Statistical Science
1999, Vol. 14, No. 1, 29–46

Confounding and Collapsibility in Causal Inference

Sander Greenland, James M. Robins and Judea Pearl

Abstract. Consideration of confounding is fundamental to the design and analysis of studies of causal effects. Yet, apart from confounding in experimental designs, the topic is given little or no discussion in most statistics texts. We here provide an overview of confounding and related concepts based on a counterfactual model for causation. Special attention is given to definitions of confounding, problems in control of confounding, the relation of confounding to exchangeability and collapsibility, and the importance of distinguishing confounding from noncollapsibility.

Key words and phrases: Bias, causation, collapsibility, confounding, contingency tables, exchangeability, observational studies, odds ratio, relative risk, risk assessment, Simpson's paradox.

Much of epidemiologic and social science research is devoted to estimation of causal effects and testing causal hypotheses using nonexperimental data. In such endeavors, issues of confounding will (or should) invariably arise. Unfortunately, the word "confounding" has been used to refer to at least three distinct concepts. In the oldest usage, confounding is a type of bias in estimating causal effects. This bias is sometimes informally described as a mixing of effects of extraneous factors (called confounders) with the effect of interest. This usage predominates in nonexperimental research, especially in epidemiology and sociology. In a second and more recent usage, "confounding" is a synonym for "*noncollapsibility*," although this usage is often limited to situations in which the parameter of interest is a causal effect. In a third usage, originating in the experimental-design literature, "confounding" refers to inseparability of main effects and interactions under a particular design. The term "*aliasing*" is also sometimes used to refer to the latter concept; this usage is common in the analysis-of-variance literature.

The three concepts of confounding are not always distinguished properly. In particular, the concepts of confounding as a bias in effect estimation and as noncollapsibility are often treated as identical. We here provide an historical overview of these two concepts and the distinctions among them. Because these distinctions require a formal model for causal effects, we begin with a discussion of the counterfactual model of causation. We then trace the history of the concept of confounding from the writings of J. S. Mill to its modern counterfactual formalization. We discuss how approaches to control of confounding fit into this formalization; we give special attention to the relation of confounding to exchangeability and randomization. We then shift our focus to concepts of collapsibility and describe how the counterfactual model distinguishes noncollapsibility from confounding. Our penultimate section covers some miscellaneous issues that arise when considering confounding in studies of interventions. We end with a recommendation to include more thorough discussion of confounding in basic statistics education, given the importance of the concept in causal inference.

1. THE COUNTERFACTUAL APPROACH TO CAUSE AND EFFECT

1.1 Overview

The concepts of cause and effect are central to most areas of scientific research. Thus, it may be surprising that consensus about basic definitions and methods for causal inference is limited, despite

Sander Greenland is Professor, Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095-1772. James M. Robins is Professor, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. Judea Pearl is Professor, Department of Computer Science, UCLA School of Engineering and Applied Science, Los Angeles, CA 90095-1596.

some three centuries of debate. A brief review cannot do justice to all the history and details of this debate, nor to all the schools of thought on causation. We therefore focus on one conceptualization that has proved useful in the analysis of confounding. This *counterfactual or potential-outcomes* approach has become common in philosophy, statistics and epidemiology.

Since the early eighteenth century, philosophers noted serious deficiencies in common definitions of causation. For example, *Webster's New Twentieth-Century Dictionary* (1979) offered "that which produces an effect or result" as a definition of "cause," but "to cause" is among the definitions of "produces." Informal definitions of "effect" suffer from the same circularity, because "effect" as a verb is merely a synonym for "cause," while "effect" as a noun is defined as a "result," which is in turn defined as an "effect" in causal contexts.

Hume (1739, 1748) offered another view of causation that pointed a way out of the circularity of common definitions:

> We may define a cause to be an object, followed by another, . . . where, if the first object had not been, the second had never existed (Hume, 1748, page 115).

