2013

Case Definition and Design Sensitivity

Dylan S. Small
University of Pennsylvania

Jing Cheng

M. Elizabeth Halloran

Paul R. Rosenbaum

Recommended Citation

Small, D. S., Cheng, J., Halloran, M., & Rosenbaum, P. R. (2013). Case Definition and Design Sensitivity. *Journal of the American Statistical Association, 108* (504), 1457-1468. http://dx.doi.org/10.1080/01621459.2013.820660

Case Definition and Design Sensitivity

Abstract

In a case-referent study, cases of disease are compared to noncases with respect to their antecedent exposure to a treatment in an effort to determine whether exposure causes some cases of the disease. Because exposure is not randomly assigned in the population, as it would be if the population were a vast randomized trial, exposed and unexposed subjects may differ prior to exposure with respect to covariates that may or may not have been measured. After controlling for measured preexposure differences, for instance by matching, a sensitivity analysis asks about the magnitude of bias from unmeasured covariates that would need to be present to alter the conclusions of a study that presumed matching for observed covariates removes all bias. The definition of a case of disease affects sensitivity to unmeasured bias. We explore this issue using: (i) an asymptotic tool, the design sensitivity, (ii) a simulation for finite samples, and (iii) an example. Under favorable circumstances, a narrower case definition can yield an increase in the design sensitivity, and hence an increase in the power of a sensitivity analysis. Also, we discuss an adaptive method that seeks to discover the best case definition from the data at hand while controlling for multiple testing. An implementation in R is available as SensitivityCaseControl.

Keywords

case-control study, matching, observational study, sensitivity analysis

Disciplines

Statistics and Probability

Case Definition and Design Sensitivity

Dylan S. Small, Jing Cheng, M. Elizabeth Halloran, Paul R. Rosenbaum[1]

University of Pennsylvania, University of California at San Francisco, University of Washington and Fred Hutchinson Cancer Research Center

Abstract. In a case-referent study, cases of disease are compared to non-cases with respect to their antecedent exposure to a treatment in an effort to determine whether exposure causes some cases of the disease. Because exposure is not randomly assigned in the population, as it would be if the population were a vast randomized trial, exposed and unexposed subjects may differ prior to exposure with respect to covariates that may or may not have been measured. After controlling for measured pre-exposure differences, for instance by matching, a sensitivity analysis asks about the magnitude of bias from unmeasured covariates that would need to be present to alter the conclusions of a study that presumed matching for observed covariates removes all bias. The definition of a case of disease affects sensitivity to unmeasured bias. We explore this issue using: (i) an asymptotic tool, the design sensitivity, (ii) a simulation for finite samples, and (iii) an example. Under favorable circumstances, a narrower case definition can yield an increase in the design sensitivity, and hence an increase in the power of a sensitivity analysis. Also, we discuss an adaptive method that seeks to discover the best case definition from the data at hand while controlling for multiple testing. An implementation in `R` is available as `SensitivityCaseControl`.

Keywords: Case-control study; matching; observational study; sensitivity analysis.

[1] *Address for correspondence:* Professor Dylan Small, Department of Statistics, The Wharton School, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. E-mail: dsmall@wharton.upenn.edu. 24 June 2013. This study was supported by grant SES-1260782 from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation, grant RC4MH092722 from the National Institute of Mental Health, grant U54DE019285 from the National Institute of Dental and Craniofacial Research and grant R37-AI032042 from the National Institute of Allergy and Infectious Diseases.

1 Introduction: Motivating example; Outline

1.1 Defining a case in a case-referent study

A case-referent study compares cases of some disease or rare event to some group of non-cases, called referents by Oli Miettenen, looking backwards in time to contrast the frequency of treatment among cases and referents. Although the non-cases or referents in a case-referent study are sometimes called "controls," this is not the best terminology, because conventionally controls did not receive the treatment, rather than not manifesting the disease outcome. In a nested or synthetic case-referent study, a single cohort yields a case-referent study by oversampling cases and undersampling referents from the cohort, and the sampling is used to reduce the costs of obtaining needed information about pretreatment covariates or exposures to treatments; see, for instance, Mantel (1973).

Randomization assigns subjects to treatments in clinical trials but not in observational studies, and in particular, not in case-referent studies. Absent randomization, there is little to ensure that treated and untreated subjects were comparable prior to treatment, so differing outcomes among treated and untreated subjects may not be effects caused by the treatment but rather may reflect differences in measured or unmeasured pretreatment covariates. After controlling for measured covariates, for instance by matching, a sensitivity analysis asks: What magnitude of bias from unmeasured covariates would need to be present to materially alter the conclusions of a naive analysis that presumes adjustments for measured covariates suffice to remove all bias?

The first step in designing a case-referent study is to define a case of disease and a referent. Our goal in this manuscript is to examine the effects of this design decision on the sensitivity of conclusions to unmeasured biases. Before discussing technical issues, it is useful to consider an example.

1.2 Example: Child abuse and adult anger

Does physical abuse by parents in childhood result in a tendency towards greater anger in adulthood? Springer et al. (2007) examined this question using the 1993-4 sibling survey of the Wisconsin Longitudinal Study (WLS). Two questions (nw036rer and nw037rer) asked: "During the first 16 years of your life, how much did your father/mother slap, shove or throw things at you?" Responses were "not at all," "a little," "some" and "a lot," and Springer et al. (2007, p. 519) defined physical abuse as a response of "a lot" or "some" to at least one of the two questions. Anger was measured using Spielberger's (1996) anger scale. Our two illustrative case-referent studies are built from this one cohort study, and they use the 2841 people with data on abuse, anger and seven covariates. Because the Wisconsin Longitudinal Study (WLS) is a cohort study, it is possible to contrast several nested case-referent studies, with different case definitions, that might be built from the WLS. A conventional case-referent study would have a single case definition, so it could not illustrate the consequences of different case definitions.

We consider two definitions of a case. The narrow definition consisted of individuals with anger scores greater than or equal to the 90% point of 18, yielding 312 cases, while the broad definition consisted of individuals with anger scores greater than or equal to the 75% point of 10, yielding 794 cases. Referents had anger scores less than 10 in both definitions. Narrow cases were matched to four referents, making $4 \times 312 = 1248$ referents or $312 + 1248 = 1560$ people, while broad cases were pair matched making 794 pairs or $2 \times 794 = 1588$ people. A conventional case-referent study would interview cases and referents to determine exposure to the treatment and covariates, and the study's cost might be the sum of certain fixed costs plus a cost that is proportional to the number of interviews. For instance, such a case-referent study might use the anger scores from the WLS to define cases, then substantially improve the measurement of parental abuse by

reinterviewing cases and referents with an appropriate structured interview. With 1560 or 1588 subjects to interview, the two designs would have similar costs. Which design is less sensitive to unmeasured biases?

We matched for sex (ssbsex), age (sa029re) at the time of interview, father' education (edfa57q), mother's education (edmo57q), parental income (bmpin1), farm background (derived from rlur57), and an indicator of parents' marital problems or single parent (nw001rer). The matching minimized a robust Mahalanobis distance with penalties to ensure an exact match for sex and for age ≤ 50 years, age strictly between 50 and 57, age ≥ 57 years; see Rosenbaum (2010, §8) for discussion of the robust distance and penalty functions. Figure 1 shows covariate balance after matching. Matched subjects are typically close on covariates. For instance, in the broad definition, the Spearman correlations between paired cases and referents is .94 for age, .96 for father's education, .96 for mother's education and .96 for parental income. Table 1 shows the frequencies of child abuse among cases and referents with broad and narrow case definitions.

Table 2 conducts four sensitivity analyses, two for the broadly defined pairs, two for the narrowly defined matched sets, all using the technique in Rosenbaum (1991; 2002, §4.4.4-§4.4.5) which is briefly reviewed in §2.2. The sensitivity analysis is indexed by a parameter $\Gamma \geq 1$: it says that, under the null hypothesis of no treatment effect, two subjects matched for observed covariates may differ in their odds of exposure to the treatment, here child abuse, by at most a factor of Γ. The situation with $\Gamma = 1$ yields the usual reference distribution for the Mantel-Haenszel statistic; more generally, at each value of observed covariates, $\Gamma = 1$ yields a randomization distribution. For $\Gamma > 1$, the distribution of treatment assignments is unknown but to a degree controlled by the magnitude of Γ. At each value of Γ, there is a range of possible values for an inference quantity, say a P-value or a point estimate, and a sensitivity analysis computes this range for several values of Γ.

Table 2 reports upper bounds on the P-value testing the null hypothesis of no effect of abuse on anger. In matched pairs, it is possible to unpack Γ into two parameters, Λ controlling the relationship between exposure to the treatment Z, here abuse, and an unobserved covariate u, and Δ controlling the relationship between the outcome r_C, here anger, and the same unobserved covariate u, where $\Gamma = (\Lambda\Delta + 1) / (\Lambda + \Delta)$; see Rosenbaum and Silber (2009, Proposition 1) for a precise statement. The values $\Gamma = 1.1$, 1.2, and 1.5 are important in Table 2. For these values, in matched pairs, an unobserved covariate that doubles the odds, $\Delta = 2$, of a positive difference in anger scores corresponds with $\Gamma = 1.1$ if the covariate increases the odds of abuse by a factor of $\Lambda = 1.333$, with $\Gamma = 1.2$ if it increases the odds of abuse by $\Lambda = 1.75$, and with $\Gamma = 1.5$ if it multiplies the odds of abuse by $\Lambda = 4$; see Figure 2. How much bias Γ would need to be present to alter the conclusions of a naive analysis ($\Gamma = 1$) that assumes matching removes all bias?

Two sensitivity analyses in Table 2 use the familiar Mantel-Haenszel test which views case-referent status as binary, so cases are not distinguished by their anger scores beyond being 10 or more for the broad definition and 18 or more for the narrow definition. Before matching, the median anger score in the broad group was 15, not 10, the median in the narrow group was 25, not 18, and the median among referents was 3. The second sensitivity analysis uses aberrant ranks as defined by Rosenbaum and Silber (2008), where cases are ranked by their anger scores and referents have rank zero. Aberrant ranks ignore variation in some normal range and measure the magnitude of the deviation from the normal range.

In Table 2, if matching had removed all bias, $\Gamma = 1$, then the null hypothesis of no effect of abuse on anger would be rejected with a small P-value for both case definitions and both test statistics, and a bias of $\Gamma = 1.1$ could not lead to acceptance. A bias of $\Gamma = 1.2$ could lead to a P-value above 0.05 for the broad definition using the Mantel-Haenszel test, and a bias of $\Gamma = 1.3$ could do the same for the broad definition using the aberrant rank test.

