

## Attributable Effects in Case<sup>2</sup>-Studies

Paul R. Rosenbaum

Statistics Department, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, U.S.A.  
*email:* rosenbaum@stat.wharton.upenn.edu

**SUMMARY.** In an effort to determine whether a particular treatment causes a particular outcome event, data are obtained from a database system that records events when they occur, and for such events, the system records exposure to the treatment. That is, the system records information about cases. The system provides no information about events that might have occurred but did not, that is, about units which are not cases. Roughly speaking, we know the number of successes for two proportions, treated and control, but not the numbers of trials or units for these proportions; indeed, the concept of a “trial” may be somewhat vague. With no further information, the situation is quite hopeless. However, an interesting strategy that is sometimes used entails identifying two types of cases whose origin is entirely different so that it is known the cases of the second type were definitely not affected by the treatment under study. This strategy—the case–case or case<sup>2</sup>-study—seems to have been reinvented independently many times, and has recently been offered as a general strategy for infectious disease epidemiology by McCarthy and Giesecke (1999, *International Journal of Epidemiology* **28**, 764–768). Can this strategy permit estimation of the number of cases caused by the treatment? Using attributable effects in a new way, a method of exact inference is proposed, along with a large sample approximation. Two examples are discussed: one concerning the effects of daytime running lights (DRLs) on the risk of multivehicle accidents; the other concerning the origin of a Salmonella infection. A counterexample with superficially similar appearance is also discussed concerning suicide rates following the publication of *Final Exit*; here, the treatment may alter the outcome, or it may alter the type, and the attributable effect cannot be estimated.

**KEY WORDS:** Attributable effect; Case<sup>2</sup>-study; Case–case study; Causal effect; Observational study; Sensitivity analysis.

### 1. Introduction, Outline, and Notation

#### 1.1 Case<sup>2</sup>-Studies

Quite often, interesting events, such as accidents or infections or crimes, are selectively recorded when they occur, but the system that records this information provides no information about events that might have occurred but, in fact, did not occur. In the language of Bernoulli trials, information is recorded about successes but not about trials that do not produce successes. If exposure to a treatment causes a unit to have an event it would not have had under control, then the treatment also causes information to be collected about this unit, which would have passed unnoticed had it received the control.

Case<sup>2</sup>-studies or case–case studies are an attempt to use such data to study the effects of a treatment that may cause some of the events. These studies compare one subtype of cases to another subtype of cases. Case<sup>2</sup>-studies have been used and reinvented independently many times by many researchers, and recently McCarthy and Giesecke (1999) have advocated case<sup>2</sup>-studies as a systematic tool in the study of infectious diseases. They focus on infectious diseases that present similar clinical symptoms but which are, on the microbial or molecular level, biologically different diseases.

My purpose here is to consider the case<sup>2</sup>-study as a research design—a design for an observational study—and to consider

what must be true if the study is to provide useful information about the treatment effects in the population. This entails describing the population carefully; then describing the relationship between the limited data in a case<sup>2</sup>-study and the population. It is not obvious—in fact, it is not true—that treatment effects can generally be estimated by comparing the different subtypes of cases. Two conditions will be crucial: (i) The treatment must not cause the second type of case, and (ii) the treatment must not cause a change in an individual’s case type. These considerations are very plausible, almost inevitable, in some settings, and quite implausible or far from certain in other settings.

Section 1.2 introduces notation for case<sup>2</sup>-studies with causal effects defined in terms of potential responses under alternative treatments, in the manner of Neyman (1923) and Rubin (1974). It is essential to describe the population, the database, and their relationship. In Section 2, two examples of case<sup>2</sup>-studies are described, including one of the type suggested by McCarthy and Giesecke (1999), and then a third study is described which appears, superficially, to be similar, but is in fact quite different. In this third study, the treatment may cause outcome events, or it may cause a change in a case’s subtype, and the logic of the case<sup>2</sup>-study no longer applies.

Section 3 discusses inference in case<sup>2</sup>-studies, which turns out to differ from inference in cohort studies. If a case of

disease is prevented in a cohort study by preventing exposure to the treatment, then the former case would have appeared in the study but as an individual free of disease. In contrast, in a case<sup>2</sup>-study, the database records only information about cases, so preventing a case would remove an individual from the study. In consequence, numerically the same  $2 \times 2$  contingency table produces different inferences in cohort and case<sup>2</sup>-studies.

1.2 Recording Cases Possibly Caused by Exposure to a Treatment

The population contains  $I$  units,  $i = 1, \dots, I$ , and unit  $i$  is either exposed to the treatment, denoted as  $Z_i = 1$ , or not exposed, denoted as  $Z_i = 0$ . A unit might be a person on a particular day, or a car in a particular hour of a car trip. Unit  $i$  would exhibit a binary response,  $r_{Ti}$ , if exposed to the treatment, and another binary response,  $r_{Ci}$ , if not exposed to the treatment, so the treatment effect is  $\delta_i = r_{Ti} - r_{Ci}$  (see Neyman, 1923; Rubin, 1974). The actual response  $R_i$  for unit  $i$  is the response to treatment,  $r_{Ti}$ , if  $i$  actually received treatment,  $Z_i = 1$ ; otherwise, if  $i$  received the control,  $Z_i = 0$ , then  $R_i$  is the response to control,  $r_{Ci}$ ; so  $R_i = Z_i r_{Ti} + (1 - Z_i) r_{Ci}$ . The effect  $\delta_i = r_{Ti} - r_{Ci}$  cannot be calculated from observable data.

