# Attributing Effects to Treatment in Matched Observational Studies

Paul R. ROSENBAUM

An effect is attributable to treatment if it would not have been observed had the individual been exposed to control instead. Extending earlier results on attributable effects in unmatched groups, a method of exact randomization inference and sensitivity analysis is developed for case-referent, case-crossover, and cohort studies with matched sets, and a large sample approximation to the exact inference is given. The unmatched case, considered previously, has certain symmetries that the matched case, considered here, does not have. As a result, approximation for the matched case requires the use of the recently developed method of asymptotic separability, which was not needed in the unmatched case. Several examples are presented, including a case-referent study of *Helicobacter pylori* infection as a cause of myocardial infarction, a case-crossover study of alcohol as a cause of injury, a cohort study of women who gave birth at home, and a study of the effects of cadmium exposure with a continuous outcome measuring kidney function. Unlike tests of no effect, inference about attributable effects has a different form in case-referent and cohort studies.

KEY WORDS: Asymptotic separability; Attributable effect; Displacement effect; Randomization inference; Sensitivity analysis.

## 1. ATTRIBUTABLE EFFECTS IN EXPERIMENTS AND OBSERVATIONAL STUDIES

### 1.1 How Does Randomization Affect Inference?

In his 1935 book *Design of Experiments*, Fisher carefully argued that the random assignment of treatments in experiments justifies certain inferences about the effects caused by those treatments—that randomization forms the "reasoned basis for inference" in experiments—and that these same inferences would not be justified by identical data obtained in a nonrandomized study. In Fisher's view, causal inference depends partly on the observed data, but also partly on how the data were obtained. This view has two desirable consequences. First, it serves to encourage randomized experimentation when randomization is ethical and feasible. Second, it forces analyses of nonrandomized or observational studies of treatment effects to explicitly acknowledge, as part of the quantitative findings, greater uncertainty about causal effects than would be present had random assignment been used.

A limitation is that many randomization tests of the hypothesis of no treatment effect are not paired with confidence intervals. If the treatment has an additive effect, $\tau$, then a randomization test of no effect can be inverted to yield confidence intervals and point estimates for $\tau$ (see, e.g., Hodges and Lehmann 1963; Moses 1965; Lehmann 1963, 1975; Rosenbaum 1995a, sec. 2). The model of an additive effect is useful in many settings but is inapplicable in many others—for instance, for binary responses, where inferences typically invoke distributional assumptions not derived from random assignment. This article extends a line of reasoning begun in earlier work (Rosenbaum 2001), in which attributable effects are used to substantially expand the collection of randomization tests that may be inverted to yield confidence intervals. The approach also yields sensitivity analyses that measure the added uncertainty present when treatments are not randomly assigned.

An attributable effect describes how treated subjects would have responded had they been spared exposure to the treatment. As a consequence, the attributable effect depends in part on the identity of the subjects exposed to treatment, and so different randomizations typically produce different attributable effects. In this sense, the attributable effect is not a parameter, but rather is an unobserved random variable. Nonetheless, inference proceeds along a relatively conventional path. Before embarking on that path, I present a useful illustration.

### 1.2 Example: Infection as a Possible Cause of Early Onset Myocardial Infarction

The example in this section is from an observational study and is intended to provide the simplest illustration of the general method developed in the current article. With this goal in mind, the analysis presumes that matching has been effective in removing bias due to nonrandom assignment, and the discussion is entirely informal. After notation is developed in Section 2 and methods in Section 3, this example and several others are discussed more formally, and some analyses done to account for possible hidden biases due to nonrandom assignment.

Motivated by earlier findings, Danesh et al. (1999) conducted a case-referent study of *Helicobacter pylori* infection as a possible cause of myocardial infarction (MI) in relatively young people, age 30–49. This thoughtful study is of interest not only for its medical conclusions, but also for its methodology. It is an epidemiologic study appended to a randomized clinical trial, in which blood samples and questionnaires were obtained from cases who participated in the trial and relatives who did not. Two matched comparisons were reported, one of the cases to their own siblings and the other to the relatives of other cases, matching for age and gender. The authors write (p. 1157):

> We report two complementary studies, one comparing young patients with acute myocardial infarction with young controls (which should maximize the strength of any association) and one comparing people with myocardial infarction at any age with an unaffected sibling (which should minimize any artifactual association due to confounding factors).

Paul R. Rosenbaum is Robert G. Putzel Professor, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *rosenbaum@stat.wharton.upenn.edu*). The author acknowledges hospitality and support of the Center for Advanced Study in the Behavioral Sciences. This work was supported by a grant from the Methodology, Measurement and Statistics Program and the Statistics and Probability Program of the National Science Foundation.

The sibling comparison was performed and reported as matched pairs, thus whereas the other analysis was unmatched. I use the sibling data to illustrate methods in this article. However, the two comparisons differ in interesting ways, and the reader should refer to Danesh et al. (1999) for the full story. In particular, the sibling analysis exhibited a smaller odds ratio and marginal significance, whereas the other comparison found a stronger and more significant association.

The matched sibling analysis is given in Table 1, which counts pairs in the manner of McNemar's test. For example, in 173 pairs, both the case and the sibling were seropositive for *H. pylori*.

The conditional maximum likelihood estimate of an odds ratio that is constant over pairs is $91/67 = 1.36$, and a one-sided deviate for McNemar's test of $(91-67)/\sqrt{91+67} = 1.91$ (MacMahon and Trichopoulos 1996, p. 288), and so would be judged significant in a one-sided .05-level test, where the critical value from the normal distribution would be 1.65. This reference distribution for McNemar's test may be derived as the normal approximation to the randomization distribution, so the reference distribution implicitly assumes that the hidden biases are absent.

Although formal justification of the analysis requires some attention to detail, an inference about attributable effects may be presented informally in an intuitive manner. Suppose that of the 91 MI cases, 4 were caused by *H. pylori* infection; that is, $A = 4$ cases of MI are attributable to infection. If this supposition were true, then without those 4 cases, the remaining 87 cases would satisfy the null hypothesis of no effect. Moreover, McNemar's test becomes $(87-67)/\sqrt{87+67} = 1.61$, so with 87 exposed cases, the null hypothesis is barely plausible. In contrast, with $3 = A$, McNemar's test is $(88-67)/\sqrt{88+67} = 1.69$, and the null hypothesis is barely rejected at the one-sided .05 level.

Although intuitive, the analysis just presented raises several questions. Is there some formal structure under which this analysis is justified? Is this a randomization inference? The analysis presumed that the MIs attributable to infection were all to be found among the discordant pairs, so that removing them converted discordant pairs into concordant pairs, reducing the effective sample size. Might some MIs attributable to infection be found among concordant pairs? If so, removing them creates discordant pairs from concordant pairs and in turn increases the effective sample size. Why does the analysis presume that the attributable MIs are among the discordant pairs? Can the analysis be extended to sensitivity analyses for hidden biases, cohort studies, matching with multiple controls, and continuous responses? If so, what changes are needed for these extensions? The general discussion in Sections 2 and 3 addresses these issues, and later sections provide practical illustrations.

*Table 1. Helicobacter Pylori Seropositivity for MI Cases and Siblings*

| | *Sibling+* | *Sibling−* |
|---|---|---|
| Case+ | 173 | 91 |
| Case− | 67 | 179 |

One of these issues merits motivating discussion. The issue of attributable MIs among concordant and discordant pairs is an aspect of the difference between the matched case, discussed in this article, and the unmatched case, discussed in earlier work (Rosenbaum 2001). In the unmatched case, all treated subjects who exhibit a response are indistinguishable, so the number of attributable responses matters but the identity of the respondents does not. This is not true in the matched case. Even with matched pairs, concordant pairs differ from discordant pairs. Moreover, in matching with multiple controls, there are several types of discordant matched sets. In saying, as before, that it is not plausible that three or fewer of the MIs are attributable to infection but that four or more is plausible, one is saying that *no* pattern yielding three attributable effects is plausible, but *some* pattern yielding four or more is plausible. In paired case-referent studies, the pattern of three attributable effects that is most difficult to reject is the pattern that assumes that all three attributable effects are among the discordant pairs, thereby reducing the effective sample size. If that most difficult pattern is rejected, then so too are all other patterns, and three MIs being attributable to infection is not plausible. Somewhat surprisingly, the situation in paired cohort studies is entirely different; see Section 6. These issues are clarified in Sections 2 and 3.