In contrast, the narrow definition is insensitive to a bias of $\Gamma = 1.5$, with similar results for the Mantel-Haenszel and aberrant rank tests.

So, in this one example, the narrow definition of a case is better, leading to less sensitivity to unmeasured biases. Using the broad definition with ranks to give greater weights to more extreme cases is better than the broad definition without ranks, but not as good as the narrow definition with or without ranks. Obviously, this is one example. How general is the phenomenon seen in Table 2? Under what circumstances should we expect less sensitivity from a narrower case definition? Are there analytical strategies that ensure opportunities for reduced sensitivity will not be missed? Are there study designs that have better prospects? These are the questions addressed in the current manuscript.

2 Notation; Review; Example

2.1 Notation for many possible case-referent studies from one population

In a population, there are L individuals, $\ell = 1, \ldots, L$, some of whom were exposed to a treatment, denoted $Z_\ell = 1$, others being spared exposure, denoted $Z_\ell = 0$. In §1.2 and Springer et al. (2007), Z_ℓ refers to physical abuse as a child by parents. If exposed to the treatment, individual ℓ exhibits response $\mathbf{r}_{T\ell}$, whereas if ℓ is spared exposure then ℓ exhibits response $\mathbf{r}_{C\ell}$, so the response actually exhibited by ℓ is $\mathbf{R}_\ell = Z_\ell \, \mathbf{r}_{T\ell} + (1 - Z_\ell) \, \mathbf{r}_{C\ell}$ and the effect caused by the treatment $\mathbf{r}_{T\ell} - \mathbf{r}_{C\ell}$ is not observed for any individual (Neyman 1923, Rubin 1974). In §1.2, $\mathbf{r}_{T\ell}$ is the anger score ℓ would exhibit if abused, $\mathbf{r}_{C\ell}$ is the anger score ℓ would exhibit if not abused, $\mathbf{r}_{T\ell} - \mathbf{r}_{C\ell}$ is the effect that abuse would have on ℓ's anger score, and $\mathbf{R}_\ell$ is the anger score observed from ℓ under the treatment Z_ℓ that ℓ actually received. Let $\mathcal{R}$ be a set containing the possible values of $\mathbf{r}_{T\ell}$ and $\mathbf{r}_{C\ell}$. In §1.2, $\mathcal{R}$ is the set of integers from 0 to 70, $\mathcal{R} = \{0, 1, \ldots, 70\}$. In addition, each individual has a vector of observed covariates $\mathbf{x}_\ell$ and there is concern about another unmeasured covariate

u_ℓ. In §1.2, $\mathbf{x}_\ell$ contained seven covariates, four of which appear in Figure 1. Write $\mathcal{F} = \{(\mathbf{r}_{T\ell}, \mathbf{r}_{C\ell}, \mathbf{x}_\ell, u_\ell), \ell = 1, \ldots, L\}$. Fisher's sharp null hypothesis of no treatment effect asserts $H_0 : \mathbf{r}_{T\ell} = \mathbf{r}_{C\ell}, \ell = 1, \ldots, L$. Using notation of this kind, Holland and Rubin (1988) discuss case-referent studies including some of their limitations.

A case definition $\kappa(\cdot)$ is a function, $\kappa : \mathcal{R} \to \{1, 0, -1\}$, where $\kappa(\mathbf{r}) = 1$ for a case, $\kappa(\mathbf{r}) = 0$ for a referent, and $\kappa(\mathbf{r}) = -1$ for all others. In §1.2, for the narrow definition, $\kappa(\mathbf{r}) = 1$ if the anger score $\mathbf{r}$ was ≥ 18, $\kappa(\mathbf{r}) = 0$ if the score was < 10, and $\kappa(\mathbf{r}) = -1$ if the score was ≥ 10 and < 18. A case-referent study entails the steps: (i) create I labels, $i = 1, \ldots, I$, (ii) define a case, that is, the function $\kappa(\cdot)$, (iii) sample at random without replacement I cases (perhaps all the cases) with $\kappa(\mathbf{R}_\ell) = 1$, assigning them the labels $i = 1, \ldots, I$ at random, noting the value of $\mathbf{x}$ for the ith case, (iv) for case i, sample at random $J - 1 \geq 1$ referents with $\kappa(\mathbf{R}_\ell) = 0$ and the same value of $\mathbf{x}$, (v) attach noninformative (perhaps random) indices $j = 1, \ldots, J$ to the J individuals in matched set i, noticing that $\kappa(\mathbf{R}_{ij}) = 0$ or $\kappa(\mathbf{R}_{ij}) = 1$ for each i, j, $1 = \sum_{j=1}^{J} \kappa(\mathbf{R}_{ij})$ for each i, and $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ for each i, j, j', (vi) draw inferences about treatment effects on the basis of $\mathbf{R}_{ij}$, Z_{ij}, $\mathbf{x}_{ij}$ among sampled cases and referents. In §1.2, $L = 2841$, all $I = 312$ narrow cases with $\kappa(\mathbf{R}_\ell) = 1$ were used, and $J - 1 = 4$ referents were matched to each narrow case. If $\mathbf{R}_\ell$ is binary, then there is effectively only one possible case definition and no possible excluded groups with $\kappa(\mathbf{r}) = -1$. This notation is fairly expressive. For instance, if $\kappa(\mathbf{r}_{Cij}) = 0$ and $\kappa(\mathbf{r}_{Tij}) = -1$, then ij would be sampled as a referent $\kappa(\mathbf{R}_{ij}) = 0$ if spared exposure, $Z_{ij} = 0$, but if exposed, $Z_{ij} = 1$, ij would not have been a candidate for the case-referent study because $\kappa(\mathbf{R}_{ij}) = -1$.

Case definition $\kappa(\cdot)$ identifies a group, $\kappa(\mathbf{R}_\ell) = 1$, that is oversampled, but a case definition does not require the outcome $\mathbf{R}_\ell$ to be analyzed as binary. The aberrant rank test in Table 2 distinguished among cases, $\kappa(\mathbf{R}_\ell) = 1$, based on the degree of their caseness,

$\mathbf{R}_\ell$, and other procedures also do this; see, for instance, Mukherjee, Liu and Sinha (2007).

As is seen in §3 and §4, the choice of case definition $\kappa(\cdot)$ can have a substantial impact on the power of a sensitivity analysis testing the null hypothesis of no effect, H_0. A change in case definition $\kappa(\cdot)$ will typically also change the magnitude and meaning of parameter estimates that characterize the magnitude of an effect when the null hypothesis is false. In §1.2, the magnitude of the effect of child abuse on extreme anger may be different from its effect on elevated anger.

Write $Y_i = \sum_{j=1}^{J} Z_{ij}\,\kappa(\mathbf{R}_{ij})$, so $Y_i = 1$ if the case in matched set i was exposed to the treatment and $Y_i = 0$ otherwise. The rows of Table 1 record Y_i. Each case or matched set i has a score d_i that is a function of $\mathcal{F}$ when H_0 is true. Write $T = \sum_{i=1}^{I} d_i Y_i$ for the total score for exposed cases. In Table 2, the Mantel-Haenszel test has $d_i = 1$ for all i so T is the number of exposed cases, while the aberrant rank test has d_i equal to the rank of the anger score of the case, which is a function of $\mathcal{F}$ when H_0 is true, so T is the total of ranks for exposed cases. Write $m_i = \sum_{j=1}^{J} Z_{ij}$ for the number exposed in set i.

The Mantel-Haenszel statistic is the large sample approximation to the uniformly most powerful unbiased test of $H_* : \varpi = 0$ against $H_A : \varpi > 0$ in a conditional logit model for exposure Z_{ij} with a parameter υ_i for each matched set representing the matched covariates and a parameter ϖ representing the effect of exposure on case-referent status, $\kappa(\mathbf{R}_{ij}) = 1$ or 0; see Cox (1970, §5.3). For matched pairs, $J = 2$, the Mantel-Haenszel statistic becomes McNemar's test; see Cox (1970, §5.2). If the conditional distribution of $\mathbf{R}_{ij}$ given $Z_{ij} = z$ were multivariate Normal with expectation $\boldsymbol{\eta}_z$ and covariance matrix $\boldsymbol{\Sigma}$, then the conditional distribution of Z_{ij} given $\mathbf{R}_{ij}$ would follow a linear logit model (Cox 1970, Problem 49, p. 121); however, this is not true for most distributions of $\mathbf{R}_{ij}$.

With a finite population of L individuals, a more restrictive case-definition $\kappa(\cdot)$ may reduce the number of cases and force I to be smaller, while a less restrictive case definition

may increase the number of cases and permit I to be larger; see Table 1. If the total cost of the study is proportional to the number IJ of individuals studied, then a more restrictive case definition may reduce the number of cases, I, but it may with the same budget permit more referents to matched to each case so that IJ remains constant. In practice, an algorithm of some sort creates a close but not an exact matching in step (iv), but issues of this sort are peripheral to the current topic so we assume step (iv) is feasible as described, as it would be if $\mathbf{x}$ were discrete taking a moderate number of values and if $(L-I)/I$ were much larger than J, as is typically true in case-referent studies.

If one assumes that there is no bias from unmeasured covariates, so that matching for $\mathbf{x}_{ij}$ effectively creates a randomized experiment, then considerations of efficiency might lead to a preference for matched pairs, $J=2$; see Ury (1975). If biases from nonrandom assignment may be present, then, in cohort studies with continuous responses, matched sets may yield greater power in a sensitivity analysis; see Rosenbaum (2013). As a consequence, we evaluate power for several values of $J \geq 2$.

2.2 Treatment assignment in the population; sensitivity analysis

The model for treatment assignment in the population asserts that treatment assignments for distinct individuals are independent and

$$\Pr\left(Z_\ell = 1 \mid \mathcal{F}\right) = \frac{\exp\left\{\lambda\left(\mathbf{x}_\ell\right) + \gamma u_\ell\right\}}{1+\exp\left\{\lambda\left(\mathbf{x}_\ell\right) + \gamma u_\ell\right\}} \text{ with } 0 \leq u_\ell \leq 1, \tag{1}$$

where $\lambda(\cdot)$ is an unknown real-valued function and $\gamma \geq 0$ is an unknown parameter. Write $\Gamma = \exp(\gamma) \geq 1$, so that (1) says that two individuals, ℓ and ℓ', with the same observed covariates, $\mathbf{x}_\ell = \mathbf{x}_{\ell'}$, may differ in their odds of exposure by at most a factor of $\Gamma \geq \exp\left(\gamma\left|u_\ell - u_{\ell'}\right|\right)$ because of differences in $\mathcal{F}$. A sensitivity analysis asks how large must Γ be to alter the conclusions of a naïve analysis that assumes adjustments for observed

covariates suffice to remove all bias. The sensitivity model (1) is related to the sensitivity analysis proposed by Cornfield et al. (1959); see also Gastwirth (1992).