Thus, by focusing on specific instances of causation, we say that an event A caused an event B if occurrence of A was necessary for occurrence of B under the observed background circumstances (e.g., see Simon and Rescher, 1966; Lewis, 1973a). Essentially the same concept of causation can be found in the works of J. S. Mill (1843, 1862) and R. A. Fisher (1918) (both quoted in Rubin, 1990), as well as in later works in statistics and related fields (Sobel, 1995). A typical example is from MacMahon and Pugh (1967, page 12), who state that ". . . an association may be classed as presumptively causal when it is believed that, *had the cause been* altered, the effect *would have been* changed" (italics added). The italicized phrases emphasize that the alteration of the antecedent condition ("cause") and the subsequent change in the outcome ("effect") are contrary to what was in fact observed; that is, they are *counterfactual*.

The preceding definition falls short of the formalism necessary for derivation of statistical methods for causal inference. Such a formalism and derivation first appeared in the statistics literature in Neyman (1923). The basic idea is as follows: suppose that N units (e.g., individuals, populations, objects) are to be observed in an experiment that will assign each unit one of $K + 1$ treatments x_0, $x_1, \dots, x_K$. The outcome of interest for unit i is the value of a response variable Y_i. Suppose that Y_i will equal y_{ik} if unit i is assigned treatment x_k. Usually, one treatment level, say x_0, is designated the reference treatment against which other treatments are to be evaluated; typically, x_0 is "no treatment," a placebo or a standard treatment. We define the *causal effect* of $x_k (k \geq 1)$ on Y_i relative to x_0 (the referent) to be $y_{ik} - y_{i0}$. (If the response variable is strictly positive, we may instead define the causal effect as y_{ik}/y_{i0} or $\log y_{ik} - \log y_{i0}$.) In words, a causal effect is a contrast between the outcomes of a *single* unit under different treatment possibilities.

Neyman's formalism is sometimes referred to as the *potential-outcomes* model of causation, and has reappeared in various guises (e.g., see Cox, 1958, Chapter 2; Copas, 1973; Rubin, 1974; Hamilton, 1979). Defining effects as contrasts of potential outcomes y_{ik} gives precise meanings to words such as "cause," "effect," and "affect." For example, "changing X from x_0 to x_k affects Y_i" is an assertion that $y_{ik} - y_{i0} \neq 0$. Note, however, that because only one of the potential outcomes y_{ik} can be observed in any one unit, an individual effect $y_{ik} - y_{i0}$ cannot be observed.

Counterfactual analysis can be viewed as a special type of latent-variable analysis, in which y_{ik} remains latent for any individual i who did not receive treatment k (e.g., see the volume edited by Berkane, 1997). The potential outcomes model can also be derived from a structural equations approach familiar in the social sciences. Here, one models the response variable Y as one output of a series of mechanisms, where each mechanism is an input–output device whose behavior follows a given equation (Simon and Rescher, 1966). The potential response y_{ik} is then simply the solution for Y of the system of equations when X is "set" to x_k; a change from x_0 to x_k has no effect if X does not appear in the system, or more generally, if the solution is the same regardless of whether X is set to x_0 or x_k (Balke and Pearl, 1994; Robins, 1995a; Pearl, 1995; Galles and Pearl, 1998).

There are several crucial restrictions that the potential outcomes definition places on the notion of causal effects (and hence, cause). Appendix 1 discusses four of them in detail. In Appendix 2, we discuss some difficulties that arise in defining potential outcomes when competing risks are present.

1.2 Probabilistic Extensions

There are several probabilistic extensions of counterfactual approaches. One is based on considering the sampling distribution of fixed potential outcomes, that is, the joint distribution $F(y_0, \dots, y_K)$ of $y_{i0}, \dots, y_{iK}$ in a population of units. We may also consider conditional distributions of potential outcomes in subpopulations defined by covariates such as age, sex and received treatment. Population effects can be defined as differences in average population response under different treatments, or more generally as differences among the marginal distributions $F(y_0), \dots, F(y_K)$. Statistical procedures for inferences about these effects follow from randomization assumptions about treatment assignment mechanisms and from assumptions of independencies between units (Cox, 1958). The basic ideas were developed by Neyman (1923) and Fisher (1935); some key elaborations were given by Copas (1973) and Rubin (1974, 1978). We will discuss these ideas below.