## 2. NOTATION AND REVIEW: MATCHED EXPERIMENTS AND OBSERVATIONAL STUDIES

### 2.1 Notation: Treatment Assignments, Treatment Effects, and Attributable Effects

The $i$th of $I$ matched sets, $i = 1, \ldots, I$, contains $n_i \geq 2$ subjects, numbered $j = 1, \ldots, n_i$, with $N = \sum n_i$ subjects in total. Subjects were matched on observed pretreatment measures or covariates, such as age. In Section 1.2 there are $I = 510$ pairs of two subjects, $n_i = 2$ for $i = 1, \ldots, 510$ and $N = 2 \times 510 = 1{,}020$. The subscript $j$ is assigned at random to subjects and therefore conveys no information about them. Write $Z_{ij} = 1$ if the $j$th subject in set $i$ was exposed to the treatment, $Z_{ij} = 0$ if this subject was not exposed, and write $Z_{i+} = \sum_{j=1}^{n_i} Z_{ij}$ for the number of exposed subjects in set $i$. In Section 1.2, $Z_{i+} = 2$ for 173 concordant pairs, $Z_{i+} = 1$ for $158 = 67 + 91$ discordant pairs, and $Z_{i+} = 0$ for 179 concordant pairs. For any integer $q$ with $n \geq q \geq 0$, write $\mathcal{B}(n, q)$ for the set containing the $\binom{n}{q}$ vectors of dimension $n$ with $q$ coordinates equal to 1 and $n - q$ coordinates equal to 0, so that, given $Z_{i+}$, it is known that $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{i,n_i})^T \in \mathcal{B}(n_i, Z_{i+})$. In Section 1.2 there are three relevant sets, namely $\mathcal{B}(2,2) = \{(1,1)^T\}$, $\mathcal{B}(2,1) = \{(1,0)^T, (0,1)^T\}$, and $\mathcal{B}(2,0) = \{(0,0)^T\}$. In a matched, randomized experiment, one treatment assignment, $\mathbf{z}_i \in \mathcal{B}(n_i, Z_{i+})$, would be picked at random, each having probability $\binom{n}{Z_{i+}}^{-1}$, with independent assignments in distinct matched sets. In an observational study, treatments are not randomly assigned, and $\Pr(\mathbf{Z}_i = \mathbf{z}_i)$ may not equal $\binom{n}{Z_{i+}}^{-1}$ for $\mathbf{z}_i \in \mathcal{B}(n_i, Z_{i+})$ and $\Pr(\mathbf{Z}_i = \mathbf{z}_i)$ may be unknown. In particular, in observational studies, there is often the concern that subjects matched for observed covariates may differ in terms of an unobserved covariate, say $u_{ij}$, related to treatment assignment. Write $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_I)$ for the $N$-dimensional vector recording all of the treatment assignments.

Each subject has two potential binary responses, $r_{Tij}$ and $r_{Cij}$, where $r_{Tij}$ would be observed if the subject were exposed to treatment and $r_{Cij}$ would be observed if not (Neyman 1923; Rubin 1974). Here a 1 response signifies that a particular event occurred, whereas a 0 response indicates that it did not. Following Hamilton (1979), it is assumed that the exposure to treatment may cause the event in a person who would not otherwise experience it, but the treatment does not prevent the event in a person who would otherwise experience it; that is, it is assumed that $r_{Tij} \geq r_{Cij}$. Write $\boldsymbol{\delta}$ for the $N$-dimensional vector of treatment effects, $\delta_{ij} = r_{Tij} - r_{Cij}$, in the lexical order. The observed response is $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$. Write $R_{i+} = \sum_{j=1}^{n_i} R_{ij}$ for the number of subjects who had events in set $i$ and $r_{Ci+} = \sum_{i=1}^{n_i} r_{Cij}$ for the number who would have had events if exposure to the treatment had been prevented. In a paired case-referent study, as in Section 1.2, each pair contains one case with $R_{ij} = 1$ and one referent with $R_{ij} = 0$, so $R_{i+} = 1$ for every pair $i$.

The attributable effect is $A = \sum_{i,j} Z_{ij}(r_{Tij} - r_{Cij}) = \sum_{i,j} Z_{ij}\delta_{ij}$. This is the number of treated subjects who experienced events caused by the treatment, that is, events that would not have occurred if the treated subjects had not been exposed to treatment. For instance, in Section 1.2, the possibility that $A = 3$ was judged implausible, but the possibility that $A = 4$ was judged barely plausible. Notice that the attributable effect is not a parameter, because assigning treatments, $Z_{ij}$, differently would change the value of $A$. Rather, $A$ is a random variable whose value is not observed, because $r_{Tij}$ and $r_{Cij}$ are never observed jointly on the same person. The data do provide some information about $A$. For instance, in a randomized experiment, if the null hypothesis of no treatment effect, $H_0 : r_{Tij} = r_{Cij}$ for all $i, j$, is rejected by a randomization test (Fisher 1935), then it is not plausible that $A = 0$. Write $T = \sum_{i,j} Z_{ij}R_{ij}$ for the number of treated individuals with events, so that $T - A = \sum_{i,j} Z_{ij}r_{Cij}$ is the number of treated subjects who would have exhibited events even in the absence of treatment. In Section 1.2, $T = 173 + 91 = 264$ of the 510 cases of MI were exposed to treatment (i.e., 264 were seropositive), but of course we cannot observe the number, $A$, of MIs that would not have occurred without exposure to treatment.

In Fisher's (1935) randomization inference, quantities that depend on the random assignment of treatments, $Z_{ij}$, were random variables, but other quantities were fixed features of the finite population of $N$ subjects. In this way, in a randomized experiment, random quantities have distributions created by randomization, so randomization forms the basis for inference. The potential responses, $r_{Tij}$ and $r_{Cij}$, and the treatment effect, $\delta_{ij} = r_{Tij} - r_{Cij}$, are fixed, but the observed response $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$ and the attributable effect $A = \sum_{i,j} Z_{ij}\delta_{ij}$ are random variables.

This article discusses the two common situations with a single notation and a single set of formulas. The first situation is a cohort study in which one treated subject is matched to $n_i - 1$ untreated controls, so that $Z_{i+} = 1$ for $i = 1, \ldots, I$. The second situation is a case-referent study in which one case with an event is matched to $n_i - 1$ noncases or referents without the event, so that $R_{i+} = 1$ for $i = 1, \ldots, I$. One of these two situations is assumed to hold throughout. These two situations differ in many important ways, but the differences do not matter for the technical details of the current argument, so it is convenient and compact to discuss both at once. In both situations, the number of exposed subjects with events, $\sum_{j=1}^{n_i} Z_{ij}R_{ij}$, is either 1 or 0 in each matched set $i$.

Attributable effects are related to measures of attributable risk, as discussed by, for example, Hamilton (1979), MacMahon and Trichopoulos (1996), Walter (1976), and Robins (1988). The continuous analogs of attributable effects, namely the displacement effects discussed in Section 6, are related to a variety of methods, including the control median test of Gart (1963) and Gastwirth (1968), the methods based on placements of Fligner and Wolfe (1976) and Orban and Wolfe (1982), and the quantile comparison function of Li, Tiwari, and Wells (1996). The attributable effect $A/\sum Z_{i+}$ is the average effect actually experienced by the $\sum Z_{i+}$ treated subjects in the current study. If one imagined that the $N$ subjects were randomly sampled from an infinite population, then the expectation of $A/\sum Z_{i+}$ averaged over the imagined random sampling is a population parameter that has been used for various purposes (see, e.g., Rosenbaum and Rubin 1985; Heckman 1997).

## 2.2 Model for Treatment Assignment

The model for treatment assignment in this section forms the basis for randomization inference in randomized experiments and for studying the sensitivity of causal inferences to hidden biases in nonrandomized experiments. In a randomized experiment, the model says simply that treatments were randomly assigned, $\Pr(\mathbf{Z}_i = \mathbf{z}_i) = \binom{n}{Z_{i+}}^{-1}$ for $\mathbf{z}_i \in \mathcal{B}(n_i, Z_{i+})$, independently in distinct matched sets. For instance, in a randomized paired experiment, every pair contains a treated subject and a control, $n_i = 2, Z_{i+} = 1$, there are two possible treatment assignments in each pair, $\mathcal{B}(2, 1) = \{(1, 0)^T, (0, 1)^T\}$, and each assignment has probability $\Pr(\mathbf{Z}_i = \mathbf{z}_i) = \binom{2}{1}^{-1} = 1/2$, yielding the randomization distribution for McNemar's test. In an observational study, the model for treatment assignment says two things: first, matching on observed covariates may have made subjects somewhat similar in their chances of exposure to treatment, but second, unlike a randomized experiment, exposure to treatment may not be completely random within matched sets, because the matching may have failed to control for an important unobserved covariate, $u_{ij}$. The formal model for treatment assignment that follows may be derived from three assumptions: (1) that in the population before matching on covariates, treatments are assigned independently with probabilities that may vary from person to person and may be unknown; (2) that the matching of treated subjects to controls or cases to referents was based solely on observed covariates; and (3) that subjects in the same matched set may differ in their odds of receiving treatment by at most a factor of $\Gamma \geq 1$. Here $\Gamma$ is an unknown sensitivity parameter whose value is varied to display the sensitivity of the inference to departures from random assignment. The derivation of the model from these assumptions has been given in earlier work (Rosenbaum 1995a, sec. 4.2; 1995b, sec. 1.2) and is not repeated here. The formal model for treatment assignment involves an unobserved covariate $u_{ij}$ with $1 \geq u_{ij} \geq 0$, perhaps an unobserved binary attribute, which was not controlled by matching. Write

$\gamma = \log(\Gamma)$ and $\mathbf{u}_i = (u_{i1}, \ldots, u_{i,n_i})^T$. For $\mathbf{z}_i \in \mathcal{B}(n_i, z_{i+})$, the model asserts that

$$\Pr(\mathbf{Z}_1 = \mathbf{z}_1, \ldots, \mathbf{Z}_I = \mathbf{z}_I | Z_{1+} = z_{1+}, \ldots, Z_{I+} = z_{I+}) = \prod_{i=1}^{I} \frac{\exp(\gamma \mathbf{z}_i^T \mathbf{u}_i)}{\sum_{\mathbf{b}_i \in \mathcal{B}(n_i, z_{i+})} \exp(\gamma \mathbf{b}_i^T \mathbf{u}_i)}. \quad (1)$$

The case $\Gamma = 1$ or $\gamma = 0$ yields random assignment of treatments. Specifically, when $\Gamma = 1$ so that $\gamma = 0$, the unobserved covariate does not matter, and $z_{i+}$ of the $n_i$ individuals in set $i$ receive the treatment at random with equal probabilities, $\Pr(\mathbf{Z}_i = \mathbf{z}_i) = \binom{n}{z_{i+}}^{-1}$, as in a randomized experiment. For instance, in a paired randomized experiment, $n_i = 2, Z_{i+} = 1$, expression (1) is simply $1/2^I$ for each of the $2^I$ possible treatment assignments in the $I$ treated/control pairs. Other models for sensitivity analysis in observational studies have been discussed by Cornfield et al. (1959), Greenhouse (1982), Rosenbaum and Rubin (1983), Rosenbaum (1986), Manski (1990), Copas and Li (1997), Lin, Psaty, and Kronmal (1998), Gastwirth, Krieger, and Rosenbaum (1998), Robins, Rotnitzky, and Scharfstein (1999), Wasserman (1999), and Robins, Greenland, and Hu (1999).

