2013

Effect Modification and Design Sensitivity in Observational Studies

Jesse Y. Hsu
University of Pennsylvania

Dylan S. Small
University of Pennsylvania

Paul R. Rosenbaum
University of Pennsylvania

Recommended Citation

Hsu, J. Y., Small, D. S., & Rosenbaum, P. R. (2013). Effect Modification and Design Sensitivity in Observational Studies. *Journal of the American Statistical Association, 108* (501), 135-148. http://dx.doi.org/10.1080/01621459.2012.742018

Effect Modification and Design Sensitivity in Observational Studies

Abstract

In an observational study of treatment effects, subjects are not randomly assigned to treatment or control, so differing outcomes in treated and control groups may reflect a bias from nonrandom assignment rather than a treatment effect. After adjusting for measured pretreatment covariates, perhaps by matching, a sensitivity analysis determines the magnitude of bias from an unmeasured covariate that would need to be present to alter the conclusions of the naive analysis that presumes adjustments eliminated all bias. Other things being equal, larger effects tend to be less sensitive to bias than smaller effects. Effect modification is an interaction between a treatment and a pretreatment covariate controlled by matching, so that the treatment effect is larger at some values of the covariate than at others. In the presence of effect modification, it is possible that results are less sensitive to bias in subgroups experiencing larger effects. Two cases are considered: (i) an a priori grouping into a few categories based on covariates controlled by matching and (ii) a grouping discovered empirically in the data at hand. In case (i), subgroup specific bounds on *p*-values are combined using the truncated product of *p*-values. In case (ii), information that is fixed under the null hypothesis of no treatment effect is used to partition matched pairs in the hope of identifying pairs with larger effects. The methods are evaluated using an asymptotic device, the design sensitivity, and using simulation. Sensitivity analysis for a test of the global null hypothesis of no effect is converted to sensitivity analyses for subgroup analyses using closed testing. A study of an intervention to control malaria in Africa is used to illustrate.

Keywords

Fisher's combination of p-values, power of a sensitivity analysis, sensitivity analysis, Stephenson's test, truncated product of p-values, U-statistic, Wilcoxon test

Disciplines

Statistics and Probability

Effect Modification and Design Sensitivity in Observational Studies

Jesse Y. Hsu[1], Dylan S. Small, Paul R. Rosenbaum

University of Pennsylvania, Philadelphia

Summary. In an observational study of treatment effects, subjects are not randomly assigned to treatment or control, so differing outcomes in treated and control groups may reflect a bias from nonrandom assignment rather than a treatment effect. After adjusting for measured pretreatment covariates, perhaps by matching, a sensitivity analysis determines the magnitude of bias from an unmeasured covariate that would need to be present to alter the conclusions of the naive analysis that presumes adjustments eliminated all bias. Other things being equal, larger effects tend to be less sensitive to bias than smaller effects. Effect modification is an interaction between a treatment and a pretreatment covariate controlled by matching, so that the treatment effect is larger at some values of the covariate than at others. In the presence of effect modification, it is possible that results are less sensitive to bias in subgroups experiencing larger effects. Two cases are considered: (i) an a priori grouping into a few categories based on covariates controlled by matching, (ii) a grouping discovered empirically in the data at hand. In case (i), subgroup specific bounds on P-values are combined using the truncated product of P-values. In case (ii), information that is fixed under the null hypothesis of no treatment effect is used to partition matched pairs hoping to identify pairs with larger effects. The methods are evaluated using an asymptotic device, the design sensitivity, and using simulation. Sensitivity analysis for a test of the global null hypothesis of no effect is converted to sensitivity analyses for subgroup analyses using closed testing. A study of an intervention to control malaria in Africa is used to illustrate.

Keywords: Design sensitivity; Fisher's combination of P-values; power of a sensitivity analysis; observational study; sensitivity analysis; Stephenson's test; truncated product of P-values; U-statistic; Wilcoxon test

[1] *Address for correspondence:* Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. E-mail: hsu9@wharton.upenn.edu. 27 September 2012.

1 Treatment Effects that Vary with Covariates

1.1 If effect size varies with covariates, does sensitivity to bias vary in parallel?

In an observational study of treatment effects, subjects are not assigned at random to treatment or control, so they may differ visibly with respect to measured pretreatment covariates, $\mathbf{x}$, and may also differ with respect to a covariate not measured, u. Visible differences in $\mathbf{x}$ are removed by adjustments, such as matching, but there is invariably concern that adjustments failed to compare comparable individuals, that differing outcomes in treated and control groups reflect neither a treatment effect nor chance but rather a systematic bias from failure to control some unmeasured covariate, u. A sensitivity analysis asks: What would u have to be like in order to materially and substantively alter the conclusions of an analysis that presumes adjustments for the observed $\mathbf{x}$ suffice to eliminate bias?

The first sensitivity analysis in an observational study was conducted by Cornfield et al. (1959) in their discussion of heavy smoking as a cause of lung cancer, concluding that only very large biases could explain away the observed association as something other than an effect caused by smoking. Since then various methods of sensitivity analysis have been proposed; e.g., Rosenbaum and Rubin (1983), Yanagawa (1984), Rosenbaum (1987; 2002, §4), Manski (1990), Gastwirth (1992), Marcus (1997), Imbens (2003), Altonji et al. (2005), Yu and Gastwirth (2005), Small (2007), Ichino et al. (2008), Hosman et al. (2010), Millimet and Tchernis (2012), Pepper (2012), and Schwartz et al. (2012). Several of these methods place bounds on inference quantities, such as P-values or point estimates, for a specified magnitude of departure from random treatment assignment. For instance, the method in Rosenbaum (1987; 2002, §4) says that two subjects with the same observed covariates $\mathbf{x}$ may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$ because of differences in u, and for several values of Γ computes the possible range of inferences; this method is briefly reviewed in §2.2.

Once one can measure sensitivity to bias, it is natural to ask: What aspects of design and analysis affect sensitivity to bias? An aid to answering this question is the power of a sensitivity analysis and a number, the design sensitivity, that characterizes the power in large samples (Rosenbaum 2004). Some test statistics tend to exaggerate the reported sensitivity to unmeasured biases (Rosenbaum 2010a), whereas some design elements tend to make studies less sensitive to bias (Rosenbaum 2004; 2010b, Part III; 2011a). Generally,

larger effects are less sensitive than smaller ones. This last point suggests that effect modification — that is, an interaction between a pretreatment covariate and the magnitude of a treatment effect — might matter for sensitivity to unmeasured biases. Unfortunately, such an interaction may be uncertain or unexpected. How should one conduct a sensitivity analysis in the absence of a priori knowledge of where the effect with turn out to be large or small? Before developing the technical aspects, it is helpful to consider a motivating example.

1.2 Motivating example: malaria in West Africa

Working with the government of Nigeria, the World Health Organization contrasted several strategies to control malaria (Molineaux et al. 1980). We will look at one of these, namely spraying with an insecticide, propoxur, together with mass administration of a drug, sulfalene-pyrimethamine at high frequency. Matching for an observed covariate $\mathbf{x}$ consisting of age and gender, we paired 1560 treated subjects with 1560 untreated controls, making $I = 1560$ matched pairs. As is typically true in statistical applications of matching, there are $1560 + 1560 = 3120$ distinct individuals in the 1560 matched pairs — that is, no one is used twice. Also, the matching used only age, gender and assigned treatment and so is "on the basis of $\mathbf{x}$ alone" in the sense of Rosenbaum and Rubin (1985); therefore, if individuals were independent prior to matching, then outcomes in distinct pairs are conditionally independent given $\mathbf{x}$, treatment assignment and the pairing.

The outcome is a measure of the frequency of plasmodium falciparum in blood samples, that is, the frequency of a protozoan parasite that causes malaria. A slide containing blood is divided into fields, and the outcome is the number of fields with plasmodium falciparum per 200 fields examined. Blood samples were collected in a series of surveys, and we computed baseline scores using the four surveys (#5-8) immediately prior to treatment and posttreatment scores using the four surveys (9-12) immediately subsequent to treatment. To be included, an individual had to have at least two measurements from the four pretreatment surveys and at least two measurements from the four post treatment surveys. The 2-to-4 pretreatment measures were summarized using Huber's m-estimate with its scale parameter fixed to trim at 100, so it is essentially a mean but with a little control of wild fluctuations within a person. In the same way, the 2-to-4 posttreatment measures were summarized into one number per person. The data appendix contains details about the data and the matching.

Figure 1 displays (i) the close match for age, (ii) after-minus-before changes in parasite frequency in treated and control groups, ignoring the matching, (iii) the matched pair treated-minus-control difference in after-minus-before changes in parasite frequencies, and (iv) a density estimate for this difference in changes. Density estimates use the default settings in `R` but with double the default bandwidth. Although declines in parasite frequency are more common in the treated group, many differences in changes are close to zero.

Columns I and II of Table 1 display two sensitivity analyses for these 1560 matched pair differences, using the method reviewed in §2.2. The first analysis in column I uses Wilcoxon's signed rank test, which has the virtue of being familiar. Using Wilcoxon's statistic, we would judge the results to be sensitive to a bias of magnitude $\Gamma = 2$, because the upper bound on the one sided P-value testing no treatment effect exceeds the conventional 0.05 level. The other analysis in column II reports insensitivity to larger biases, and theory suggests this is to be expected, because Wilcoxon's statistic is an unwise choice in sensitivity analyses; see Rosenbaum (2010a, 2011b). In Table 1, the "U-statistic" $(m_1, m_2, m) = (7, 8, 8)$ is one of the more attractive members of the family of U-statistics discussed in Rosenbaum (2011b). The general statistic (m_1, m_2, m), with $1 \leq m_1 \leq m_2 \leq m$, looks at all $\binom{I}{m}$ subsets of m of the I pairs, sorting these m pairs into increasing order of the absolute pair difference in responses, and then totals the number of positive response differences among the pairs holding positions $m_1, \ldots, m_2$ in this order. The statistic $(m_1, m_2, m) = (1, 1, 1)$ is the sign statistic, and $(m_1, m_2, m) = (2, 2, 2)$ is the U-statistic that is virtually identical to Wilcoxon's signed rank statistic. The statistic $(m_1, m_2, m) = (m, m, m)$ was proposed by Stephenson (1981), and it approximates the locally optimal ranks for detecting a treatment effect that benefits some treated people and has no effect on many others; see Conover and Salsburg (1988) and Rosenbaum (2007a). Many members of this class of U-statistics report less sensitivity to unmeasured biases than does Wilcoxon's statistic. In particular, if the treatment effect shifts the distribution of differences, then $(m_1, m_2, m) = (7, 8, 8)$ has greater power in a sensitivity analysis than Wilcoxon's statistic for errors from the Normal, the logistic, and the t-distribution with 3 or 4 degrees of freedom. In Table 1, the U-statistic $(7, 8, 8)$ is insensitive at $\Gamma = 2.6$, in contrast to Wilcoxon's statistic which is sensitive at $\Gamma = 2$. Stephenson's statistic with $(m_1, m_2, m) = (m, m, m)$ is superior to $(7, 8, 8)$ when only some units respond to treatment, in the sense discussed by Conover and Salsburg (1988), and $(m_1, m_2, m) = (6, 7, 8)$ is superior for the longer tailed t-distribution with 2 degrees of freedom; see Rosenbaum (2011b, Table 3). In the current paper, we

do not want to focus attention on the choice of test statistic, so we use only Wilcoxon's statistic with $(m_1, m_2, m) = (2, 2, 2)$ and the U-statistic with $(m_1, m_2, m) = (7, 8, 8)$, often referring to the latter briefly as "the U-statistic".

Figure 2 splits the 1560 pairs into two groups, 447 pairs of young children aged 10 or less, and 1113 pairs of individuals older than ten years. The impression from Figure 2 is that the treatment was of much greater benefit to young children than to older individuals. Columns III-VI of Table 1 repeat the sensitivity analyses separately for young children and for older individuals. Despite the reduced sample size, the 447 pairs of young children exhibit an association with treatment that is far less sensitive to unmeasured bias than the full sample of 1560 pairs, with the U-statistic being insensitive at $\Gamma = 6$ even if the Bonferroni inequality is applied to double the smaller of two P-values. What is a good strategy for conducting a sensitivity analysis when the treatment effect may or may not vary across two or a few pretreatment subgroups?