Another extension considers parameters of probability distributions (rather than events) as the potential outcomes; this extension addresses objections to treating the unit outcomes as deterministic entities once treatment is given (Greenland, 1987; Robins, 1988; Robins and Greenland, 1989). For example, we could consider the difference between the probability that a given atom emits a photon in the second following absorption of a photon ("treatment 1") and the probability of emission in the same second if no photon had been absorbed ("treatment 0"). This probability difference is the effect of photon absorption on the atom relative to no absorption. In quantum mechanics, this probability difference (effect) is well defined whether or not a photon is actually emitted (e.g., see Feynman, 1963). Yet, according to the Bohr–Heisenberg ("Copenhagen") interpretation of quantum theory, the emission indicator ($Y_i = 1$ if the atom emits a photon in the following second; 0 if not) is undefined under counterfactual alternatives to the actual history of the atom and is not even defined under the actual history of the atom if no emission measurement is made.

1.3 Objections to Counterfactuals

Counterfactual approaches are sometimes criticized because, in considering causes of past events, they invoke distributions for events that never occurred and hence cannot be observed. As a consequence, some important features of these distributions remain empirically untestable, and thus some causal inferences based on counterfactuals will depend entirely on untestable assumptions (Dawid, 1998).

It is our view that this property of counterfactual inferences reflects a strength of counterfactual approach, rather than a weakness. It is an unfortunate but true fact that many important causal questions are simply not answerable, at least not without employing assumptions that are untestable given current technology. Examples of such assumptions include assumptions of no confounding, as discussed in the following sections, assumptions about independence of unit-specific susceptibilities or responses, and various distributional assumptions (Copas, 1973; Rubin, 1978, 1991; Holland, 1986; Heckman and Hotz, 1989; Robins and Greenland, 1989; Sobel, 1995; Rosenbaum, 1995; Copas and Li, 1997). Inferences from counterfactual approaches properly reflect this harsh epistemic reality when they display sensitivity to such assumptions.

More constructively, the counterfactual approach also aids in precise formulation of assumptions needed to identify causal effects statistically, which in turn can aid in developing techniques for meeting those assumptions. The basic example on which we will focus is the assumption of exchangeability of response distributions under homogeneous treatment assignment, which is met when treatment is successfully randomized, or, more generally, when treatment assignment is independent of the potential outcomes y_{ik}.

2. CONFOUNDING

2.1 Background

Counterfactual approaches to causal inference emphasize the importance of randomization in assuring identifiability of causal effects (Neyman, 1923; Rubin, 1978, 1990, 1991; Greenland and Robins, 1986; Robins, 1986; Greenland, 1990; Rosenbaum, 1995). In observational studies, however, no such assurance is available, and issues of confounding become paramount.

One of the earliest systematic discussions of "confounded effects" is Chapter X of Mill (1843), "Of Plurality of Causes, and the Intermixture of Effects" (although in Chapter III Mill lays out the primary issues and acknowledges Francis Bacon as a forerunner in dealing with them). There, Mill listed a requirement for an experiment intended to determine causal relations:

> ... none of the circumstances [of the experiment] that we do know shall have effects susceptible of being *confounded with* those of the agents whose properties we wish to study [emphasis added].

It should be noted that, in Mill's time, the word "experiment" referred to an observation in which some circumstances were under the control of the observer, as it still is used in ordinary English, rather than to the notion of a comparative trial. Nonetheless, Mill's requirement suggests that a comparison is to be made between the outcome of his experiment (which is, essentially, an uncontrolled trial) and what we would expect the outcome to be if the agents we wish to study had been absent. If the outcome is not that which one would expect in the absence of the study agents, his requirement insures that the unexpected outcome was not brought about by extraneous circumstances. If, however, those circumstances do bring about the unexpected outcome, and that outcome is mistakenly attributed to effects of the study agents, the mistake is one of confounding (or confusion) of the extraneous effects with the agent effects.

