
 311 dependence of  $R$  on  $Z$ . If the groups  $\mathcal{G}$  are random quantities depending indirectly on  $Z$ , then  
 312 taking the groups to be fixed, conditioning on  $\mathcal{G}$ , may alter the distribution of  $Z$ . Propositions 1  
 313 and 2 say that this genuine problem can be kept under control.

314 Let  $h \subseteq \{1, \dots, I\}$  be the union of all the groups  $s_g$  for which there is no treatment effect,  
 315 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,  $h = \emptyset$ .  
 316 Obviously, the investigator does not know  $h$ .

317 As  $\mathcal{G}$  is a random quantity,  $h$  is also a random quantity because it is a union of some of the  
 318  $s_g$ . 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.  
 319 Conditionally given  $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ , if  $h = \emptyset$  then there are affected pairs in every group  $s \in \mathcal{G}$ , so  
 320 every  $H_{0s}$  is false, and false rejection of a true  $H_{0s}$  cannot occur. Conversely, conditionally given  
 321  $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ , if  $h \neq \emptyset$  then some group or groups  $s \in \mathcal{G}$  contain no affected individuals, and false  
 322 rejection of a true  $H_{0s}$  is possible. Proposition 1 and its corollary concern the distribution of  
 323 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  
 324 treatment but the grouping  $\mathcal{G}$  itself, and hence also  $h$ , may have been affected by the treatment.  
 325 Stated informally, Proposition 1 says that the data-dependent grouping  $\mathcal{G}$  does not alter the null  
 326 distribution of  $T_h$  even when  $H_0$  is false so that  $\mathcal{G}$  changes with  $Z$ : the null distribution of  $T_h$   
 327 is still bounded by (3) with  $s = h$ . To emphasize this point,  $T_h$  is computed from the union  $h$   
 328 of all groups  $s_g$  for which there is no treatment effect, and because the investigator does not  
 329 know  $h$ , she cannot know when she has computed  $T_h$ . Proposition 1 is a step in the development  
 330 of a multiple inference procedure that strongly controls false rejections, as discussed in § 3.2.  
 331 Although the proof of Proposition 1 requires some attention to detail, the idea is not difficult:  
 332  $Z$  affects the groups  $\mathcal{G}$  only indirectly by affecting  $R$ , but for  $i \in h$  the treatment assignment  $Z_{ij}$   
 333 does not affect  $R$ , so the distribution of  $Z_{ij}$  for  $i \in h$  is unchanged by conditioning on  $(\mathcal{F}, \mathcal{Z}, \mathcal{G})$ .

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

340 *Proof.* Assume  $h \neq \emptyset$ . Let  $\mathcal{N} \subseteq \{1, \dots, I\}$  be the set of pairs with no treatment effect, so  
 341  $r_{Ti1} = r_{Ci1}$  and  $r_{Ti2} = r_{Ci2}$  if and only if  $i \in \mathcal{N}$ ; let  $\mathcal{E} \subseteq \{1, \dots, I\}$  be the complementary set of  
 342 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  $2|h|$ -  
 343 dimensional vector  $z = (z_{11}, z_{12}, \dots, z_{\ell j}, \dots, z_{|h|,1}, z_{|h|,2})^\top$  with  $z_{\ell j} = 1$  or  $z_{\ell j} = 0$  and  $z_{\ell 1} +$   
 344  $z_{\ell 2} = 1$  for each  $\ell$ . Write  $\mathcal{D}$  for the combination of the data  $\{(r_{Ci1}, r_{Ci2}, x_i), i \in \mathcal{N}\}$  and the data  
 345  $\{(R_{i1}, R_{i2}, Z_{i1}, Z_{i2}, x_i), i \in \mathcal{E}\}$ . Because pairs  $i \in \mathcal{N}$  are unaffected, with  $R_{ij} = r_{Cij}$  for  $i \in \mathcal{N}$   
 346 and  $j = 1, 2$ , the grouping  $\mathcal{G} = \{s_1, \dots, s_G\}$  is a function of  $\mathcal{D}$ . Because the grouping  $\mathcal{G}$  is a  
 347 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  $i \in \mathcal{E}$ ,  
 348 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  $i \in$   
 349  $\mathcal{E}$ ; 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  
 350 this all together under (2), the  $Z_{i1} = 1 - Z_{i2}$  for  $i \in \mathcal{N}$  satisfy  
 351

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

352 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  
 353  $Z_{i'1}$  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  $i' \in \mathcal{E}$   
 354 is already fixed by conditioning on  $(\mathcal{F}, \mathcal{Z})$ . Using (2) again together with the fact that  $h$  is fixed  
 355 by conditioning on  $(\mathcal{G}, \mathcal{F}, \mathcal{Z})$  yields  
 356

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

357 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  
 358 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 distri-  
 359 bution of  $Z_h$ , namely  $\text{pr}(Z_h = z \mid \mathcal{F}, \mathcal{Z}, \mathcal{G})$ , and hence also the distribution  $\text{pr}(T_h \geq k \mid \mathcal{F}, \mathcal{Z}, \mathcal{G})$ ,  
 360 is identical to the distribution that produced the bounds in (3) with  $s = h$ , proving the result.  $\square$

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

### 3.2. Closed testing with groups built from the data

364 Let  $\mathcal{K} \subseteq \{1, \dots, G\}$  be a nonempty subset of the groups. If the groups  $\mathcal{G}$  were fixed a priori,  
 365 then the hypothesis  $H_{\mathcal{K}}$  could be defined to say that there is no treatment effect in the pairs  
 366  $i \in \bigcup_{g \in \mathcal{K}} s_g$ ; 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  
 367 the a priori hypothesis  $H_{\mathcal{K}}$  with a priori groups  $\mathcal{G}$  could be based on  $T_s$  in § 2.3 with  $s = \bigcup_{g \in \mathcal{K}} s_g$   
 368 and, in particular, for each fixed  $\Gamma = \exp(\gamma) \geq 1$  a level- $\alpha$  test could be constructed using the  
 369 upper bound in (3), and this would be a conventional randomization test if  $\Gamma = 1$ . With a priori  
 370 groups  $\mathcal{G}$ , the closed testing procedure of Marcus et al. (1976) would reject  $H_{\mathcal{K}}$  at level  $\alpha$  if and  
 371 only if it had rejected at level  $\alpha$  all  $H_{\mathcal{L}}$  with  $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \dots, G\}$ , and it would strongly control  
 372 the familywise error rate, i.e., it would falsely reject at least one true  $H_{\mathcal{K}}$  with probability at most  
 373  $\alpha$  no matter which hypotheses  $H_{\mathcal{M}}$  are true for  $\mathcal{M} \subseteq \{1, \dots, G\}$ . See Hochberg & Tamhane  
 374 (1987, ch. 1) for discussion of the familywise error rate, and see Rosenbaum & Silber (2009b)