In the case-referent study under Fisher's null hypothesis H_0 and (1), the conditional distribution of the Z_{ij} given the m_i is free of the unknown $\lambda(\cdot)$ because $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, and it is not difficult to show that the Y_i are conditionally independent given $\mathcal{F}$ and $\mathbf{m} = (m_1, \ldots, m_I)^T$, with

$$\overline{p}_i = \frac{m_i}{m_i + \Gamma(J - m_i)} \leq \Pr(Y_i = 1 \mid \mathbf{m}, \mathcal{F}) \leq \frac{\Gamma m_i}{\Gamma m_i + (J - m_i)} = \overline{\overline{p}}_i, \tag{2}$$

so that $T = \sum d_i Y_i$ is the sum of I conditionally independent random variables taking the value d_i with probabilities bounded by (2) and otherwise taking the value 0; see Rosenbaum (1991; 2002, §4.4.4). Let $\overline{\overline{T}}$ be the sum of I independent random variables taking the value d_i with probability $\overline{\overline{p}}_i$ and the value 0 with probability $1 - \overline{\overline{p}}_i$, and define $\overline{T}$ in the same way but with $\overline{p}_i$ in place of $\overline{\overline{p}}_i$. If $\gamma = 0$ so $\Gamma = 1$, then $\overline{p}_i = \overline{\overline{p}}_i = m_i/J$, and in a one sided test of H_0, the quantity

$$1 - \Phi\left(\frac{T - \sum d_i m_i/J}{\sqrt{\sum d_i^2 m_i (J - m_j)/J^2}}\right) \tag{3}$$

is the approximate one-sided P-value for a within-set randomization test, for instance the Mantel-Haenszel statistic with $d_i = 1$ and no continuity correction. (We omit the continuity correction commonly associated with the Mantel-Haenszel statistic because the correction is inappropriate with certain types of scores d_i and has only a minor effect otherwise, and we do not want to present two sets of formulas, with and without correction. A user can apply a continuity correction if desired.) Under H_0 with $\Gamma \geq 1$, the null

distribution $\Pr(T \geq k \mid \mathbf{m}, \mathcal{F})$ may be bounded using (2), yielding the exact bounds

$$\Pr\left(\overline{T} \geq k \,\middle|\, \mathbf{m}, \mathcal{F}\right) \leq \Pr(T \geq k \mid \mathbf{m}, \mathcal{F}) \leq \Pr\left(\overline{\overline{T}} \geq k \,\middle|\, \mathbf{m}, \mathcal{F}\right) \tag{4}$$

and the corresponding large sample approximate bounds

$$1 - \Phi\left(\frac{k - \sum d_i \, \overline{p}_i}{\sqrt{\sum d_i^2 \, \overline{p}_i \, (1 - \overline{p}_i)}}\right) \leq \Pr(T \geq k \mid \mathbf{m}, \mathcal{F}) \leq 1 - \Phi\left(\frac{k - \sum d_i \, \overline{\overline{p}}_i}{\sqrt{\sum d_i^2 \, \overline{\overline{p}}_i \left(1 - \overline{\overline{p}}_i\right)}}\right), \tag{5}$$

so that (5) evaluated at $k = T$ gives the approximate bounds on the one-sided P-value for each specific Γ; see Rosenbaum (1991; 2002, §4.4.4). In particular, both bounds in (5) equal (3) when $\Gamma = 1$ in (2). Table 2 reports the upper bound on the right in (5).

The sensitivity analysis based on model (1) is quite general and applies to many kinds of response — binary, ordinal, continuous, censored — and to many study designs (e.g., Rosenbaum 2002, §4; Small and Rosenbaum 2008); however, in the special situation of a case-referent study with binary outcomes and $d_i = 1$ for all i, the bounds on P-values for general Γ in (5) are closely related to familiar confidence limits for a common odds ratio under $\Gamma = 1$. Specifically, the upper bound in (5) equals 0.05 at the value of Γ which forms the endpoint of a one-sided 95% confidence interval for a common odds ratio for a $2 \times 2 \times I$ table; see Rosenbaum (1991) for discussion. This simplification again reinforces the connection to the method of Cornfield et al. (1959). Various methods of sensitivity analysis in observational studies are discussed by Diprete and Gangl (2004), Egleston et al. (2009), Hosman et al. (2010), Imbens (2003), Gastwirth (1992), Lin et al. (1998), Marcus (1997), McCandless et al. (2007), Yanagawa (1984) and Yu and Gastwirth (2005).

2.3 Exact sensitivity bounds

The exact bound based on $\overline{\overline{T}}$ in (4) for integer d_i is easily obtained by convolving probability generating functions (Rosenbaum 2010, §3.9). Although useful in general, exact bounds are used here in the adaptive procedure in §5. If $d_i = k$, the generating function for the ith summand is a vector with $k+1$ coordinates, the first coordinate being $1-\overline{\overline{p}}_i$, the $k+1$ coordinate being $\overline{\overline{p}}_i$, the remaining $k-1$ coordinates being 0, $g_i = \left(1-\overline{\overline{p}}_i, 0, 0, \ldots, 0, \overline{\overline{p}}_i\right)$, which signifies that 0 occurs with probability $1-\overline{\overline{p}}_i$ and k occurs with probability $\overline{\overline{p}}_i$. In `R` (R Core Team 2012) define: `gconv <- function(g1,g2){convolve(g1,rev(g2), type="o")}`. If $g^{(i)}$ is the probability generating function for the sum of the first i summands, then $g^{(i+1)} = \texttt{gconv}\left(g^{(i)}, g_{i+1}\right)$. For instance, with pair matching, $J = 2$, if we retain only discordant pairs with $m_i = 1$ exposed subject in the pair, then $\overline{\overline{p}}_i = \Gamma/\left(\Gamma+1\right)$. For $\Gamma = 3$, $I = 3$ discordant pairs and the Mantel-Haenszel statistic with $d_i = 1$ for all i,

```
gconv(gconv(c(1/4,3/4),c(1/4,3/4)),c(1/4,3/4))
```

yields `0.015625, 0.140625, 0.421875, 0.421875` as $\Pr\left(\overline{\overline{T}} = k \,\middle|\, \mathbf{m}, \mathcal{F}\right)$ for $k = 0$, 1, 2, 3. For $\Gamma = 3$, $I = 3$ discordant pairs and the aberrant rank statistic with $d_i = i$ for all i,

```
gconv(gconv(c(1/4,3/4),c(1/4,0,3/4)),c(1/4,0,0,3/4))
```

yields `0.015625, 0.046875, 0.046875, 0.187500, 0.140625, 0.140625, 0.421875` as $\Pr\left(\overline{\overline{T}} = k \,\middle|\, \mathbf{m}, \mathcal{F}\right)$ for $k = 0,\ 1,\ 2,\ \ldots,\ 6$. Matched sets with $J > 2$ and concordant matched sets with $m_i = 0$ or $m_i = J$ change the value of $\overline{\overline{p}}_i$ but otherwise require no special treatment; however, removal of concordant sets saves computation without altering significance levels. This method may be applied when matched sets vary in size, with a different set size J_i for each set i. For $\Gamma = 1$, $d_i = 1$ for all i, $J_i \geq 2$, this procedure yields the familiar exact null distribution of $\sum Y_i$; see Cox (1970, §5.3).

3 Design sensitivity in case-referent studies

3.1 What is design sensitivity?

If a case-referent study were actually free of bias from unmeasured covariates, we could not recognize this from the data, and the best we could hope to say is that the study is insensitive to small and moderate biases. The power of a sensitivity analysis is the probability that we will be able to say this, where the probability is computed under a model for $\mathbf{R}_{ij}$ with a treatment effect and no bias. It is particularly in the case of a treatment effect without bias that we hope the sensitivity analysis will reassure us by saying that only large biases could explain away the ostensible treatment effect, so power is computed with a treatment effect and no bias. More precisely, the power of a one-sided level α sensitivity analysis is the probability that the upper bound in (5) is less than or equal to α when, in fact, there is no bias from u_{ij} so $\gamma = 0$ in (1) but there is a treatment effect; this is the *favorable situation* in which we would like to report insensitivity to unmeasured bias. In many situations, there is a value, $\widetilde{\Gamma}$, called the design sensitivity, such that, as the number of matched sets increases, $I \rightarrow \infty$, the power of the sensitivity analysis goes to 1 if the sensitivity analysis is performed with $\Gamma < \widetilde{\Gamma}$ and it tends to 0 for $\Gamma > \widetilde{\Gamma}$, so in the limit we can distinguish a particular treatment effect without bias from all biases smaller than $\widetilde{\Gamma}$ but not from some biases larger than $\widetilde{\Gamma}$. Figures 14.2 and 14.3 in Rosenbaum (2010) depict this convergence with increasing but finite I. For calculations of $\widetilde{\Gamma}$ and the power of a sensitivity analysis, see Rosenbaum (2004, 2010, 2012a, 2013), Small and Rosenbaum (2008), and Heller et al. (2009).

The goal of §3 is to study the relationship between case-definition $\kappa(\cdot)$ and design sensitivity $\widetilde{\Gamma}$. In contrast to the asymptotic results in §3, in §4 the effect of case definition $\kappa(\cdot)$ on the finite-population, finite-sample power of a sensitivity analysis is examined.

The asymptotics of design sensitivity are intended to indicate how changing the definition of a case, $\kappa(\cdot)$, affects sensitivity to bias, and for this limited purpose it is reasonable to employ slightly stylized assumptions such as exactly matched, independent and identically distributed (*iid*) observations from an infinite population. Asymptotic results will assume that the finite population of L individuals is a simple random sample from an infinite population, that I cases with $\kappa(\mathbf{R}) = 1$ are then randomly sampled from the cases among these L individuals and matched exactly for the observed covariates $\mathbf{x}$ with $J-1$ distinct referents with $\kappa(\mathbf{R}) = 0$, so that the finite population is an *iid* sample of L individuals from the infinite population, and the I case-referent matched sets are an *iid* sample from the infinite population of case-referent sets that can be constructed from the infinite population of individuals. When reference is made to the distribution of a quantity in the infinite population, the subscript ℓ is omitted.