Much of the modern literature follows the same informal conceptualization given by Mill. Terminology is now more specific, with "treatment" used to refer to an agent administered by the investigator and "exposure" often used to denote an unmanipulated agent. The chief development beyond Mill is that the expectation for the outcome in absence of the study exposure is now almost always explicitly derived from observation of a control group that is untreated or unexposed. For example, Clayton and Hills (1993, page 133) state that, in observational studies,

> ... there is always the possibility that an important influence on the outcome ... differs systematically between the comparison [exposed and unexposed] groups. It is then possible [that] part of the apparent effect of exposure is due to these differences, [in which case] the comparison of the exposure groups is said to be *confounded*. [emphasis in the original]

As discussed below, confounding is also possible in randomized experiments, because of systematic elements in treatment allocation, administration, and compliance and because of random differences between comparison groups (Fisher, 1935, page 49; Rothman, 1977; Greenland and Robins, 1986; Greenland, 1990).

2.2 Formalization

Attempts to quantify the above notion of confounding can be traced at least as far back as the work of Karl Pearson and George Yule on spurious correlation, but these attempts ran afoul of the absence of a formal model for causal effects; see Aldrich (1995) for a review of this work. Various mathematical formalizations of confounding have since been proposed. Perhaps the one closest to Mill's concept is based on the counterfactual model for effects. Suppose our objective is to determine the effect of applying a treatment or exposure x_1 on a parameter μ of the distribution of the outcome y in population A, relative to applying treatment or exposure x_0. That is, we wish to contrast the marginal distributions $F_A(y_1)$ and $F_A(y_0)$ of the potential outcomes under treatments 1 and 0, using some parameter (summary) μ of the distributions. For example, population A could be a cohort of breast-cancer patients, treatment x_1 could be a new hormone therapy, x_0 could be a placebo therapy, and the parameter μ could be the expected survival or the five-year survival probability in the cohort; μ could also be a vector or even a function, such as an entire survival curve. The population A is sometimes called the *target population* or *index population*; the treatment x_1 is sometimes called the *index* treatment and the treatment x_0 is sometimes called the *control* or *reference* treatment.

Suppose that μ will equal μ_{A1} if x_1 is applied to population A, and will equal μ_{A0} if x_0 is applied to that population; the causal effect of x_1 relative to x_0 is defined as the change from μ_{A0} to μ_{A1}, which could be measured by $\mu_{A1} - \mu_{A0}$ (or by μ_{A1}/μ_{A0} if μ is strictly positive). If A is observed under treatment x_1, μ will equal μ_{A1}, which is observable or estimable, but μ_{A0} will be unobserved. Suppose, however, we expect μ_{A0} to equal μ_{B0}, where μ_{B0} is the value of the outcome μ observed or estimated for a population B that was administered treatment x_0. The latter population is sometimes called a *control* or *reference* population. We say *confounding* is present if, in fact, $\mu_{A0} \neq \mu_{B0}$, for then there must be some difference between populations A and B (other than treatment) that is responsible for the discrepancy between μ_{A0} and μ_{B0}.

If confounding is present, a naive (crude) association parameter obtained by substituting μ_{B0} for μ_{A0} in the effect measure will not equal the causal parameter, and the association parameter is said to be *confounded*. For example, if $\mu_{B0} \neq \mu_{A0}$, then $\mu_{A1} - \mu_{B0}$, which measures the *association* of treatments with outcomes *across* the populations, is confounded for $\mu_{A1} - \mu_{A0}$, which measures the *effect* of treatment x_1 on population A. Thus, saying an association parameter such as $\mu_{A1} - \mu_{B0}$ is confounded for a causal parameter such as $\mu_{A1} - \mu_{A0}$ is synonymous with saying the two parameters are not equal.

The above formalization has several interesting implications. One is that confounding depends on the outcome parameter. For example, suppose populations A and B would have different five-year survival probabilities μ_{A0} and μ_{B0} under placebo treatment x_0, so that $\mu_{A1} - \mu_{B0}$ is confounded for the actual effect $\mu_{A1} - \mu_{A0}$ of treatment on five-year survival. It is then still possible that ten-year survival ν under the placebo would be identical in both populations; that is, ν_{A0} could still equal ν_{B0}, so that $\nu_{A1} - \nu_{B0}$ is not confounded for the actual effect of treatment on ten-year survival. (We should generally expect no confounding for 200-year survival, because no treatment is likely to raise the 200-year survival probability of human patients above zero.)