2 Notation and Review: Experiments and Observational Studies

2.1 Randomization inference in experiments

There are I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated indicated by $Z_{ij} = 1$, the other control indicated by $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. The pairs are matched for observed covariates, $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_i$, say, but may possibly differ in terms of an unobserved covariate, $u_{i1} \neq u_{i2}$. A subject ij exhibits response r_{Tij} if treated, $Z_{ij} = 1$, or response r_{Cij} if control, $Z_{ij} = 0$, so ij actually exhibits response $R_{ij} = Z_{ij}\, r_{Tij} + (1 - Z_{ij})\, r_{Cij}$, and the effect of the treatment $r_{Tij} - r_{Cij}$ is not observed for any subject; see Neyman (1923), Welch (1937) and Rubin (1974). Fisher's (1935) sharp null hypothesis of no effect is $H_0 : r_{Tij} = r_{Cij}$, $i = 1, \ldots, I$, $j = 1, 2$. Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}),\, i = 1, \ldots, I,\, j = 1, 2\}$, and write $\mathbf{Z} = (Z_{11}, \ldots, Z_{I2})^T$, $\mathbf{R} = (R_{11}, \ldots, R_{I2})^T$, $\mathbf{r}_C = (r_{C11}, \ldots, r_{CI2})^T$, for the $2I$-dimensional vectors, with a similar notation for $\mathbf{r}_T$ and $\mathbf{u}$. Let $\mathcal{Z}$ be the set containing the 2^I possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \mathcal{Z}$ if $z_{ij} \in \{0, 1\}$ with $z_{i1} + z_{i2} = 1$ for each i, and in conditional probabilities abbreviate conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ as conditioning on $\mathcal{Z}$. Write $|A|$ for the number of elements in a finite set A, so $|\mathcal{Z}| = 2^I$. In a randomized paired experiment, $\Pr(\mathbf{Z} = \mathbf{z} \,|\, \mathcal{F},\, \mathcal{Z}) = 2^{-I}$ for each $\mathbf{z} \in \mathcal{Z}$. If H_0 is true, then $\mathbf{R} = \mathbf{r}_C$, and in a randomized experiment the null distribution $\Pr(T \geq v \,|\, \mathcal{F},\, \mathcal{Z})$ of any test statistic $T = t(\mathbf{Z}, \mathbf{R}) = t(\mathbf{Z}, \mathbf{r}_C)$ is simply the proportion of treatment assignments $\mathbf{z} \in \mathcal{Z}$ with $t(\mathbf{z}, \mathbf{r}_C) \geq v$ — that is, $\Pr(T \geq v \,|\, \mathcal{F},\, \mathcal{Z}) = |\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq v\}| \,/\, |\mathcal{Z}|$ — because $\mathbf{r}_C$ is

fixed by conditioning on $\mathcal{F}$ and $\mathbf{Z}$ is uniform on $\mathcal{Z}$.

Let Y_i be the treated-minus-control difference in observed responses in pair i, $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$, so that $Y_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) = \pm(r_{Ci1} - r_{Ci2})$ if H_0 is true. Let $q_i \geq 0$ be a function of $|Y_i|$ such that $q_i = 0$ if $|Y_i| = 0$, and let $\text{sgn}(y) = 1$ if $y > 0$ and $\text{sgn}(y) = 0$ if $y \leq 0$, so that under H_0 in a paired randomized experiment, the test statistic $T = \sum \text{sgn}(Y_i)\, q_i$ is the sum of I independent random variables, $i = 1, \ldots, I$, taking the value 0 with probability 1 if $q_i = 0$ and otherwise taking the values q_i and 0 with equal probabilities 1/2. For instance, if q_i is the rank of $|Y_i|$, then this yields the familiar null distribution of Wilcoxon's signed rank statistic, and many other statistics may be expressed in this form, including the permutational t-statistic (Welch 1937), tests based on order statistics (Noether 1973, Brown 1981), M-statistics (Maritz 1979), and various U-statistics (Stephenson 1981, Brown and Hettmansperger 1994, Rosenbaum 2011b).

2.2 Sensitivity analysis in observational studies: bounds on inferences for biases of limited magnitude

A simple model for sensitivity analysis in an observational study says that in the population before matching, treatment assignment probabilities $\pi_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$ are unknown but two subjects, ij and $i'j'$ with the same observed covariates may differ in their odds of treatment by at most $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\left(1-\pi_{i'j'}\right)}{\pi_{i'j'}\left(1-\pi_{ij}\right)} \leq \Gamma \text{ whenever } \mathbf{x}_{ij} = \mathbf{x}_{i'j'}, \tag{1}$$

and then returns the distribution of $\mathbf{Z}$ to $\mathcal{Z}$ by conditioning on $Z_{i1} + Z_{i2} = 1$ for all i in pairs matched for $\mathbf{x}$. Write $\mathcal{U} = [0,1]^{2I}$ for the $2I$-dimensional unit cube. It is easy to check that (1) and conditioning on $\mathbf{Z} \in \mathcal{Z}$ is the same as assuming

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

where $\gamma = \log(\Gamma)$. See Rosenbaum (2002, §4) for the straightforward derivation, in which u_{ij} is obtained from π_{ij} in (1) as $u_{ij} = [\log(\pi_{ij}) - \min\{\log(\pi_{i1}), \log(\pi_{i2})\}]/\gamma$. Let $\overline{\overline{T}}$ be the sum of I independent random variables taking the value 0 with probability 1 if $q_i = 0$ and otherwise the value q_i with probability $\Gamma/(1+\Gamma)$ and the value 0 with probability $1/(1+\Gamma)$, and define $\overline{T}$ similarly but with $\Gamma/(1+\Gamma)$ and $1/(1+\Gamma)$ interchanged. It is

straightforward to show that under (2), if H_0 is true, then

$$\Pr\left(\overline{T} \geq v \mid \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(T \geq v \mid \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(\overline{\overline{T}} \geq v \mid \mathcal{F}, \mathcal{Z}\right) \text{ for all } \mathbf{u} \in \mathcal{U}, \tag{3}$$

and, as $I \rightarrow \infty$, the upper bound $\Pr\left(\overline{\overline{T}} \geq v \mid \mathcal{F}, \mathcal{Z}\right)$ in (3) may be approximated by

$$\Pr\left(\overline{\overline{T}} \geq v \mid \mathcal{F}, \mathcal{Z}\right) \doteq 1 - \Phi\left[\frac{v - \{\Gamma/(1+\Gamma)\}\sum q_i}{\sqrt{\left\{\Gamma/(1+\Gamma)^2\right\}\sum q_i^2}}\right]; \tag{4}$$

see Rosenbaum (1987; 2002, §4; 2007b). The upper bounds on the P-values in Table 1 are obtained from (4) with v replaced by the observed value of the test statistic T.

The U-statistic (m_1, m_2, m) is a signed rank statistic with $q_i = 0$ if $|Y_i| = 0$ and otherwise

$$q_i = \binom{I}{m}^{-1} \sum_{\ell=m_1}^{m_2} \binom{a_i - 1}{\ell - 1}\binom{I - a_i}{m - \ell} \tag{5}$$

where a_i is the rank of $|Y_i|$, and $\binom{A}{B}$ is defined to equal zero for $B < 0$; see Rosenbaum (2011b, §3.1).

2.3 Sensitivity analyses for point estimates, confidence intervals and equivalence tests

As is commonly done, point estimates, confidence intervals and equivalence tests are formed by inverting tests of the hypothesis H_0 of no effect. For instance, the hypothesis of an additive treatment effect, $H_{\tau_0} : r_{Tij} = r_{Cij} + \tau_0$, $\forall ij$, implies the treated-minus-control pair difference, Y_i, equals $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2}) = \tau_0 + \epsilon_i$ where $\epsilon_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2})$. If H_{τ_0} were true, then $Y_i^{'} = Y_i - \tau_0$ would satisfy the null hypothesis of no treatment effect, so H_{τ_0} may be tested in a randomized experiment by applying the methods in §2.1 to $Y_i^{'}$, and sensitivity to bias may be examined by applying the methods in §2.2 to $Y_i^{'}$.

The hypothesis of a Tobit effect ϱ_0 asserts $H_{\varrho_0} : r_{Tij} = \max(r_{Cij} - \varrho_0, 0)$ and it is more appropriate for a response, such as the frequency of plasmodium falciparum, that can equal zero but cannot be negative: it says the treatment may drive a positive response under control, $r_{Cij} > 0$, to zero under treatment, $r_{Tij} = 0$, but not beyond zero. In parallel with the hypothesis of an additive effect, if H_{ϱ_0} were true, then

$R_{ij}^{'} = \max\{R_{ij} - (1 - Z_{ij})\,\varrho_0,\, 0\} = r_{Tij}$ satisfies the null hypothesis of no treatment effect, and $Y_i^{''} = (Z_{i1} - Z_{i2})\left(R_{i1}^{'} - R_{i2}^{'}\right) = (Z_{i1} - Z_{i2})\,(r_{Ti1} - r_{Ti2})$, so again the methods in §2.1 and §2.2 may be applied to $Y_i^{''}$ to test H_{ϱ_0}. See Rosenbaum (2010b, §2.4.5) for discussion, an example, and analogous developments for other hypotheses about treatment effects.

By the duality of confidence intervals and hypothesis tests (e.g., Lehmann and Romano 2005, §3.5), a $1-\alpha$ confidence interval for an additive effect τ or a Tobit effect ϱ is formed by testing each hypothesis, H_{τ_0} or H_{ϱ_0}, and retaining for the confidence interval the values not rejected at level α. See, for instance, Maritz (1979). The sensitivity of confidence intervals to unmeasured biases is analogous: if, for a specific $\Gamma \geq 1$, the upper bound on the P-value testing H_{τ_0} is $\leq \alpha$ then at this Γ, the value τ_0 is excluded from the $1-\alpha$ confidence interval for τ in the presence of a bias no larger than Γ; see Rosenbaum (2002, §4.3.5) for details and a numerical example.

In a randomized experiment, point estimates of τ or ϱ are obtained from tests by the device of Hodges and Lehmann (1963): the hypothesis, H_{τ_0} or H_{ϱ_0}, which equates the test statistic T to its null expectation under randomization is the point estimate. In a sensitivity analysis for fixed $\Gamma \geq 1$, there is not one null expectation of T but rather an interval of possible expectations, and this yields an interval of point estimates for each Γ, the interval becoming longer as Γ increases; see Rosenbaum (1993; 2002, §4.3.4; 2007b).

An equivalence test is a test of the null hypothesis that a treatment effect is of substantial magnitude, so rejection in an equivalence test is evidence that the treatment effect is not of substantial magnitude; see, for instance, Bauer and Kieser (1996) or Berger and Hsu (1996). An equivalence test correctly replaces the common error of taking failure to reject a null hypothesis of no effect as evidence that the null hypothesis is approximately true. With an additive treatment effect, $r_{Tij} = r_{Cij} + \tau$, the null hypothesis of inequivalence asserts $H_\iota : |\tau| \geq \iota$ and rejection of H_ι is evidence in favor of $|\tau| < \iota$, where $\iota > 0$ defines a negligible effect. As discussed by Bauer and Kieser (1996) or Berger and Hsu (1996), the two-one-sided-test procedure may be used: specifically H_ι may be rejected at level α if both (i) for all $\tau_0 \geq \iota$, the null hypothesis H_{τ_0} is rejected at level α in a one-sided test against the alternative that $H_{\tau_0}^{'} : \tau > \tau_0$, and (ii) for all $\tau_0 \leq -\iota$, the null hypothesis H_{τ_0} is rejected at level α in a one-sided test against the alternative that $H_{\tau_0}^{'} : \tau < \tau_0$. In an observational study, an apparent absence of treatment effect may be highly insensitive to unmeasured biases, so that only a large bias, measured by Γ, could have masked a large treatment effect as an apparent absence of effect. Again, an analogous procedure is used to

conduct a sensitivity analysis for an equivalence test: for a specific $\Gamma \geq 1$, the hypothesis of inequivalence $H_\iota : |\tau| \geq \iota$ is rejected at level α if the upper bounds on the P-values testing H_{τ_0} would lead to rejection in an equivalence test; see Rosenbaum and Silber (2009) where an absence of cost-savings from a certain medical practice was found to be insensitive to unmeasured biases.