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  $\alpha$  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  $\mathcal{G} = \{s_1, \dots, s_G\}$  by a method that satisfies Condition 2. Fix  $\Gamma \geq 1$ , and for each  $\mathcal{L} \subseteq \{1, \dots, G\}$  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  $\text{pr}(\bar{T}_{\Gamma s} \geq k_{\Gamma, \mathcal{L}} \mid \mathcal{F}, \mathcal{Z}) \leq \alpha$  for a fixed  $\alpha$  with  $0 < \alpha < 1$ . Reject the null hypothesis  $H_{\mathcal{K}}$  if  $\bar{T}_{\Gamma s} \geq k_{\Gamma, \mathcal{L}}$  with  $s = \bigcup_{g \in \mathcal{L}} s_g$  for all  $\mathcal{L}$  such that  $\mathcal{K} \subseteq \mathcal{L} \subseteq \{1, \dots, G\}$ .

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

*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  $T_h \geq k_{\Gamma, \mathcal{T}}$ . 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  $\text{pr}(T_h \geq k_{\Gamma, \mathcal{T}} \mid \mathcal{F}, \mathcal{Z}, \mathcal{G}) \leq \text{pr}(\bar{T}_{\Gamma s} \geq k_{\Gamma, \mathcal{T}} \mid \mathcal{F}, \mathcal{Z}, \mathcal{G}) \leq \alpha$ .  $\square$

**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  $\alpha$ .*

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  $\psi$ -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

433 Table 1. Comparison of two sensitivity analyses for the change in work income following the  
 434 Chilean earthquake, where the second sensitivity analysis pools subgroup analyses and yields  
 435 simultaneous inferences about subgroups. The tabulated values are 100 times the upper bounds  
 436 on one-sided  $p$ -values for a given value of  $\Gamma$ ; in each column, the least sensitive  $p$ -value bound  
 437 significant at the 0.05 level in closed testing is starred

439 $\Gamma$	438 Overall tests			438 Two groups			438 Individual groups		
	439 Combined	439 $zf \times zm \times p$	439 $zf \times zm$	439 $zf \times p$	439 $zm \times p$	439 $zf$	439 $zm$	439 $p$	
440 1	0	0	0	0	0	0	0	0.9*	
441 1.1	0.1	0	0	0.6	0.4	0.2	0.1	7.9	
442 1.2	2.3*	0	0	1.8	0.7	0.8	0.2	28.8	
443 1.25	6.7	0.1	0	3.1*	1.0	1.5*	0.4	43.7	
444 1.3	15.1	0.2	0.1	5.1	1.3	2.6	0.6	58.9	
444 1.35	28.0	0.4	0.2	8.0	1.8	4.1	0.8	72.4	
445 1.4	44.0	3.7	2.4	100	2.4	6.2	1.1	82.9	
446 1.45	60.6	4.8*	3.1*	100	3.1*	8.9	1.5*	90.3	
447 1.5	74.9	6.1	4.0	100	4.0	12.3	2.0	94.8	
448 1.6	92.5	9.6	6.5	100	6.5	21.0	3.3	98.8	

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

455  
 456 but computes three  $p$ -value bounds, one for each subgroup in Fig. 1 (denoted by ‘ $p$ ’, ‘ $zm$ ’ and  
 457 ‘ $zf$ ’ in the table), and then pools the  $p$ -value bounds using the truncated product of  $p$ -values due  
 458 to Zaykin et al. (2002) with  $\tilde{\alpha} = 0.1$ .

459 The second sensitivity analysis in Table 1 tests for no effect at all in any of the subgroups,  
 460 i.e.,  $H_0$ , as suggested in Hsu et al. (2013), and if  $H_0$  is rejected it then tests hypotheses about  
 461 subgroups using Algorithm 1. The three tests within groups use the same test in each of the  
 462 groups defined by the regression tree, namely ‘ $zf$ ’ for zero-work-income females, ‘ $zm$ ’ for zero-  
 463 work-income males, and ‘ $p$ ’ for positive-work-income individuals. These individual  $p$ -values  
 464 are combined using the truncated product of  $p$ -values truncated at 0.1; for example,  $zf \times zm$   
 465 combines the two  $p$ -values for the pairs with zero work income before the earthquake. Closed  
 466 testing starts with  $zf \times zm \times p$ , continuing to subhypotheses only if certain rejections take place.  
 467 When testing the null hypothesis  $H_0$  of no effect at all, the combined test is sensitive at  $\Gamma = 1.3$ ,  
 468 while the truncated product  $zf \times zm \times p$  is insensitive at  $\Gamma = 1.45$ . Although the null hypothesis  
 469 of no effect is rejected for all groups at  $\Gamma = 1$ , at  $\Gamma = 1.45$  the null hypothesis of no effect is  
 470 rejected only for men with no work income prior to the earthquake.

471 In Table 1, the truncated product test of no effect at all in any of the subgroups is less sensitive  
 472 to bias than the combined test, the former being insensitive to  $\Gamma = 1.45$  and the latter being  
 473 sensitive to  $\Gamma = 1.25$ . There is a substantial difference between  $\Gamma = 1.45$  and  $\Gamma = 1.25$ , as can  
 474 be seen using the device in Rosenbaum & Silber (2009a); specifically, an unobserved covariate  
 475 that doubled the odds of exposure to the treatment (i.e., doubled the odds of  $Z_{i1} - Z_{i2} = 1$ ) and  
 476 doubled the odds of a positive pair difference in outcomes (i.e., doubled the odds of  $Y_i > 0$ )  
 477 corresponds to  $\Gamma = 1.25$ , whereas an unobserved covariate that doubled the odds of exposure to  
 478 the treatment and tripled the odds of a positive pair difference in outcomes corresponds to  $\Gamma =$   
 479  $1.4$ . Proposition 2 permits more to be said. In the absence of bias in exposure to the earthquake,  
 480  $\Gamma = 1$ , the hypothesis of no effect on change in work income is rejected in all three subgroups.