Another important implication is that confounding depends on the target population of inference. The preceding example, with A as the target, had different five-year survivals μ_{A0} and μ_{B0} for A and B under placebo therapy, and hence $\mu_{A1} - \mu_{B0}$ was confounded for the effect $\mu_{A1} - \mu_{A0}$ of treatment on population A. A lawyer or ethicist may also be interested in what effect the treatment would have had on population B. Writing μ_{B1} for the (unobserved) outcome of B under treatment, this effect on B may be measured by $\mu_{B1} - \mu_{B0}$. Substituting μ_{A1} for the unobserved μ_{B1} yields $\mu_{A1} - \mu_{B0}$. This measure of association is confounded for $\mu_{B1} - \mu_{B0}$ (the effect of treatment x_1 on five-year survival in population B) if and only if $\mu_{A1} \neq \mu_{B1}$. Thus, the same measure of association $\mu_{A1} - \mu_{B0}$ may be confounded for the effect of treatment on neither, one or both of populations A and B.

A third implication is that absence of confounding ($\mu_{A0} = \mu_{B0}$), which is a population condition, is not sufficient to identify the sharp null hypothesis of no causal effects at the unit level ($y_{i1} = y_{i0}$ for all units i) because causal effects of treatment may cancel out (Greenland and Robins, 1986). For example, suppose the outcome parameter μ is expected response and response is binary, with half of units in A and half in B having $y_{i1} = 1$, $y_{i0} = 0$ and half having $y_{i1} = 0$, $y_{i0} = 1$. Then $\mu_{A1} = \mu_{A0} = \mu_{B0} = 1/2$, so that there is no confounding and no identifiable effect of treatment on the outcome distribution; nonetheless, *every* unit is affected by treatment. Neyman (1935) and Stone (1993) make the analogous point that randomization does not identify the sharp null hypothesis.

A noteworthy aspect of the above definition of confounding is that it does *not* involve the notion of probabilistic independence, and makes no reference to individual units or probability distributions other than through the summary μ. This is in sharp contrast to concepts of randomization, exchangeability and ignorability, as well as certain definitions of "no confounding," which we will discuss in Sections 3.1, 3.3 and 6.4.

2.3 Components of Associations

We may write the difference in the outcome parameters of populations A and B as

$$(1) \quad \mu_{A1} - \mu_{B0} = (\mu_{A1} - \mu_{A0}) + (\mu_{A0} - \mu_{B0}),$$

which shows that $\mu_{A1} - \mu_{B0}$ is a mix of the true treatment effect $\mu_{A1} - \mu_{A0}$ and a bias term $\mu_{A0} - \mu_{B0}$ (Groves and Ogburn, 1928; Kitagawa, 1955). Nonidentifiability of the true effect $\mu_{A1} - \mu_{A0}$ follows if the bias $\mu_{A0} - \mu_{B0}$ is not identifiable, as is the case in typical epidemiologic studies (Greenland and Robins, 1986).

By rearranging (1) we may obtain $\mu_{A0} - \mu_{B0}$ as a measure of confounding in $\mu_{A1} - \mu_{B0}$:

$$(2) \quad \mu_{A0} - \mu_{B0} = (\mu_{A1} - \mu_{B0}) - (\mu_{A1} - \mu_{A0}).$$

When the outcome parameters μ are risks (probabilities), epidemiologists use instead the analogous ratio

$$(3) \qquad \frac{\mu_{A1}/\mu_{B0}}{\mu_{A1}/\mu_{A0}} = \frac{\mu_{A0}}{\mu_{B0}}$$

as a measure of confounding (Cornfield et al., 1959; Bross, 1967; Miettinen, 1972); μ_{A0}/μ_{B0} is sometimes called the *confounding risk ratio*. The latter term is somewhat confusing, as it is sometimes misunderstood to refer to the effect of a particular confounder on risk. This is not so, although the ratio does reflect the net effect of the differences in the confounder distributions of populations A and B.