2.4 Design sensitivity: the sensitivity of a data generating process

If after matching for observed covariates, an observational study were free of bias from unmeasured covariates u in the sense that $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 2^{-I}$, and if the association between treatment Z and response R were the consequence of a treatment effect, not bias, then there would be no way to know this from the data. Call the situation just described, with an effect and no unmeasured bias, the "favorable situation." An investigator cannot know if she is in the favorable situation, and the best she can hope to say is that the conclusions are insensitive to small and moderate biases as measured by Γ. The power of a sensitivity analysis is the probability that she will be able to say this when she is indeed in the favorable situation. That is, for a specific Γ, the power of an α-level sensitivity analysis is the probability that the upper bound on the P-value in (4) is less than or equal to α, this probability being computed in the favorable situation.

In the favorable situation, there is typically a value $\widetilde{\Gamma}$ called the design sensitivity such that, as the sample size increases, $I \rightarrow \infty$, the upper bound on the P-value in (4) tends to zero when the analysis is performed with $\Gamma < \widetilde{\Gamma}$ and it tends to 1 when the analysis is performed with $\Gamma > \widetilde{\Gamma}$. Somewhat more precisely, if the Y_i are independent and identically distributed observations from some distribution, and if H_0 is rejected for a specific $\Gamma \geq 1$ when the upper bound on the P-value in (4) is $\leq \alpha$, conventionally $\alpha = 0.05$, then the probability of rejection or the power of the sensitivity analysis is tending to 1 for $\Gamma < \widetilde{\Gamma}$ and to 0 for $\Gamma > \widetilde{\Gamma}$ as $I \rightarrow \infty$; see Rosenbaum (2004; 2010b, Part III). For example, if $Y_i = \tau + \varepsilon_i$ where the ε_i are sampled from the standard Normal distribution and $\tau = 1/2$, then $\widetilde{\Gamma} = 3.2$ for Wilcoxon's signed rank statistic and $\widetilde{\Gamma} = 5.1$ for the U-statistic $(m_1, m_2, m) = (7, 8, 8)$, whereas if $\tau = 1$ and the ε_i are sampled from the t-distribution on 3 degrees of freedom, the corresponding design sensitivities are $\widetilde{\Gamma} = 6.0$ for Wilcoxon's statistic and $\widetilde{\Gamma} = 6.8$ for the U-statistic; see Rosenbaum (2011b, Table 3).

3 Effect Modification in a Few Nonoverlapping Prespecified Groups

3.1 Combining independent P-values using their truncated product

Let P_ℓ, $\ell = 1, \ldots, L$, be valid, statistically independent P-values testing hypotheses H_ℓ, $\ell = 1, \ldots, L$, respectively, so $\Pr(P_\ell \leq \alpha) \leq \alpha$ for all $\alpha \in [0, 1]$ if H_ℓ is true. In the context of the current paper, the I pairs have been partitioned into L nonoverlapping groups of pairs based on a pretreatment covariate controlled by matching, and H_ℓ asserts that there is no treatment effect in group ℓ. The conjunction $H_\wedge = H_1 \wedge H_2 \wedge \cdots \wedge H_L$ asserts that all L hypotheses are true, so $H_\wedge$ is Fisher's H_0 in §2.1. Fisher (1932) proposed testing the conjunction $H_\wedge$ using minus twice the log of the product of the P-values, $\prod_{\ell=1}^{L} P_\ell$, which is stochastically smaller than the chi-square distribution on $2L$ degrees of freedom if $H_\wedge$ is true, and is exactly chi-square distributed on $2L$ degrees if additionally $\Pr(P_\ell \leq \alpha) = \alpha$ when H_ℓ is true.

In general, a valid P-value must satisfy so $\Pr(P_\ell \leq \alpha) \leq \alpha$ if H_ℓ is true, but it need not satisfy $\Pr(P_\ell \leq \alpha) = \alpha$. If H_ℓ is a composite hypothesis, then the composite hypothesis may be true in a manner such that $\Pr(P_\ell \leq \alpha) < \alpha$. For instance, if H_ℓ asserts that the expectation is nonpositive for independent Normal random variables with constant variance, then H_ℓ is true if the expectation is strictly negative, but in this case the P-value from the t-test satisfies $\Pr(P_\ell \leq \alpha) < \alpha$. As a consequence, Fisher's method may be quite conservative when $H_\wedge$ is false but many H_ℓ are true with $\Pr(P_\ell \leq \alpha) < \alpha$, because this makes $\prod_{\ell=1}^{L} P_\ell$ excessively large.

This phenomenon occurs in an acute fashion in sensitivity analyses. If the L tests have different design sensitivities, then for certain values of Γ some of the L P-values P_ℓ are tending to zero with increasing sample size while others are tending to one. As a consequence, for some values of Γ, one may see several very small P_ℓ and many others that are near 1. The relevant question is whether there are an excess of very small P_ℓ's.

With general issues of this sort in mind, Zaykin, Zhivotovsky, Westfall and Weir (2002) proposed testing $H_\wedge$ using a truncated product of P-values, $P_\wedge = \prod_{\ell=1}^{L} P_\ell^{\chi(P_\ell \leq \widetilde{\alpha})}$, where $\chi(E) = 1$ if event E occurs and $\chi(E) = 0$ otherwise, so $P_\wedge$ is the product of the P-values that are less than or equal to $\widetilde{\alpha}$. Taking $\widetilde{\alpha} = 1$ yields Fisher's statistic, but taking $\widetilde{\alpha} = \alpha = 0.1$ computes the product of those P-values less than or equal to 0.1.

Zaykin et al. (2002, p. 173) give the distribution of $P_\wedge$ when $\Pr(P_\ell \leq \alpha) = \alpha$ and this yields the needed null reference distribution of $P_\wedge$ when $\Pr(P_\ell \leq \alpha) \leq \alpha$. For $\widetilde{\alpha} < 1$, the

truncated product $P_\wedge$ is a larger number than Fisher's $\prod_{\ell=1}^{L} P_\ell$, so the null distribution of $P_\wedge$ is stochastically larger than the gamma distribution for Fisher's procedure, and the P-value from $P_\wedge$ can be either larger or smaller than the P-value from Fisher's $\prod_{\ell=1}^{L} P_\ell$. As an example, consider the case of $L = 2$ hypotheses with $\widetilde{\alpha} = 0.05$, so only P-values ≤ 0.05 are included in $P_\wedge$. The Bonferroni inequality would reject at level 0.05 with $L = 2$ hypotheses if $\min(P_1, P_2) \leq 0.05/2 = 0.025$, and in this case $P_\wedge \leq 0.025$ and $\prod_{\ell=1}^{L} P_\ell \leq 0.025$; however, Fisher's method gives P-value 0.1172 if $\prod_{\ell=1}^{L} P_\ell = 0.025$ whereas the method of Zaykin et al. (2002) gives P-value 0.05 if $P_\wedge = 0.025$, so $P_\wedge$ rejects whenever the Bonferroni inequality rejects and would also reject if $P_1 = P_2 = 0.05$, but Fisher's method may not reject when $\min(P_1, P_2) \leq 0.025$.

Zaykin et al. obtain the distribution of $P_\wedge$ by a calculus argument, but it may alternatively but equivalently be written as a binomial mixture of gamma distributions. In this paragraph, the exponential and gamma distributions refer to their standard forms with scale parameter equal to one. Let $F_k(\cdot)$ be the cumulative gamma distribution with shape parameter k, so that, in particular, $F_k(w) = 0$ for $w < 0$, and recall that the sum of k independent exponential random variables has distribution $F_k(\cdot)$. If the P_ℓ are independent uniform random variables, then for $0 < w \leq 1$

$$\Pr(P_\wedge \leq w) = \sum_{k=1}^{L} \binom{L}{k} \widetilde{\alpha}^k (1-\widetilde{\alpha})^{L-k} \left[1 - F_k\left\{-\log\left(\frac{w}{\widetilde{\alpha}^k}\right)\right\}\right] \tag{6}$$

or in `R`,

$$\Pr(P_\wedge \leq w) = \texttt{sum}(\texttt{dbinom}(\texttt{1:L}, \texttt{L}, \widetilde{\alpha}) * (\texttt{1} - \texttt{pgamma}(-\texttt{log}(\texttt{w}/(\widetilde{\alpha}\hat{}(\texttt{1:L}))), \texttt{1:L}))).$$

To see (6), recall that: (i) $-\log(P)$ is exponential if P is uniform, (ii) by the memoryless property of the exponential distribution, the conditional distribution of $E = -\log(P/\widetilde{\alpha})$ given $P \leq \widetilde{\alpha}$ is exponential, (iii) the probability that exactly k of the L independent uniforms are less than or equal to $\widetilde{\alpha}$ is $\binom{L}{k}\widetilde{\alpha}^k (1-\widetilde{\alpha})^{L-k}$, and (iv) the conditional cumulative distribution of $P_\wedge$ given exactly k of the L independent uniforms are less than or equal to $\widetilde{\alpha}$ is $1 - F_k\left\{-\log\left(w/\widetilde{\alpha}^k\right)\right\}$. If the P_ℓ are independent and stochastically larger than the uniform, $\Pr(P_\ell \leq \alpha) \leq \alpha$, then $\Pr(P_\wedge \leq w)$ is less than or equal to the right side of (6). Indeed, if the P_ℓ are dependent but $(P_1, \ldots, P_L)^T$ is stochastically larger than the uniform distribution on the L-dimensional unit cube $[0,1]^L$, then $\Pr(P_\wedge \leq w)$ is less than or equal to the right side of (6); see Brannath, Posch and Bauer (2002) and Rosenbaum (2011c, §2)

for discussion of dependent P-values of this form.

3.2 Using the truncated product in sensitivity analysis

Suppose the I pairs are divided to L groups, $\ell = 1, \ldots, L$, based on mutually exclusive and exhaustive categories formed from an observed covariate, $\mathbf{x}_{ij}$, controlled by matching, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i. In §1.2, the pairs were divided into $L = 2$ nonoverlapping groups based on an age less than or equal to ten years or an age greater than ten years, where the pairs were matched for age. Because these categories do not overlap and distinct pairs are independent, analyses performed separately in each of the L groups are independent. Let H_ℓ be the hypothesis asserting that there is no treatment effect in the ℓth group of pairs, so Fisher's hypothesis of no effect asserts that H_ℓ is true for every ℓ, $\ell = 1, \ldots, L$, or $H_0 = H_1 \wedge H_2 \wedge \cdots \wedge H_L$. For $\ell = 1, \ldots, L$, using just the pairs in group ℓ, let P_ℓ be a P-value testing H_ℓ using the pairs in group ℓ and computed from (2) for a specific unknown $\mathbf{u}$ and $\gamma = \log(\Gamma)$, and let $\overline{\overline{P}}_{\Gamma\ell}$ be the corresponding upper bound in (3). For instance, in Table 1, $\overline{\overline{P}}_{\Gamma 1}$ and $\overline{\overline{P}}_{\Gamma 2}$ were computed separately for those under and over ten years of age for several Γ.

If the bias is at most Γ, then (3) implies $P_\ell \leq \overline{\overline{P}}_{\Gamma\ell}$ and $\Pr\left(\overline{\overline{P}}_{\Gamma\ell} \leq \alpha \,|\, \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(P_\ell \leq \alpha \,|\, \mathcal{F}, \mathcal{Z}\right) \leq \alpha$ for $\alpha \in [0,1]$, $\ell = 1, \ldots, L$. Because the truncated product is a monotone increasing function, it follows that $\overline{\overline{P}}_{\Gamma\wedge} = \prod_{\ell=1}^{L} \left(\overline{\overline{P}}_{\Gamma\ell}\right)^{\chi\left(\overline{\overline{P}}_{\Gamma\ell} \leq \widetilde{\alpha}\right)}$ is an upper bound for $P_\wedge = \prod_{\ell=1}^{L} P_\ell^{\chi(P_\ell \leq \widetilde{\alpha})}$. Combining these two facts, if the bias is at most Γ then $\Pr\left(\overline{\overline{P}}_{\Gamma\wedge} \leq w \,|\, \mathcal{F}, \mathcal{Z}\right) \leq \Pr\left(P_\wedge \leq w \,|\, \mathcal{F}, \mathcal{Z}\right)$ where $\Pr\left(P_\wedge \leq w \,|\, \mathcal{F}, \mathcal{Z}\right)$ is at most (6). If we calculate w such (6) equals α, conventionally $\alpha = 0.05$, and if we reject H_0 when $\overline{\overline{P}}_{\Gamma\wedge} \leq w$, then we will falsely reject H_0 with probability at most α if the bias is at most Γ.