481 For people with work income prior to the earthquake, this rejection is sensitive to a small bias  
 482 of  $\Gamma = 1.1$ ; for women without work income it is insensitive to a bias of  $\Gamma = 1.25$ ; and for men  
 483 without work income it is insensitive to a bias of  $\Gamma = 1.45$ . Although the  $p$ -value bound is 0.026  
 484 for women without work income at  $\Gamma = 1.3$ , Algorithm 1 never performs this test, as it is stopped  
 485 by the  $p$ -value bound of 0.051 for the group consisting of women without work income as well  
 486 as for the group of all people with positive income. The strongest evidence of an effect of the  
 487 earthquake on work income is among men without work income prior to the earthquake: those  
 488 exposed to the earthquake were less likely to find jobs and have work income after the earthquake  
 489 than similar men located far from the earthquake.

490 A novel aspect of Table 1 is that the three groups  $\mathcal{G} = \{s_1, s_2, s_3\}$  were constructed using the  
 491 data at hand, yet Proposition 2 implies that the familywise error rate has been controlled with data-  
 492 dependent groups and multiple tests in a sensitivity analysis that allows for a bias of  $\Gamma = 1.45$ . In  
 493 this specific sense, we can discover groups exhibiting effect modification using the data, yet act  
 494 as if those groups were specified a priori in closed testing of subgroup hypotheses, while strongly  
 495 controlling the familywise error rate.

## 497 5. SIMULATION

### 498 5.1. Structure of the simulation

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

509 The simulation has two versions, a limited version presented in the main article and a more  
 510 extensive version reported in the Supplementary Material. A description of the limited version  
 511 follows. In parallel with the example in § 1.2, the simulation considers six covariates  $x$  as potential  
 512 effect modifiers, and there are  $I = 2000$  pair differences  $Y_i$ . Each of these covariates is binary,  
 513 and they are six independent Bernoulli trials with probability of success 1/2. Of these, at most  
 514 two of the covariates interact with the treatment to modify the effect on the pair differences  $Y_i$ , but  
 515 it is left to the regression tree to discover which covariates matter. For the two levels of the two  
 516 active covariates,  $a = 0, 1$  and  $b = 0, 1$ , the pair differences are distributed as  $Y_i \sim N(\mu_{ab}, 1)$ .  
 517 In the null case,  $\mu_{ab} = 0$  for all  $a$  and  $b$ . In all other cases, the expected effect or the average  
 518 of the four  $\mu_{ab}$  is 1/2. In the case of a constant effect without effect modification,  $\mu_{ab} = 1/2$   
 519 for all  $a$  and  $b$ , and in this case it is a mistake to split the pairs into groups. How frequently  
 520 does this mistake occur and how harmful is it when it does occur? In the case of slight effect  
 521 modification,  $\mu_{00} = \mu_{01} = 0.6$  and  $\mu_{10} = \mu_{11} = 0.4$ . Are there benefits to grouping when the  
 522 effect modification is so slight? In the case of complex effect modification,  $\mu_{00} = 1.5$ ,  $\mu_{01} =$   
 523  $\mu_{10} = 0$  and  $\mu_{11} = 0.5$ , so a tree-based procedure must split on both covariates to succeed in  
 524 separating all pairs with different expectations. In this case, the null hypothesis of no effect is  
 525 false, but there are subgroups with no effect, so falsely rejecting a true subgroup hypothesis is now  
 526 possible. While  $\sigma_{ab}^2 = 1$  for all  $a$  and  $b$  in the simulation presented here in the main article, the  
 527 extensive simulation in the Supplementary Material varies both  $\mu_{ab}$  and  $\sigma_{ab}^2$  in a variety of other  
 528 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  $\sqrt{(0.5 \times 0.5/5000)} = 0.0071$ .

### 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  $\mu_i$  and  $\sigma_i^2$ . In all simulated cases,  $\mu_i$  and  $\sigma_i^2$  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  $s_1 = \{1, \dots, I\}$ .

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  $\alpha$  will falsely reject with probability at most  $\alpha$ , 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  $\alpha$ , the familywise error rate is strongly controlled at  $\alpha$ : the chance of falsely rejecting at least one true hypothesis is at most  $\alpha$ . 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.

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

Scenario ( $\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}$ )	Perfect MSE ( $t_G, t_A$ )	# 1-leaf trees	Leaves Null (Total)	$\Gamma$	False rejections		Power to reject $H_0$		Reject false $H_0$
					All	Family	One	Trunc	
Null case, no effect (0, 0, 0, 0)	1 (1, 1)	4939	1.01 (1.01)	1 1.01 1.1 1.2 1.3	52 35 0 0 0	52 34 0 0 0	51 34 0 0 0	52 34 0 0 0	
Constant effect without effect modification (0.5, 0.5, 0.5, 0.5)	1 (1, 1)	4945	0 (1.01)	1 2.8 3 3.2 3.4			1000 807 378 77 7	1000 805 378 78 8	1000 803 377 77 7
Slight effect modification (0.6, 0.6, 0.4, 0.4)	0.18 (1.01, 1.01)	4063	0 (1.19)	1 2.8 3 3.2 3.4			1000 796 322 59 5	1000 803 438 211 131	1000 711 347 128 67
Complex effect modification (1.5, 0, 0, 0.5)	0.18 (1.03, 1.38)	0	1.18 (3.19)	1 2.3 2.5 15 30	48 0 0 0 0	48 0 0 0 0	1000 822 1000 0 0	1000 1000 1000 999 64	1000 574 553 499 32

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

612 Here too, the 0.05 familywise level appears to have been preserved, consistent with the claim of  
 613 Proposition 2.

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

618 5.4. Power of the tests

619 The two columns in Table 2 headed ‘Power to reject  $H_0$ ’ give the power of two sensitivity  
 620 analyses when testing for no effect at all,  $H_0$ . Here, the column labelled ‘One’ is for the combined  
 621 test in Table 1, which performs a single test using all  $I = 2000$  pairs. In contrast, the column  
 622 labelled ‘Trunc’ performs a separate test for each subgroup and combines their  $p$ -values using  
 623 the truncated product of Zaykin et al. (2002) truncated at 0.05. The Supplementary Material  
 624

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

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

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

640

641

642

## 6. STRENGTH- $k$ MATCHING: NEAR-EXACT MATCHING

643

644

645

646

647

648

649

650

651

652

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.

653

654

655

656

657

658

659

660

661

Strength- $k$  balance means that every subset of  $k$  covariates is exactly balanced. Here, bal-  
 ance 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, 2015): essentially, one requests balance for  $C(V, k) = 20$  covariates formed as direct prod-  
 ucts of  $k$  of the  $V$  covariates, ignoring the origin of these 20 covariates as built from six covariates.