The favorable situation is defined by a treatment effect and no unmeasured bias. That is, in the favorable situation, $\gamma = 0$ in (1), so that by Bayes' theorem applied to (1),

$$\begin{aligned}
\Pr(\mathbf{R} \mid Z = 1, \mathbf{x}) &= \Pr(\mathbf{r}_T \mid Z = 1, \mathbf{x}) = \Pr(\mathbf{r}_T \mid \mathbf{x}) \\
\Pr(\mathbf{R} \mid Z = 0, \mathbf{x}) &= \Pr(\mathbf{r}_C \mid Z = 0, \mathbf{x}) = \Pr(\mathbf{r}_C \mid \mathbf{x}),
\end{aligned}$$

and a cohort study of the finite population of L individuals could consistently estimate (as $L \rightarrow \infty$ with I/L fixed) treatment effects defined in terms of $\Pr(\mathbf{r}_T \mid \mathbf{x})$, $\Pr(\mathbf{r}_C \mid \mathbf{x})$ and $\Pr(\mathbf{x})$ such as the average treatment effect $\mathrm{E}(\mathbf{r}_T - \mathbf{r}_C)$; see Rubin (1974). Here, we are interested in the power $\Psi_{\Gamma,I,L}$ of a one-sided α-level sensitivity analysis testing Fisher's H_0 in a case-referent study with sensitivity parameter Γ; that is, $\Psi_{\Gamma,I,L}$ is the probability that the upper bound in (5) is less than or equal to α in a *favorable situation* $\mathcal{S}$ that specifies there is no unmeasured bias, $\gamma = 0$ in (1), and a treatment effect expressed in terms of specific distributions $\Pr(\mathbf{r}_T \mid \mathbf{x})$, $\Pr(\mathbf{r}_C \mid \mathbf{x})$ and $\Pr(\mathbf{x})$ with $\Pr(\mathbf{r}_T \mid \mathbf{x}) \neq \Pr(\mathbf{r}_C \mid \mathbf{x})$. The

Mantel-Haenszel test is discussed in §3.2 and the aberrant rank statistic in §3.3.

3.2 Design sensitivity of the Mantel-Haenszel test

Proposition 1 determines the design sensitivity for (5) in the special case with $d_i = 1$ for all i, so T is the Mantel-Haenszel statistic. Write $\rho = \mathrm{E}(Y)$ for the proportion of exposed cases in the infinite population when $\mathcal{S}$ is true. Here, ρ depends upon the definition of a case, $\kappa(\cdot)$, but the notation does not indicate this explicitly. Also, when $\mathcal{S}$ is true, for $w \in \{0, 1, \ldots, J\}$, write $\pi_w = \Pr(m = w)$ for the probability that the number m of exposed individuals equals w in a set of J individuals formed by picking a case at random from the infinite population and picking $J - 1$ referents with the same $\mathbf{x}$ at random, and write $\mu = \sum_{w=0}^{J} w\,\pi_w$ for the expected value of m. To avoid degenerate cases, we assume that when $\mathcal{S}$ is true $\pi_w > 0$ for each $w \in \{0, 1, \ldots, J\}$. Here, π_w and μ depend upon both the definition of a case, $\kappa(\cdot)$ and the investigator's choice of J. In Proposition 1, the condition $\rho > \mu/J$ says, in effect, that the treatment effect is such that it yields a higher frequency of exposed cases than of exposed referents.

Proposition 1 *As $L \to \infty$ and $I \to \infty$ in the favorable situation $\mathcal{S}$, if $\rho > \mu/J$ then the power $\Psi_{\Gamma,I,L}$ of a one-sided α-level sensitivity analysis satisfies $\Psi_{\Gamma,I,L} \to 1$ if $\Gamma < \widetilde{\Gamma}$ and $\Psi_{\Gamma,I,L} \to 0$ if $\Gamma > \widetilde{\Gamma}$, where $\widetilde{\Gamma}$ is the unique ω that solves the equation*

$$\rho = \varphi(\omega) \text{ where } \varphi(\omega) = \sum_{w=0}^{J} \pi_w \left\{ \frac{\omega w}{\omega w + (J - w)} \right\}. \tag{6}$$

Proof. First we show that $\rho = \varphi(\omega)$ has a unique solution. If $Y = 1$ then $m \geq 1$ so $\Pr(Y = 1) = \rho < \Pr(m \geq 1) = 1 - \pi_0$. The function $\varphi(\omega)$ has $\varphi(1) = \mu/J < \rho$ and $\lim_{\omega \to \infty} \varphi(\omega) = 1 - \pi_0 > \rho$. Also, $\varphi(\omega)$ is continuous and strictly increasing, so $\rho = \varphi(\omega)$ does indeed have a unique solution, say $\widetilde{\Gamma}$. By the weak law of

large numbers, $T/I = I^{-1}\sum_{i=1}^{I} Y_i$ converges in probability to $\rho = \varphi\left(\widetilde{\Gamma}\right)$. Also, $\overline{\overline{p}}_i = \Gamma m_i / \{\Gamma m_i + (J - m_i)\}$ and in the infinite population $\pi_w = \Pr(m_i = w)$, so $\overline{\overline{p}}_i$ has expectation $\sum_{w=0}^{J} \pi_w [\Gamma w / \{\Gamma w + (J - w)\}]$, and again the weak law of large numbers implies

$$\frac{1}{I}\sum_{i=1}^{I} \overline{\overline{p}}_i \xrightarrow{P} \sum_{w=0}^{J} \pi_w \left\{ \frac{\Gamma w}{\Gamma w + (J - w)} \right\} = \varphi(\Gamma).$$

Because $\varphi(\cdot)$ is strictly increasing, the difference of these two limits, namely $\varphi\left(\widetilde{\Gamma}\right) - \varphi(\Gamma)$, has the same sign as $\widetilde{\Gamma} - \Gamma$. Because $\overline{\overline{p}}_i\left(1 - \overline{\overline{p}}_i\right) \leq 1/4$, it follows that $I^{-1}\sum_{i=1}^{I} \overline{\overline{p}}_i\left(1 - \overline{\overline{p}}_i\right) \leq 1/4$, and in (5)

$$\Pr\left\{ \frac{T - \sum \overline{\overline{p}}_i}{\sqrt{\sum \overline{\overline{p}}_i\left(1 - \overline{\overline{p}}_i\right)}} \geq t \right\} = \Pr\left\{ \frac{\sqrt{I}\left(T/I - I^{-1}\sum \overline{\overline{p}}_i\right)}{\sqrt{I^{-1}\sum \overline{\overline{p}}_i\left(1 - \overline{\overline{p}}_i\right)}} \geq t \right\}$$

tends to 1 for every t if $\Gamma < \widetilde{\Gamma}$ and to 0 if $\Gamma > \widetilde{\Gamma}$. ■

Example 2 *Consider a simple example in which a third of the population is exposed to the treatment,* $\Pr(Z_\ell = 1) = 1/3$, *responses if exposed are* $r_{T\ell} \sim N(1, 1)$, *whereas if not exposed responses are* $r_{C\ell} \sim N(0, 1)$, *the covariates are irrelevant so cases and referents are paired at random. The upper 10% point of* $N(0, 1)$ *is* $\Phi^{-1}(1 - .1) = 1.28$ *and the upper 20% point is* $\Phi^{-1}(1 - .2) = 0.84$. *If a case is narrowly defined by* $\kappa(R) = 1$ *if* $R > 1.28$, $\kappa(R) = 0$ *otherwise, then for* $J = 4$, *we obtain by simulation* ρ *and* $(\pi_0, \pi_1, \pi_2, \pi_3, \pi_4)$ *such that the probability of an exposed case is* $\rho = 0.66 = \varphi(5.7)$ *so* $\widetilde{\Gamma} = 5.7$. *This says that a sensitivity analysis using the Mantel-Haenszel test with this case definition has power tending to 1 for* $\Gamma < 5.7$ *and power tending to 0 for* $\Gamma > 5.7$. *Creating one large simulated study with* $I = 60,000$ *matched sets yields an upper bound (5) on the one-sided P-value of* 8.3×10^{-12} *at* $\Gamma = 5.5$ *and 0.999 at* $\Gamma = 5.9$. *If the case definition is broadened to* $\kappa(R) = 1$ *if* $R > 0.84$, $\kappa(R) = 0$ *otherwise, then* $\widetilde{\Gamma} = 5.2$ *yielding greater sensitivity to*

unmeasured biases.

Example 2 has $\Pr(R_\ell | Z_\ell = z, \mathbf{x}_\ell)$ as $N(\eta_z, 1)$ for $z = 0$, 1 and then cuts R_ℓ to form narrow and broad cases and referents. For use in the Mantel-Haenszel test, the same result would be obtained if R_ℓ were a three-category ordinal variable obtained from the corresponding ordinal probit model.

Table 3 calculates the design sensitivity using Proposition 1 for a binary covariate, $x = 0$ or $x = 1$, for various odds ratios, Ω_x for $x = 0, 1$, linking case-referent status to exposure to the treatment. In Table 3, when the odds ratio does not change with x, $\Omega_0 = \Omega_1$, the design sensitivity $\widetilde{\Gamma}$ equals the common value of the odds ratio. In Table 3, when exposure and disease are unrelated for $x = 0$ but related for $x = 1$, $\Omega_0 = 1 < \Omega_1$, the design sensitivity equals the marginal odds ratio in the table that collapses over x. In the other situations in Table 3, the design sensitivity is between Ω_0 and Ω_1, sometimes increasing with J, sometimes decreasing with J. Presumably, the situation in the example in Table 2 is more complex: there were seven covariates, three of which were binary.

3.3 Design sensitivity of the aberrant rank test

Let $\delta(\mathbf{R}_{ij})$ be a real-valued function of the observed response $\mathbf{R}_{ij}$ which is used to make quantitative distinctions among cases with $\kappa(\mathbf{R}_{ij}) = 1$. For the narrow cases in Table 2, $\delta(\mathbf{R}_{ij})$ is the actual anger score for cases who, by definition, have anger scores of at least 18, $\kappa(\mathbf{R}_{ij}) = 1$. For theoretical calculations of the design sensitivity, it is convenient to assume that the $\delta(\mathbf{R}_{ij})$ are from a continuous distribution and hence are untied; however, using average ranks for ties in (5) is appropriate in data analysis when ties are present. Each matched set contains one case, $1 = \sum_{j=1}^{J} \kappa(\mathbf{R}_{ij})$ and the score for this case is $\Delta_i = \sum_{j=1}^{J} \kappa(\mathbf{R}_{ij})\ \delta(\mathbf{R}_{ij})$. Let d_i be the rank of Δ_i among the I cases, so the d_i are a permutation of 1, 2, ..., I. Let $V_{ik} = 1$ if $\Delta_i \geq \Delta_k$, $V_{ik} = 0$ otherwise, where $V_{ii} = 1$, so

that the rank of Δ_i among cases is $d_i = \sum_{k=1}^{I} V_{ik}$. Finally, let $\theta = \mathrm{E}\left(Y_i\, V_{ik}\right)$ for $k \neq i$.