Columns VII and VIII of Table 1 perform these calculations for the malaria data using $\widetilde{\alpha} = 0.05$. In 1, $\overline{\overline{P}}_{\Gamma\ell}$, $\ell = 1, 2$ are computed for young and old pairs, and these are combined into the truncated product $\overline{\overline{P}}_{\Gamma\wedge}$, whose P-value is determined from (6). The results in columns VII and VIII Table 1 testing H_0 using $\overline{\overline{P}}_{\Gamma\wedge}$ are much less sensitive to bias than the results in columns I and II using all of the pairs in a single analysis. To emphasize, combining two independent sensitivity analyses yields less sensitivity to unmeasured bias than a single sensitivity analysis that uses all of the data, and this occurred because the treatment effect appears to be much larger for children aged 10 or less. Indeed, the sensitivity Γ for $\overline{\overline{P}}_{\Gamma\wedge}$ is only slightly worse than knowing a priori that attention should focus on the young pairs in Table 1.

The degree of sensitivity to unmeasured bias is a measurable fact in the data — for example, H_0 becomes marginally plausible in the last column of Table 1 for $\Gamma > 6$ — but the actual degree of bias in treatment assignment probabilities is not known. A bias of $\Gamma = 6$ is a large bias, sufficient in magnitude to explain away the effects of heavy smoking on lung cancer in Hammond's (1964) study; see Rosenbaum (2002, §4.3.2). The last column of Table 1 says that if the bias from unobserved covariates in (1) is at most $\Gamma = 6$ for all values of the observed covariates $\mathbf{x}$, then such a bias is too small to render plausible the null hypothesis H_0 of no effect. It is, of course, possible that the maximum degree of bias in (1) is different for different values of $\mathbf{x}$, but it would have to exceed $\Gamma = 6$ for at least some value of $\mathbf{x}$ to render H_0 plausible. If one had reason to believe that departures from random assignment were smaller at some values of $\mathbf{x}$ than at others, then a more complex sensitivity analysis could have a different $\Gamma_{\mathbf{x}}$ for each $\mathbf{x}$. In the example with $L = 2$, $\ell = 1$ for young, $\ell = 2$ for old, Table 1 would then vary a 2-dimensional sensitivity parameter, (Γ_1, Γ_2), comparing $\prod_{\ell=1}^{L}\left(\overline{\overline{P}}_{\Gamma_\ell,\ell}\right)^{\chi\left(\overline{\overline{P}}_{\Gamma_\ell,\ell}\leq\widetilde{\alpha}\right)}$ to (6). The last column in Table 1 would then correspond with $\Gamma_1 = \Gamma_2 = \Gamma$.

3.3 Design sensitivity of the truncated product of P-values

Proposition 1 indicates that the pattern seen in Table 1 is the pattern expected in general if the sample size, I, is sufficiently large. Although Proposition 1 is stated in terms of matched pairs, a similar result and proof would apply for many situations that yield upper bounds on P-values.

Proposition 1 *Suppose there are L nonoverlapping subgroups of pairs defined by a covariate controlled by matching, and allow the number of pairs I to increase, $I \rightarrow \infty$, with L fixed, in such a way that the fraction of pairs in each subgroup is tending to a nonzero constant. Suppose that the test in subgroup ℓ has design sensitivity $\widetilde{\Gamma}_\ell$. Then for any $\widetilde{\alpha}$ with $0 < \widetilde{\alpha} \leq 1$, the design sensitivity of the truncated product $P_\wedge$ is $\widetilde{\Gamma}_{\max} = \max\left(\widetilde{\Gamma}_1, \ldots, \widetilde{\Gamma}_L\right)$.*

Proof. If the sensitivity analysis is performed at a $\Gamma > \widetilde{\Gamma}_{\max}$, then all L bounds $\overline{\overline{P}}_{\Gamma\ell}$ on P-values are tending to 1 as $I \rightarrow \infty$, so $\overline{\overline{P}}_{\Gamma\wedge}$ is tending to 1. Let ℓ' be any ℓ such that $\widetilde{\Gamma}_{\ell'} = \widetilde{\Gamma}_{\max}$. If the sensitivity analysis is performed with $\Gamma < \widetilde{\Gamma}_{\max} = \widetilde{\Gamma}_{\ell'}$, then $\overline{\overline{P}}_{\Gamma\ell'}$ is tending to zero as $I \rightarrow \infty$, and $\overline{\overline{P}}_{\Gamma\wedge}$ is also tending to zero. So the power of the sensitivity analysis using $\overline{\overline{P}}_{\Gamma\wedge}$ is tending to 1 for $\Gamma < \widetilde{\Gamma}_{\max}$ and to zero for $\Gamma > \widetilde{\Gamma}_{\max}$, proving the proposition. ■

3.4 Testing hypotheses about subgroups using closed testing

Rejecting H_0 in §3.2 suggests there is an effect in at least one subgroup ℓ, but it does not provide an inference about specific subgroups. Of course, it would be interesting to know which subgroups are affected.

Closed testing was proposed by Marcus, Peritz and Gabriel (1976) as a general method for converting a test of a global null hypothesis into a multiple inference procedure for subhypotheses. Let $\mathcal{L} \subseteq \{1, 2, \ldots, L\}$ be a nonempty subset, and let $H_{\mathcal{L}} = \bigwedge_{\ell \in \mathcal{L}} H_\ell$ be the subhypothesis that asserts H_ℓ is true for $\ell \in \mathcal{L}$, so $H_{\mathcal{L}}$ asserts that a specific set of $|\mathcal{L}|$ hypotheses are all true. Let $P_{\mathcal{L}} = \prod_{\ell \in \mathcal{L}} P_\ell^{\chi(P_\ell \le \widetilde{\alpha})}$ and $\overline{\overline{P}}_{\Gamma\mathcal{L}} = \prod_{\ell \in \mathcal{L}} \left(\overline{\overline{P}}_{\Gamma\ell}\right)^{\chi\left(\overline{\overline{P}}_{\Gamma\ell} \le \widetilde{\alpha}\right)}$ be the truncated products of P-values for these hypotheses, these being compared to (6) with $|\mathcal{L}|$ in place of L. Closed testing rejects $H_{\mathcal{L}}$ at level α if $P_{\mathcal{L}'} \le \alpha$ for all $\mathcal{L}'$ such that $\mathcal{L} \subseteq \mathcal{L}'$, and Marcus et al. (1976) show that the chance that closed testing rejects at least one true hypothesis is at most α. Of course, $P_{\mathcal{L}}$ depends upon the unknown $\mathbf{u}$ and $\gamma = \log(\Gamma)$ in (2), but if the bias is at most Γ, then $P_{\mathcal{L}} \le \overline{\overline{P}}_{\Gamma\mathcal{L}}$, and a procedure that rejects $H_{\mathcal{L}}$ at level α if $\overline{\overline{P}}_{\Gamma\mathcal{L}'} \le \alpha$ for all $\mathcal{L}'$ such that $\mathcal{L} \subseteq \mathcal{L}'$ will falsely reject a true null hypothesis with probability at most α.

Using the U-statistic in Table 1 with $\Gamma = 1$, closed testing rejects no effect $H_0 = H_1 \wedge H_2$, and then rejects both H_1 and H_2. Using the U-statistic in Table 1 with $\Gamma = 6$, closed testing rejects no effect $H_0 = H_1 \wedge H_2$, and then rejects H_1 but not H_2. In words, there is some evidence of a treatment effect for both those under and over ten years of age, the evidence about the young children being insensitive to a large bias of $\Gamma = 6$, while the evidence for older individuals is sensitive to some biases smaller than $\Gamma = 2$.

3.5 Simulation when several groups are defined a priori

Table 2 simulates the power of a 0.05-level sensitivity analysis testing the null hypothesis of no effect, H_0, that is, the probability that the upper bound (3) on the one-sided P-value is at most 0.05. In Table 2, four sampling situations are considered, in which the Y_i are independent and $Y_1, \ldots, Y_{500} \sim N(\delta_1, 1)$ and $Y_{501}, \ldots, Y_{1000} \sim N(\delta_2, 1)$. Two test statistics are compared, Wilcoxon's statistic with $(m_1, m_2, m) = (2, 2, 2)$ and the U-statistic with $(m_1, m_2, m) = (7, 8, 8)$. Eight testing methods are used. The one-test method uses all 1000 matched pairs. The other methods compute two bounds $\overline{\overline{P}}_{\Gamma 1}$ and $\overline{\overline{P}}_{\Gamma 2}$, on two P-values and combine them. The truncated product of P-values is used with

$\widetilde{\alpha} = 0.05$, 0.10, 0.015 and 0.20, and Fisher's method is used, the last being the same as $\widetilde{\alpha} = 1$. The Bonferroni inequality rejects if $\min\left(\overline{\overline{P}}_{\Gamma 1}, \overline{\overline{P}}_{\Gamma 2}\right) \leq 0.05/2$. The Simes (1986) procedure rejects if either $\min\left(\overline{\overline{P}}_{\Gamma 1}, \overline{\overline{P}}_{\Gamma 2}\right) \leq 0.05/2$ or $\max\left(\overline{\overline{P}}_{\Gamma 1}, \overline{\overline{P}}_{\Gamma 2}\right) \leq 0.05$. Each sampling situation is replicated 10,000 times, so the power is estimated with a standard error of at most $\sqrt{0.25/10,000} = 0.005$. By definition, the Simes procedure is always at least as powerful as the Bonferroni procedure, and as expected from Rosenbaum (2011b), $(m_1, m_2, m) = (7, 8, 8)$ yields more power than Wilcoxon's statistic.

Table 2 exhibits some notable patterns. In Table 2, there are 12 contests among methods defined by four patterns of (δ_1, δ_2) and three values of Γ. The single test based on the U-statistic with all pairs is best when there is no effect modification, $\delta_1 = \delta_2$, but it is only slightly better in this case than Fisher's combination ($\widetilde{\alpha} = 1$) or the truncated product $P_\wedge$ with $\widetilde{\alpha} = .2$ when applied to the U-statistic. However, even for slightly unequal effects, $\delta_1 = .6 > .4 = \delta_2$, the single test is inferior to all methods that combine two independent P-values, and the gap in performance widens as $|\delta_1 - \delta_2|$ increases. In the cases covered by Table 2, Fisher's method offers little advantage over the truncated product $P_\wedge$ with $\widetilde{\alpha} = .2$, and Fisher's method is inferior in several cases. In Table 2, the truncated product $P_\wedge$ with $\widetilde{\alpha} = .05$ is similar to the Bonferroni and Simes procedures, though perhaps ever so slightly more powerful. As suggested by considerations of the design sensitivity in Proposition 1 and related results for the Bonferroni inequality (Rosenbaum and Silber 2009), the power of the single test may be tending to zero while methods that combine P-values may have power tending to one as $I \rightarrow \infty$. The one disaster in Table 2 is not competitive in terms of power in any of the 12 contests: it is Wilcoxon's statistic applied to all I pairs. Moreover, with $\Gamma = 5$, $\delta_1 = 1 > 0 = \delta_2$, both the Wilcoxon method and the U-statistic have power 0.00 using the one-test method and power 1.00 using all methods that combine two-P-values. Reducing $\widetilde{\alpha}$ increases power when $\delta_1 \geq 0.6 > 0.4 \geq \delta_2$ and reduces power when $\delta_1 = \delta_2$. Table 3 is similar, except there are now five groups rather than two groups, 200 pairs per group rather than 500 pairs, with logistic errors rather than Normal errors.

3.6 Designs other than matched pairs

Although the example in §1.2 involves matched pairs, because (6) combines P-values, it may be used with any design and method of sensitivity analysis that yields upper bounds on P-values. For instance, if each treated person had been matched to three untreated controls for age and gender, then the matched sets could be grouped into age ≤ 10 and

> 10, as in §3.2, yielding for each $\Gamma \geq 1$ a pair $\left(\overline{\overline{P}}_{\Gamma 1}, \overline{\overline{P}}_{\Gamma 2}\right)$ of upper bounds on the two P-values derived from separate sensitivity analyses in the two age groups, and these could be combined as before using (6). The computation of $\overline{\overline{P}}_{\Gamma 1}$ and $\overline{\overline{P}}_{\Gamma 2}$ for matched sets with more than one control is not difficult, though it is slightly different than (4); see, for instance, Rosenbaum (2007b, §4). With either pairs or sets, the pairing may be supplemented by covariance adjustment for an observed covariate incompletely controlled by matching; see Rosenbaum (2007b, §5).