662

663

664

665

666

By definition, strength- $k$  matching means that, in addition to the constraint of strength- $k$  bal-  
 ance, 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.

667

668

669

670

671

672

The tree in Fig. 1 was built using the 1978 exactly matched pairs. Once Fig. 1 had selected  
 $2 \leq k$  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

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

#### 677 678 679 ACKNOWLEDGEMENT

680 We thank the referees, associate editor, and editor for their insightful comments. This work was  
 681 supported in part by the U.S. National Science Foundation and the Alfred P. Sloan Foundation.  
 682

#### 683 684 SUPPLEMENTARY MATERIAL

685 Supplementary material available at *Biometrika* online includes the details of implementing  
 686 the strength- $k$  match in § 6 using the R package ‘mipmatch’ (Zubizarreta, 2012, 2015), together  
 687 with a more extensive version of the simulation in § 5.  
 688

#### 689 690 691 REFERENCES

- 692 BENJAMINI, Y. & HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64**, 1215–22.  
 693 BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. & STONE, C. J. (1984). *Classification and Regression Trees*. New York:  
 694 Chapman and Hall/CRC.  
 695 CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M. & WYNDER, E. (1959). Smoking and  
 696 lung cancer. *J. Nat. Cancer Inst.* **22**, 173–203.  
 697 EGGLESTON, B. L., SCHARFSTEIN, D. O. & MACKENZIE, E. (2009). On estimation of the survivor average causal effect  
 698 in observational studies when important confounders are missing due to death. *Biometrics* **65**, 497–504.  
 699 FISHER, R. A. (1935). *Design of Experiments*. Edinburgh: Oliver & Boyd.  
 700 GASTWIRTH, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to  
 701 omitted variables. *Jurimetrics* **33**, 19–34.  
 702 HEDAYAT, A. S., SLOANE, N. J. A. & STUFKEN, J. (1999). *Orthogonal Arrays*. New York: Springer.  
 703 HOCHBERG, Y. & TAMHANE, A. C. (1987). *Multiple Comparison Procedures*. New York: Wiley.  
 704 HOSMAN, C. A., HANSEN, B. B. & HOLLAND, P. W. H. (2010). The sensitivity of linear regression coefficients’  
 705 confidence limits to the omission of a confounder. *Ann. Appl. Statist.* **4**, 849–70.  
 706 HSU, J. Y., SMALL, D. S. & ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational  
 707 studies. *J. Am. Statist. Assoc.* **108**, 135–48.  
 708 IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93**, 126–32.  
 709 JOGDEO, K. (1977). Association and probability inequalities. *Ann. Statist.* **5**, 495–504.  
 710 LIN, D. Y., PSATY, B. M. & KRONMAL, R. A. (1998). Assessing the sensitivity of regression results to unmeasured  
 711 confounders in observational studies. *Biometrics* **54**, 948–63.  
 712 LIU, W., KURAMOTO, S. J. & STUART, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding  
 713 in nonexperimental prevention research. *Prev. Sci.* **14**, 570–80.  
 714 MARCUS, R., ERIC, P. & GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis  
 715 of variance. *Biometrika* **63**, 655–60.  
 716 MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66**, 163–6.  
 717 MCCANDLESS, L. C., GUSTAFSON, P. & LEVY, A. (2007). Bayesian sensitivity analysis for unmeasured confounding  
 718 in observational studies. *Statist. Med.* **26**, 2331–47.  
 719 NEYMAN, J. (1923). On the application of probability theory to agricultural experiments (1990 translation). *Statist.*  
 720 *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.

- 721 ROSENBAUM, P. R. (2013a). Impact of multiple matched controls on design sensitivity in observational studies.  
722 *Biometrics* **69**, 118–27.
- 723 ROSENBAUM, P. R. (2013b). *Sensitivitymv: Sensitivity Analysis in Observational Studies*. R package version 1.2. Q1
- 724 ROSENBAUM, P. R. & SILBER, J. H. (2009a). Sensitivity analysis for equivalence and difference in an observational  
725 study of neonatal intensive care units. *J. Am. Statist. Assoc.* **104**, 501–11.
- 726 ROSENBAUM, P. R. & SILBER, J. H. (2009b). Amplification of sensitivity analysis in observational studies. *J. Am.*  
727 *Statist. Assoc.* **104**, 1398–405.
- 728 RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ.*  
729 *Psych.* **66**, 688–701.
- 730 WANG, L. & KRIEGER, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statist.*  
731 *Med.* **25**, 2257–71.
- 732 YANAGAWA, T. (1984). Case-control studies: Assessing the effect of a confounding factor. *Biometrika* **71**, 191–4.
- 733 YU, B. B. & GASTWIRTH, J. L. (2005). Sensitivity analysis for trend tests: Application to the risk of radiation exposure.  
734 *Biostatistics* **6**, 201–9.
- 735 ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. & WEIR, B. S. (2002). Truncated product method of combining  
736 *P*-values. *Genet. Epidemiol.* **22**, 170–85.
- 737 ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure  
738 after surgery. *J. Am. Statist. Assoc.* **107**, 1360–71.
- 739 ~~ZUBIZARRETA, J. R. (2015). *mipmatch: Optimal Matching in Observational Studies Using Mixed Integer Program-*  
740 *ming*. R package version 1.0, available at <http://www.columbia.edu/~jz2313/>.~~
- 741 ZUBIZARRETA, J. R., CERDA, M. & ROSENBAUM, P. R. (2013). Effect of the 2010 Chilean earthquake on posttraumatic  
742 stress: Reducing sensitivity to unmeasured bias through study design. *Epidemiol.* **24**, 79–87.

739 [Received May 2014. Revised April 2015]

740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768



## Supplementary material for ‘Strong control of the family-wise 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, and Department of Statistics, 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 of the University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.*

dsmall@wharton.upenn.edu    rosenbaum@wharton.upenn.edu

### 1. SPECIFICS OF STRENGTH $k$ MATCHING

#### 1.1. *Strength $k$ matching for studying effect modification*

In studying effect modification, it is convenient to have treatment-control pairs matched to have the same values of the effect modifiers. If a covariate is not exactly matched, if men are sometimes matched to women, then the treated-minus-control pair difference in outcomes  $Y_i$  may be associated with gender because of the mismatch on gender rather than because the treatment effect is different for men than for women. Expressed in familiar if imprecise terms, gender may have a main effect and an interaction with the treatment, and when pairs are exactly matched for gender the main effect is removed so the interaction can be seen clearly. In the example of the Chilean earthquake, it was not possible to match exactly for all  $V = 6$  candidate effect modifiers  $v_{ij}$ , yet we did not want to lose any pairs because of this.