Because $r_{Tij} \geq r_{Cij}$, a treated subject without an event would not have an event under control, and a control subject with an event would also have an event under treatment. The hypothesis $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is *compatible* with the data if $\delta_{ij} = 0$ whenever either ($Z_{ij} = 1$ and $R_{ij} = 0$) or ($Z_{ij} = 0$ and $R_{ij} = 1$); otherwise, the hypothesis is *incompatible* (see Rosenbaum 2001 for use of compatible hypotheses). An incompatible hypothesis can be rejected with certainty, that is, with type 1 error rate of 0. If the hypothesis is true, then it is compatible for every $\mathbf{Z}$, whereas a false hypothesis may be compatible for some $\mathbf{Z}$ and not for others.

If the null hypothesis $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ were true, then $r_{Cij}$ would be known for all $i, j$, because $r_{Cij}$ would equal the observed quantity, $R_{ij} - Z_{ij}\delta_{0ij}$. Write $A_0 = \sum_{i,j} Z_{ij}\delta_{0ij}$, so that under the null hypothesis, $T - A_0 = \sum_{i,j} Z_{ij} r_{Cij}$. Now $\sum_{i,j} Z_{ij} r_{Cij}$ is the sum of $I$ independent binary random variables, where $B_i = \sum_j Z_{ij} r_{Cij}$. Write $\pi_i = \Pr(B_i = 1)$, which is unknown in observational studies because $u_{ij}$ is unknown. Keeping in mind that either $Z_{i+} = 1$ for $i = 1, \ldots, I$ for a cohort study or $R_{i+} = 1$ for $i = 1, \ldots, I$ for a case-referent study, one can show that $\pi_i$ is bounded above by a quantity $\bar{\pi}_i$, where

$$\bar{\pi}_i = \frac{\Gamma Z_{i+} r_{Ci+}}{\Gamma Z_{i+} r_{Ci+} + n_i - Z_{i+} r_{Ci+}} \geq \Pr(B_i = 1) \geq \frac{Z_{i+} r_{Ci+}}{Z_{i+} r_{Ci+} + \Gamma(n_i - Z_{i+} r_{Ci+})}, \quad (2)$$

and, in particular, in a randomized experiment, $\Gamma = 1$, so that

$$\Pr(B_i = 1) = \pi_i = \bar{\pi}_i = \frac{Z_{i+} r_{Ci+}}{n_i}. \quad (3)$$

(See Rosenbaum 1988 for proof in a cohort study and Rosenbaum 1991 for the quite different proof in a case-referent study.) For example, in Section 1.2, if infection never causes MI, so $r_{Tij} = r_{Cij}$ for every $i, j$, then the $67 + 91$ discordant case-referent pairs all have $n_i = 2, Z_{i+} = 1$, and $r_{Ci+} = 1$, and $\bar{\pi}_i = \Gamma/(\Gamma + 1)$; in particular, if there is no hidden bias, so $\Gamma = 1$, then $\bar{\pi}_i = 1/2$ as in a randomized experiment.

It is important to note that when the hypothesis $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is true, $\bar{\pi}_i$ can be calculated from the observed data and the hypothesis. In Section 3.1, this means that it is straightforward to test any one hypothesis, $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$, because the needed quantities can be computed by combining the information given by the data with the information given by the hypothesis.

## 3. INFERENCE

### 3.1 Testing a Specific Hypothesis About $\boldsymbol{\delta}$

Consider testing the null hypothesis $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ against the alternative $H_1$: $\boldsymbol{\delta} \geq \boldsymbol{\delta}_0$, $\boldsymbol{\delta} \neq \boldsymbol{\delta}_0$. If the null hypothesis is incompatible, then reject it immediately with type 1 error rate of 0. Otherwise, if the null hypothesis were true, then it would follow that $r_{Cij} = R_{ij} - Z_{ij}\delta_{0ij}$, and $T - A_0 = \sum_{i,j} Z_{ij} r_{Cij} = \sum_i B_i$ would be distributed as the sum of $I$ independent binary trials with $\Pr(B_i = 1) = \pi_i$. Write $\beta(k, \boldsymbol{\pi})$ for the probability of at least $k$ successes in $I$ independent binary trials, where trial $i$ has probability of success $\pi_i$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_I)^T$. If $\boldsymbol{\pi}$ were known, the tail probability $\Pr(\sum_i B_i \geq k) = \beta(k, \boldsymbol{\pi})$ would be the one-sided significance level for testing the null hypothesis when it is computed at the observed value of the statistic, $k = T - A_0$.

In a randomized experiment, under the null hypothesis $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$, one computes first $r_{Cij} = R_{ij} - Z_{ij}\delta_{0ij}$, next $r_{Ci+}$, then $\pi_i = \bar{\pi}_i = Z_{i+} r_{Ci+}/n_i$ using (3), and finally the exact tail area $\Pr(\sum_i B_i \geq k) = \beta(k, \boldsymbol{\pi})$. In a sensitivity analysis with $\Gamma > 1$ in an observational study, the tail area $\Pr(\sum_i B_i \geq k) = \beta(k, \boldsymbol{\pi})$ cannot be computed because the $u_{ij}$ are unknown; however, a sharp upper bound on the tail area is given by $\beta(k, \bar{\boldsymbol{\pi}}) \geq \beta(k, \boldsymbol{\pi})$. In either case, as $I \to \infty$, $\beta(k, \bar{\boldsymbol{\pi}})$ may be approximated by

$$\beta(k, \bar{\boldsymbol{\pi}}) \to 1 - \Phi\left( \frac{k - \sum_i \bar{\pi}_i}{\sqrt{\sum \bar{\pi}_i (1 - \bar{\pi}_i)}} \right). \quad (4)$$

To illustrate, return again to the example in Section 1.3 and assume there is no hidden bias, so $\Gamma = 1$. There were $T = 173 + 91$ MIs among infected cases, and 91 of these are in discordant pairs. Suppose that the hypothesis $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ results in exactly $A_0 = \sum_{i,j} Z_{ij}\delta_{0ij} = 3$ MIs attributable to infection, and all 3 are among 91 discordant pairs. Under this hypothesis, $R_{i+} = r_{Ci+} + 1$ for these 3 pairs and $R_{i+} = r_{Ci+}$ for the remaining 507 pairs, and in a paired case-referent study, $R_{i+} = 1$ for every pair. From (2) or (3), there are four situations: (1) $\bar{\pi}_i = 1$ for the 173 concordant pairs with $Z_{i+} = 2$, (2) $\bar{\pi}_i = 0$ for the 179 concordant pairs with $Z_{i+} = 0$, (3) $\bar{\pi}_i = 0$ for the three discordant pairs with $Z_{i+} = 1$ and $r_{Ci+} = 0$, and (4) $\bar{\pi}_i = \Gamma/(\Gamma + 1) = 1/2$ for the remaining 155 discordant pairs with $Z_{i+} = 1$ and $r_{Ci+} = 1$. Then $T - A_0 = 173 + 91 - 3$ has null expectation $\sum_i \bar{\pi}_i = 173 + 155/2$ and null variance $\sum \bar{\pi}_i\ (1 - \bar{\pi}_i) = 155 \times 1/2 \times 1/2$, yielding the standardized deviate

$$\frac{T - A_0 - \sum_i \bar{\pi}_i}{\sqrt{\sum \bar{\pi}_i (1 - \bar{\pi}_i)}} = \frac{(173 + 91 - 3) - (173 + (155/2))}{\sqrt{155 \times (1/2) \times (1/2)}} = \frac{88 - 77.5}{\sqrt{38.75}} = 1.69,$$

which is identical to the result in Section 1.2. In this very simple case, the exact test of $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ based on $\beta(k, \bar{\boldsymbol{\pi}})$ can be computed from the binomial distribution with 155 trials and probability $\bar{\pi}_i = 1/2$ of success. Notice also that a different result might be obtained with a different hypothesis, and, of course, there are many possible hypotheses. The goal in Section 3.4 is to identify a single hypothesis to test.

If $\boldsymbol{\delta}_0$ and $\tilde{\boldsymbol{\delta}}_0$ are two compatible hypotheses with $\boldsymbol{\delta}_0 - \tilde{\boldsymbol{\delta}}_0 \geq \mathbf{0}$, then say that $\boldsymbol{\delta}_0$ is *higher* than $\tilde{\boldsymbol{\delta}}_0$, so $\boldsymbol{\delta}_0$ attributes more events to treatment than $\tilde{\boldsymbol{\delta}}_0$ does. If $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is rejected as too small—that is, if it is rejected against the alternative $H_1$: $\boldsymbol{\delta} \geq \boldsymbol{\delta}_0, \boldsymbol{\delta} \neq \boldsymbol{\delta}_0$, and if $\boldsymbol{\delta}_0$ is higher than $\tilde{\boldsymbol{\delta}}_0$—then $\tilde{\boldsymbol{\delta}}_0$ is also rejected. The proof of this important but intuitive fact is given in the Appendix.