Proposition 3 *As $L \to \infty$ and $I \to \infty$ in the favorable situation $\mathcal{S}$, if $\rho > \mu/J$ then the power $\Psi_{\Gamma,I,L}$ of a one-sided α-level sensitivity analysis using the aberrant rank test satisfies $\Psi_{\Gamma,I,L} \to 1$ if $\Gamma < \widetilde{\Gamma}$ and $\Psi_{\Gamma,I,L} \to 0$ if $\Gamma > \widetilde{\Gamma}$, if $\widetilde{\Gamma}$ solves the equation $\theta = \zeta\left(\widetilde{\Gamma}\right)$ with*

$$\zeta(\omega) = \mathrm{E}\left\{\frac{V_{ik}\,\omega\, m_i}{\omega\, m_i + (J - m_i)}\right\} \text{ with } k \neq i.$$

Proof. Consider first the behavior of $\zeta(\omega)$ as ω changes. The quantity $V_{ik}\, m_i / \{m_i + (J - m_i)/\omega\}$ is nondecreasing in ω, and is strictly increasing for some values of k if $0 < m_i < J$. Because $\rho > \mu/J$, it follows that $\Pr(0 < m_i < J) > 0$ and so $\zeta(\omega)$ is strictly increasing in ω. As in the proof of Proposition 1, the proof concerns the behavior of the upper bound in (5), or more precisely the random variable

$$\frac{\sum_{i=1}^{I} d_i\, \left(Y_i - \overline{\overline{p}}_i\right)}{\sqrt{\sum d_i^2\, \overline{\overline{p}}_i \left(1 - \overline{\overline{p}}_i\right)}}. \tag{7}$$

The statistic $T = \sum_{i=1}^{I} d_i\, Y_i$ equals $\sum_{i=1}^{I} Y_i \left(1 + \sum_{k \neq i} V_{ik}\right)$ and has expectation $I\rho + I(I-1)\,\theta$, and $T/\{I(I-1)\}$ converges in probability to θ. Also, for each $\Gamma \geq 1$,

$$\begin{aligned}\{I(I-1)\}^{-1} \sum_{i=1}^{I} d_i\, \overline{\overline{p}}_i &= \{I(I-1)\}^{-1} \sum_{i=1}^{I} \overline{\overline{p}}_i \left(1 + \sum_{k \neq i} V_{ik}\right) \\ &= \{I(I-1)\}^{-1} \sum_{i=1}^{I} \frac{\Gamma m_i}{\Gamma m_i + (J - m_i)} \left(1 + \sum_{k \neq i} V_{ik}\right)\end{aligned}$$

converges in probability to $\zeta(\Gamma)$. Let

$$\nu_\Gamma = \max_{w \in \{0,\ldots,J\}} \frac{\Gamma w\,(J - w)}{\{\Gamma w + (J - w)\}^2} \text{ so that } \overline{\overline{p}}_i \left(1 - \overline{\overline{p}}_i\right) \leq \nu_\Gamma \text{ for all } i. \tag{8}$$

Now $\sum d_i^2 = 1^2 + 2^2 + \cdots + I^2 = I\left(I+1\right)\left(2I+1\right)/6$, so that using (8)

$$\left\{I\left(I-1\right)\right\}^{-1}\sqrt{\sum d_i^2\,\overline{\overline{p}}_i\left(1-\overline{\overline{p}}_i\right)} \leq \sqrt{\frac{\nu_\Gamma I\left(I+1\right)\left(2I+1\right)/6}{\left\{I\left(I-1\right)\right\}^2}} \rightarrow 0 \text{ as } I \rightarrow \infty.$$

As $I \rightarrow \infty$, the numerator of (7) multiplied by $\left\{I\left(I-1\right)\right\}^{-1}$ converges in probability to $\theta - \zeta\left(\Gamma\right)$ while the denominator multiplied by $\left\{I\left(I-1\right)\right\}^{-1}$ converges to zero for all Γ. ■

Example 4 *Continuing Example 2 with cases narrowly defined by* $\kappa\left(R\right) = 1$ *if* $R > 1.28$, $\kappa\left(R\right) = 0$ *otherwise, we obtain by simulation* $\theta = 0.36 = \zeta\left(7.5\right)$ *so* $\widetilde{\Gamma} = 7.5$ *which is substantially higher than 5.7 for the Mantel-Haenszel test. If one sample with* $I = 60,000$ *matched sets is drawn and the upper bound (5) on the one-sided P-value is computed for the aberrant rank statistic, the bound is 0.018 for* $\Gamma = 7.3$ *and 0.96 for* $\Gamma = 7.7$. *Using the aberrant rank test with the broad definition from Example 2 yields* $\widetilde{\Gamma} = 7.2$. *So in this one stylized Gaussian example, in large samples, the narrow definition together with the aberrant rank test is least sensitive to unmeasured biases.*

4 Simulated Power of a Sensitivity Analysis

We examine the power of sensitivity analyses in a simulation study. That is, we estimate the probability that a sensitivity analysis performed with its sensitivity parameter set to Γ will produce an upper bound on the one-sided P-value of at most $\alpha = 0.05$ in the favorable situation with a genuine treatment effect and no unmeasured biases. We consider a setting similar to Example 2 in which a third of the population is exposed to the treatment, $\Pr(Z_\ell = 1) = 1/3$. The responses if exposed to the treatment, $r_{T\ell}$, have mean 0.5 and the responses if not exposed, $r_{C\ell}$, have mean 0. The distributions of $r_{T\ell}$ and $r_{C\ell}$ are either a normal distribution with standard deviation 1 or a shifted t distribution with 3 or 5 degrees of freedom divided by its standard deviation of $\sqrt{3}$ or $\sqrt{5/3}$ respectively; therefore each

distribution has standard deviation 1, and the effect size is half of the standard deviation in all sampling situations. We refer to the two t distributions as standardized and write t_3 or t_5. The cut off for the broad definition of a case is chosen so that the population is 20% broad cases and 80% referents. For instance, when $r_{T\ell}$ and $r_{C\ell}$ have normal distributions, the cutoff point is the 0.8 quantile of the mixture of a $N(0,1)$ with probability $2/3$ and a $N(0.5,1)$ with probability $1/3$, which equals 1.031. A total of 2500 broad cases are sampled from their conditional distribution, and the required number of referents are sampled from their conditional distribution, 2500 or 1250 or 3750. The largest half of the 2500 broad cases are the narrow cases, so there are always 1250 narrow cases. There are three study designs. One has 2500 matched pairs of a broad case and a referent, making a total sample size of 5000. Another has 1250 matched pairs of a narrow case and a referent, excluding the remaining broad cases, making a total sample size of 2500. The third design has 1250 narrow cases, each matched to three referents, making a total sample size of 5000. In many contexts, the two designs with a total sample size of 5000 would have similar costs of data collection. Two test statistics are used, the Mantel-Haenszel statistic and the aberrant rank statistic. The power of sensitivity analysis is estimated based on 1000 simulations for each setting. Table 5 reports results to three decimals, so a power of 0.898 means that 898 of 1000 samples led to rejection. Table 5 reports the design sensitivity $\widetilde{\Gamma}$ for each sampling situation, but recall that $\widetilde{\Gamma}$ refers to the limiting power for large I. Not shown in Table 5 is a repetition of the simulation doubling all of the sample sizes. Generally, the patterns in that simulation were qualitatively similar to Table 5. Specifically, for $\Gamma < \widetilde{\Gamma}$ the powers were higher, closer to 1, and for $\Gamma > \widetilde{\Gamma}$ the powers were smaller, closer to zero, consistent with the general fact that the power is tending as I increases to a step function with a single step down from power 1 to power 0 at the design sensitivity $\widetilde{\Gamma}$. However, the relative performance of different procedures in different settings was similar with double the sample

size.

In the Normal case in Table 5, the highest power in a sensitivity analysis is obtained using the aberrant rank statistic on narrow cases matched to three referents, whereas the Mantel-Haenszel statistic with broad case-referent pairs has the lowest power. At $\Gamma = 2.5$, 7 of 1000 samples rejected using broad cases and the Mantel-Haenszel statistic, but 922 of 1000 samples rejected using narrow cases, 1-3 matching and the aberrant rank statistic. For the t_5 distribution, the highest power is attained using the Mantel-Haenszel test with narrow cases in 1-3 matched sets. For the t_3 distribution, the highest power is attained using the Mantel-Haenszel test with broad cases in matched pairs. Although often not the best, the Mantel-Haenszel statistic used with narrow cases matched to three referents is never very bad. Except for the t_3 distribution, the less expensive design with 1250 narrow pairs has better power in a sensitivity analysis using the Mantel-Haenszel test than does the more expensive design with 2500 broad pairs.

Because no one procedure is uniformly best, in §5 adaptive inference is proposed to always attain the design sensitivity of the better of the broad and narrow case definitions.

5 Adaptive Inference

5.1 Using two Mantel-Haenszel statistics with different case definitions

In Table 2, the least sensitive results were obtained using the narrow case definition and the Mantel-Haenszel statistic applied to a study with $I = 312$ cases having anger scores of at least 18 each matched to four referents with anger scores of less than 10. However, it is not possible to be certain of this before looking at the data. An adaptive sensitivity analysis selects the less sensitive analysis based on the data while correcting the significance level for performing more than one analysis (Rosenbaum 2012a). The adaptive procedure has the design sensitivity $\widetilde{\Gamma}$ of the Mantel-Haenszel statistic applied with the better of these

two case definitions, so in sufficiently large samples it sorts things out correctly; however, in finite samples, a price is paid for adaptation.