It is easy to match to balance many covariates, perhaps by matching on the propensity score (Rosenbaum and Rubin 1985), or perhaps using fine balance (Rosenbaum et al. 2007), but even with nominal covariates, it is not possible to match everyone exactly for more than a few covariates because the number of combinations grows exponentially with the number of covariates. This is a common problem when more than a few covariates are candidates as effect modifiers.

In the earthquake example, there were  $V = 6$  candidate covariates  $v_{ij}$  that were plausible effect modifiers, yielding  $|\mathcal{V}| = 2 \times 2 \times 3 \times 3 \times 5 \times 2 = 360$  types of individuals, yet there were only  $I = 2106$  pairs of an exposed and an unexposed individual. In fact, only  $1978/2106 = 94\%$  of the pairs are exactly matched for all six basic covariates. However, by design, the six covariates exhibit strength 3 balance. In a strength 3 match, each of the  $C(V, 3) = C(6, 3) = 20$  combinations of 3 of the  $V$  covariates is exactly balanced: the marginal distribution is the same in treated and control matched pairs. Table 1 illustrates this in the case of gender, marital status

and self-rated health, but the same balance occurs for all 20 groups of three of the  $V = 6$  basic covariates. One of these 20 choices of 3 covariates had  $3 \times 3 \times 5 = 45$  categories, where all six covariates had  $2 \times 2 \times 3 \times 3 \times 5 \times 2 = 360$  categories.

How can the marginal distributions be identical with pairs that are not perfectly matched? A mismatch in one pair counterbalances a mismatch in another. Subject to the requirement (or constraint) of strength  $k$  balance, plus balance requirements on  $x_{ij}$ , the matching algorithm maximized the number of pairs that were exactly matched for  $v$ , so  $v_{i1} = v_{i2}$  as often as possible. Specifically,  $1978/2106 = 94\%$  of the pairs are exactly matched for all  $V = 6$  basic covariates  $v_{ij}$ , although 128 pairs could be balanced but not exactly matched. Because the six covariates are exactly balanced, whenever an inexact match does occur in the 128 inexact pairs, the mismatch is counterbalanced in another inexact pair. For example, there were  $15/2106 = 0.007$  pairs in which a treated individual with a health problem was paired to a control without a health problem, but this was counterbalanced by  $15/2106 = 0.007$  pairs in which a control with a health problem was paired to a treated individual without one.

With a strength  $k$  match, we build groups using just the exactly matched pairs, say the pairs  $i$  in  $\mathcal{I} = \{i : v_{i1} = v_{i2}\} \subseteq \{1, \dots, I\}$  where  $|\mathcal{I}| = 1978$ . If the resulting groups end up being defined using  $k$  or fewer variables, then the  $I - |\mathcal{I}| = 128$  inexact pairs can be rearranged to be exactly matched for all of the  $k$  or fewer variables that define the groups. In the example, the three groups were defined by two covariates, gender and income quartile, and the match was strength 3, so it was possible to break the original pairing of the 128 inexact pairs, and pair these  $2 \times 128$  individuals again to be exactly matched for the groups defined by the regression tree, with the consequence that all of the individuals in the original 2106 pairs were retained. The marginal distributions showing covariate balance in Figure 1 are unchanged because they do not depend upon who is paired with whom.

### 1.2. Implementing strength $k$ matching

Implementing strength  $k$  matching is straightforward using Zubizarreta (2012)'s `mipmatch` package in R. First, the  $C(V, k)$  nominal variables formed from  $k$  of the  $V$  basic variables  $v_{ij}$  are determined, and the match is constrained to perfectly balance all of these. Second, additional balance constraints are imposed on the remaining observed covariates  $x_{ij}$ . At this stage, the problem becomes an integer program, a constrained combinatorial optimization problem that `mipmatch` solves. The match maximizes the number of pairs subject to covariate balance constraints. Optionally, one may also use other standard matching techniques also available in `mipmatch`. In the current paper, as in Zubizarreta et al. (2014), we view matching and pairing as separate tasks: matching selects treated and control groups that exhibit covariate balance, whereas in the current paper the pairing in §4.2 focuses attention on the  $V$  nominal candidate effect modifiers.

The remainder of §1 describes the construction of the matched earthquake data; however, this material is not used later in the paper and may be skipped. The match was the largest possible match that exhibited certain stipulated and desired properties of covariate balance, a process called “cardinality matching” (Zubizarreta et al., 2014). We matched:

- (i) with exact pair matching for age groups, using 6 age groups, namely [15, 25), [25, 35), [35, 45), [45, 55), [55, 65), and [65, );
- (ii) to exactly balance the 20 possible 3-way interactions of sex, married or cohabitant, number of persons in the household (1, 2, 3, 4 or 5, 6 or more), health problem, health perception (poor, fair, good), and quartile of work income  $\{0, (0, 150000], (150000, \infty)\}$ ;

(iii) to force very similar means or proportions for age, marital status (divorced or widow, single), health problem, hospitalized, psychological problem, disability (self sufficient or low, moderate or severe, no, unknown), health insurance (public, private, other, no, unknown), years of education, employment status (employed, unemployed, inactive), work income, total income, poverty status, housing status (own housing or paying to own it, rented housing, ceded housing, irregular use of housing), housing structure (acceptable, reparable, irreparable), overcrowding (no, medium, critical), and an estimated propensity score. Here, we constrained the differences in means or proportions to be at most 0.1 times their standard deviations before matching; see Rosenbaum and Rubin (1985). This produced 2106 matched pairs meeting the covariate balance properties.

The re-pairing of the  $2 \times 128 = 2 \times (I - |\mathcal{I}|)$  individuals inexactly matched for  $v_{ij}$  used a distance that severely penalized mismatches for the three groups defined by Figure 1, but otherwise simply counted the number of mismatches, 0 to 6, on coordinates of  $v_{ij}$ ; see Rosenbaum (2010, §9.2). Among these 128 re-paired inexact pairs, none differed on the groups in Figure 1, 126 pairs differed on exactly one of the six basic covariates in  $v_{ij}$ , and 2 pairs differed on two of the basic covariates. Now, all  $I = 2106$  pairs are exactly matched for the variables that define the groups,  $\mathcal{G} = \{s_1, s_2, s_3\}$ .