A database system records the events when they occur, but it provides no information about units that did not have an event. That is, if  $R_i = 1$ , if  $i$  is a case, then information is obtained about unit  $i$ , in particular, the treatment indicator  $Z_i$  is observed, but if  $R_i = 0$  no information is obtained, and even the total number of units,  $I$ , is often not available. The number of treated units with events,  $\sum_{i=1}^I R_i Z_i$ , and the number of untreated or control units with events,  $\sum_{i=1}^I R_i (1 - Z_i)$ , are observed, but the number of treated units,  $\sum_{i=1}^I Z_i$ , and the number of control units,  $\sum_{i=1}^I (1 - Z_i)$ , are not observed. Moreover, if  $\delta_i = 1$ , then applying the treatment to unit  $i$  causes  $i$  to have an event, so that the information about  $i$  is recorded, while withholding the treatment from  $i$  causes  $i$  not to have an event, so that no record of unit  $i$  is obtained. Issues of this sort are sometimes called “problems of ascertainment” (see Fisher, 1934).

In the simplest situation, units are assigned to treatment ( $Z_i = 1$ ) or control ( $Z_i = 0$ ) at random, as in a randomized experiment, by the flip of a coin that comes up heads with constant probability  $\theta$ . Quite plausibly, however, different units  $i$  have different chances, say  $\theta_i$ , of exposure to the treatment because the units differ prior to treatment with respect to covariates, which may or may not be observed for cases,  $R_i = 1$ , but are not observed for units  $i$  which are not cases,  $R_i = 0$ .

The situation just described is quite common. For instance, a number of databases record information about certain types

of motor vehicle accidents but provide no information about cars or driving trips that produce no accident. For instance, this is true of the U.S. National Highway Traffic Safety Administration’s Fatal Accident Reporting System and of the State Data System. Similarly, information about some cases of certain infectious diseases is either recorded in databases or available from hospitals, with no information about individuals without disease.

Although this situation is quite common, it is also quite hopeless, unless additional information of one form or another is found. One interesting idea that has been used independently by various researchers and has been offered as a strategy for infectious disease epidemiology by McCarthy and Giesecke (1999, p. 767) entails dividing cases into two subtypes,  $h = 1$  and  $0$ , whose origins are entirely different. Specifically, the cases of one subtype,  $h = 1$ , may or may not have been caused by the treatment under study, but the cases of the other subtype,  $h = 0$ , because of some evident aspect of their origin, were definitely not caused by this treatment. That is,  $h_i = 0$  implies  $r_{Ti} = r_{Ci}$  and  $\delta_i = 0$ . If  $h_i = 1$ , then we simply do not know: Case  $i$  might or might not have been affected by the treatment. The database permits the construction of the  $2 \times 2$  table crossclassifying type by treatment,  $h_i \times Z_i$ , for cases,  $R_i = 1$ , and it has the form of Table 1. Note that the sums in Table 1 are over the whole population, from  $i = 1$  to  $I$ , but only cases,  $R_i = 1$ , are counted in these sums, so the total count in the table is  $\sum_{i=1}^I R_i \leq I$ .

1.3 Attributable Effect: A Discrete Pivot

Randomization or permutation tests of no treatment effect are sometimes inverted to yield exact, distribution-free confidence intervals for the magnitude of an additive treatment effect (see Lehmann, 1998). This convenient technique does not apply with treatments that are not additive, for example, to Fisher’s exact test for a  $2 \times 2$  table. Attributable effects are a device that permits a much larger class of permutation tests to be inverted to yield confidence intervals (see Rosenbaum, 2001, 2002, 2003). An attributable effect describes how the responses of treated units would have been different under control, and typically it is not a fixed parameter but rather an unobserved random variable. In particular, in a randomized experiment with two treatments and binary outcome, it is possible to draw inferences about the number of events caused by the treatment by inverting Fisher’s exact test (see Rosenbaum, 2001).

In Section 1.2, the attributable effect is the net increase in the number of cases recorded in the database that were caused by the treatment, that is,  $A = \sum_{i=1}^I Z_i \delta_i = \sum_{i=1}^I Z_i (r_{Ti} - r_{Ci})$ , which cannot be observed because  $(r_{Ti}, r_{Ci})$  are not jointly observed for any unit  $i$ . If the treatment has no effect,

Table 1  
Distribution type by treatment for cases

	Treated $Z_i = 1$	Control $Z_i = 0$	Total
$h_i = 1$	$\sum Z_i h_i R_i$	$\sum (1 - Z_i) h_i R_i$	$\sum h_i R_i$
$h_i = 0$	$\sum Z_i (1 - h_i) R_i$	$\sum (1 - Z_i) (1 - h_i) R_i$	$\sum (1 - h_i) R_i$
Total	$\sum Z_i R_i$	$\sum (1 - Z_i) R_i$	$\sum R_i$

$r_{Ti} = r_{Ci}$  for every  $i$ , then  $A = 0$ . Because  $\delta_i$  takes values in  $\{-1, 0, 1\}$ , the attributable effect  $A$  can be positive or negative, and it can be zero even if some  $\delta_i \neq 0$ . As the second subtype is unaffected by the treatment,  $h_i = 0$  implies  $\delta_i = 0$ , so  $A = \sum_{i=1}^I h_i Z_i \delta_i$ .