The $I = 794$ broadly defined cases in Table 1 may be divided into $I_1 = 312$ narrowly defined cases with anger scores of 18 or more, and $I_2 = 482$ marginal cases who are cases by the broad but not the narrow definition with anger scores of at least 10 but less than 18; see Table 4. Renumber the $I = 794$ pairs so the first $I_1 = 312$ pairs are the narrowly defined pairs. Then the Mantel-Haenszel statistic for the $I_1 = 312$ narrow cases is $T_1 = \sum_{i=1}^{I_1} Y_i$, for the marginal cases is $T_2 = \sum_{i=I_1+1}^{I} Y_i$, and for all cases is $T = T_1 + T_2 = \sum_{i=1}^{I} Y_i$. The adaptive inference will use both T_1 and T. For a fixed $\Gamma \geq 1$, let $\overline{\overline{T}}_1$ and $\overline{\overline{T}}_2$ and $\overline{\overline{T}} = \overline{\overline{T}}_1 + \overline{\overline{T}}_2$ be the upper bounding random variables defined in §2.2. We may determine the independent distributions of $\overline{\overline{T}}_1$ and $\overline{\overline{T}}_2$ exactly using the methods in §2.3, their joint distribution by multiplication, and the joint distribution of $\left(\overline{\overline{T}}_1, \overline{\overline{T}}\right)$ by rearranging the joint distribution of $\left(\overline{\overline{T}}_1, \overline{\overline{T}}_2\right)$. Select a value $0 < \alpha < 1$, conventionally $\alpha = 0.05$, and find k_1 and k such that under H_0

$$\Pr\left(\overline{\overline{T}}_1 \geq k_1 \text{ or } \overline{\overline{T}} \geq k \,\middle|\, \mathbf{m}, \mathcal{F}\right) \leq \alpha, \tag{9}$$

$$\Pr\left(\overline{\overline{T}}_1 \geq k_1 - 1 \text{ or } \overline{\overline{T}} \geq k \,\middle|\, \mathbf{m}, \mathcal{F}\right) > \alpha \text{ and } \Pr\left(\overline{\overline{T}}_1 \geq k_1 \text{ or } \overline{\overline{T}} \geq k - 1 \,\middle|\, \mathbf{m}, \mathcal{F}\right) > \alpha, \tag{10}$$

and subject to (9) and (10),

$$\left|\Pr\left(\overline{\overline{T}}_1 \geq k_1 \,\middle|\, \mathbf{m}, \mathcal{F}\right) - \Pr\left(\overline{\overline{T}} \geq k \,\middle|\, \mathbf{m}, \mathcal{F}\right)\right| \text{ is minimized.} \tag{11}$$

Here, (9) says that an adaptive procedure that rejects if either $\overline{\overline{T}}_1 \geq k_1$ or $\overline{\overline{T}} \geq k$ will falsely reject H_0 with probability at most α if (1) is true. Condition (10) says the pair of critical values (k_1, k) cannot be improved. Finally, condition (11) says that among pairs of critical values that cannot be improved, (k_1, k) are the most equitable, dividing the chance

of false rejection as equally as possible between the events $\overline{\overline{T}}_1 \geq k_1$ and $\overline{\overline{T}} \geq k$.

In Table 4, $T_1 = 60 = 51 + 9$, $T_2 = 63 = 50 + 13$, $T = 123 = 60 + 63$, where there are $9 + 13 = 22$ cases exposed in concordant pairs with $m_i = 2$. For the m_i in Table 4, with $\Gamma = 1.366$, under H_0, we find $\Pr\left(\overline{\overline{T}}_1 \geq 60 \text{ or } \overline{\overline{T}} \geq 133 \,\middle|\, \mathbf{m}, \mathcal{F}\right) = 0.04989485 \leq 0.05$, $\Pr\left(\overline{\overline{T}}_1 \geq 60 \,\middle|\, \mathbf{m}, \mathcal{F}\right) = 0.032152261$, $\Pr\left(\overline{\overline{T}} \geq 133 \,\middle|\, \mathbf{m}, \mathcal{F}\right) = 0.026721734$, so that

$$\left|\Pr\left(\overline{\overline{T}}_1 \geq 60 \,\middle|\, \mathbf{m}, \mathcal{F}\right) - \Pr\left(\overline{\overline{T}} \geq 133 \,\middle|\, \mathbf{m}, \mathcal{F}\right)\right| = 0.005430527,$$

which is the minimal value in (11).

By the calculation just performed, using the $I = 794$ case-referent pairs based on the broad case definition, but using the adaptive procedure to distinguish narrow and marginal cases, the null hypothesis H_0 of no treatment effect is rejected for all $\Gamma \leq 1.366$. In (9), it is the narrow cases that lead to rejection, but we could not be sure of this before looking at the data. By comparison with Table 2, the adaptive approach is not as insensitive as using only the narrow cases with four matched referents, but it is considerably less sensitive than the Mantel-Haenszel-McNemar statistic applied to the $I = 794$ broadly defined pairs. The adaptive procedure made better use of the $I = 794$ broadly defined pairs than did the Mantel-Haenszel procedure.

By the argument in Rosenbaum (2012a), the adaptive procedure has the larger design sensitivity of T_1 and T, so in sufficiently large samples, the adaptive procedure exhibits the same sensitivity to unmeasured biases as the better of the two case definitions. Essentially, because the upper bound on the significance level for each statistic separately is tending to zero for Γ less than the design sensitivity for that statistic, eventually it is less than $\alpha/2$ which suffices for rejection by (9).

An alternative approach, similar to that in Rosenbaum (2012a), uses $T_1 = 60$ and $T^* = 2T_1 + T_2 = 183$ as a pair of test statistics, so the narrow cases receive twice the

emphasis of the marginal cases. The exact distributions are again determined as in §2.3. The critical values at $\Gamma = 1.395$ are $\Pr\left(\overline{\overline{T}}_1 \geq 60 \text{ or } \overline{\overline{T}}^* \geq 192 \,\middle|\, \mathbf{m}, \mathcal{F}\right) = 0.04952414$, so this is slightly less sensitive than the unweighted adaptive statistic; however, this small difference reflects the discreteness of the distributions, because it is still $T_1 = 60$ for the narrow cases that leads to rejection.

An alternative design matches narrow cases to two referents and marginal cases to one referent. The 80% point of the anger scores is 12. The alternative design has 312 matched triples, $J_i = 3$, with narrow cases and 313 matched pairs, $J_i = 2$, with marginal cases having anger scores of at least 12 and less than 18, making $3 \times 312 + 313 \times 2 = 1562$ subjects in total, so this design has about the same total number of subjects as the designs considered previously. In the example, the adaptive procedure applied to this alternative design became sensitive at $\Gamma = 1.32$ barely rejecting the null hypothesis H_0 of no effect because of the 312 narrow cases, T_1.

The adaptive procedure is available in `R` in the package `SensitivityCaseControl`. A separate function in the same package implements the adaptive method in Rosenbaum (2012a).

5.2 Simulated performance of the adaptive procedure

Table 6 continues the simulation from §4, now evaluating the adaptive test. The Normal and t_3 distributions are as in §4, with a treatment effect that is half the standard deviation. In Table 6, there are either $I = 1000$ or $I = 2000$ broad cases, and $I_1 = I/2$ narrow cases. In each situation, the sensitivity analysis is performed with two values of Γ that yield powers far from 0 and 1. The lowest power in each situation is in bold.

In each of the eight columns of Table 6, the adaptive procedure is never best, never worst, whereas each of the other procedures is sometimes worst. In each case, the adaptive

procedure has power closer to the best power than to the worst. In Table 6, the best power in a column is never more than 0.1 more than the power of the adaptive procedure, and the power of the adaptive procedure is always at least 0.1 more than the power of the worst procedure. Arguably, the adaptive procedure pays only a small price to avoid a case definition that exaggerates sensitivity to unmeasured biases.

6 Discussion: summary; other methods; other applications

6.1 Summary

Under certain simple models, a narrower case definition, with fewer cases, yields a larger design sensitivity, and hence less sensitivity in large samples. Matching narrow cases each to several referents may be no more costly than pair matching with a broader case definition, yet it may be less sensitive to unmeasured biases. When in doubt about the best case definition, the adaptive approach of §5 uses both definitions with a correction for multiple testing, and it has design sensitivity associated with the better case definition.

6.2 Other methods of sensitivity analysis

A natural question is whether our findings about reduced sensitivity with narrower case definitions would also be found using other methods of sensitivity analysis besides the sensitivity analyses discussed in this paper. It is not possible to directly compare our sensitivity analysis to other sensitivity analyses that involve different assumptions and parameters for unmeasured covariates, but it is possible to compare broad and narrow case definitions using another method of sensitivity analysis.

In an interesting paper, Lin, Psaty and Kronmal (1998) proposed sensitivity analyses for various types of logistic regression and in particular they discuss matched case-referent studies. We applied their method of sensitivity analysis (specifically their (2.9)) to our

example in exact parallel to their analysis of their case-referent example in their section 4.1. That is, we applied their method twice to the data in §1.2, once to 794 pairs of a broad case definition and a referent, a second time to 312 narrow cases matched to 4 referents. Their sensitivity analysis involves three parameters: the prevalence P_1 of an unobserved binary covariate among the exposed, the prevalence P_0 of the unobserved binary covariate among those not exposed, and the odds ratio Ξ linking the unobserved covariate with the binary outcome conditional on observed covariates. To create unmeasured bias, one assumes $P_1 \neq P_0$. Their analysis also assumes the unobserved binary covariate is independent of observed covariates. (They use the symbol Γ for Ξ, but it is different from our Γ, so we have used a different symbol to avoid confusion.) We tried values of P_1 in the set $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and values of P_0 in the set $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ in all combinations with $P_1 > P_0$; then we determined the largest Ξ that failed to make the null hypothesis of no effect plausible in the sense that the confidence interval for the treatment effect odds ratio excluded 1. In each of these 24 analyses, for each of the 24 pairs $P_1 > P_0$, the results were less sensitive with the narrow case definition than with the broad definition. For instance, with $(P_1, P_0) = (0.8, 0.1)$, the narrow case definition was sensitive at $\Xi = 1.57$ and the broad definition at $\Xi = 1.08$, whereas with $(P_1, P_0) = (0.6, 0.4)$, the narrow case definition was sensitive at $\Xi = 8.48$ and the broad definition at $\Xi = 1.35$. So, although their method is formulated with different assumptions about unobserved covariates and correspondingly different sensitivity parameters, it is still true that the narrow case definition lead to less sensitivity to unmeasured biases than the broad definition in §1.2.

6.3 Other applications

Narrow case definitions have been used in genetic studies. For instance, in searching for rare genetic variants that protect against LDL cholesterol ("bad cholesterol"), Cohen et al. (2005) employed a narrow case definition of extremely low LDL cholesterol, below the 5th percentile. Cohen et al.'s reason for using this narrow case definition is that they were looking for rare gene variants that strongly perturb biology, and they expected these rare gene variants to stand out most starkly in the extreme cases.