In full matching, a matched set has either: (i) one treated subject and one or more controls, or (ii) one control and one or more treated subjects; see Rosenbaum (1991) for motivation, and see Hansen and Klopher (2006) and Hansen (2007) for optimal full matching as implemented in `R`. Full matching can use every available subject, and it is the structure of the closest match that does use every subject. Sensitivity analysis for full matching has the same form as sensitivity analysis for matching with multiple controls, as discussed in the previous paragraph.

4 Effect Modification Discovered Using the Data

4.1 Locating possible effect modification by grouping pairs

In §3, the investigator came to the data with an a priori partition of the pairs based on observed covariate $\mathbf{x}_i$. Can one obtain appropriate inferences using a partition discovered with the aid of the data at hand?

If Fisher's hypothesis H_0 of no effect were true, then $|Y_i| = |r_{Ci1} - r_{Ci1}|$ is a function of $\mathcal{F}$, and hence is fixed by conditioning on $\mathcal{F}$ in (2). In testing H_0 using (2), we may, therefore, select a specific test statistic having examined $|Y_i|$ while obtaining the same level α for this test as if we had selected the test based on a priori considerations. For instance, in a different context, Jones (1979) uses this approach to improve efficiency. This would not be true if we used Y_i rather than $|Y_i|$ to select the test, because unlike $|Y_i|$, the signed difference Y_i depends on $\mathbf{Z}$ even under H_0 and it is not a function of $\mathcal{F}$, so it is not fixed by conditioning on $\mathcal{F}$. Because $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_i$ is in $\mathcal{F}$, $\mathbf{x}_i$ is also fixed by conditioning on $\mathcal{F}$ in (2).

We would like to select a test statistic that would yield good power in a sensitivity analysis in the favorable situation, that is, if H_0 were false because the treatment actually had an effect and if Y_i were correctly describing that effect because there actually is no

unmeasured bias, $\Pr(\mathbf{Z}=\mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 2^{-I}$ for each $\mathbf{z} \in \mathcal{Z}$. In this case, with good power, we would be likely to be able to assert that the observed association between treatment Z_{ij} and outcome R_{ij} cannot easily be dismissed as noncausal — that only a moderately large bias could explain it. In the current paragraph only, suppose that we are in the favorable situation and before conditioning on $\mathcal{F}$, the treated-minus-control difference in outcomes Y_i in pairs matched for $\mathbf{x}_i$ was generated by $Y_i = \rho(\mathbf{x}_i) + \xi_i$ where $\rho(\mathbf{x}_i) \geq 0$ and all of the ξ_i were sampled independently from the same continuous, unimodal density symmetric about 0. For instance, in Table 2, this model applied with $\rho(\mathbf{x}_i) = \delta_1 \geq 0$ or $\rho(\mathbf{x}_i) = \delta_2 \geq 0$ and with ξ_i having the standard Normal distribution. We cannot estimate $\rho(\mathbf{x}_i)$ from $|Y_i|$ and $\mathbf{x}_i$ under this model, but we do know that $|Y_i|$ is stochastically larger than $|Y_{i'}|$ if $\rho(\mathbf{x}_i) > \rho(\mathbf{x}_{i'})$; see Jogdeo (1977, Theorem 2.2). This suggests defining groups in terms of $\mathbf{x}_i$ empirically using the $|Y_i|$'s so that, at certain $\mathbf{x}_i$ values, the distribution of $|Y_i|$'s appears to be larger. A simple strategy takes the ranks of the $|Y_i|$'s, regresses these ranks in some fashion on the $\mathbf{x}_i$, identifies a group of $\mathbf{x}_i$'s associated with large ranks of $|Y_i|$'s, and then performs the test and sensitivity analysis in this group.

Several difficulty suggest themselves. If the distribution of ξ_i were not the same, but became more unstable at some values of $\mathbf{x}_i$, then large $|Y_i|$ at these $\mathbf{x}_i$ might indicate greater instability rather than a higher median $\rho(\mathbf{x}_i)$, and there is no point in giving greater weight to more unstable Y_i. Similarly, if some $\rho(\mathbf{x}_i)$'s were positive, other $\rho(\mathbf{x}_i)$'s were negative, then $|Y_i|$ would not reveal this. Ultimately, despite these difficulties, one would like a procedure that is never much worse than the aggregate test, but sometimes much better. We proceed naively at first in §4.2, ignoring these potential difficulties, and then fully address them in §4.3.

4.2 Example: Values of $\mathbf{x}_i$ associated with large $|Y_i|$

In §1.2, the matching controlled for age and gender. Using the regression trees of Breiman et al. (1983), as implemented in `R` with default settings in the `rpart` package of Therneau and Atkinson (1997), we regressed the ranks of $|Y_i|$ on age and gender. Working with these ranks of absolute differences in outcomes, the algorithm ignored gender and split age into four bins with three cuts at 7.5 years, 17.5 years and 32.25 years. Beginning with the youngest, the bins contained 340 individuals under age 7.5, 243 between 7.5 and 17.5, 413 between 17.5 and 32.25, and 564 at least 32.5 years old, where $1560 = 340+243+413+564$. The ranks, of course, ranged from 1 to 1560, but the mean ranks were 1241 below age 7.5,

992 between ages 7.5 and 17.5, 659 between 17.5 and 32.5, and 501 above 32.5. This partition turned out to be fairly good advice.

If one confines attention to the 340 children under age 7.5 with the largest mean ranks of $|Y_i|$, then the sensitivity analysis for the U-statistic with $(m_1, m_2, m) = (7, 8, 8)$ yields an upper bound (4) on the one-sided P-value of 0.030 at $\Gamma = 6$, similar to Table 1 which cut at 10 years and used 447 pairs. If the split is made at 17.5 years, using $340 + 243 = 583$ pairs, the results are more sensitive to bias, with an upper bound on the one-sided P-value of 0.16 at $\Gamma = 6$.

To repeat, we may select pairs to analyze based on the relationship between $|Y_i|$ and $\mathbf{x}_i$, because under H_0 these quantities are functions of $\mathcal{F}$, which is fixed by conditioning in (2), so this process of selecting pairs does not affect the level α of the test, although it is expected to affect the power and the design sensitivity. In the example, using the youngest pairs in this way yielded much less sensitivity to bias than using either all pairs in Table 1 or the two youngest groups combined, those under age 17.5.

There are several difficulties with the approach just described. First, as mentioned in §4.1, large values of $|Y_i|$ at certain values of $\mathbf{x}_i$ are compatible with either a large typical effect, $\rho(\mathbf{x}_i)$, or with greater instability at this $\mathbf{x}_i$, and we cannot distinguish these before looking at Y_i and $\mathbf{Z}$, which we cannot do without affecting the level of the test. Second, combining a few leaves of a small tree may produce higher expected ranks of $|Y_i|$, but it also a lowers the sample size, and one cannot shop around for the most favorable of several analyses without paying a price for multiple testing. Section 4.3 provides a solution that always yields the highest design sensitivity.

4.3 Multiple analyses derived from regression trees

Instead of performing one test of H_0, as in §4.2, while being uncertain as to which one test to perform, the current section performs four tests of one H_0, adjusting for multiple testing using the technique in Rosenbaum (2012). The four tests concern hypotheses H_1, H_{12}, H_{123}, and H_{1234} in Figure 3, where H_{1234} is the same as Fisher's hypothesis H_0 of no effect. Because these are four tests of one null hypothesis, all computed from the same data, the tests are highly correlated, and a correction for multiple testing that takes the high correlation into account is a small correction. Specifically, the problem is represented as picking the largest of four standardized deviates computed from statistics that score the $|Y_i|$'s differently. Under mild conditions, for each Γ, the large sample joint null distribution

of these four standardized deviates is a correlated multivariate Normal distribution, with known correlation, from which the relevant correction for multiple testing is derived; see Rosenbaum (2012) for specifics. It is shown there that this multiple test procedure has the largest design sensitivity of the four component tests. The number of tests depends upon the number of leaves of the regression tree, namely four in §4.2, but under H_0 the $|Y_i|$'s and $\mathbf{x}_i$'s are fixed by conditioning on $\mathcal{F}$, so the regression tree is fixed, and so is the number of tests.

The first test uses the one bin with the highest fitted $|Y_i|$ in §4.2, namely the 340 pairs with age below 7.5 years, the second test combines the two highest bins, that is the $340 + 243 = 583$ pairs with age below 17.5 years, the third test uses the three highest bins, and the fourth test uses all the pairs. In (5), the U-statistic $(m_1, m_2, m) = (7, 8, 8)$ is computed using $I = 340$ pairs, using $I = 340+243 = 583$ pairs, using $I = 340+243+413 = 996$ pairs, and using all $I = 1560$ pairs.

Taking this approach, the largest of the four standardized deviates is found using the children under age 7.5 years, and after correcting for multiple testing, the upper bound on the one-sided P-value for the four-test procedure is 0.0475 at $\Gamma = 5.8$. So the four-test procedure is almost as insensitive as knowing a priori that the test should focus on children under age 7.5, because in §4.2 that single test had a P-value bound of 0.030 at $\Gamma = 6$. Again, this is expected in large samples because the combination of four tests has the same design sensitivity as the best of the four individual tests (Rosenbaum 2012, Proposition 2).

The computations are straightforward in `R`. The $L = 4$ bins in Figure 3 were obtained by regressing the ranks of $|Y_i|$ on age and gender using the `rpart` package. Other methods of regression might be substituted. The trade-off is between larger predicted ranks of $|Y_i|$ and a smaller sample size, so the four hypotheses involve the one, two, three or four groups with the largest predicted ranks of $|Y_i|$ as indicted by the horizontal lines in Figure 3, with sample sizes $I = 340$, $I = 583 = 340 + 243$, $I = 996 = 340 + 243 + 413$, and $I = 1560$. The on-line appendix to Rosenbaum (2011b) at the journal's web-page contains the few lines of `R`-code computing q_i for fixed I in (5) with a numerical example. The rank scores q_i of $|Y_i|$ are scored for the U-statistic $(m_1, m_2, m) = (7, 8, 8)$ using (5) separately for each group with its own value of I, then q_i is set to 0 for pairs not in this group; e.g., q_i is determined from (5) for the $I = 340$ children in the youngest group under age 7.5 years, with a_i ranging from 1 to 340; then q_i is set to zero for the $1560 - 340 = 1220$ individuals over age 7.5. This creates four signed rank statistics with the same $\text{sgn}(Y_i)$ but with four

different rank scores q_i. As $I \rightarrow \infty$, there is a single $\mathbf{u}$ in (2) that provides the upper bound (3) for all $L = 4$ signed rank statistics, and the $L = 4$ upper bound statistics $\overline{\overline{T}}$ tend to an $L = 4$ dimensional Normal distribution as $I \rightarrow \infty$ with easily computed covariance matrix given by Lemma 1 in Rosenbaum (2012). The relevant tail-probability is one minus the probability in a lower quadrant of this L-dimensional Normal distribution and is computed using the `pmvnorm` function in the `mvtnorm` package in `R`; see Rosenbaum (2012, §2.2).

4.4 Simulation with groups discovered using regression trees

The simulation compares the power of a sensitivity analysis with one test using all I pairs (called "one-test") to tree-based grouping and multiple testing (called "tree"). The tree procedure is the same as in §4.2: it (i) finds the tree-based regression of the ranks of $|Y_i|$ on a covariate $\mathbf{x}_i$, yielding say L leaves, (ii) forms L overlapping groupings of the I pairs, the first consisting of all I pairs, the second omitting the one leaf with the lowest mean rank of $|Y_i|$, the third omitting the two lowest leaves, and so on, until the Lth which uses the one leaf with the highest mean rank, (iii) calculates L test statistics for the L groups, and corrects the smallest upper bound on the L P-values using the method in Rosenbaum (2012).

The sampling situation is the same as in Table 2, but permits unequal variances: there are $I = 1000$ independent pairs, the first 500 pairs having $Y_i = \delta_1 + \sigma_1\epsilon_i$, $i = 1, \ldots, 500$, the last 500 pairs having $Y_i = \delta_2 + \sigma_2\epsilon_i$, $i = 501, \ldots, 1000$, where $\epsilon_i \sim N(0, 1)$. Because $(\delta_1 + \delta_2)/2 = 0.5$ and $\sigma_1^2 + \sigma_2^2 = 2$ in all nine columns of Table 4, the mean pair difference $\overline{Y} = I^{-1}\sum Y_i$ is $N(0.5, 1/1000)$ in all nine columns.