## 2. SIMULATION: AN EXTENSIVE VERSION

In this more extensive version of the simulation, we consider (1) the effect of not being able to exactly match all pairs for all of the candidate effect modifiers and building the regression tree using only those pairs that can be exactly matched; (2) the effect of both the mean and variance differing among groups; and (3) different patterns of variation among the groups. A description of the extensive version of the simulation follows the notation and definitions in §5 of the paper. (i) Six potential effect modifiers: Each of these covariates is binary and independently follows six Bernoulli trials with probability of success 1/2. Of these, at most two of the covariates interact with the treatment to affect the responses, affecting either the mean or the variance of the pair differences  $Y_i$ , but it is left to the regression tree to discover which covariates affect the response. (ii) Building trees: In the limited version, all  $I = 2000$  pairs are used to build the regression tree. In this extensive version, only  $|\mathcal{I}| = 1000$  pairs are used to build the tree. In comparison with the earthquake data in §1, i.e., among  $I = 2106$  pairs,  $|\mathcal{I}| = 1978$  exactly matched pairs were used to build the tree. The tree is built from the CART regression to the ranks of 1,000 absolute pair differences  $|Y_i|$  on  $x_i$  for the 1,000 exact matched pairs. The remaining 1,000 pairs of the  $I = 2000$  pairs were classified using the tree constructed from the first 1,000 pairs. (iii) Sampling situations: Table 2 describes eight sampling situations with Normal errors and constant variance of matched pair differences,  $Y_i$ , whereas Table 3 permits the variance of  $Y_i$  to change with the covariates among eleven sampling situations. In Table 2, cases  $\mathbb{A}$ ,  $\mathbb{B}$ ,  $\mathbb{C}$ , and  $\mathbb{D}$  are the same scenarios as in Table 2 of the paper, except (ii) above. We also include four scenarios for strong effect modification,  $\mathbb{E}$ -1 to  $\mathbb{E}$ -4. In Table 3, the means for the different cases are the same as for the cases in Table 2, but the variances vary across four cells defined by the first two covariates and the average variance over the active cases is 1. For example, in case  $\mathbb{G}$  in Table 3, which is analogous to case  $\mathbb{B}$  in Table 2, the expected effect is constant but the variance changes.

To evaluate the groups, we quantify departures from perfection by  $\iota_{\mathcal{G}}$ , the fractional increase in the mean square error from grouping by  $\mathcal{G} = \{s_1, \dots, s_G\}$ . A perfect tree has  $\iota_{\mathcal{G}} = 1$ . We also compute  $\iota_{\mathcal{A}}$  where  $\mathcal{A}$  is a single group of all the pairs,  $\mathcal{A} = \{s_1\}$  with  $s_1 = \{1, \dots, I\}$ . In Tables 2 and 3, the mean of  $\iota_{\mathcal{G}}$  and  $\iota_{\mathcal{A}}$  is reported for nineteen sampling scenarios, each replicated 5,000

times. Without groups, the increase in mean square error ranges from  $\iota_A = 1.0$  for cases of  $\mathbb{A}$ ,  $\mathbb{B}$ ,  $\mathbb{F}-1$ ,  $\mathbb{F}-2$ , and  $\mathbb{G}$ , to  $\iota_A > 3$  in several other cases. In contrast, the groups formed from the tree  $\iota_G$  are typically much better; i.e.,  $1 \approx \iota_G \ll \iota_A$ . Tables 2 and 3 also record the number of trees, out of 5,000 trees, that had a single leaf, so CART produced just one group consisting of all  $I$  pairs. For instance, in cases of  $\mathbb{A}$  and  $\mathbb{B}$ , the pairs are homogeneous, and more than 4,000 of the 5,000 trees had a single leaf. The good power of CART groups in homogeneous cases like  $\mathbb{B}$  partly reflects CART's typical decision not to form subgroups in homogeneous cases.

In Tables 2 and 3, the column '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. The column 'Null rejects, family' is the proportion of applications of closed testing that issued in at least one false rejection, that is, a null leaf declared to have been affected. Here too, the 0.05 family-wise level appears to have been preserved. In addition to the two columns 'one' and 'trunc' under 'Rejecting  $H_0$ ', which appeared in Table 2 of the paper, the columns 'Fisher' and 'Simes' were added. 'Fisher' is Fisher's combination of  $P$ -values used in place of the truncated product. 'Simes' is the Simes method for combining independent  $P$ -values, and it is, by definition, a uniform improvement on use of the Bonferroni inequality. The truncated product and Simes method are similar and often best in terms of power, but they are not uniformly best; see, for instance, the homogeneous case  $\mathbb{B}$  where the combined method and Fisher's method win by a small margin.

The final column, 'Reject false  $H_0$ ', is the average over 5,000 replicates of the proportion of 1's among the affected pairs. For example, if this number were 0.5, then we expect half of the affected pairs to be in groups successfully identified by closed testing as 'non-null groups'. This section is blank when there is no expected effect, as in cases  $\mathbb{A}$ ,  $\mathbb{F}-1$ , and  $\mathbb{F}-2$ . In case  $\mathbb{E}-4$ , at  $\Gamma = 9$ , 56.2% of affected pairs were in groups where an effect was found, whereas the combined test found no evidence of an effect.

Particularly interesting is case  $\mathbb{G}$ . In case  $\mathbb{G}$ , the first two covariates affect the variance of  $Y_i$  but not its mean. Despite this, and consistent with results in Rosenbaum (2010, §15) and Zubizarreta et al. (2014), the single test has inferior power when compare to all of the tree-based methods that focus on subgroups.

As in Table 2 in the paper, Tables 2 and 3 of this supplementary materials shows that a single test for all pairs is substantially 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 in case  $\mathbb{B}$ . 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.

## REFERENCES

- ROSENBAUM, P. R. & RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Statist.* **39**, 33–8.
- ROSENBAUM, P. R., ROSS, R. N., SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Am. Statist. Assoc.* **102**, 75–83.
- ROSENBAUM, P. R. (2010). *Design of Observational Studies*. New York: Springer.
- 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 (R software `mipmatch` at <http://www-stat.wharton.upenn.edu/~josezubi/>)
- ZUBIZARRETA, J. R., PAREDES, R. D. & ROSENBAUM, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. App. Statist.* **8**, 204–31.