When exposures to treatment strike units in the population at random, with a constant probability, Section 3.1 will show that inference about  $A$  is possible in a case<sup>2</sup>-study by inverting Fisher's exact test, although the way this is done is technically quite different from inverting Fisher's exact test in a cohort study. The difference is that, in a case<sup>2</sup>-study, a case  $i$  caused by the treatment,  $Z_i \delta_i = 1$ , would have entirely disappeared from the database had exposure to the treatment been prevented, whereas in a cohort study, this unit would have remained in the study but would no longer be a case. In Sections 3.2 and 3.3, the assumption that treatments strike units at random is dropped. In Section 3.3, the chance of exposure to the treatment may not be constant, and instead may vary with the observed covariates that are controlled by matching; here, the McNemar–Mantel–Haenszel test replaces Fisher's exact test. In Section 3.2, the chance of exposure to the agent may vary with unobserved covariates, which cannot be controlled by matching, and the sensitivity of inferences to such hidden biases is investigated. The matched situation in Section 3.3 requires slight changes in notation, so it is discussed later.

Recall that in Fisher's theory of randomization inference, probability enters through the random assignment of treatments,  $Z_i$ , so randomization creates the distributions needed for inference and forms its "reasoned basis" in Fisher's phrase. Quantities which depend on treatment assignment,  $Z_i$ , like the observed response,  $R_i = Z_i r_{Ti} + (1 - Z_i) r_{Ci}$ , are random variables, whereas quantities that do not depend upon  $Z_i$ , like  $(r_{Ti}, r_{Ci})$  and  $h_i$ , are fixed features of the finite population of  $I$  units. To say that a quantity is fixed is to say that all probabilities are implicitly conditional probabilities given the values of fixed quantities. For instance, in an experiment in which treatments are assigned by the flip of a coin, the chance that  $i$  is exposed to treatment does not depend on any attribute that  $i$  might have,  $\frac{1}{2} = \Pr(Z_i = 1) = \Pr(Z_i = 1 | r_{Ti}, r_{Ci}, h_i)$ . This need not be true in an observational study where treatments are not randomly assigned.

**2. Two Examples and a Counterexample**

*2.1 Example: Do Daytime Running Lights Prevent Crashes?*

DRLs are always on but are only about  $\frac{1}{10}$  as bright as a car's headlights. DRLs have been required on new cars in Norway since 1985 and in Canada since 1989. By making a car more visible to other drivers, DRLs might prevent some multiple-vehicle accidents. DRLs have no effect at night, when much brighter ordinary headlights are in use, but DRLs may have an effect in daylight, when most headlights are off, or in twilight, when only some drivers have turned on headlights.

Table 2 is a small piece of an interesting study by Farmer and Williams (2002) concerning the effects of DRLs on multiple-vehicle crashes. Certain models of cars, pickups, and SUVs added DRLs in their 1996 model year, including various Cadillacs, Oldsmobiles, Volkswagen Passats, and Suburbans, among others. In Table 2, the treated group consists of the 1995 models of these cars, without DRLs, and the control

**Table 2**

*Distribution of multiple-vehicle crashes in Texas 1996–1998 by time of day and model year for cars that added DRLs in 1996. Twilight is excluded.*

	Model year 1995 $Z_i = 1$	Model year 1996 $Z_i = 0$	Total
Daylight $h_i = 1$	11,190	10,683	21,873
Darkness $h_i = 0$	2,836	2,876	5,712
Total	14,026	13,559	27,585

Source: Farmer and Williams (2002).

group consists of the 1996 models, with DRLs, so that  $A > 0$  would be expected if DRLs prevented more crashes than they caused. Table 2 is based on the State Data System and it counts multivehicle crashes in Texas between 1996 and 1998 in daylight and darkness with twilight excluded. Notice carefully that both the 1995 cars and the 1996 cars are being studied in the same years, 1996–1998. Only crashes are counted in Table 2; many cars of the type under study were not involved in crashes during 1996–1998 and Table 2 provides no information about such cars. Farmer and Williams's (2002) study is more extensive, involving other states, other model years, and some other data, and a reader interested in the effects of DRLs will want to examine the entire study; however, for the statistical method I wish to discuss, the small piece in Table 2 suffices. Because much brighter headlights are turned on in darkness, DRLs will have no effect in darkness,  $\delta_i = 0$  when  $h_i = 0$ , whereas in daylight,  $h_i = 1$ , it is possible that DRLs affect some crashes,  $\delta_i \neq 0$ .

*2.2 Example: Molecular Subtyping of Bacterial Diseases*

The cause of an outbreak of bacterial disease has often been determined using a case–control study that compares cases of the disease to healthy controls or noncases (e.g., Hennessy et al., 1996). Recently, McCarthy and Giesecke (1999, p. 767) suggested that the use of healthy controls in outbreak studies might be avoided by comparing cases of a disease to other cases of what is clinically the same disease, but is biologically a recognizably different strain of that disease. As they observe, this strategy reduces cost and increases speed, because the microbial or molecular subtyping of cases is necessary in any event. Moreover, because cases of disease are not typically aware of their subtype, various biases of recall and ascertainment may be avoided. For instance, to the extent that suffering from a gastrointestinal infection affects one's recall of foods eaten, different subtypes are likely to be similarly affected. Seeking medical attention is far from a universal response to a gastrointestinal infection, and to the extent that this leads to a usual selection of cases, it is likely to affect the two subtypes in a similar way.