Unlike Table 2, the procedure regresses the rank of $|Y_i|$ on the 'covariate' i and must discover the single step from δ_1 to δ_2 at $i = 500$. In Table 4, there is no step to discover when $\delta_1 = \delta_2$. The situations $\delta_1 = \delta_2 = 0.5$, $\sigma_1^2 = 1.5$, $\sigma_2^2 = 0.5$ (column 5) and $\delta_1 = 0.75$, $\delta_2 = 0.25$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 1.5$ (column 7) are unfavorable to the tree-based test: in both situations, the $|Y_i|$ are elevated where δ_ℓ is not elevated because of a larger variance σ_ℓ^2, so the tree-procedure emphasizes the wrong stratum.

The sensitivity analysis is done using the U-statistic with $(m_1, m_2, m) = (7, 8, 8)$ or Wilcoxon's statistic both with $\Gamma = 4$. The design sensitivity using $(7, 8, 8)$ and all pairs is $\widetilde{\Gamma} = 5.1$ for $\delta_1 = \delta_2 = 0.5$, and is $\widetilde{\Gamma} = 40.5$ for $\delta_1 = \delta_2 = 1$, whereas for Wilcoxon's statistic it is $\widetilde{\Gamma} = 3.2$ for $\delta_1 = \delta_2 = 0.5$ and $\widetilde{\Gamma} = 11.7$ for $\delta_1 = \delta_2 = 1$; see Rosenbaum (2011b, Table 3). If the U-statistic is used and $\delta_1 = 1$, $\delta_2 = 0$, then the design sensitivity in group 1

would be $\widetilde{\Gamma} = 40.5$ if groups 1 and 2 are successfully distinguished; however, the average effect if the groups are merged remains $(\delta_1 + \delta_2)/2 = 0.5$.

In Table 4, each sampling situation is replicated 10,000 times. For $\Gamma = 4$, Table 4 reports the power as the proportion of times the upper bound on the P-value was less than or equal to $\alpha = 0.05$. Table 4 also reports the median upper bound on the P-value for $\Gamma = 4$. The parameters of the `rpart` function in `R` are set to require each leaf of the tree to include at least 50 pairs (`minsplit`=100, `minbucket`=50).

If the tree-based splitting makes no split, then the one-test and tree-test procedures are the same, yielding the same upper bound on the P-value. The frequency of ties in the P-values is recorded in Table 4, as is the number of times the one-test procedure had the smaller P-value, and the number of times that the tree-procedure had the smaller P-value. Table 4 also reports the number of trees of 10,000 with 1, 2, 3 or > 3 leaves. The ideal tree would have 1 leaf if $\delta_1 = \delta_2$ and two leaves if $\delta_1 \gg \delta_2$.

When $\sigma_1^2 = \sigma_2^2$ in Table 4, the tree-procedure performs either acceptably or extremely well. In column 1 of Table 4, when $\delta_1 = \delta_2 = 0.5$ with $\sigma_1^2 = \sigma_2^2$, the tree procedure does little harm, because in more than 95% of cases no split is made. In contrast, when $\delta_1 = 1$ and $\delta_2 = 0$ with $\sigma_1^2 = \sigma_2^2$, the gains from the tree procedure are enormous. When $\delta_1 = 0.6$ and $\delta_2 = 0.4$ with $\sigma_1^2 = \sigma_2^2$, there are small gains. The tree procedure does some harm in the case $\delta_1 = \delta_2 = 0.5$ with $\sigma_1^2 = 1.5$, $\sigma_2^2 = 0.5$, doing two tests in about 95% of cases and correcting for multiple testing when it would be better to do one test. Because the tree-based procedure does a test using all I pairs, at worst it pays an unnecessary price for multiple testing.

Not shown in Table 4 but as expected from theory, under the null hypothesis of no effect in a randomization test with $\Gamma = 1$, the tree-based splitting procedure had the correct level. In 10,000 simulated situations with $\delta_1 = \delta_2 = 0$, $\sigma_1^2 = \sigma_2^2 = 1$, the tree-based splitting procedure using Wilcoxon's test falsely rejected no effect in 474 cases (4.74%) whereas a conventional single application of Wilcoxon's test to all pairs falsely rejected in 476 cases (4.76%). Using the U-statistic with $(m_1, m_2, m) = (7, 8, 8)$, the tree-based splitting procedure falsely rejected in 4.86% of cases whereas a single application of U-statistic to all pairs falsely rejected in 4.90% of cases. In 96% of cases, the tree-based splitting procedure did not split, so only one test using all pairs was actually performed.

5 Discussion: summary; sample spitting to identify effect modification; application to evidence factors

Other things being equal, larger effects are less sensitive to unmeasured biases than smaller effects. When the magnitude of effect varies with an observed covariate $\mathbf{x}_{ij}$ controlled by matching but the stability of the responses does not vary with $\mathbf{x}_{ij}$, the results may be less sensitive to bias for a subset of pairs defined by $\mathbf{x}_{ij}$. Both an a priori grouping of pairs and a grouping discovered in the data have been considered in §3 and §4, respectively. Use of closed testing converts a test of the global null hypothesis H_0 of no effect at all into a multiple testing procedure that provides separate sensitivity analyses within subgroups defined by $\mathbf{x}_{ij}$.

When the sample size I is large, an alternative to the tree-procedure in §4 is sample splitting, as discussed by Heller et al. (2009). In this case, a fraction, perhaps 10%, of the pairs are sampled at random. On the basis of this 10% planning sample, the study is designed, perhaps identifying effect modification in certain groups defined by $\mathbf{x}_{ij}$. The planning sample is then discarded, and there is now an a priori plan for the analysis of the remaining 90% of the pairs. Although sample splitting discards some data, it can use the Y_i's to form groups, thereby avoiding certain potential errors discussed in §4 that can arise when $|Y_i|$'s are used to form groups. In a different context, Zhang et al (2011) used sample splitting to increase design sensitivity in an observational study.

As discussed in §3, when compared to Fisher's method, Zaykin et al. (2002)'s truncated product of P-values pays little attention to the very large P-values that can arise in sensitivity analysis, focusing instead on the number and size of the small P-values. This is relevant to effect modification as seen in §3, but also to other sensitivity analyses that combine P-values, for instance with evidence factors (Rosenbaum 2011c, Zhang et al 2012).

Data Appendix

The GARKI project is described by Molineaux and Gramiccia (1980). Assignment to treatment or control groups was based on the judgement and convenience of the investigators (Molineaux and Gramiccia 1980, pages 28-30). Issues that weighed on the investigators minds in assigning treatments included the practical aspects of spraying the insecticide, frequent data collection about mosquitoes, obtaining repeated blood samples, geography and logistics. Of 7777 study participants, 2599 did not have two blood measurements

before treatment and two after treatment, leaving 5178 subjects for our analysis, of whom 1560 were treated and 3618 were control. The 1560 pairs were formed by matching for age and gender using the `pairmatch` function in the `optmatch` package in `R` (Hansen 2007) applied to a distance matrix that combined a caliper on an estimated propensity score with a rank based Mahalanobis distance (Rosenbaum 2010b, §8).

References

Altonji, J. G., Elder, T. E. and Taber, C. R. (2005), "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151-184.

Bauer, P. and Kieser, M. (1996), "A unifying approach for confidence intervals and testing of equivalence and difference," *Biometrika*, 83, 934-7.

Berger, R. L. and Hsu, J. C. (1996), "Bioequivalence trials, intersection-union tests and equivalence confidence sets," *Statistical Science*, 11, 283-319.

Brannath, W., Posch, M. and Bauer, P. (2002), "Recursive combination tests," *Journal of the American Statistical Association*, 97, 236-244.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1983), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Brown, B. M. (1981), "Symmetric quantile averages and related estimators," *Biometrika*, 68, 235-242.

Brown, B. M. and Hettmansperger, T. P. (1994), "Regular redescending rank estimates," *Journal of the American Statistical Association*, 89, 538-542.

Conover, W. J. and Salsburg, D. S. (1988), "Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment," *Biometrics*, 44, 189-196.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E. (1959), "Smoking and lung cancer," *Journal of the National Cancer Institute*, 22, 173-203.

Fisher, R.A. (1932), *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Fisher, R.A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Gabriel, K. R. (1969), "Simultaneous test procedures — some theory of multiple comparisons," *Annals of Mathematical Statistics*, 40, 224-250.

Gastwirth, J. L. (1992), "Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables," *Jurimetrics* 33, 19-34.

Hammond, E. C. (1964), "Smoking in relation to mortality and morbidity," *Journal of the National Cancer Institute*, 32, 1161–1188.

Hansen, B. B. and Klopfer, S. O. (2006), "Optimal full matching and related designs via network flows," *Journal of Computational and Graphical Statistics*, 15, 609–627.

Hansen, B. B. (2007), "Optmatch: flexible, optimal matching for observational studies," *R News*, 7, 18-24. `R`-package `optmatch`.

Heller, R., Rosenbaum, P. R. and Small, D. S. (2009), "Spit samples and design sensitivity in observational studies," *Journal of the American Statistical Association*, 104, 1090-1101.

Hodges, J. L. and Lehmann, E. L. (1963), Estimates of location based on ranks, *Annals of Mathematical Statistics*, 34, 598-611.

Hosman, C. A., Hansen, B. B., and Holland, P. W. H. (2010), "The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder," *Annals of Applied Statistics*, 4, 849-870.

Ichino, A., Mealli, F. and Nannicini, T. (2008), "From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity?" *Journal of Applied Econometrics*, 23, 305-327.

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

Jogdeo, K. (1977), "Association and probability inequalities," *Annals of Statistics*, 5, 495-504.

Jones, D. H. (1979), "An efficient adaptive distribution-free test for location," *Journal of the American Statistical Association*, 74, 822-828.

Lehmann, E. L. and Romano, J. P. (2005), *Testing Statistical Hypotheses* (3^{rd} edition), New York: Springer.

Manski, C. (1990), "Nonparametric bounds on treatment effects," *American Economic Review*, 319-323.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976), "On closed testing procedures with special reference to ordered analysis of variance," *Biometrika*, 63, 655-660.

Marcus, S. M. (1997), "Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect," *Journal of Educational Statistics*, 22, 193-201.

Maritz, J. (1979), "Exact robust confidence intervals for location," *Biometrika*, 66, 163-166.

Millimet, D.L. and Tchernis, R. (2012), "Estimation of Treatment Effects without an Exclusion Restriction: with an Application to the Analysis of the School Breakfast

Program," *Journal of Applied Econometrics*, published on-line early.

Molineaux, L. and Gramiccia, G. (1980), *The GARKI Project: Research on the epidemiology and control of Malaria in the Sudan Savanna of West Africa*, Geneva: World Health Organization.' http://www.swisstph.ch/fr/ressources/epidemiological-databases.html

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 463-480.

Noether, G. (1973), "Some distribution-free confidence intervals for the center of a symmetric distribution," *Journal of the American Statistical Association*, 68, 716-719.

Pepper, J. (2012), "The impact of the National School Lunch Program on child health: A nonparametric bounds analysis," *Journal of Econometrics*, 166, 79-91.

Rosenbaum, P.R. and Rubin, D.B. (1985), "The bias due to incomplete matching," *Biometrics*, 41, 103-116.

Rosenbaum, P. and Rubin, D. (1983), "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society* B, 45, 212-218.

Rosenbaum, P. R. (1987), "Sensitivity analysis for certain permutation inferences in matched observational studies," *Biometrika*, 74, 13-26.

Rosenbaum, P. R. (1991), "A characterization of optimal designs for observational studies," *Journal of the Royal Statistical Society*, B 53, 597-610.

Rosenbaum, P. R. (1993), "Hodges-Lehmann point estimates of treatment effect in observational studies," *Journal of the American Statistical Association*, 88, 1250-1253.

Rosenbaum, P. R. (2002), *Observational Studies* (2^{nd} edition), New York: Springer.

Rosenbaum, P. R. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-164.

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

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

Rosenbaum, P. R. and Silber, J. H. (2009), "Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units," *Journal of the American Statistical Association*, 104, 501-511.

Rosenbaum, P. R. (2010a), "Design sensitivity and efficiency in observational studies," *Journal of the American Statistical Association*, 105, 692-702.

Rosenbaum, P. R. (2010b), *Design of Observational Studies*, New York: Springer.

Rosenbaum, P. R. (2011a), "What aspects of the design of an observational study affect its sensitivity to bias from covariates that were not observed?" In N. Dorans and S. Sinharay, eds., *Looking Back: Proceedings of a Conference in Honor of Paul W. Holland*, New York: Springer, pp. 87-114.