A narrow case definition may oversample extreme responses, $\mathbf{R}_{ij}$. An alternative or complementary strategy seeks to oversample extreme doses of exposure to treatment, and this also affects design sensitivity; see Rosenbaum (2010, §17.3, 2012b) for some numerical results.

References

[1] Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Kotowski, I.K., Graham, R., Kim Garcia, C. and Hobbs, H.H. (2005), "Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9," *Nature Genetics*, 37, 161-165.

[2] 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.

[3] Cox, D. R. (1970), *Analysis of Binary Data*, London: Methuen.

[4] Diprete, T. A. and Gangl, M. (2004), "Assessing bias in the estimation of causal effects," *Sociological Methodology*, 34, 271-310.

[5] Egleston, B. L., Scharfstein, D. O., MacKenzie, E. (2009), "On estimation of the

survivor average causal effect in observational studies when important confounders are missing due to death," *Biometrics*, 65, 497-504.

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

[7] Holland, P. W. H. and Rubin, D. B. (1988), "Causal inference in retrospective studies," *Evaluation Review*, 12, 203-231.

[8] 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.

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

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

[11] Lin, D. Y., Psaty, B. M., and Kronmal, R. A. (1998), "Assessing the sensitivity of regression results to unmeasured confounders in observational studies," *Biometrics*, 54, 948-963.

[12] Mantel, N., (1973), "Synthetic retrospective studies and related topics," *Biometrics*, 29, 479-486.

[13] 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.

[14] McCandless, L. C., Gustafson, P. and Levy, A. (2007), "Bayesian sensitivity analysis for unmeasured confounding in observational studies," *Statistics in Medicine*, 26,

2331-2347.

[15] Mukherjee, B., Liu, I., and Sinha, S. (2007), "Analysis of matched case-control data with multiple ordered disease states," *Statistics in Medicine*, 26, 3240-3257.

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

[17] R Core Team (2012), "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org/.

[18] Rosenbaum, P. R. (1991), "Sensitivity analysis for matched case-control studies," *Biometrics*, 47, 87-100.

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

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

[21] Rosenbaum, P. R. and Silber, J. H. (2008), "Aberrant effects of treatment," *Journal of American Statistical Association*, 103, 240-247.

[22] Rosenbaum, P. R. and Silber, J. H. (2009), "Amplification of sensitivity analysis in observational studies," *Journal of American Statistical Association*, 104, 1398-1405.

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

[24] Rosenbaum, P. R. (2012a), "An exact, adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer," *Annals of Applied Statistics*, 6, 83-105.

[25] Rosenbaum, P. R. (2012b), "Nonreactive and purely reactive doses in observational studies," in C. Berzuini, A. P. Dawid, and L. Bernardinelli, eds., *Causality: Statistical Perspectives and Applications*, New York: John Wiley, pp. §19, 272-289.

[26] Rosenbaum, P. R. (2013), "Impact of multiple matched controls on design sensitivity in observational studies," *Biometrics*, 69, 118-127.

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

[28] Small, D. and Rosenbaum, P. R. (2008), "War and wages: the strength of instrumental variables and their sensitivity to unobserved biases," *Journal of the American Statistical Association*, 103, 924-933.

[29] Spielberger, C. D. (1996), *State-Trait Anger Expression Inventory Professional Manual*. Odessa, FL: Psychological Assessment Resources.

[30] Springer, K. W., Sheridan, J., Kuo, D., and Carnes, M. (2007), "Long term phsyical and mental health consequences of childhood physical abuse," *Child Abuse and Neglect*, 31, 517-530.

[31] Ury, H. (1975), "Efficiency of case-control studies with multiple controls pre case," *Biometrics*, 31, 643-649.

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

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

Table 1: Frequency of exposure to childhood abuse in two case-referent studies, one with 794 pairs, the other with 312 matched sets containing one case and four referents. The Mantel-Haenszel estimate of a common odds ratio is 1.46 for the 794 broadly defined matched pairs and it is 2.06 for the 312 narrowly defined $1-4$ matched sets.

Broad Case Definition			Narrow Case Definition					
$I = 794$ Pairs, $J = 2$			$I = 312$ matched sets, $J = 5$					
$I \times J = 1588$ People			$I \times J = 1560$ People					
	Referent			# Referents Exposed				
Case	Not Exposed	Exposed	Case	0	1	2	3	4
Not exposed	602	69	Not Exposed	174	60	14	4	0
Exposed	101	22	Exposed	34	19	6	1	0

Table 2: Sensitivity analysis with two case definitions (broad or narrow) and two test statistics (Mantel-Haenszel or Aberrant Rank). The tabled values are sharp upper bounds on the one-sided P-value for biases Γ of various magnitudes. In each column, the largest P-value less than or equal to 0.05 is in bold.

	Broad Case Definition		Narrow Case Definition	
Γ	Mantel-Haenszel	Aberrant Rank	Mantel-Haenszel	Aberrant Rank
1	0.00706	0.00128	0.00001	0.00006
1.1	**0.03322**	0.00666	0.00014	0.00040
1.2	0.1013	**0.0235**	0.0008	0.0018
1.3	0.2236	0.0617	0.0037	0.0060
1.4	0.388	0.129	0.012	0.016
1.5	0.562	0.227	**0.031**	**0.036**
1.6	0.716	0.347	0.066	0.069

Table 3: Design sensitivity $\widetilde{\Gamma}$ with a binary covariate, $x = 0$ or $x = 1$. At x, the probability of exposure to the treatment is η_{cx} for cases and η_{rx} for referents, so the odds ratio is $\Omega_x = \eta_{cx}(1-\eta_{rx})/\{\eta_{rx}(1-\eta_{cx})\}$. The proportion of cases with $x = 0$ is λ. The marginal odds ratio Ω_M is for the 2×2 table collapsed over x. In situations A-C, $\Omega_M < \Omega_0 = \Omega_1 = \widetilde{\Gamma}$ for $J = 2$, 5, 10. In situations D-E, $\Omega_0 = 1$ and $\Omega_1 > \Omega_M = \widetilde{\Gamma}$ for $J = 2$, 5, 10. In case F, $\Omega_0 < \Omega_M = \widetilde{\Gamma} < \Omega_1$ for $J = 2$, 5, 10. In cases G-K, the design sensitivity $\widetilde{\Gamma}$ is between Ω_0 and Ω_1, but $\widetilde{\Gamma}$ varies with J, sometimes increasing, sometimes decreasing.

Situation	Covariate $x=0$			Covariate $x=1$					Design Sensitivity $\widetilde{\Gamma}$		
	η_{r0}	η_{c0}	Ω_0	η_{r1}	η_{c1}	Ω_1	λ	Ω_M	$J=2$	$J=5$	$J=10$
A	.4	.7	3.50	.4	.7	3.50	.5	3.50	3.50	3.50	3.50
B	.1	.28	3.50	.72	.9	3.50	.5	2.07	3.50	3.50	3.50
C	.1	.28	3.50	.72	.9	3.50	.8	2.35	3.50	3.50	3.50
D	.3	.3	1	.3	.9	21	.5	3.50	3.50	3.50	3.50
E	.3	.3	1	.3	.9	21	.8	1.69	1.69	1.69	1.69
F	.4	.6	2.25	.4	.8	7	.5	3.50	3.50	3.50	3.50
G	.4	.7	3.50	.5	.9	9	.5	4.89	5.12	5.02	4.98
H	.1	.7	21	.8	.9	2.25	.5	4.89	7.36	10.73	12.66
I	.1	.7	21	.8	.9	2.25	.8	9.01	13.50	16.88	18.11
J	.1	.2	2.25	.8	.99	24.75	.5	1.80	4.30	3.54	3.35
K	.1	.2	2.25	.8	.99	24.75	.8	1.77	2.80	2.62	2.58

Table 4: Frequency of exposure to childhood abuse in a case-referent study with $794 = 312 + 482$ pairs, with the pairs separated by whether the case is a narrow case (anger score $A \geq 18$) or a marginal case ($10 \leq A < 18$).

Narrow Case, $A \geq 18$			Marginal Case, $10 \leq A < 18$		
$I = 312$ Pairs			$I = 482$ matched pairs		
	Referent			Referent	
Case	Not Exposed	Exposed	Case	Not Exposed	Exposed
Not exposed	229	23	Not Exposed	373	46
Exposed	51	9	Exposed	50	13

Table 5: Simulated power of a sensitivity analysis in three case-referent designs, with 2500 broad cases of whom 1250 are also narrow cases. One design uses 2500 pairs of a broad case and a referent. One design uses 1250 pairs of a narrow cases and a referent. One design uses 1250 narrow cases each matched to 3 referents. When the sensitivity parameter Γ is less than the design sensitivity $\widetilde{\Gamma}$, the power is tending to 1 with increasing sample size, but when $\Gamma > \widetilde{\Gamma}$ the power is tending to zero.