Kist and Freitag (2000) used a case<sup>2</sup>-design in their study of Salmonella infections in Germany. Table 3 is their case<sup>2</sup>-comparison for the consumption of raw or undercooked eggs. Elsewhere in their study, Kist and Freitag (2000) also included noncases selected from local telephone directories, as might be done in a conventional case–referent study. The frequency of consumption of raw or undercooked eggs was similar

**Table 3**

*Salmonella* infections by subtype and consumption of raw or undercooked eggs

	Raw or undercooked eggs $Z_i = 1$	No raw or undercooked eggs $Z_i = 0$	Total
<i>Salmonella enteritidis</i> $h_i = 1$	267	26	293
Other <i>Salmonella</i> serovars $h_i = 0$	85	39	124
Total	352	65	417

Source: Kist and Freitag (2000).

for noncases and for cases of “other *Salmonella* serovars,” and that frequency was much lower than among cases of *Salmonella enteritidis*.

Other applications of the case<sup>2</sup>-design are given by de Valk et al. (2001), Zock et al. (2002), and the Campylobacter Sentinel Surveillance Scheme Collaborators (2002). Of course, some caution is needed, because a single source of infection could, in principle, provide several different types of bacteria, and a single type of bacterium might be distributed by several sources. The use of microbial or molecular subtyping does not, by any means, remove all ambiguities; see Kool et al. (2000) for a discussion of difficulties in the context of their study of an outbreak of Legionnaires’ disease in Los Angeles in 1997.

2.3 Counterexample: Suicides after the Publication of *Final Exit*

A key feature of the case<sup>2</sup>-design is that the second subtype of cases,  $h_i = 0$ , are, by their nature, entirely unaffected by the treatment. Daytime running lights may or may not prevent a particular vehicle-mile  $i$  driven in daylight from resulting in an accident, but it will not move that vehicle-mile to darkness. Avoiding an egg containing *S. enteritidis* may or may not prevent infection from *S. enteritidis*, but avoiding that egg will not cause infection with *S. typhimurium*. The situation is entirely different if the subtype,  $h$ , can be affected by the treatment, so that there is an  $h_{Ti}$  and an  $h_{Ci}$  that may be different, so the treatment can change the type of case  $i$  is without changing whether  $i$  is a case.

An example of an interesting comparison of two types of cases which is not a case<sup>2</sup>-study concerns the effects of a book, *Final Exit*, by Derek Humphry (1991), intended to provide advice about the practical aspects of suicide for the terminally ill. Following its publication in 1991, *Final Exit* was discussed in articles in the *New York Times* and the *Wall Street Journal*, and was on the *New York Times* bestseller list for 18 weeks. The book was highly controversial, in part because the ethics of suicide in general are controversial, and in part because the book might be used by individuals who were not terminally ill but merely depressed. In their article in the *New England Journal of Medicine*, Marzuk et al. (1993) “sought to determine whether the number of suicides involving methods recommended in *Final Exit* increased in New York City during the year after its publication.” The method recommended in *Final Exit* involves a plastic bag and lethal doses of medications (PB + M).

**Table 4**

Counterexample: Suicides in New York City before and after publication of *Final Exit*

	Year after <i>Final Exit</i> $Z_i = 1$	Year before <i>Final Exit</i> $Z_i = 0$	Total
PB + M	33	8	41
All other suicides	630	664	1294
Total	663	672	1335

Source: Marzuk et al. (1993).

Using data from the New York City medical examiner, Marzuk et al. (1993) constructed Table 4, which counts suicides by method in the year before and the year after publication of *Final Exit*, specifically March 1, 1990 to February 28, 1991 and March 1, 1991 to February 28, 1992. The first row of Table 4 refers to the method of suicide recommended in *Final Exit*, and the next row refers to all other methods, including falls from a height, firearms, hanging, and poisoning. In principle, the medical examiner should see all suspicious deaths, including all suicides, and should determine the cause of death by direct examination, so these may be more accurate than on death certificates. The total number of suicides is slightly lower after publication, but the number by the recommended method is four times higher.

Table 4 is not a case<sup>2</sup>-study because there is no subtype of cases, no row of the table, known to be unaffected by the publication of *Final Exit*. In particular, the publication of *Final Exit* might have caused some individuals  $i$  to commit suicide by PB + M who would not otherwise have committed suicide,  $\delta_i = 1$ , and *Final Exit* would cause them to be added to Table 4 and counted among the 33 in the upper left corner cell. Alternatively, the publication of *Final Exit* might have caused individuals who would have committed suicide in any event ( $\delta_i = 0$ ) to change to the recommended method PB + M,  $h_{Ti} - h_{Ci} = 1$ , and these cases would not be added to Table 4, but simply moved from the second row to the first, and again counted among the 33 in the upper left corner cell. In this sense, Table 4 is very different from Tables 2 and 3, and the methods proposed here do not apply to Table 4.

3. Inference about Attributable Effects

3.1 Inference with Random Exposure to Treatment

In this section, inferences are drawn about the net change in the number of cases in the database caused by exposure to the treatment,  $A = \sum_{i=1}^I Z_i \delta_i = \sum_{i=1}^I Z_i (r_{Ti} - r_{Ci})$ , under the assumption, made throughout this section, that treatments strike units at random, independently, with unknown constant probability  $\theta$ . This entails inverting Fisher’s exact test for a  $2 \times 2$  table in a new way reflecting the nature of the ascertainment of cases in a case<sup>2</sup>-study.