Rosenbaum, P. R. (2011b), "A new u-statistic with superior design sensitivity in observational studies," *Biometrics*, 67, 1017-1027. R-session at http://www-stat.wharton.upenn.edu/~rosenbap/rsession.txt

Rosenbaum, P. R. (2011c), "Some approximate evidence factors in observational studies," *Journal of the American Statistical Association*, 106, 285-295.

Rosenbaum, P. R. (2012), "Testing one hypothesis twice in observational studies," *Biometrika*, to appear and available on-line early.

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

Schwartz, S., Li, F. and Reiter, J. (2012), "Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables," *Statistics in Medicine*, 31, 949-962.

Simes, R. J. (1986), "An improved Bonferroni procedure for multiple tests of significance," *Biometrika*, 73, 751-754.

Small, D. (2007), "Sensitivity analysis for instrumental variables regression with overidentifying restrictions," *Journal of the American Statistical Association*, 102, 1049-1058.

Stephenson, W. R. (1981), "A general class of one-sample nonparametric test statistics based on subsamples," *Journal of the American Statistical Association*, 76, 960-966.

Therneau, T. M. and Atkinson, E. J. (1997), "An introduction to recursive partitioning using the RPART routines," Technical Report, Mayo Foundation.

Wang, L. and Krieger, A. M. (2006), "Causal conclusions are most sensitive to unobserved binary covariates," *Statistics in Medicine*, 25, 2257-2271.

Welch, B. L. (1937), "On the z-test in randomized blocks," *Biometrika*, 29, 21-52.

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

Yu, B. B., Gastwirth, J. L. (2005), "Sensitivity analysis for trend tests: application to the risk of radiation exposure," *Biostatistics*, 6, 201-209.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002), "Trucated product method of combining P-values," *Genetic Epidemiology*, 22, 170-185.

Zhang, K., Small, D. S., Lorch, S., Srinivas, S., and Rosenbaum, P. R. (2011), "Using split

samples and evidence factors in an observational study of neonatal outcomes," *Journal of the American Statistical Association*, 106, 511-524.

Table 1: Various sensitivity analyses for the treated-minus-control difference in after-minus-before changes in parasite frequency in blood samples. The aggregate analyses use all $I = 1560$ matched pairs. There are 447 pairs in the young group, age 10 or under, and 1113 pairs in the older group, greater than age 10. The table gives the upper bound on the one-sided P-value. Sensitivity analyses using the truncated product of subgroup specific P-values use $\widetilde{\alpha} = 0.05$. In each column, the largest P-value less than or equal to 0.05 is in **bold**.

Analysis	Aggregate		Subgroup Specific				Trucated Product	
Label	I	II	III	IV	V	VI	VII	VIII
Statistic	Wilcoxon	U-statistic	Wilcoxon		U-statistic		Wilcoxon	U-statistic
(m_1, m_2, m)	(2,2,2)	(7,8,8)	(2,2,2)		(7,8,8)		(2,2,2)	(7,8,8)
Sample Size	1560	1560	447	1113	447	1113	1560	1560
Γ	All	All	Young	Old	Young	Old	P-values combined	
1	0.000	0.000	0.000	**0.000**	0.000	**0.000**	0.000	0.000
1.9	**0.011**	0.000	0.000	1.000	0.000	0.967	0.000	0.000
2	0.072	0.000	0.000	1.000	0.000	0.991	0.000	0.000
2.6	0.996	**0.034**	0.000	1.000	0.000	1.000	0.000	0.000
3	1.000	0.365	**0.000**	1.000	0.000	1.000	**0.001**	0.000
4	1.000	1.000	0.071	1.000	0.000	1.000	1.000	0.000
5	1.000	1.000	0.524	1.000	0.004	1.000	1.000	0.010
6	1.000	1.000	0.905	1.000	0.025	1.000	1.000	**0.049**
6.5	1.000	1.000	0.969	1.000	**0.048**	1.000	1.000	0.094

Table 2: Power of a 0.05-level sensitivity analysis testing the null hypothesis of no treatment effect when the analysis is performed with $\Gamma = 3$, 4, or 5. There are 500 matched pairs with $Y_i \sim N(\delta_1, 1)$ and 500 matched pairs with $Y_i \sim N(\delta_2, 1)$. The statistics are either Wilcoxon's (W) statistic or the U-statistic (U) with $(m_1 = 7, m_2 = 8, m = 8)$. The combined test uses all 1000 pairs, whereas the other four methods each combine two P-values. Each situation is sampled 10,000 times.

	$\Gamma = 3$		$\Gamma = 4$		$\Gamma = 5$	
	W	U	W	U	W	U
Equal effects	$\delta_1 = 0.5$ and $\delta_2 = 0.5$					
Combined	0.15	1.00	0.00	0.53	0.00	0.04
Truncated $\widetilde{\alpha} = 0.05$	0.11	0.97	0.00	0.37	0.00	0.04
Truncated $\widetilde{\alpha} = 0.10$	0.11	0.98	0.00	0.41	0.00	0.03
Truncated $\widetilde{\alpha} = 0.15$	0.12	0.99	0.00	0.45	0.00	0.04
Truncated $\widetilde{\alpha} = 0.20$	0.13	0.99	0.00	0.48	0.00	0.04
Fisher	0.13	0.99	0.00	0.48	0.00	0.04
Bonferroni	0.10	0.97	0.00	0.35	0.00	0.03
Simes	0.11	0.97	0.00	0.36	0.00	0.03
Slightly unequal effects	$\delta_1 = 0.6$ and $\delta_2 = 0.4$					
Combined	0.13	0.99	0.00	0.51	0.00	0.04
Truncated $\widetilde{\alpha} = 0.05$	0.66	1.00	0.02	0.85	0.00	0.39
Truncated $\widetilde{\alpha} = 0.10$	0.65	1.00	0.02	0.84	0.00	0.35
Truncated $\widetilde{\alpha} = 0.15$	0.60	1.00	0.01	0.80	0.00	0.30
Truncated $\widetilde{\alpha} = 0.20$	0.57	1.00	0.01	0.77	0.00	0.26
Fisher	0.51	1.00	0.01	0.76	0.00	0.22
Bonferroni	0.67	1.00	0.02	0.85	0.00	0.38
Simes	0.67	1.00	0.02	0.85	0.00	0.38
Unequal effects	$\delta_1 = 0.75$ and $\delta_2 = 0.25$					
Combined	0.06	0.98	0.00	0.37	0.00	0.02
Truncated $\widetilde{\alpha} = 0.05$	1.00	1.00	0.82	1.00	0.17	0.99
Truncated $\widetilde{\alpha} = 0.10$	1.00	1.00	0.80	1.00	0.16	0.99
Truncated $\widetilde{\alpha} = 0.15$	1.00	1.00	0.76	1.00	0.12	0.98
Truncated $\widetilde{\alpha} = 0.20$	1.00	1.00	0.73	1.00	0.11	0.98
Fisher	1.00	1.00	0.66	1.00	0.07	0.96
Bonferroni	1.00	1.00	0.82	1.00	0.17	0.99
Simes	1.00	1.00	0.82	1.00	0.17	0.99
Effect only in stratum 1	$\delta_1 = 1.00$ and $\delta_2 = 0.00$					
Combined	0.00	0.89	0.00	0.09	0.00	0.00
Truncated $\widetilde{\alpha} = 0.05$	1.00	1.00	1.00	1.00	1.00	1.00
Truncated $\widetilde{\alpha} = 0.10$	1.00	1.00	1.00	1.00	1.00	1.00
Truncated $\widetilde{\alpha} = 0.15$	1.00	1.00	1.00	1.00	1.00	1.00
Truncated $\widetilde{\alpha} = 0.20$	1.00	1.00	1.00	1.00	1.00	1.00
Fisher	1.00	1.00	1.00	1.00	1.00	1.00
Bonferroni	1.00	1.00	1.00	1.00	1.00	1.00
Simes	1.00	1.00	1.00	1.00	1.00	1.00

Table 3: Power of a 0.05-level sensitivity analysis testing the null hypothesis of no treatment effect when the analysis is performed with $\Gamma = 4$. For $j = 1, ..., 5$, there are 200 matched pairs with $Y_i = \delta_j + \epsilon_i$, making 1000 pairs in total, where ϵ_i are independent observations from the standard logistic distribution. The statistics are either Wilcoxon's (W) statistic or the U-statistic (U) with $(m_1 = 7, m_2 = 8, m = 8)$. The combined test uses all 1000 pairs, whereas the other methods each combine five P-values. Each situation is sampled 10,000 times.

$(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5)$	(1,1,1,1,1)		(1.5,1,1,1,.5)	
Statistic	W	U	W	U
Combined	0.02	0.76	0.01	0.71
Truncated $\widetilde{\alpha} = 0.10$	0.01	0.52	0.46	0.86
Truncated $\widetilde{\alpha} = 0.15$	0.01	0.59	0.40	0.87
Truncated $\widetilde{\alpha} = 0.20$	0.01	0.62	0.37	0.88
Fisher	0.01	0.68	0.34	0.88
Bonferroni	0.02	0.25	0.68	0.83
Simes	0.02	0.28	0.68	0.85
$(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5)$	(1.2,1.2,1.2,0,0)		(1.2,1.2,0,0,0)	
Statistic	W	U	W	U
Combined	0.00	0.00	0.00	0.00
Truncated $\widetilde{\alpha} = 0.10$	0.31	0.84	0.15	0.59
Truncated $\widetilde{\alpha} = 0.15$	0.31	0.86	0.12	0.54
Truncated $\widetilde{\alpha} = 0.20$	0.30	0.85	0.09	0.48
Fisher	0.18	0.75	0.03	0.25
Bonferroni	0.21	0.64	0.15	0.50
Simes	0.22	0.68	0.15	0.52
$(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5)$	(1.2,1,1,.8,0)		(1.5,0,0,0,0)	
Statistic	W	U	W	U
Combined	0.00	0.01	0.00	0.00
Truncated $\widetilde{\alpha} = 0.10$	0.04	0.49	0.38	0.47
Truncated $\widetilde{\alpha} = 0.15$	0.03	0.50	0.28	0.34
Truncated $\widetilde{\alpha} = 0.20$	0.03	0.50	0.21	0.25
Fisher	0.01	0.44	0.07	0.05
Bonferroni	0.08	0.36	0.69	0.80
Simes	0.08	0.38	0.69	0.80

Table 4: Power of a 0.05-level sensitivity analysis testing the null hypothesis of no treatment effect with tree-based testing and with one test using all pairs. There are $I = 1000$ pairs, $Y_i = \delta_1 + \sigma_1\epsilon_i$ for $i = 1, ..., 500$ and $Y_i = \delta_2 + \sigma_2\epsilon_i$ for $i = 501, ..., 1000$ where ϵ_i's are independently drawn from the standard Normal distribution. In all 9 sampling situations, the mean of the 1000 pairs is Normal with expectation 0.5 and variance 1/1000. The one-test P-value uses all 1000 pairs. The tree-testing P-value performs L component tests if there are L leaves, adjusting for multiple testing, and one of these L component tests uses all 1000 pairs. The sensitivity analysis is done at $\Gamma = 4$.