Table 1. *Covariate balance for individuals exposed to severe shaking from the earthquake and matched controls*

Gender	Marital status Self-rated health	Married/cohabitating			Other		
		Poor	Fair	Good	Poor	Fair	Good
Male	Exposed	18	299	167	5	145	47
	Control	18	299	167	5	145	47
Female	Exposed	40	542	339	21	280	203
	Control	40	542	339	21	280	203

One of 20 strength-3 tables of covariate balance, this one for the 3 covariates gender, marital status and self-rated health. In all cells, the count in the exposed group equals the count in the control group, and the same is true for the other 30 tables (not shown) describing 3 of the 6 balanced covariates.

Table 2. Summary of evaluating the groups, level of the tests, and power of the tests for the null hypothesis of no treatment effect with various  $\Gamma$  when matched pair differences have Normal errors and constant variance. Sensitivity analyses in the last seven columns are reported in per 1,000.

Scenario ( $\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}$ )	Perfect MSE ( $\iota_{\mathcal{G}}, \iota_{\mathcal{A}}$ )	# Trees 1-leaf	Leaves Null (Total)	$\Gamma$	False Rejections		Power to Reject $H_0$				Reject False $H_0$	
					All	Family	one	Fisher	Simes	trunc		
A: Null case, no effect (0, 0, 0, 0)	1 (1, 1)	4128	1-28 (1-28)	1	49	48	49	49	48	48		
				1-1	2	1	1	1	1	1		
				1-2	0	0	0	0	0	0	0	
				1-3	0	0	0	0	0	0	0	
B: A constant effect (0.5, 0.5, 0.5, 0.5)	1 (1, 1)	4122	0 (1-28)	1			1000	1000	1000	1000	1000	
				2-8			809	801	777	782	738	
				3			375	365	347	348	327	
				3-2			81	81	79	79	72	
C: Slight effect modification (0.6, 0.6, 0.4, 0.4)	0.26 (1.01, 1.01)	3094	0 (1.59)	1			1000	1000	1000	1000	1000	
				2-8			767	822	810	812	640	
				3			321	477	487	487	331	
				3-2			65	245	277	278	148	
D: Complex effect modification (1.5, 0, 0, 0.5)	0.29 (1.02, 1.37)	0	1.4 (3.58)	1	47	43	1000	1000	1000	1000	1000	
				2-3	0	0	817	1000	1000	1000	615	
				2-5	0	0	290	1000	1000	1000	581	
				15	0	0	0	803	993	997	487	
E-1: Stronger effect modification (2, 0, 0, 0)	1 (1, 1.75)	0	2-11 (3-13)	1	47	47	1000	1000	1000	1000	1000	
				2	0	0	721	1000	1000	1000	1000	
				2-1	0	0	381	1000	1000	1000	1000	
				30	0	0	0	858	999	999	994	
E-2: Stronger effect modification (1.5, 0.5, 0, 0)	0.93 (1, 1.37)	0	1-1 (3-3)	1	54	54	1000	1000	1000	1000	1000	
				2-3	0	0	818	1000	1000	1000	873	
				2-5	0	0	281	1000	1000	1000	759	
				20	0	0	0	222	826	822	398	
E-3: Stronger effect modification (1.2, 0.8, 0, 0)	0.9 (1, 1.27)	0	1-11 (3-31)	1	44	44	1000	1000	1000	1000	1000	
				2-3	0	0	967	1000	1000	1000	1000	
				2-5	0	0	631	1000	1000	1000	999	
				10	0	0	0	564	849	849	409	
E-4: Stronger effect modification (1, 1, 0, 0)	1 (1, 1.25)	0	1-11 (2-35)	1	48	47	1000	1000	1000	1000	1000	
				2-5	0	0	701	1000	1000	1000	1000	
				2-6	0	0	436	1000	1000	1000	1000	
				9	0	0	0	388	591	598	562	

NOTE: There are six potential effect modifiers following six independent Bernoulli trials with probability of success 1/2. At most two of the covariates interact with the treatment to affect the response. There are  $I = 2000$  matched pairs,  $Y_i \sim N(\mu_{ab}, 1)$ ,  $a = 0, 1$  and  $b = 0, 1$ . The statistic is  $M$ -statistic. Each scenario is sampled 5,000 times. The  $\iota_{\mathcal{G}}$  and  $\iota_{\mathcal{A}}$  quantify departures from perfection, where  $\mathcal{G} = \{s_1, \dots, s_G\}$  and  $\mathcal{A} = \{s_1\}$ . A perfect tree has  $\iota_{\mathcal{G}} = 1$ . ‘# Trees, 1-leaf’ is number of single-leaf tree among 5,000 replicates. ‘Leaves, Null, (Total)’ is averaged null (total) leaves over 5,000 replicates. ‘False Rejections, All’ is the proportion of null leaves in which the hypothesis of no effect was falsely rejected. ‘False Rejections, Family’ is the proportion of applications of closed testing that issued in at least one false rejection. ‘Power to Reject  $H_0$ ’ gives the power of four sensitivity analyses when testing no effect at all,  $H_0$ . Here, ‘one’ is the combined test, ‘Fisher’ is Fisher’s combination of  $P$ -values, ‘Simes’ is the Simes method for combining independent  $P$ -values, and ‘trunc’ is the truncated product. Finally, ‘Reject, False  $H_0$ ’ is 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 5,000 replicates.

Table 3. Summary of evaluating the groups, level of the tests, and power of the tests for the null hypothesis of no treatment effect with various  $\Gamma$  when matched pair differences have Normal errors and different variances. Sensitivity analyses in the last seven columns are reported in per 1,000.