The counts in Table 1 may be rewritten as the counts in Table 5. The unobservable distribution in Table 6 describes the responses these same groups of units would have exhibited if the treatment had been withheld from all units, that is, it describes the  $r_{Ci}$ . What distinguishes Tables 5 and 6 is precisely the attributable effect  $A = \sum_{i=1}^I Z_i \delta_i = \sum_{i=1}^I Z_i h_i \delta_i$ . Specifically, if  $A$  is subtracted from the upper left corner of

**Table 5**  
*Distribution of case type by treatment*

	Treated $Z_i = 1$	Control $Z_i = 0$	Total
$h_i = 1$	$\sum Z_i h_i (r_{Ci} + \delta_i)$	$\sum (1 - Z_i) h_i r_{Ci}$	$\sum h_i (r_{Ci} + Z_i \delta_i)$
$h_i = 0$	$\sum Z_i (1 - h_i) r_{Ci}$	$\sum (1 - Z_i) (1 - h_i) r_{Ci}$	$\sum (1 - h_i) r_{Ci}$
Total	$\sum Z_i (r_{Ci} + h_i \delta_i)$	$\sum (1 - Z_i) r_{Ci}$	$\sum r_{Ci} + h_i Z_i \delta_i$

**Table 6**  
*Unobserved potential responses that would have been seen if the treated subjects had been spared the treatment*

	Treated $Z_i = 1$	Control $Z_i = 0$	Total
$h_i = 1$	$\sum Z_i h_i r_{Ci}$	$\sum (1 - Z_i) h_i r_{Ci}$	$\sum h_i r_{Ci}$
$h_i = 0$	$\sum Z_i (1 - h_i) r_{Ci}$	$\sum (1 - Z_i) (1 - h_i) r_{Ci}$	$\sum (1 - h_i) r_{Ci}$
Total	$\sum Z_i r_{Ci}$	$\sum (1 - Z_i) r_{Ci}$	$\sum r_{Ci}$

Table 5 and the marginal totals are adjusted, the result is Table 6. For instance, if  $A = 172$  in Table 3, then the four interior entries in Table 6 would be  $95 = 267 - 172$ , 26, 85, and 39.

The row marginal totals in Table 6 are fixed because they do not depend on  $Z_i$  (see Section 1.3). Given the assumption of this section that the  $\Pr(Z_i = 1) = \theta$  and the  $Z_i$  are independent, conditioning on  $\sum Z_i r_{Ci} = k$  fixes the column marginal totals of Table 6 and eliminates the unknown parameter  $\theta$ , so that the resulting conditional distribution is the hypergeometric distribution, and the unobservable  $Q = \sum Z_i h_i r_{Ci}$  has null distribution associated with Fisher's exact test for a  $2 \times 2$  table. If the treatment has no effect, so  $\delta_i = 0$  for every  $i$ , then Tables 5 and 6 are equal, so the hypothesis of no effect is tested by applying Fisher's exact test to the observed Table 5.

Write  $\delta = (\delta_1, \dots, \delta_I)$  and consider testing the hypothesis,  $H_0 : \delta = \delta_0$  for some specified  $\delta_0$ . The hypothesis  $H_0 : \delta = \delta_0$  is called incompatible if it is logically impossible given what is observed and assumed; otherwise, the hypothesis is compatible. Specifically,  $H_0 : \delta = \delta_0$  is incompatible if for some  $i$ , one of the following three conditions holds: (1)  $\delta_{0i} = 1$  for an  $i$  with  $Z_i = 0$ ,  $R_i = 1$ , or (2)  $\delta_{0i} = -1$  for an  $i$  with  $Z_i = 1$ ,  $R_i = 1$ , or (3)  $\delta_i \neq 0$  for an  $i$  with  $R_i = 1$ ,  $h_i = 0$ . An incompatible hypothesis can be rejected with certainty, that is, with type one error rate of 0. If the hypothesis  $H_0 : \delta = \delta_0$  is compatible, use the hypothesis to calculate  $A_0 = \sum_{i=1}^I Z_i h_i \delta_{0i}$ , and compute Table 7, whose corner cell and marginal totals differ from the

observed Table 1. If the hypothesis  $H_0 : \delta = \delta_0$  is true, then Table 7 equals Table 6, which has the hypergeometric distribution, so the hypothesis can be tested by applying Fisher's exact test to Table 7, that is, by comparing this table to the hypergeometric distribution.

The set of compatible hypotheses  $H_0 : \delta = \delta_0$  not rejected at level  $\alpha$  forms a  $100(1 - \alpha)\%$  confidence set  $\mathcal{C}$  for  $\delta$  (see Lehmann, 1986, Section 3.5). Because  $\delta$  is  $I$ -dimensional, the confidence set  $\mathcal{C}$  is awkward to inspect, but it is easy to describe  $\mathcal{C}$  in terms of  $A$ . Specifically,  $\delta_0 \in \mathcal{C}$  if and only if  $A_0 = \sum_{i=1}^I Z_i h_i \delta_{0i}$  is not rejected by Fisher's exact test, so that it is plausible that there were a net increase of  $A_0$  events in the database due to effects caused by the treatment.