Parameters	Sampling Situation: δ_1, δ_2, σ_1^2, σ_2^2								
δ_1	0.50	0.60	0.75	1.00	0.50	0.75	0.75	0.75	1.00
δ_2	0.50	0.40	0.25	0.00	0.50	0.25	0.25	0.25	0.00
σ_1^2	1.00	1.00	1.00	1.00	1.50	1.50	0.50	1.00	1.00
σ_2^2	1.00	1.00	1.00	1.00	0.50	0.50	1.50	1.00	1.00
δ_1/σ_1	0.50	0.60	0.75	1.00	0.41	0.61	1.06	0.75	1.00
δ_2/σ_2	0.50	0.40	0.25	0.00	0.71	0.35	0.20	0.25	0.00
Statistic	U statistic with $(m_1 = 7, m_2 = 8, m = 8)$							Wilcoxon	
Procedure	Power of a 0.05-level Sensitivity Analysis								
One Test (O)	0.53	0.51	0.37	0.08	0.26	0.58	0.02	0.00	0.00
Tree Testing (T)	0.53	0.56	0.91	1.00	0.20	0.92	0.02	0.66	1.00
Higher power	Tie	T	T	T	O	T	Tie	T	T
Procedure	Median Upper Bound on the P-value								
One Test	0.043	0.048	0.086	0.308	0.143	0.036	0.626	1.000	1.000
Tree Testing	0.043	0.038	0.000	0.000	0.195	0.003	0.657	0.015	0.000
Procedure	Which procedure had the smaller P-value?								
Methods Tied	9543	8066	1074	0	0	0	6622	1065	0
One test wins	334	401	27	0	9957	213	3279	0	0
Tree testing wins	123	1533	8899	10000	43	9787	99	8935	10000
	Description of the tree								
Leaves = 1	9543	8066	1074	0	0	0	6622	1065	0
Leaves = 2	387	1742	8424	9514	9462	9573	3100	8492	9495
Leaves = 3	65	179	449	425	465	374	253	399	448
Leaves > 3	5	13	53	61	71	53	25	44	57

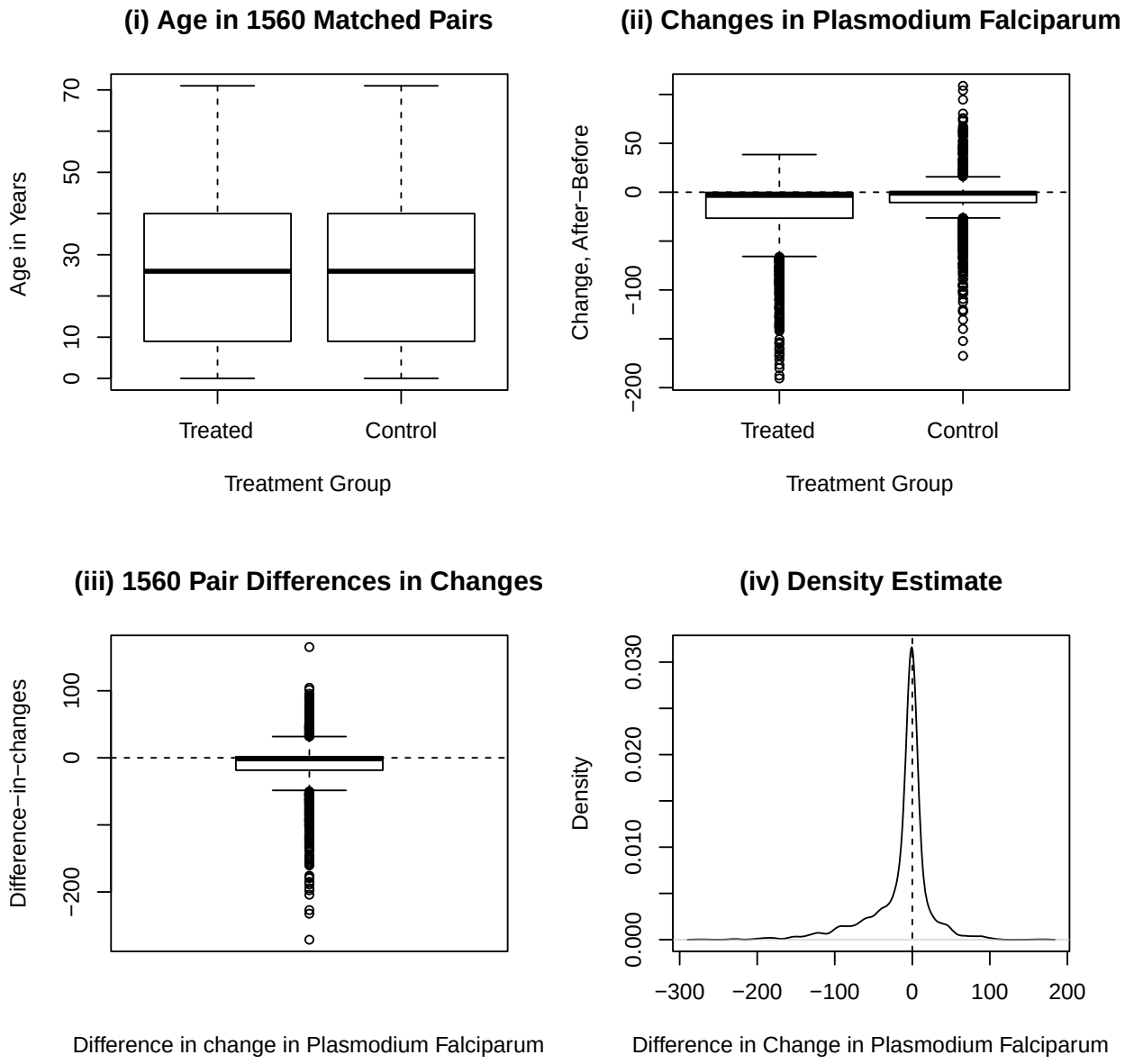

Figure 1: Age and parasite density in 1560 treated/control pairs matched for age and sex. After matching, the distribution of ages is similar, whereas the after-minus-before changes in parasite density exhibit a greater decline in the treated group. The treated-minus-control pair differences in changes in parasite density are typically negative, though many are near zero, with a long thick negative tail to their density.

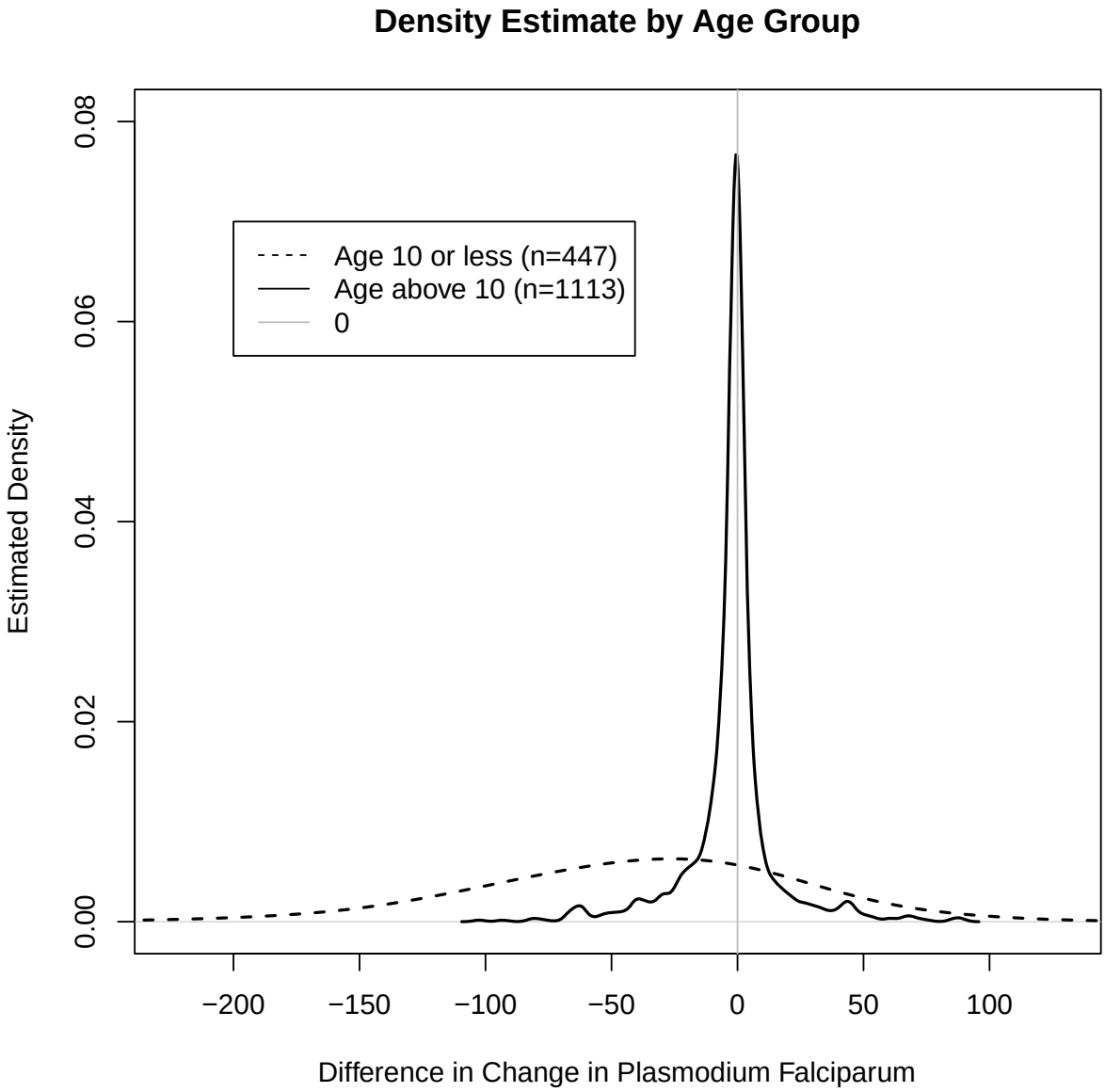

Figure 2: Density of the treated-minus-control difference in changes in parasite density separately for pairs of children 10 years old or younger and for individuals older than 10 years.

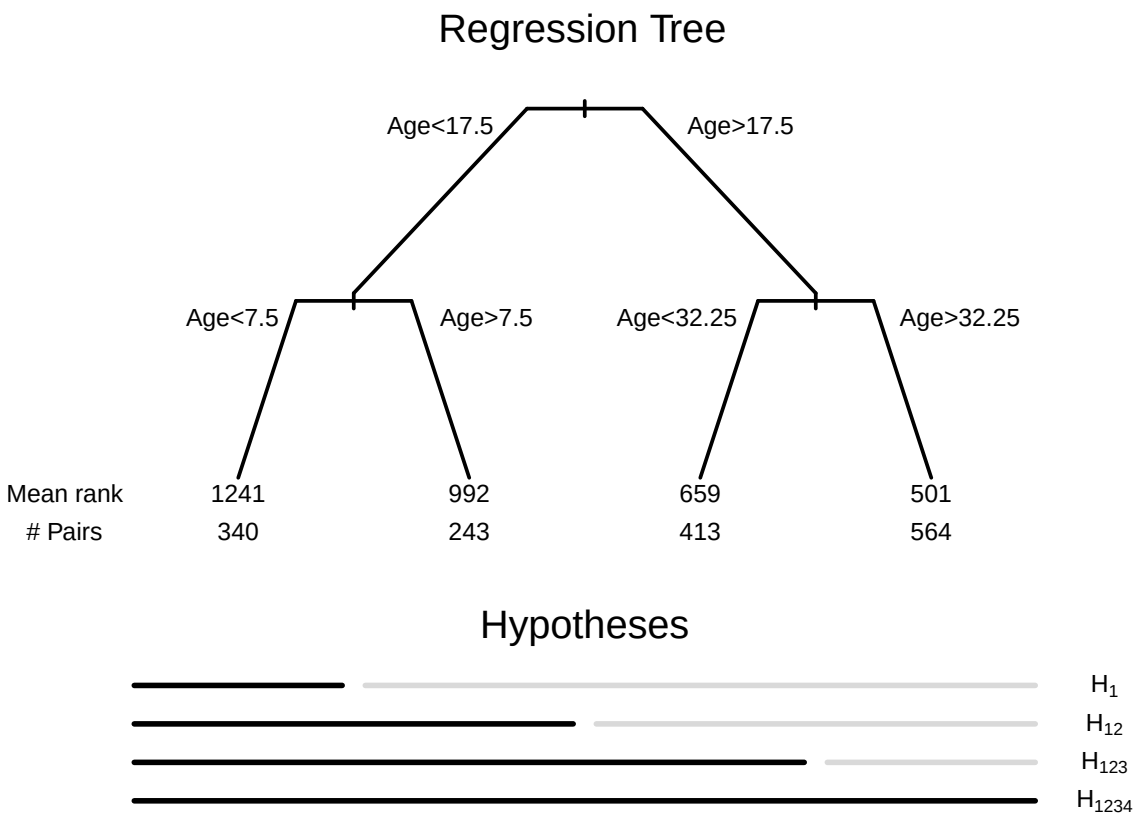

Figure 3: The regression tree formed four groups of matched pairs fitting rank($|Y_i|$) using age and gender, where it decided to ignore gender. As it turned out, the fitted means of rank($|Y_i|$) were decreasing in age, so four hypotheses were tested simultaneously: (i) H_1 no effect for age < 7.5, (ii) H_{12} no effect for age < 17.5, (iii) H_{123} no effect for age < 32.5, (iv) $H_0 = H_{1234}$ no effect for every age. The smallest *P*-value was for H_1 at Γ=5.8, and after correcting for multiple testing it was 0.0475.