Test Statistic	Mantel-Haenszel			Aberrant		
Case definition	Broad	Narrow	Narrow	Broad	Narrow	Narrow
Design	Pairs	Pairs	1-to-3	Pairs	Pairs	1-to-3
Cases	2500	1250	1250	2500	1250	1250
Referents	2500	1250	3750	2500	1250	3750
Sample size	5000	2500	5000	5000	2500	5000
			Standard Normal			
Design Sensitivity	$\widetilde{\Gamma} = 2.4$	$\widetilde{\Gamma} = 2.8$	$\widetilde{\Gamma} = 2.8$	$\widetilde{\Gamma} = 2.7$	$\widetilde{\Gamma} = 3.2$	$\widetilde{\Gamma} = 3.2$
$\Gamma = 2.00$	.898	.993	1	.991	.999	1
$\Gamma = 2.25$	.228	.845	.967	.827	.964	.999
$\Gamma = 2.50$	.007	.450	.621	.280	.772	.922
$\Gamma = 2.75$	0	.110	.143	.030	.406	.595
$\Gamma = 3.00$	0	.022	.009	0	.134	.199
$\Gamma = 3.25$	0	.001	0	0	.033	.041
			Standardized t_5			
Design Sensitivity	$\widetilde{\Gamma} = 2.6$	$\widetilde{\Gamma} = 2.8$	$\widetilde{\Gamma} = 2.8$	$\widetilde{\Gamma} = 2.7$	$\widetilde{\Gamma} = 2.8$	$\widetilde{\Gamma} = 2.8$
$\Gamma = 2.00$	.998	.992	1	.998	.966	1
$\Gamma = 2.25$	.779	.827	.956	.863	.708	.891
$\Gamma = 2.50$	.176	.382	.533	.330	.316	.443
$\Gamma = 2.75$	.006	.099	.102	.040	.076	.081
$\Gamma = 3.00$	0	.014	.004	.004	.015	.009
$\Gamma = 3.25$	0	.003	0	0	.002	0
			Standardized t_3			
Design Sensitivity	$\widetilde{\Gamma} = 3.2$	$\widetilde{\Gamma} = 3.1$	$\widetilde{\Gamma} = 3.1$	$\widetilde{\Gamma} = 3.1$	$\widetilde{\Gamma} = 3.1$	$\widetilde{\Gamma} = 3.1$
$\Gamma = 2.00$	1	1	1	1	.984	1
$\Gamma = 2.25$	1	.986	.999	.998	.815	.959
$\Gamma = 2.50$	.983	.806	.954	.923	.450	.626
$\Gamma = 2.75$	.718	.420	.609	.536	.144	.202
$\Gamma = 3.00$	.210	.109	.159	.133	.031	.023
$\Gamma = 3.25$	.019	.015	.015	.009	.002	.001

Table 6: Simulated power of a sensitivity analysis in a paired case-referent study with I broad cases and $I_1 = I/2$ narrow cases. The table contrasts the power of three methods: (i) the Mantel-Haenszel (MH) or McNemar test using all I case-referent pairs, (ii) the MH test using the narrow $I_1 = I/2$ case-referent pairs, and (iii) the adaptive method that adapts between (i) and (ii). Within each column, the lowest power is in **bold**.

Distribution	Normal				t_3			
Number of Cases I	1000 cases		2000 cases		1000 cases		2000 cases	
Γ	2	2.25	2	2.25	2.5	2.75	2.5	2.75
MH, all cases	**.540**	**.129**	**.807**	**.214**	.771	.412	.962	.639
MH using $I_1 = I/2$ cases	.826	.492	.980	.781	**.482**	**.236**	**.733**	**.372**
Adaptive	.775	.407	.965	.697	.706	.360	.930	.564

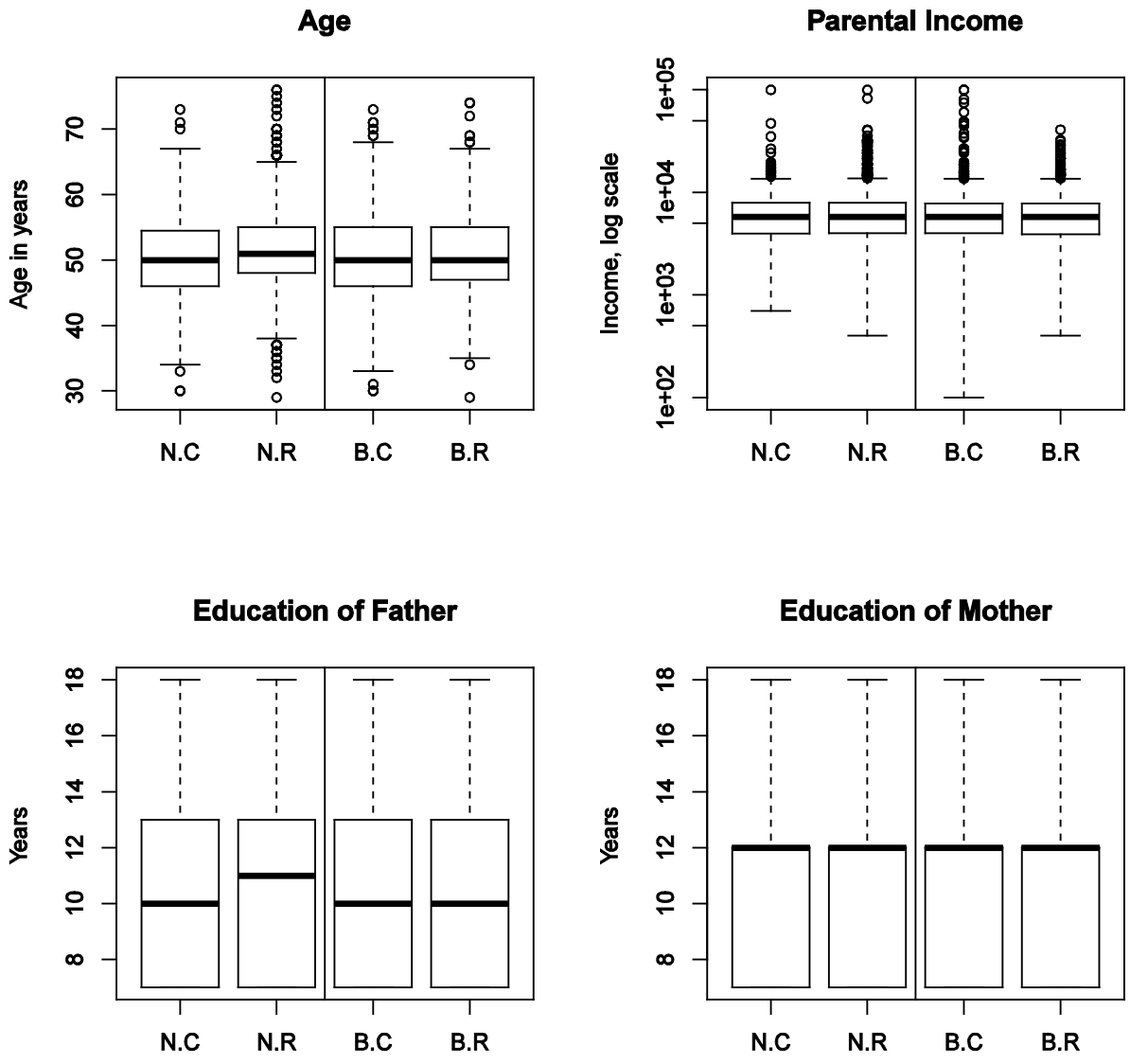

Figure 1: Covariate balance after matching for age, parental income, and years of education of father and mother. The figure compares cases (C) and referents (R) using either the narrow (N) or the broad (B) definition of a case. Narrow cases had anger scores at or above the 90% point of 18, broad cases had anger scores at or above the 75% point of 10, and referents had anger scores below 10.

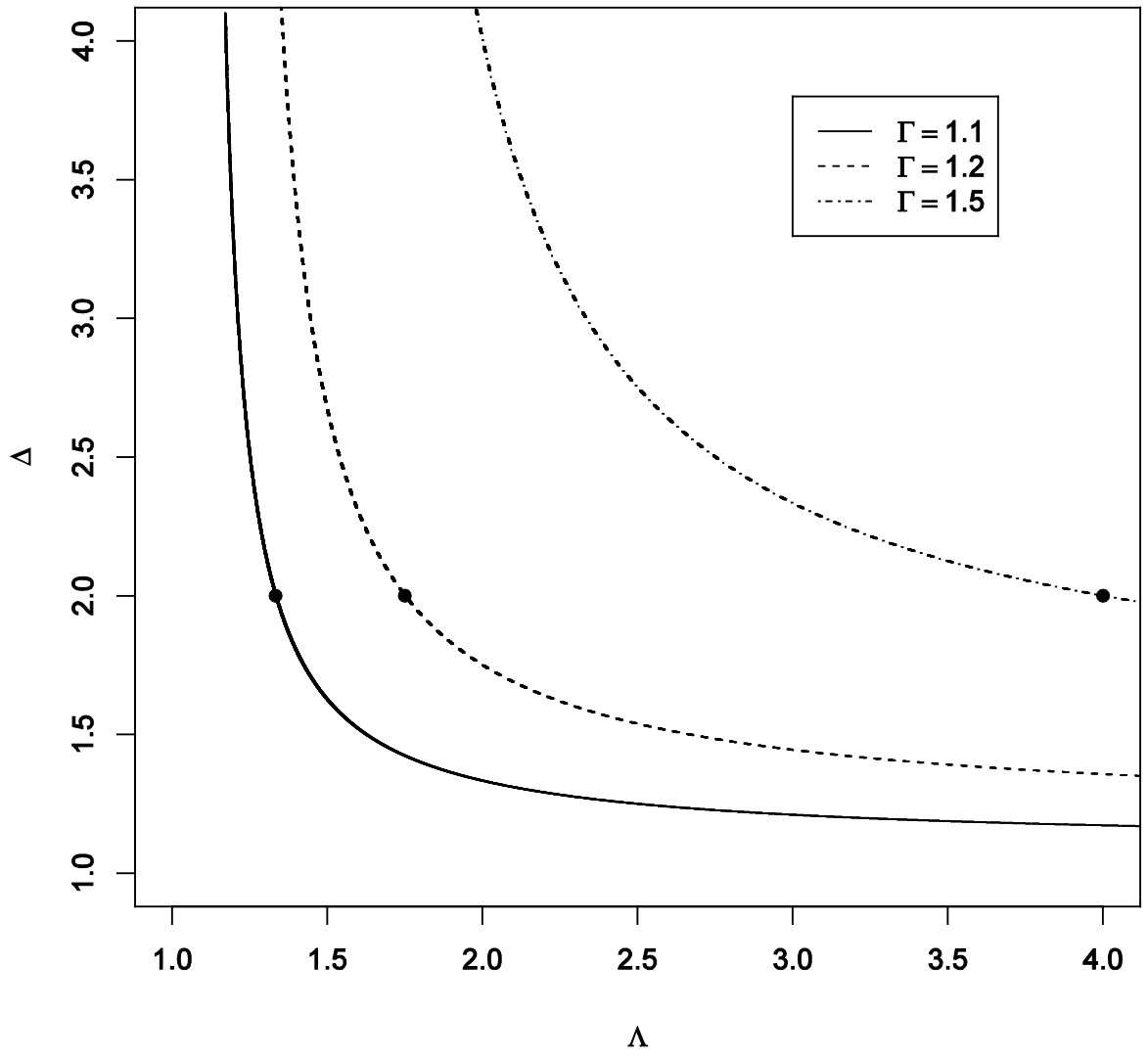

Figure 2: Amplification of Γ into two parameters, where Λ controls the association between the unobserved covariate *u* and treatment assignment *Z*, and Δ controls the association between u and r_C. The curves are $\Gamma = (\Lambda\Delta+1)/(\Lambda + \Delta)$ and the dots are the values quoted in the text for (Λ,Δ) = (1.333, 2), (1.75, 2) and (4, 2).