The procedure just described for case<sup>2</sup>-studies differs in an important respect from the corresponding procedure for cohort studies described in Rosenbaum (2001). Specifically, in cohort studies, the formal hypothesis test leads the  $A_0$  units to be moved, not removed, as in Table 7. In a cohort study, the  $A_0$  units are moved from the upper left corner cell and placed into the lower left corner cell, altering the row margins but not the column margins. The difference is that, in a case<sup>2</sup>-study, an observed case  $i$  caused by the treatment,  $R_i = 1$  and  $Z_i \delta_i = 1$ , would have been missing from the database had this unit escaped the treatment.

To illustrate the method, suppose in Table 3 that exposure to raw or undercooked eggs struck the  $I$  individuals in the population at random with unknown but constant probability  $\theta$ , so that the chance of exposure  $Z_i = 1$  did not vary

**Table 7**  
*Observed distribution adjusted for the hypothesis*

	Treated $Z_i = 1$	Control $Z_i = 0$	Total
$h_i = 1$	$\sum Z_i h_i R_i - A_0$	$\sum (1 - Z_i) h_i R_i$	$\sum h_i R_i - A_0$
$h_i = 0$	$\sum Z_i (1 - h_i) R_i$	$\sum (1 - Z_i) (1 - h_i) R_i$	$\sum (1 - h_i) R_i$
Total	$\sum Z_i R_i - A_0$	$\sum (1 - Z_i) R_i$	$\sum R_i - A_0$

with an individual's potential responses  $(r_{Ti}, r_{Ci})$  to exposure. The hypothesis of no effect,  $H_0: r_{Ti} = r_{Ci}$  for  $i = 1, \dots, I$  is tested by applying Fisher's exact test to Table 3, yielding a one-sided significance level of  $2.8 \times 10^{-8}$ . In fact, any compatible hypothesis  $H_0: \delta = \delta_0$  that attributes fewer than  $A_0 = \sum_{i=1}^I Z_i h_i \delta_{0i} = 171$  extra infections to raw or undercooked eggs is rejected with one-sided significance level  $\leq 0.05$ , whereas any compatible hypothesis that attributes at least  $A_0 = \sum_{i=1}^I Z_i h_i \delta_{0i} = 172$  extra infections to this exposure is accepted with one-sided significance level  $> 0.05$ . That is, assuming constant risk  $\theta$  of exposure, we are 95% confident that at  $172/267 = 64\%$  of the cases of *S. enteritidis* were extra cases caused by raw or undercooked eggs. Specifically, if Table 3 is adjusted, as in Table 7, with  $A_0 = 171$ , the one-sided significance level from Fisher's exact test is 0.048, whereas with  $A_0 = 172$  it is 0.052.

A table with large counts, such as Table 2, may be adjusted, as in Table 7, and the large sample normal approximation to the hypergeometric distribution, with continuity correction, may be used to approximate the tail probability. In Table 2, each compatible hypothesis  $H_0: \delta = \delta_0$  with  $A_0 = 127$  yields a one-sided significance level of 0.04999, whereas each hypothesis with  $A_0 = 128$  yields 0.05031, so the one-sided 95% confident set has  $A_0 \geq 128$  daytime accidents due to the absence of DRLs, or  $128/11,190 = 1.1\%$  of such crashes.

3.2 Sensitivity to Hidden Bias

The confidence statements in Section 3.1 were based on the premise that exposure to treatment occurred at random with constant probability  $\theta$ . In this section, it is assumed instead that each unit  $i$  has a different probability,  $\theta_i$ , of exposure to treatment, say consumption of raw eggs, which may differ because the units are heterogeneous in ways relevant to  $(r_{Ti}, r_{Ci}, h_i)$ . The first sensitivity analysis in an observational study was conducted by Cornfield et al. (1959) to clarify conflicting claims about the effects of smoking on lung cancer. The specific method discussed here is described in detail in Rosenbaum (1995, 2002b, Section 4.4).

The model for sensitivity analysis says that, because the units under study are not homogeneous, two units  $i$  and  $j$  may differ in their odds of exposure to treatment by a factor of  $\Gamma \geq 1$ ,

$$\frac{1}{\Gamma} \leq \frac{\theta_i(1 - \theta_j)}{\theta_j(1 - \theta_i)} \leq \Gamma \quad \text{for all } 1 \leq i \leq j \leq I. \quad (1)$$

If  $\Gamma = 1$ , then the  $\theta_i$  are all equal, leading to the analysis in Section 3.1, with a single significance level or a single lower endpoint for a confidence interval. If  $\Gamma > 1$ , then the chances of exposure to treatment vary from person to person in ways that are unknown but are bounded in magnitude, and this leads, not to a single inference, but to a bounded range of inferences, for instance, a range of possible significance levels, or a range of endpoints for a confidence interval. The sensitivity bounds are obtained by comparing the adjusted Table 7 to two extended hypergeometric distributions, with parameters  $\frac{1}{\Gamma}$  and  $\Gamma$  (see Rosenbaum, 1995, 2002b, Section 4.4.1).