### 3.2 Confidence Sets for $\boldsymbol{\delta}$

In principle, a one-sided $100(1-\alpha)\%$ confidence set $\mathcal{C}$ for $\boldsymbol{\delta}$ may be obtained by testing each hypothesis of the form $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ against the one-sided alternative using the procedure in Section 3.1, retaining in $\mathcal{C}$ all values of $\boldsymbol{\delta}_0$ not rejected by the test. There is no difficulty with this in principle. However, the set estimate $\mathcal{C}$ of $\boldsymbol{\delta}$ obtained in this way is a set, perhaps quite a large set, of $N$-dimensional vectors of 1 and 0, so $\mathcal{C}$ is not easy to describe and interpret.

Notice that if $\boldsymbol{\delta}_0$ is higher than $\tilde{\boldsymbol{\delta}}_0$, and if $\tilde{\boldsymbol{\delta}}_0 \in \mathcal{C}$, then $\boldsymbol{\delta}_0 \in \mathcal{C}$.

### 3.3 Attributable Effects

Is it plausible that $a$ or fewer of the treated subjects had events caused by their exposure to the treatment? In other words, is it plausible that for $a$ or fewer of the treated subjects, the effect is $\delta_{ij} = r_{Tij} - r_{Cij} = 1$, so that $a \geq A = \sum_{i,j} Z_{ij}\delta_{ij}$? Is there a $\boldsymbol{\delta}_0 \in \mathcal{C}$ such that $a \geq \sum_{i,j} Z_{ij}\delta_{0ij}$? Or, alternatively, is $a \geq \sum_{i,j} Z_{ij}\delta_{0ij}$ only for values of $\boldsymbol{\delta}_0$ that have been rejected as implausible and are not in $\mathcal{C}$?

There are three ways to answer this question. One way is of theoretical importance, because the inference is an exact randomization inference, but it can be tedious to use. A second way is a large-sample approximation to the exact inference, which is almost as tedious to use. The third way is also a large-sample approximation to the exact inference, this time using asymptotic separability; it is easy to use, requiring only simple arithmetic.

The exact inference uses the exact confidence set $\mathcal{C}$ in Section 3.2 derived from the exact bound $\beta(k, \bar{\boldsymbol{\pi}})$ in Section 3.1. One simply checks whether there is a $\boldsymbol{\delta}_0 \in \mathcal{C}$ such that $a \geq \sum_{i,j} Z_{ij}\delta_{0ij}$. If there is no $\boldsymbol{\delta}_0 \in \mathcal{C}$ such that $a \geq \sum_{i,j} Z_{ij}\delta_{0ij}$, then one has rejected as implausible every possible pattern $\boldsymbol{\delta}_0$ of treatment effects with $a$ or fewer events caused by the treatment. The tedium arises because $\mathcal{C}$ must be computed explicitly and each $\boldsymbol{\delta}_0 \in \mathcal{C}$ must be checked.

The conceptually simple but nonetheless tedious approximation uses the same procedure, but replaces $\beta(k, \bar{\boldsymbol{\pi}})$ by the approximation derived from the normal distribution. This saves a little effort by avoiding the computation of $\beta(k, \bar{\boldsymbol{\pi}})$, but leaves the tedious task of constructing $\mathcal{C}$ and checking each $\boldsymbol{\delta}_0 \in \mathcal{C}$.

It would be convenient if, instead of checking every $\boldsymbol{\delta}_0 \in \mathcal{C}$, one could check a single $\boldsymbol{\delta}_0$ and on the basis of that one $\boldsymbol{\delta}_0$ reach a conclusion about all of $\mathcal{C}$. This is possible using the notion of asymptotic separability (Gastwirth, Krieger, and Rosenbaum 2000), which is simply an approximation theorem for a sensitivity bound that is similar to an approximation for a single distribution based on a central limit theorem.

### 3.4 Approximation Using Asymptotic Separability

The approximation will find one $\boldsymbol{\delta}_0$ with $a = \sum_{i,j} Z_{ij}\delta_{0ij}$ that is hardest to reject; it then must be checked whether this $\boldsymbol{\delta}_0$ is in $\mathcal{C}$. If this $\boldsymbol{\delta}_0 \in \mathcal{C}$, then it is plausible that $a \geq \sum_{i,j} Z_{ij}\delta_{0ij}$, and it is plausible that $a$ or fewer events among treated subjects were caused by treatment; however, if this $\boldsymbol{\delta}_0 \notin \mathcal{C}$, then this $\boldsymbol{\delta}_0$ and all others with $a \geq \sum_{i,j} Z_{ij}\delta_{0ij}$ have been rejected, so it is not plausible that $a$ or fewer events are attributable to treatment. This is a large-sample approximation based on a limiting normal distribution characterized by its expectation and variance.

The approximation will construct one $\boldsymbol{\delta}_0$ such that in $a$ of the matched sets with $\sum_{j=1}^{n_i} Z_{ij}R_{ij} = 1$, the event was indeed caused by the treatment, so that $a = \sum_{i,j} Z_{ij}\delta_{0ij}$. Which $a$ sets should be affected? The approximation considers each matched set, one at a time, and determines by how much the maximum expected contribution $\bar{\pi}_i$ from this set would decline if it were assumed that a treated subject in this set had an event caused by the treatment. The decline is given by $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$, where $\bar{\bar{\lambda}}_i$ is the value of $\bar{\pi}_i$ in (2) if the event in matched set $i$ was not attributable so that $\sum_{j=1}^{n_i} Z_{ij}\delta_{0ij} = 0$ and $R_{i+} = r_{Ci+}$, whereas $\bar{\lambda}_i$ is the value of $\bar{\pi}_i$ in (2) if the event in matched set $i$ was attributable so that $\sum_{j=1}^{n_i} Z_{ij}\delta_{0ij} = 1$ and $R_{i+} = r_{Ci+} + 1$; specifically,

$$\bar{\bar{\lambda}}_i = \frac{\Gamma Z_{i+}R_{i+}}{\Gamma Z_{i+}R_{i+} + n_i - Z_{i+}R_{i+}}$$

and

$$\bar{\lambda}_i = \frac{\Gamma Z_{i+}(R_{i+}-1)}{\Gamma Z_{i+}(R_{i+}-1) + n_i - Z_{i+}(R_{i+}-1)}.$$

To illustrate these formulas, return to the example of Section 1.2, where $n_i = 2$ and $R_{i+} = 1$ for every $i$, and assume that there is no hidden bias, so $\Gamma = 1$. There were $173+91$ matched pairs in which the MI case was infected, so $\sum_{j=1}^{n_i} Z_{ij}R_{ij} = 1$ in these pairs. In the 173 concordant pairs with $Z_{i+} = 2$, if $R_{i+} = r_{Ci+}$, then $\bar{\bar{\lambda}}_i = 1$, whereas if $R_{i+} = r_{Ci+} + 1$, then $\bar{\lambda}_i = 0$, so that $\bar{\bar{\lambda}}_i - \bar{\lambda}_i = 1 - 0 = 1$. In a such a concordant pair, attributing the MI to infection reduces $\bar{\pi}_i$ from 1 to 0. In the 91 discordant pairs with $Z_{i+} = 1$ and $\sum_{j=1}^{n_i} Z_{ij}R_{ij} = 1$, if $R_{i+} = r_{Ci+}$, then $\bar{\bar{\lambda}}_i = 1/2$ if $\Gamma = 1$, whereas if $R_{i+} = r_{Ci+} + 1$, then $\bar{\lambda}_i = 0$, so that $\bar{\bar{\lambda}}_i - \bar{\lambda}_i = 1/2 - 0 = 1/2$. In such a discordant pair, attributing the MI to infection reduces $\bar{\pi}_i$ from $1/2$ to 0.

The $\boldsymbol{\delta}_0$ formed using the $a$ smallest declines, $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$, has the highest expected number of events among treated subjects. Ties are broken by picking the sets that do the least to decrease the variance, the decrease being $\bar{\bar{\omega}}_i - \bar{\omega}_i$, where $\bar{\bar{\omega}}_i = \bar{\bar{\lambda}}_i(1 - \bar{\bar{\lambda}}_i)$ and $\bar{\omega}_i = \bar{\lambda}_i(1 - \bar{\lambda}_i)$. Proposition 1 of Gastwirth et al. (2000) says that as $I \to \infty$, the largest approximate significance level $\beta(k, \bar{\boldsymbol{\pi}})$ is obtained from the normal reference distribution with the highest expectation, and where there are

several normal distributions with the same highest expectation, the one among these with the highest variance. A normal distribution with a high expectation and variance attaches high probabilities to large values. Asymptotic separability is an approximation that works as $I \to \infty$.

The following steps find the $\boldsymbol{\delta}_0$ with $a = \sum_{i,j} Z_{ij}\delta_{0ij}$ that maximizes the expectation $\sum_i \bar{\pi}_i$, and if several $\boldsymbol{\delta}_0$ do this, then finds among these the one that also maximizes the variance $\sum \bar{\pi}_i(1-\bar{\pi}_i)$:

1. If $a \geq \sum_{i,j} Z_{ij}R_{ij}$, then $a$ or fewer of the treated subjects had events, whether caused by exposure or not, so it is certain that $a$ or fewer had events caused by the exposure; stop. Otherwise, if $\sum_{i,j} Z_{ij}R_{ij} > a$, then continue with step 2.