Scenario ( $\mu_{00}, \mu_{01}, \mu_{10}, \mu_{11}$ ) ( $\sigma_{00}^2, \sigma_{01}^2, \sigma_{10}^2, \sigma_{11}^2$ )	Perfect MSE ( $\iota_G, \iota_A$ )	# Trees 1-leaf	Leaves Null (Total)	$\Gamma$			Power to Reject $H_0$				Reject False $H_0$
				All	Family	one	Fisher	Simes	trunc		
F-1: Null scenario, no effect (0, 0, 0, 0) (0.67, 1.33, 0.67, 1.33)	1 (1, 1)	0	2.38 (2.38)	1	51	49	57	56	51	51	
				1.1	3	1	1	1	1	1	
				1.2	0	0	0	0	0	0	
				1.3	0	0	0	0	0	0	
F-2: Null scenario, no effect (0, 0, 0, 0) (0.67, 1.33, 1.33, 0.67)	1 (1, 1)	1286	4.21 (4.21)	1	49	40	51	51	53	54	
				1.1	9	1	0	0	4	1	
				1.2	2	0	0	0	0	0	
				1.3	0	0	0	0	0	0	
G: A constant effect (0.5, 0.5, 0.5, 0.5) (0.25, 0.75, 1, 2)	1 (1, 1)	0	0 (4.29)	1			1000	1000	1000	1000	1000
				3			830	1000	1000	1000	390
				3.2			445	1000	1000	1000	332
				3.4			127	1000	1000	1000	297
H-1: Slight effect modification (0.75, 0.75, 0.25, 0.25) (1.33, 1.33, 0.67, 0.67)	1 (1, 1.06)	0	0 (2.33)	1			1000	1000	1000	1000	1000
				2.8			561	1000	1000	1000	495
				3			147	995	997	998	489
				3.2			19	972	987	990	477
H-1: Slight effect modification (0.75, 0.75, 0.25, 0.25) (0.67, 0.67, 1.33, 1.33)	0.03 (1.06, 1.06)	4074	0 (1.29)	1			1000	1000	1000	1000	1000
				2.8			785	791	768	771	715
				3			359	385	373	375	332
				3.2			70	113	113	113	82
I-1: Strong effect modification (1.5, 0.5, 0, 0) (1.8, 0.2, 1, 1)	0.95 (1, 1.37)	0	2.12 (4.26)	1	51	50	1000	1000	1000	1000	1000
				2.3	0	0	879	1000	1000	1000	980
				2.5	0	0	380	1000	1000	1000	978
				20	0	0	0	0	0	0	0
I-2: Strong effect modification (1.5, 0.5, 0, 0) (1.33, 0.67, 1, 1)	0.38 (1.02, 1.37)	0	1.11 (3.3)	1	51	51	1000	1000	1000	1000	1000
				2.3	0	0	816	1000	1000	1000	688
				2.5	0	0	279	1000	1000	1000	687
				20	0	0	0	0	49	43	21
I-3: Strong effect modification (1, 1, 0, 0) (1.33, 1.33, 0.67, 0.67)	1 (1, 1.25)	0	1.05 (2.3)	1	45	45	1000	1000	1000	1000	1000
				2.5	0	0	485	1000	1000	1000	1000
				2.6	0	0	230	1000	1000	1000	1000
				9	0	0	0	0	1	1	1
I-4: Strong effect modification (1, 1, 0, 0) (0.67, 0.67, 1.33, 1.33)	0.97 (1.01, 1.25)	98	1.16 (2.35)	1	50	49	1000	1000	1000	1000	1000
				2.5	0	0	884	993	993	993	990
				2.6	0	0	700	987	986	986	983
				9	0	0	0	975	976	976	969
J-1: Complex effect modification (1.5, 0, 0, 0.5) (1.8, 1, 1, 0.2)	0.92 (1, 1.37)	0	2.09 (4.22)	1	50	55	1000	1000	1000	1000	1000
				2.3	0	0	891	1000	1000	1000	963
				2.5	0	0	392	1000	1000	1000	962
				15	0	0	0	1	10	8	3
J-2: Complex effect modification (1.5, 0, 0, 0.5) (1.33, 1, 1, 0.67)	0.01 (1.03, 1.37)	0	1.08 (3.28)	1	52	52	1000	1000	1000	1000	1000
				2.3	0	0	820	1000	1000	1000	505
				2.5	0	0	280	1000	1000	1000	505
				15	0	0	0	100	582	568	276

NOTE: There are  $I = 2000$  matched pairs,  $Y_i \sim N(\mu_{ab}, \sigma_{ab}^2)$ ,  $a = 0, 1$  and  $b = 0, 1$ . The remaining settings are the same with scenarios in Table 2.

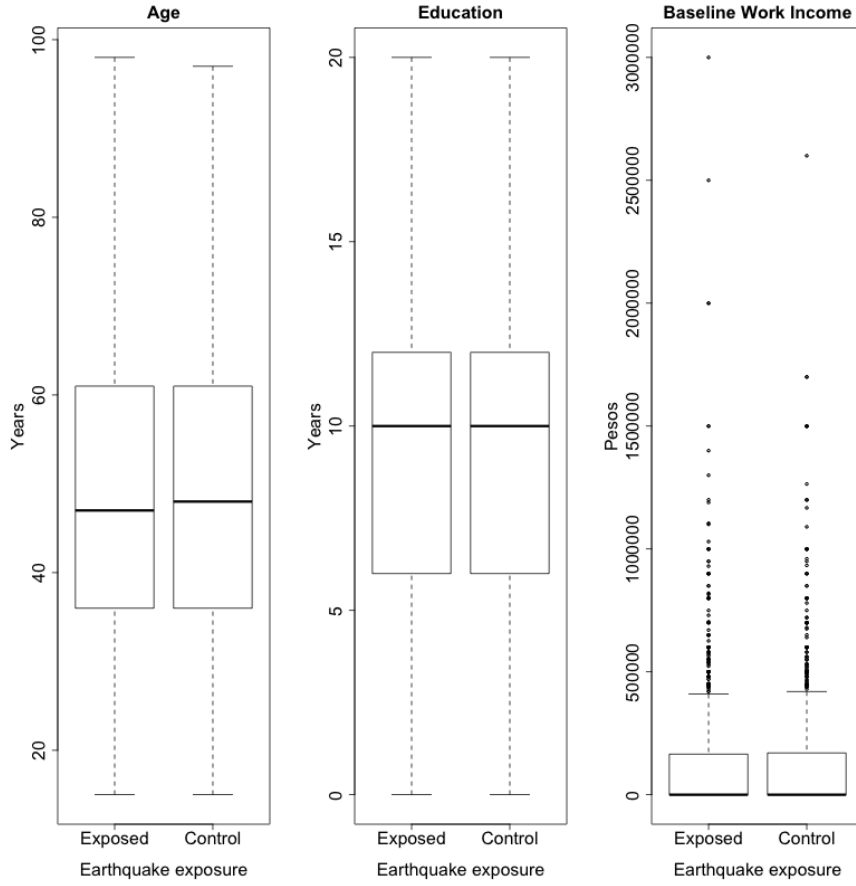


Fig. 1. Balance for three continuous covariates in 2009, before the earthquake, in  $I = 2016$  matched pairs containing one individual from a severely shaken region of Chile and one control from a region barely touched by the earthquake.