Table 8 displays the sensitivity analysis for the data in Table 3 concerning infection from raw eggs. For  $\Gamma = 1$ , there is the single confidence interval,  $A \geq 172$ , discussed in Section 3.1. For  $\Gamma > 1$ , the inequality (1) produces a range of con-

Table 8

Sensitivity analysis for *Salmonella enteritidis* infections: minimum endpoints for confidence interval for the number of cases attributable to raw or undercooked eggs

$\Gamma$	Minimum 95% interval
1	$A \geq 172$
1.5	$A \geq 126$
2	$A \geq 79$
2.5	$A \geq 33$
3	n.s.

fidence intervals, and Table 8 reports the smallest of these. That is, different unknown  $\theta_i$ 's produce different confidence intervals, but every pattern of  $\theta_i$ 's satisfying (1) with  $\Gamma = 2$  yields a confidence interval whose lower endpoint is at least 79, and one pattern of  $\theta_i$ 's satisfying (1) with  $\Gamma = 2$  yields exactly the interval  $A \geq 79$ . If one person is at most twice as likely as another to be exposed to raw eggs, we remain 95% confident that at least  $A \geq 79$  of the 293 *S. enteritidis* infections, or 27%, were added cases caused by the raw eggs. A larger bias,  $\Gamma = 3$ , just barely explains away the observed association: The maximum  $P$ -value for testing no effect for  $\theta_i$ 's satisfying (1) with  $\Gamma = 3$  is 0.072. Compared to examples in Rosenbaum (2002, Section 4), Table 8 is insensitive to moderately large biases, but more sensitive than studies of heavy smoking as cause of lung cancer.

3.3 Matched Case<sup>2</sup>-Studies

In Table 2, one might wish to compare Passats to other Passats, Suburbans to other Suburbans, highway accidents to other highway accidents, and so on, and this could be done by matching daylight accidents to similar darkness accidents. That is, the chance of exposure,  $\theta_i$ , might vary from one unit  $i$  to another as a function of observed covariates, say  $\mathbf{x}_i$ , but this might be controlled by matching on  $\mathbf{x}_i$ .

In Section 3, Fisher's exact test was inverted in a different way in a case<sup>2</sup>-study than in a cohort study, because preventing a case deletes the case from the database system. For a technical reason, the situation is simpler with matching, and the methods in Rosenbaum (2002a) apply directly. This concise paragraph explains why. In the database, there are  $J$  possibly affected cases,  $j = 1, \dots, J$ , of the first type,  $h = 1$ , and the  $j$ th such case is matched on  $\mathbf{x}$  to  $n_j - 1 \geq 1$  unaffected cases of type  $h = 0$  from the database. In set  $j$ , the first case,  $k = 1$ , is the possibly affected case  $h_{j1} = 1$ , the rest  $k = 2, 3, \dots, n_j$  are the unaffected cases of the second type,  $h_{jk} = 0$ . Quantities in Section 1.2 are unchanged in meaning but now have two subscripts:  $Z_{jk}, h_{jk}, (r_{Tjk}, r_{Cjk}), \delta_{jk} = r_{Tjk} - r_{Cjk}, R_{jk}, \theta_{jk}$ , and  $m_j = \sum_{k=1}^{n_j} Z_{jk}$  cases were exposed in set  $j$ . There are two assumptions: (i) As in Section 3,  $h_{jk} = 0$  implies  $\delta_{jk} = 0$ , so  $R_{jk} = r_{Cjk}$  if  $h_{jk} = 0$  and  $R_{jk} = r_{Cjk} + Z_{jk}\delta_{jk}$  if  $h_{jk} = 1$ , and (ii) to be compatible with Rosenbaum (2002a),  $\delta_{jk} \geq 0$  or  $r_{Tjk} \geq r_{Cjk}$ , so preventing an exposure would never cause an additional case. A compatible hypothesis  $H_0: \delta = \delta_0$  is tested by assuming the hypothesis for the purpose of testing it, and applying the McNemar–Mantel–Haenszel test  $T = \sum_{j=1}^J Z_{j1}r_{Cj1}$  to the adjusted responses  $r_{Cjk} = R_{jk} - Z_{jk}\delta_{0jk}$ . Under  $H_0$ , if the chance of exposure is constant within matched sets,  $\theta_{jk} = \theta_{j'k}$

for each  $j, k, k'$ , then the conditional distribution of  $Z_{jk}$  given  $m_j$  does not depend on  $\theta_{jk}$ , and  $T$  is that of the sum of  $J$  independent Bernoulli variables with probabilities of success  $\Pr(r_{Cj1}Z_{j1} = 1 | m_j) = r_{Cj1}m_j/n_j = \pi_j$ , say (see Cox, 1966). The key technical point is the following: If a compatible hypothesis,  $H_0: \delta = \delta_0$ , attributes the first case in matched set  $j$  to an effect of the treatment—that is, if  $Z_{jk}\delta_{0jk} = 1$  so that  $R_{jk} = 1$  but  $r_{Cjk} = R_{jk} - Z_{jk}\delta_{0jk} = 0$ —then matched set  $j$  becomes concordant because  $\pi_j = r_{Cj1}m_j/n_j = 0$ , which is, for all practical purposes, the same as deleting matched set  $j$ .

#### ACKNOWLEDGEMENT

Supported by grant SES-0345113 from the U.S. National Science Foundation.