2. For each matched set $i$ with $\sum_{j=1}^{n_i} Z_{ij}R_{ij} = 1$, calculate $\bar{\bar{\lambda}}_i$, $\bar{\lambda}_i$, $\bar{\bar{\omega}}_i$, and $\bar{\omega}_i$.

3. Select exactly $a$ of the matched sets $i$ with $\sum_{j=1}^{n_i} Z_{ij}R_{ij} = 1$ having the smallest values of $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$. If ties among the $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$ mean that several different groups of $a$ matched sets all have the smallest values of $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$, then among these several groups with the smallest $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$, pick any one group with the smallest values of $\bar{\bar{\omega}}_i - \bar{\omega}_i$. For the selected $a$ matched sets, let $\bar{\pi}_i = \bar{\lambda}_i$, whereas for the remaining $I - a$ matched sets, let $\bar{\pi}_i = \bar{\bar{\lambda}}_i$. If $a + \sum_i \bar{\pi}_i \geq \sum_{i,j} Z_{ij}R_{ij}$, then there is one $\mathbf{u}$ and one $\boldsymbol{\delta}_0$ with $a = \sum_{i,j} Z_{ij}\delta_{0ij}$ events caused by treatment that would lead us to expect more than the observed number $\sum_{i,j} Z_{ij}R_{ij}$ of events among treated subjects; conclude that this $a$ is plausible and stop; otherwise, continue with step 4.

4. Calculate the large-sample approximation to the upper bound on the significance level $\beta(k, \bar{\boldsymbol{\pi}})$ in Section 3.1. If $\beta(k, \bar{\boldsymbol{\pi}})$ is small (say, less than .05), then conclude that every compatible $\boldsymbol{\delta}_0$ with $a \geq \sum_{i,j} Z_{ij}\delta_{0ij}$ may be rejected as implausible at the $\beta(k, \bar{\boldsymbol{\pi}})$ level; that is, conclude that it is not plausible that exposure to treatment caused $a$ or fewer events.

## 4. EXAMPLE: CASE-REFERENT STUDY

The informal analysis in Section 1.2 is the simplest illustration of the general procedure developed in Section 3. The purpose of this section is to derive the calculations in Section 1.2 from the general theory in Section 3, thereby justifying Section 1.2 and illustrating Section 3.

McNemar's test applies when there is no hidden bias, $\Gamma = 1$. Each matched set contains $n_i = 2$ subjects, one of whom is a case, $R_{i+} = 1$. Table 2 presents the formal details of the procedure in Section 3.4, which exhibit many simplifications in this special case. If $0 = \sum_j Z_{ij}R_{ij}$, then the case in pair $i$ was not exposed, so the MI could not have been caused by the exposure, and such a matched pair cannot contribute to the attributable effect $A$; hence the "na" in the table signifying not applicable.

We are considering the possibility that $A = 4$ cases had MIs caused by exposure. These 4 could come from the 91 discordant pairs with an exposed case or from the 173 concordant pairs with an exposed case. If the 4 MIs caused by exposure were deleted from the 91 discordant pairs with an exposed case, then for each of these 4 pairs, the expected number of exposed cases would drop from $\bar{\bar{\lambda}}_i = Z_{i+}R_{i+}/2 = 1/2$ to $\bar{\lambda} = Z_{i+}(R_{i+}-1)/2 = 0$. If instead the 4 MIs caused by exposure were deleted from the 173 concordant pairs with an exposed case, then the expected number of exposed cases in the pair would drop from $\bar{\bar{\lambda}}_i = Z_{i+}R_{i+}/2 = 1$ to $\bar{\lambda} = Z_{i+}(R_{i+}-1)/2 = 0$. It follows that $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$ is 1/2 in the discordant pairs and 1 in the concordant pairs, so step 3 in Section 3.4 always assigns the attributable effects to discordant pairs. In Step 4, the 179 concordant pairs with no exposure have no chance of exposure $\bar{\pi}_i = 0$, whereas 173 concordant pairs with two exposures have $\bar{\pi}_i = 1$; these have the effect of deleting concordant pairs from $(\sum B_i - \sum_i \bar{\pi}_i)/\sqrt{\sum \bar{\pi}_i(1-\bar{\pi}_i)}$. Four of the 91 discordant pairs with an exposed case have $R_{ij} - Z_{ij}\delta_{ij} = 1 - 1 \times 1 = 0$; they reduce $\sum B_i$ by 4, $\sum_i \bar{\pi}_i$ by $4 \times 1/2 = 2$, and $\sum \bar{\pi}_i(1-\bar{\pi}_i)$ by $4 \times 1/4 = 1$, with the consequence that $(\sum B_i - \sum_i \bar{\pi}_i)/\sqrt{\sum \bar{\pi}_i(1-\bar{\pi}_i)}$ drops from 1.91 to 1.61. The remaining $67 + (91 - 4)$ discordant pairs have $\bar{\pi}_i = 1/2$ as in McNemar's test. In short, after simplification, the procedure in Section 3.4 yields exactly the informal analysis.

*Table 2. Formal Calculations for a Case-Referent Study*

| Count | 179 | 67 | 91 | 173 |
|---|---|---|---|---|
| $\sum_j Z_{ij}R_{ij}$ | 0 | 0 | 1 | 1 |
| $R_{i+}$ | 1 | 1 | 1 | 1 |
| $Z_{i+}$ | 0 | 1 | 1 | 2 |
| $\bar{\bar{\lambda}}_i$ | 0 | 1/2 | 1/2 | 1 |
| $\bar{\lambda}_i$ | na | na | 0 | 0 |
| $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$ | na | na | 1/2 | 1 |
| $\bar{\bar{\omega}}_i - \bar{\omega}_i$ | na | na | 1/4 | 0 |

## 5. EXAMPLE: A CASE-CROSSOVER STUDY

### 5.1 Alcohol as an Immediate Cause of Injury

Case-crossover studies were proposed by Maclure (1991) to study treatments that have immediate, transient effects (see also Maclure and Mittleman 2000). In such studies, a case's recent exposure to treatment is compared to the case's own previous history of exposure to the treatment, assuming that long past exposures do not have current effects. For instance, case-crossover studies have been used to study the effects of anger on risk of myocardial infarction (Mittleman et al. 1995). Use of sensitivity analysis in related contexts was mentioned by Mittleman, Maldonado, Gerberich, Smith, and Sorock (1997).

Vison et al. (1995) conducted a case-crossover study of alcohol consumption as a proximate cause of injury. Table 3 describes 340 patients in their study, all of whom were treated for acute trauma at an emergency room of one of two mid-

*Table 3. Alcohol and Injury*

| | | *Injury day* | |
|---|---|---|---|
| | | *0–3 drinks* | *≥ 4 drinks* |
| Day before | 0–3 *drinks* | 320 | 15 |
| | ≥4 *drinks* | 3 | 2 |

western hospitals. The table records alcohol consumption during the six hours before injury and the corresponding six hours on the previous day. McNemar's one-sided deviate is $(15-3)/\sqrt{15+3} = 2.83$, with approximate significance level .0023 and odds ratio $15/3 = 5$. One premise is that a drink taken more than 24 hours ago does not increase the current risk of injury.

### 5.2 A Sensitivity Analysis

Of the $17 = 15+2$ injuries following 4 drinks, how many are attributable to drinking? In the absence of hidden bias, that is, with $\Gamma = 1$, it is not plausible that 6 or fewer injuries are due to drinking as $(15-6-3)/\sqrt{15-6+3} = 1.73 > 1.65$, but it is just plausible that 7 or more are attributable to drinking as $(15-7-3)/\sqrt{15-7+3} = 1.50$. If there is an unobserved time-dependent covariate, $u_{ij}$, strongly associated with injury and $\Gamma = 1.5$ times more common on heavy drinking days, then it is still not plausible that $A = 2$ or fewer of the 17 injuries following heavy drinking are attributable to drinking, because the deviate for $A = 2$ is 1.74, but it is plausible that $A = 3$ or more are attributable, because the deviate for $A = 3$ is 1.58. If the unobserved covariate is twice as common on heavy injury days, $\Gamma = 2$, then it is plausible that none of the injuries is attributable to drinking, because the deviate for $A = 0$ is 1.5.

## 6. EXAMPLE: A COHORT STUDY

When testing the null hypothesis of no effect, McNemar's test handles matched pairs in a similar manner in case-referent, case-crossover, and cohort studies. This is no longer true when the null hypothesis is false and inferences are made about attributable effects. The reason is that a matched-pair case-referent study always has one case or outcome event in each pair, $R_{i+} = 1$ for each $i$, but a cohort study may have $R_{i+} = 0, 1$, or 2 outcome events in a matched pair. In a case-referent study, the separable algorithm attributes effects to discordant pairs $(R_{i+} = 1, Z_{i+} = 1)$, whereas in a cohort study, it turns out that effects are first attributed to certain concordant pairs $(R_{i+} = 2, Z_{i+} = 1)$.

Table 4 is from a matched pairs cohort study reported by Ackermann-Liebrich et al. (1996) comparing women who planned to deliver children at home with women who planned to deliver in a hospital. Between 1989 and 1991, 214 women in Zurich who planned to deliver at home were matched to 214 women planning to deliver in a hospital. Note that the groups are defined by where they *planned* to deliver. The matching was based on age, parity, medical history, partner situation, social class, and nationality. In the months before delivery, a few women miscarried, leaving 207 pairs in which both women delivered. Table 4 describes induction of labor in the planned-home/planned-hospital pairs. [Table 4 is deduced from table 5 in Ackermann-Liebrich et al. 1996, where an odds ratio based on discordant pairs is given as .18 together total frequencies of induced labor of 7 at home and 35 in the hospital, so $((7-x)/(35-x)) = .18$ has solution $x = 1$. The $166 = 207-(6+34+1)$ may be off by one or two because of deliveries in taxis, etc.; however, the 166 figure does not affect the analysis.] Women who planned to deliver in the hospital were more likely to be induced. In principle, either planned-home or planned-hospital delivery could be the treatment; however, because the notation assumes a nonnegative effect, $r_{Tij} \geq r_{Cij}$, no change in notation is required if hospital delivery is taken to be the treatment.

*Table 4. Induction of Labor in Home/Hospital Pairs*

| | | Planned hospital | |
|---|---|---|---|
| | | *Induced* | *Not induced* |
| Planned home | Induced | 1 | 6 |
| | Not induced | 34 | 166 |

Table 5 shows the separable calculations, which differ in important ways from those in Table 2. Notice that among pairs with an induced hospital patient—that is, among pairs with $\sum_j Z_{ij}R_{ij} = 1$—the change in expectations is the same, $\bar{\bar{\lambda}}_i - \bar{\lambda}_i = 1/2$, for both for the one concordant induced/induced pair, and for the 34 discordant pairs with $\sum_j Z_{ij}R_{ij} = 1$. This tie is broken by looking at the change in variances, $\bar{\bar{\omega}}_i - \bar{\omega}_i$. Changing any of the 34 discordant pairs into a concordant pair reduces its variance from 1/4 to 0, whereas changing the concordant pair into a discordant pair increases the variance from 0 to 1/4. If the change in expectations is the same, then it will be harder to reject a $\boldsymbol{\delta}_0$ that increases the variance.

Suppose that we wish to ask whether $A = 1$ is plausible assuming no hidden bias, $\Gamma = 1$. There are $\sum_{i,j} Z_{ij}R_{ij} = 35$ induced births in the hospital, and if $A = 1$ is attributable to the hospital, then 34 would have happened anyway; that is, $\sum_{i,j} Z_{ij}r_{Cij} = 34$. There are seven induced births at home, six in discordant pairs and one in a concordant pair. If $A = 1$, then $\sum_{i,j} Z_{ij}r_{Cij}$ has expectation $20.5 = 1+6\times 1/2+33\times 1/2$ if the $A = 1$–attributable induction is among the 34 discordant pairs, and expectation $20.5 = 7\times 1/2+34\times 1/2$ if the $A = 1$–attributable induction is in the single concordant pair with two inductions, so the expectations are the same. However, if $A = 1$, then $\sum_{i,j} Z_{ij}r_{Cij}$ has variance $9.75 = 0+6\times 1/4+33\times 1/4$ if the $A = 1$–attributable induction is among the 34 discordant pairs, and variance $10.25 = 7\times 1/4+34\times 1/4$ if the $A = 1$ attributable induction is in the single concordant pair with two inductions. The standardized deviates for the corresponding two McNemar test statistics are $(35-1-20.5)/\sqrt{9.75} = 4.36$ and $(35-1-20.5)/\sqrt{10.75} = 4.22$, so it is clear in either case that $A = 1$ is too low, but the significance level is a little higher—less significant—when the attributable induction is in the one concordant pair. Again, this is simply the separable algorithm at work—it looked at the expectations and variances and approximated the largest significance level using the

*Table 5. Formal Calculations for a Cohort Study*

| | | | | |
|---|---|---|---|---|
| Count | 166 | 6 | 34 | 1 |
| $\sum_j Z_{ij}R_{ij}$ | 0 | 0 | 1 | 1 |
| $R_{i+}$ | 0 | 1 | 1 | 2 |
| $Z_{i+}$ | 1 | 1 | 1 | 1 |
| $\bar{\bar{\lambda}}_i$ | 0 | $\frac{1}{2}$ | $\frac{1}{2}$ | 1 |
| $\bar{\lambda}_i$ | na | na | 0 | $\frac{1}{2}$ |
| $\bar{\bar{\lambda}}_i - \bar{\lambda}_i$ | na | na | $\frac{1}{2}$ | $\frac{1}{2}$ |
| $\bar{\bar{\omega}}_i - \bar{\omega}_i$ | na | na | $\frac{1}{4}$ | $-\frac{1}{4}$ |

normal distribution. In testing $A = 2$, the separable algorithm would take as the null hypothesis that the 1 concordant pair and 1 of the 34 discordant pairs had the inductions attributable to the hospital, and so forth.

## 7. DISPLACEMENT EFFECTS

### 7.1 Effects of Treatments on Quantiles

Suppose that the $j$th subject in matched set $i$ would exhibit ordered response $y_{Tij}$ under treatment and $y_{Cij}$ under control, where it is assumed that the treatment may increase but that this will not decrease the response, so $y_{Tij} \geq y_{Cij}$. Arranging the $N$ potential responses to control, $y_{Cij}$, into decreasing order gives the order statistics of potential responses to control, namely $y_{C(N)} \geq y_{C(N-1)} \geq \cdots \geq y_{C(1)}$. Notice that, because some of the $N$ subjects did not receive the control, $y_{C(k)}$ cannot be determined from observed data.

Fix a $k$ so that $y_{C(k)}$ is the $k/N$ quantile of responses that would be seen if all subjects received the control. Let $\theta$ be a value strictly between $y_{C(k)}$ and $y_{C(k+1)}$, so that $y_{C(k+1)} > \theta > y_{C(k)}$. The data will indicate whether such a $\theta$ exists. For instance, if $N$ were even and $k = N/2$, then $\theta$ would be a median, that is, a value strictly between the two middle-order statistics that would have been observed had all subjects received the control. Subject $j$ in set $i$ has a *displacement* around $\theta$ if $y_{Tij} > \theta > y_{Cij}$. For instance, with $N$ even and $k = N/2$, there is a displacement if the subject would have had a response below the median $\theta$ under control but would have had a response above the control median $\theta$ had the subject received treatment instead. Alternatively, if $k/N = .95$, then a displacement might be described as signifying a subject whose response would have been fairly typical under control and usually high under treatment. In the unmatched situation, displacement effects were discussed in earlier work (Rosenbaum 2001). Let $Y_{ij} = Z_{ij}y_{Tij} + (1 - Z_{ij})y_{Cij}$ be the response observed from the $j$th subject in matched set $i$, and let the observed order statistics of the values of the $Y_{ij}$ be $Y_{(N)} \geq Y_{(N-1)} \geq \cdots \geq Y_{(1)}$.

Write $r_{Cij} = 1$ if $y_{Cij} > \theta$ and $r_{Cij} = 0$ otherwise. In parallel, write $r_{Tij} = 1$ if $y_{Tij} > \theta$ and $r_{Tij} = 0$ otherwise. Then there is a displacement if $\delta_{ij} = r_{Tij} - r_{Cij} = 1$ and no displacement if $\delta_{ij} = r_{Tij} - r_{Cij} = 0$. Notice carefully that because $y_{C(k)}$ cannot be observed, neither $r_{Cij}$ nor $r_{Tij}$ can be calculated, unlike the situation in Section 3. The number of displacements attributable to treatment is $A = \sum_{i,j} Z_{ij}(r_{Tij} - r_{Cij}) = \sum_{i,j} Z_{ij}\delta_{ij}$ and $A = 0$ under the null hypothesis of no effect, $H_0$: $y_{Tij} = y_{Cij}$. The following result is a trivial extension of an earlier result (Rosenbaum 2001), where there are no matched sets. Because the proof is brief and provides insight, it is repeated here.

*Proposition 1.* If $a = \sum_{i,j} Z_{ij}\delta_{ij}$, then $Y_{(k+1-a)} > \theta > Y_{(k-a)}$.

*Proof.* There are exactly $N - k$ subjects with $y_{Cij} > \theta$, and because $y_{Tij} \geq y_{Cij}$, it follows that these $N - k$ subjects all have $Y_{ij} > \theta$. Because $a = \sum_{i,j} Z_{ij}\delta_{ij}$, there are exactly $a$ other subjects not included among the $N - k$ subjects, with $Y_{ij} = y_{Tij} > \theta > y_{Cij}$. For the remaining $k - a$ subjects, $\theta > Y_{ij}$. So there are exactly $N - k + a$ subjects with $Y_{ij} > \theta$ and exactly $k - a$ subjects with $\theta > Y_{ij}$. This means that $Y_{(N)} \geq Y_{(N-1)} \geq \cdots \geq Y_{(k+1-a)} > \theta > Y_{(k-a)} \geq \cdots \geq Y_{(1)}$.

Notice that if $a = \sum_{i,j} Z_{ij}\delta_{ij}$ but $Y_{(k+1-a)} = Y_{(k-a)}$, then it must be the case that $y_{C(k+1)} = y_{C(k)}$ and no $\theta$ strictly between $y_{C(k+1)}$ and $y_{C(k)}$ exists.

If $a = \sum_{i,j} Z_{ij}\delta_{ij}$, then it follows from the proposition that $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij} = 1$ if $Y_{ij} > Y_{(k-a)}$ and $R_{ij} = 0$ when $Y_{(k-a)} \geq Y_{ij}$. Hence for each particular $a$, the $R_{ij}$ may be calculated and the method of Section 3.4 applied. In particular, for matched pairs, $n_i = 2$ for all $i$, the calculations reduce to those in Table 5.