#### RÉSUMÉ

On extrait d'une base de données des événements et la présence ou l'absence d'exposition antérieure à un traitement défini, dans le but de déterminer si l'exposition provoque l'événement. Autrement dit, le système n'enregistre l'information que sur les cas. Il ne fournit aucune information sur les événements potentiels qui ne sont finalement pas produits, c'est-à-dire sur les unités qui ne sont pas des cas. On connaît donc le nombre de succès dans deux groupes, traités et témoins, mais non le nombre d'essais au dénominateur des proportions correspondantes; en fait, le concept d'essai peut être vague. Sans autre information, la situation est tout à fait désespérée. Cependant une stratégie fructueuse consiste à identifier deux types de cas dont l'origine est totalement différente, de sorte qu'il est certain que ceux du second type ne peuvent en aucune manière être affectés par le traitement étudié. Cette stratégie - l'étude cas-cas ou cas<sup>2</sup> - semble avoir été réinventée indépendamment à de multiples reprises et McCarthy et Giesecke (1999) en ont fait une stratégie générale en épidémiologie des maladies infectieuses. Permet-elle d'estimer le nombre de cas dus au traitement? nous proposons une méthode d'inférence exacte, ainsi qu'une approximation pour de grand échantillons, qui repose sur une nouvelle utilisation des effets attribuables. Nous présentons deux exemples : l'un concerne l'effet des phares diurnes sur le risque de collision entre plusieurs véhicules; l'autre concerne l'origine des infections à salmonelles. Nous discutons également un contre-exemple d'apparence superficiellement similaire; il concerne les taux de suicides après la sortie de «last exit» et, ici, le traitement peut affecter le nombre d'événements ou il peut en modifier le type, d'où l'impossibilité d'estimer un effet attribuable.

#### REFERENCES

- Campylobacter Sentinel Surveillance Scheme Collaborators. (2002). Ciprofloxacin resistance in *Campylobacter jejuni*: Case-case analysis as a tool for elucidating risks at home and abroad. *Journal of Antimicrobial Chemotherapy* **50**, 561–568.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959). Smoking and lung cancer. *Journal of the National Cancer Institute* **22**, 173–203.
- Cox, D. R. (1966). A simple example of a comparison involving quantal data. *Biometrika* **53**, 215–220.
- de Valk, H., Vaillant, V., Jacquet, C., Rocourt, J., Le Querrec, F., Stainer, F., Quelquejeu, N., Pierre, O., Pierre, V., Desenclos, J. C., and Goulet, V. (2001). Two consecutive nationwide outbreaks of listeriosis in France. *American Journal of Epidemiology* **154**, 944–950.
- Farmer, C. M. and Williams, A. F. (2002). Effects of daytime running lights on multiple-vehicle daylight crashes in the United States. *Accident Analysis and Prevention* **34**, 197–203.
- Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics* **6**, 13–25.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.
- Hamilton, M. A. (1979). Choosing the parameter for  $2 \times 2$  or  $2 \times 2 \times 2$  table analysis. *American Journal of Epidemiology* **109**, 362–375.
- Hennessy, T. W., Hedberg, C. W., Slutsker, L., White, K. E., Besser-Wiek, J. M., Moen, M. E., Feldman, J., Coleman, W. W., Edmonson, L. M., MacDonald, K. L., and Osterholm, M. T. (1996). A national outbreak of *Salmonella enteritidis* infections from ice cream. *New England Journal of Medicine* **334**, 1281–1286.
- Humphry, D. (1991). *Final Exit: The Practicalities of Self-Deliverance and Assisted Suicide for the Dying*. Eugene, Oregon: Hemlock Society.
- Kist, M. J. and Freitag, S. (2000). Serovar specific risk factors and clinical features of *Salmonella enterica* ssp. *enterica* serovar enteritidis: A study in South-West Germany. *Epidemiology and Infection* **124**, 383–392.
- Kool, J. L., Buchholz, U., Peterson, C., et al. (2000). Strengths and limitations of molecular subtyping in a community outbreak of Legionnaires' disease. *Epidemiology and Infection* **125**, 599–608.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edition. New York: Wiley.
- Lehmann, E. L. (1998). *Nonparametrics*. Upper Saddle River, New Jersey: Prentice Hall.
- Marzuk, P. M., Tardiff, K., Hirsch, C. S., Leon, A. C., Stajic, M., Hartwell, N., and Portera, L. (1993). Increase in suicide by asphyxiation in New York City after the publication of *Final Exit*. *New England Journal of Medicine* **329**, 1508–1510.
- McCarthy, N. and Giesecke, J. (1999). Case-case comparisons to study causation of common infectious diseases. *International Journal of Epidemiology* **28**, 764–768.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9 (in Polish), *Roczniki Nauk Rolniczych*, Tom X, p. 1–51, reprinted in English with discussion in *Statistical Science* 1990, **5**, 463–480.
- Rosenbaum, P. R. (1995). Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association* **90**, 1424–1431.
- Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika* **88**, 219–231.
- Rosenbaum, P. R. (2002a). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association* **97**, 183–192.

- Rosenbaum, P. R. (2002b). *Observational Studies*. New York: Springer-Verlag.
- Rosenbaum, P. R. (2003). Exact confidence intervals for nonconstant effects by inverting the signed rank test. *American Statistician* **57**, 132–138.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.
- Zock, J. P., Kogevinas, M., Sunyer, J., Jarvis, D., Toren, K., and Anto, J. M. (2002). Asthma characteristics in cleaning workers, workers in other risk jobs and office workers. *European Respiratory Journal* **20**, 679–685.

*Received September 2003. Revised March 2004.*

*Accepted April 2004.*