### 7.2 Kidney Function in Cadmium Workers

In an effort to estimate the effects of cadmium exposure on kidney function, Thun et al. (1989) compared workers exposed to cadmium with unexposed controls in terms of $\beta$-2-microglobulin in $\mu$g/g of creatinine (see also Thun 1993). Male workers at a cadmium recovery plant in Colorado were compared to unexposed male workers at a Colorado hospital, after "frequency matching" for an important covariate, age. As in Rosenbaum (1996, sec. 4.3), frequency matching is replaced by pair matching for age, yielding 23 pairs, one cadmium worker, one hospital control. Their data, given in Table 6, show much higher levels of $\beta$-2-microglobulin among some of the cadmium workers than among controls.

There are $46 = 2 \times 23$ subjects, each of whom has a potential response, $y_{Cij}$, in the absence of cadmium exposure. Because $46 \times .8 = 36.8$, the 80th percentile of these potential responses is any value $\theta$ strictly between $r_{C(36)}$ and $r_{C(37)}$, neither of which is observed, because only 23 of the 46 values of $y_{Cij}$ were observed. It is assumed that cadmium exposure does not improve kidney function, so $y_{Tij} \geq y_{Cij}$ for all $i, j$. A displacement around the 80th percentile point means that $y_{Tij} > \theta > y_{Cij}$, and the task is to draw inferences about the number of displacements among the 23 treated subjects. A treated subject with a displacement had a $\beta$-2-microglobulin above 80% of the responses that would have been seen without

*Table 6. Kidney Function of Cadmium Workers and Unexposed Controls*

| Pair | Cadmium worker | Hospital control |
|---|---|---|
| 1 | 107,143 | 311 |
| 2 | 33,679 | 338 |
| 3 | 18,836 | 159 |
| 4 | 173 | 110 |
| 5 | 389 | 226 |
| 6 | 1144 | 305 |
| 7 | 513 | 222 |
| 8 | 211 | 242 |
| 9 | 24,288 | 250 |
| 10 | 67,632 | 256 |
| 11 | 488 | 135 |
| 12 | 700 | 96 |
| 13 | 328 | 142 |
| 14 | 98 | 120 |
| 15 | 122 | 376 |
| 16 | 2302 | 173 |
| 17 | 10,208 | 178 |
| 18 | 892 | 213 |
| 19 | 2803 | 257 |
| 20 | 201 | 81 |
| 21 | 148 | 199 |
| 22 | 522 | 114 |
| 23 | 941 | 247 |

cadmium exposure, and would have had a $\beta$-2-microglobulin below this 80th percentage point had he been spared cadmium exposure.

If there were no displacements attributable to treatment, so $A = 0$, then the proposition implies $941 = Y_{(37)} > \theta > Y_{(36)} = 892$. In this case there are 10 matched pairs discordant for having $Y_{ij} > 892$, and in all 10 pairs, it is the cadmium worker who has $Y_{ij} > 892$ (Table 7). In the absence of hidden bias, $\Gamma = 1$, this yields a standardized deviate of $(10 - 10/2)/\sqrt{10 \times \frac{1}{4} + 13 \times 0} = 3.16$ with an approximate one-sided significance level of .00078. If there was one displacement attributable to treatment, $A = 1$, then exactly one of the observed responses among treated subjects is above $\theta$ because of the treatment and would have been below $\theta$ under control, so that $892 = Y_{(36)} > \theta > Y_{(35)} = 700$. In this case, there are 11 discordant pairs for $Y_{ij} > 700$, and in all 11 pairs, it is the cadmium worker who has $Y_{ij} > 700$. In the absence of hidden bias, $\Gamma = 1$, the statistic $T - A = 11 - 1$ has null expectation $\sum \bar{\pi}_i = (11-1) \times 1/2 + (12+1) \times 0 = 5$ and null variance $\sum \bar{\pi}_i(1 - \bar{\pi}_i) = (11-1) \times 1/4 + (12+1) \times 0 = 2.5$, giving the same standardized deviate of $(10-5)/\sqrt{2.5} = 3.16$ with an approximate one-sided significance level of .00078. Without hidden bias, $\Gamma = 1$, when $A = 9$, the one-sided significance level is .0289, but when $A = 10$, it is .103, so it is not plausible that 9 or fewer displacements are attributable to treatment, but it is plausible that 10 or more are.

Although in practice one simply applies the separable algorithm, it is useful to inspect what this algorithm does in one particular case. When $A = 9$, there are 9 displacements above $\theta$ among the treated subjects, so $328 = Y_{(36+1-9)} > \theta > Y_{(36-9)} = 311$. Now, in 1 pair, namely pair $i = 2$, both responses are above 311; in 1 pair, namely pair $i = 15$, the hospital worker is above 311 but the cadmium worker is not; and in 16 pairs only the cadmium worker is above 311. One could attribute all $A = 9$ displacements to the 16 discordant pairs with high values for the cadmium worker, or attribute one displacement to concordant pair $i = 2$ and the remaining 8 displacements to the 16 discordant pairs. In either case, $T - A = 17 - 9 = 8$ with null expectation $(18-8)/2 = 5$ when $\Gamma = 1$; however, in the first case, there would be $17 - 9 = 8$ discordant pairs and $T - A$ would have variance $8 \times 1/4 + 15 \times 0 = 2$, whereas in the second case, there would be $18 - 8 = 10$ discordant pairs, because one new discordant pair is created and one fewer is removed, so $T - A$ would have variance $10 \times 1/4 + 13 \times 0 = 2.5$. It is harder to reject the hypothesis when the variance is larger, so the separable algorithm attributes one displacement to the concordant pair, yielding a deviate of $(8-5)/\sqrt{2.5} = 1.897$ with approximate one-sided significance level of $1 - \Phi(1.897) = .0289$. The sensitivity analysis for $\Gamma \geq 1$ uses the steps in Section 3.4 applied to the same figures as in Table 7. For $\Gamma = 1, 2$, and 3, the smallest plausible numbers of displacements $A$ are 10, 7, and 7, whereas for $\Gamma = 4$, no displacements, $A = 0$, becomes barely plausible with an upper bound on the significance level of .057.

Table 6 exhibits a common pattern. Some cadmium workers exhibit extremely high $\beta$-2-microglobulin levels not seen among any hospital controls, but other cadmium workers exhibit values in the normal range. When this is true, the sensitivity to hidden bias may vary with $k$, so that displacements about upper quantiles are less sensitive to bias than displacements about lower quantiles. For instance, for the median, $k = 23$, the smallest plausible number of displacements for $\Gamma = 1, 2$, and 3 is $A = 3$, 1, and 0, so there is less evidence of large numbers of displacements about the median than about the 80th percentile. A similar pattern was found, under a different model, using different methods in earlier work (Rosenbaum 1999).

*Table 7. Calculations for Displacement Effects in Matched Pairs*

| | | | |
|---|---|---|---|
| *A = 0* | | | |
| *Cadmium worker* | | | |
| | | $Y_{ij} > 892$ | $892 \geq Y_{ij}$ |
| Control | $Y_ij > 892$ | 0 | 0 |
| | $892 \geq Y_{ij}$ | 10 | 13 |
| | $\frac{10-5}{\sqrt{2.5}} = 3.16, \quad 1 - \Phi(3.16) = .00078$ | | |
| *A = 1* | | | |
| *Cadmium worker* | | | |
| | | $Y_{ij} > 700$ | $700 \geq Y_{ij}$ |
| Control | $Y_{ij} > 700$ | 0 | 0 |
| | $700 \geq Y_{ij}$ | 11 | 12 |
| | $\frac{10-5}{\sqrt{2.5}} = 3.16, \quad 1 - \Phi(3.16) = .00078$ | | |
| *A = 9* | | | |
| *Cadmium worker* | | | |
| | | $Y_{ij} > 311$ | $311 \geq Y_{ij}$ |
| Control | $Y_{ij} > 311$ | 1 | 1 |
| | $311 \geq Y_{ij}$ | 16 | 5 |
| | $\frac{8-5}{\sqrt{2.5}} = 1.897, \quad 1 - \Phi(1.897) = .0289$ | | |
| *A = 10* | | | |
| *Cadmium worker* | | | |
| | | $Y_{ij} > 305$ | $305 \geq Y_{ij}$ |
| Control | $Y_{ij} > 305$ | 2 | 1 |
| | $305 \geq Y_{ij}$ | 15 | 5 |
| | $\frac{7-5}{\sqrt{2.5}} = 1.265, \quad 1 - \Phi(1.265) = .103$ | | |

## 8. SUMMARY

The attributable effect is the number of events among treated subjects that were actually caused by the treatment. The attributable effect is an unobserved random variable, not a parameter, because its value changes as the treatment assignment changes. Exact and approximate inferences about attributable effects have been discussed, where the approximation used the technique of asymptotic separability. Randomization inferences for experiments and sensitivity analyses for observational studies have been developed in parallel.

## APPENDIX: PROOF OF A DETAIL

Suppose that $\boldsymbol{\delta}_0$ and $\tilde{\boldsymbol{\delta}}_0$ are two compatible null hypotheses, where $\boldsymbol{\delta}_0 = \tilde{\boldsymbol{\delta}}_0 + (0, 0, \ldots, 0, 1, 0, \ldots, 0)^T$, so $\boldsymbol{\delta}_0$ has one more nonzero effect than $\tilde{\boldsymbol{\delta}}_0$ does. This Appendix shows that if $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is rejected in favor of $H_1$: $\boldsymbol{\delta} \geq \boldsymbol{\delta}_0$, $\boldsymbol{\delta} \neq \boldsymbol{\delta}_0$, then $H_0$: $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}_0$ is also rejected. There are two cases. In the first case, if $A_0 = \sum_{i,j} Z_{ij}\delta_{0ij} = \sum_{i,j} Z_{ij}\tilde{\delta}_{0ij}$, then $T - A_0 = \sum_{i,j} Z_{ij} r_{Cij} = \sum_i B_i$ is unchanged, and the same inferences result. In the second case, $\sum_{i,j} Z_{ij}\delta_{0ij} = 1 + \sum_{i,j} Z_{ij}\tilde{\delta}_{0ij}$; assume this to be true for the remainder of this paragraph. In the second case, several things change. For convenience of

notation but without loss of generality, assume that the change affects the first person, so $\delta_{011} = 1$, $\tilde{\delta}_{011} = 0$, and $Z_{11} = 1$. Because both hypotheses are compatible, it must be the case that $0 = R_{11} - Z_{11}\delta_{011}$ and $1 = R_{11} - Z_{11}\tilde{\delta}_{011} = R_{11}$, so $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ hypothesizes $r_{C11} = 0$ but $H_0$: $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}_0$ hypothesizes $r_{C11} = 1$. This means that when testing $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$, $B_1 = \sum_{j=2}^{n_1} Z_{1j} r_{C1j}$, whereas when testing $H_0$: $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}_0$, $B_1 = Z_{11} + \sum_{j=2}^{n_1} Z_{1j} r_{C1j}$. Also, the observed value of the test statistic $T - A_0$ is $T - \sum_{i,j} Z_{ij}\delta_{0ij} = k$, say, when testing $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$, but is $T - \sum_{i,j} Z_{ij}\tilde{\delta}_{0ij} = T - \sum_{i,j} Z_{ij}\delta_{0ij} + 1 = k + 1$ when testing $H_0$: $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}_0$. Now, trivially,

$$\Pr\left(\sum_{j=2}^{n_1} Z_{1j} r_{C1j} + \sum_{i=2}^{I} B_i \geq k\right) \geq \Pr\left(Z_{11} + \sum_{j=2}^{n_1} Z_{1j} r_{C1j} + \sum_{i=2}^{I} B_i \geq k+1\right);$$

however, these are the two significance levels. It follows that if $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is rejected because $\Pr(\sum_{j=2}^{n_1} Z_{1j} r_{C1j} + \sum_{i=2}^{I} B_i \geq k)$ is small, then $H_0$: $\boldsymbol{\delta} = \tilde{\boldsymbol{\delta}}_0$ is also rejected because $\Pr(Z_{11} + \sum_{j=2}^{n_1} Z_{1j} r_{C1j} + \sum_{i=2}^{I} B_i \geq k+1)$ is smaller still.

More generally, if $\boldsymbol{\delta}_0$ and $\tilde{\boldsymbol{\delta}}_0$ are two compatible hypotheses with $\delta_0 - \tilde{\boldsymbol{\delta}}_0 \geq 0$, then $\boldsymbol{\delta}_0$ is *higher* than $\tilde{\boldsymbol{\delta}}_0$. By induction on the argument just given, if $H_0$: $\boldsymbol{\delta} = \boldsymbol{\delta}_0$ is rejected against the alternative $H_1$: $\boldsymbol{\delta} \geq \boldsymbol{\delta}_0$, $\boldsymbol{\delta} \neq \boldsymbol{\delta}_0$, and $\boldsymbol{\delta}_0$ is higher than $\tilde{\boldsymbol{\delta}}_0$, then $\tilde{\boldsymbol{\delta}}_0$ is also rejected.

*[Received November 2000. Revised August 2001.]*

## REFERENCES


Ackermann-Liebrich, U., Voegeli, T., Gunter-Witt, K., Kunz, I., Zullig, M., Schindler, C., and Maurer, M. (1996), "Home Versus Hospital Deliveries: Follow-Up Study of Matched Pairs for Procedures and Outcome," *British Medical Journal*, 313, 1313–1318.

Copas, J. B., and Li, H. G. (1997), "Inference for Non-Random Samples" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 55–96.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959), "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions," *Journal of the National Cancer Institute*, 22, 173–203.

Danesh, J., Youngman, L., Clarck, S., Parish, S., Peto, R., Collins, R., and The International Studies of Infarct Survival Collaborative Group (1999), "*Helicobacter Pylori* Infection and Early Onset Myocardial Infarction: Case-Control and Sibling Pairs Study," *British Medical Journal*, 319, 1157–1162.

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

Fligner, M. A., and Wolfe, D. A. (1976), "Some Applications of Sample Analogues to the Probability Integral Transformation and a Coverage Property," *The American Statistician*, 30, 78–85.

Gart, J. (1963), "A Median Test With Sequential Application," *Biometrika,* 50, 55–62.

Gastwirth, J. L. (1968), "The First-Median Test: A Two-Sided Version of the Control Median Test," *Journal of the American Statistical Association,* 63, 692–706.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998), "Dual and Simultaneous Sensitivity Analysis for Matched Pairs," *Biometrika*, 85, 907–920.

———(2000), "Asymptotic Separability in Sensitivity Analysis," *Journal of the Royal Statistical Society*, Ser. B, 62, 545–556.

Greenhouse, S. (1982), "Jerome Cornfield's Contributions to Epidemiology," *Biometrics* (suppl.), 33–45.

Hamilton, M. A. (1979), "Choosing the Parameter for $2 \times 2$ and $2 \times 2 \times 2$ Table Analysis," *American Journal of Epidemiology,* 109, 362–375.

Heckman, J. (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441–462.

Li, G., Tiwari, R. C., and Wells, M. T. (1996), "Quantile Comparison Functions in Two-Sample Problems, With Application to Comparisons of Diagnostic Markers," *Journal of the American Statistical Association,* 91, 689–698.

Lin, D., Psaty, B., and Kronmal, R. (1998), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics*, 54, 948–963.

Maclure, M. (1991), "The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events," *American Journal of Epidemiology,* 133, 144–152.

Maclure, M., and Mittleman, M. A. (2000), "Should We Use a Case-Crossover Design?," *Annual Review of Public Health*, 21, 193–221.

MacMahon, B., and Trichopoulos, D. (1996), *Epidemiology: Principles and Methods*, Boston: Little, Brown.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review*, 319–323.

Mittleman, M. A., Maclure, M., Sherwood, J. B., Mulry, R. P., Tofler, G. H., Jacobs, S. C., Friedman, R., Benson, H., Muller, J. E., and The Determinants of Myocardial Infarction Onset Study Investigators (1995), "Triggering of Acute Myocardial Infarction Onset by Episodes of Anger," *Circulation*, 92, 1720–1725.

Mittleman, M. A., Maldonado, G., Gerberich, S. G., Smith, G. S., and Sorock, G. S. (1997), "Alternative Approaches to Analytical Designs in Occupational Injury Epidemiology," *American Journal of Industrial Medicine*, 32, 129–141.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9" (in Polish), *Roczniki Nauk Roiniczych,* Tom X, pp. 1–51. Reprinted in English in *Statistical Science*, 1990, 5, 463–480, with discussion by T. Speed and D. Rubin.

Orban, J., and Wolfe, D. A. (1982), A Class of Distribution-Free Two-Sample Tests Based on Placements," *Journal of the American Statistical Association,* 77, 666–672.

Robins, J. M. (1988), "Confidence Intervals for Causal Parameters," *Statistics in Medicine*, 7, 773–785.

Robins, J. M., Greenland, S., and Hu, F. C. (1999), "Rejoinder," *Journal of the American Statistical Association,* 94, 708–712.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Models in Epidemiology*, eds. E. Halloran and D. Berry, New York: Springer-Verlag, pp. 1–94.

Rosenbaum, P. R. (1986), "Dropping Out of High School in the United States: An Observational Study," *Journal of Educational Statistics*, 11, 207–224.

———(1988), "Sensitivity Analysis for Matching With Multiple Controls," *Biometrika*, 75, 577–581.

———(1991), "Senstivity Analysis for Matched Case-Control Studies," *Biometrics*, 47, 87–100.

———(1995a), *Observational Studies,*New York: Springer-Verlag.

———(1995b), "Quantiles in Nonrandom Samples and Observational Studies," *Journal of the American Statistical Association,* 90, 1424–1431.

———(1996), "Observational Studies and Nonrandomized Experiments," in *Handbook of Statistics, Design and Analysis of Experiments,* Vol. 13, eds. S. Ghosh and C. R. Rao, New York: Elsevier.

———(1999), "Reduced Sensitivity to Hidden Bias at Upper Quantiles in Observational Studies With Dilated Effects," *Biometrics*, 55, 560–564.

———(2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies With a Discrete Pivot," *Biometrika*, 88, 219–231.

Rosenbaum, P. R., and Rubin, D. B. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

———(1985), "The Bias due to Incomplete Matching," *Biometrics*, 41, 103–116.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Thun, M. (1993), "Kidney Dysfunction in Cadmium Workers," in *Case Studies in Occupational Epidemiology,* ed. K. Steenland, New York: Wiley, pp. 105–126.

Thun, M., Osorio, A., Schober, S., Hannon, W. H., Lewis, B., and Halporn, W. (1989), "Nephropathy in Cadmium Workers: Assessment of Risk From Airborne Occupational Exposure to Cadmium," *British Journal of Industrial Medicine*, 46, 689–697.

Vison, D., Mabe, N., Leonard, L., Alexander, J., Becker, J., Boyer, J., and Moll, J. (1995), "Alcohol and Injury: A Case-Crossover Study," *Archives of Family Medicine*, 4, 505–511.

Wasserman, L. (1999), "Comment," *Journal of the American Statistical Association,* 94, 704–706.

Walter, S. D. (1976), "The Estimation and Interpretation of Attributable Risk in Health Research," *Biometrics,* 32, 829–49.