2.4 Confounders

The above formalization of confounding invokes no explicit differences (imbalances) between populations A and B with respect to circumstances or covariates that might affect μ (Greenland and Robins, 1986). It seems intuitively clear that, if μ_{A0} and μ_{B0} differ, then A and B must differ with respect to factors that affect μ. This intuition has led some authors to define confounding in terms of differences in covariate distributions among the compared populations (e.g., Stone, 1993). Nonetheless, confounding as we have defined it is not an inevitable consequence of covariate differences; A and B may differ profoundly with respect to covariates that affect μ, and yet confounding may be absent. In other words, a covariate difference between A and B is a necessary but not sufficient condition for confounding, because the effects of the various covariate differences may balance out in such a way that no confounding is present.

Suppose now that populations A and B differ with respect to certain covariates that affect μ and that these differences have led to confounding of an association measure for the effect measure of interest. The responsible covariates are then termed "confounders" of the association measure. In the above example, with $\mu_{A1} - \mu_{B0}$ confounded for the effect $\mu_{A1} - \mu_{A0}$, the factors that led to $\mu_{A0} \neq \mu_{B0}$ are the confounders. A variable cannot be a confounder (in this sense) unless (1) it can causally affect the outcome parameter μ within treatment groups, and (2) it is distributed differently among the compared populations (e.g., see Yule, 1903, who however uses terms such as "fictitious association" rather than "confounding"). The two necessary conditions (1) and (2) are sometimes offered together as a definition of a confounder. Nonetheless, counterexamples show that the two conditions are not sufficient for a variable with more than two levels to be a confounder as defined above; one such counterexample is given below.

While definitions of "confounder" similar to that just given are common in epidemiology texts (e.g., see Kelsey, Whittemore, Evans and Thompson, 1996; Rothman and Greenland, 1998), they are not universal. Some authors (e.g., Miettinen and Cook, 1981; Robins and Morgenstern, 1987) define a confounder more broadly, as any variable for which adjustment is helpful in reducing bias in effect estimation; variables that are confounders by virtue of their effects on the outcome parameter are then called *causal confounders*. Such broad definitions of confounders stem from recognition that confounding may be dealt with by stratification on variables that are not themselves causes of the outcome. Examples include surrogates for such causes (Kelsey et al., 1996) and determinants of treatment (Rosenbaum and Rubin, 1983).

2.5 Regression Formulations

For simplicity, the above presentation has focused on comparing two groups and two treatments. The basic concepts extend immediately to consideration of multiple groups and treatments. Pairwise comparisons may be represented using the above formalization without modification. Although the comparisons may be made nonparametrically, it is instructive to examine their representation in terms of familiar regression models.

As an illustration, suppose that the treatment level x may range over a continuum or a multidimensional space (in the latter case x and β are row and column vectors), and that population j is given treatment x_j, even though it could have been given some other treatment. Under the causal model

$$\mu_j(x) = \alpha_j + x\beta \quad \text{for all } j, \tag{4}$$

the absolute effect of x_1 versus x_0 on μ in population 1 is

$$\mu_1(x_1) - \mu_1(x_0) = (x_1 - x_0)\beta. \tag{5}$$

(In the earlier notation, $j = A, B$, so that $\mu_1(x_1)$ was μ_{A1} and $\mu_1(x_0)$ was μ_{A0}.) Substitution of $\mu_0(x_0)$, the value of μ in population 0 under treatment x_0, for $\mu_1(x_0)$ yields

$$\mu_1(x_1) - \mu_0(x_0) = \alpha_1 - \alpha_0 + (x_1 - x_0)\beta, \tag{6}$$

which is biased by the amount

$$\mu_1(x_0) - \mu_0(x_0) = \alpha_1 - \alpha_0. \tag{7}$$

Thus, under this model, no confounding for β will occur if the intercepts α_j are constant across populations, so that $\mu_j(x) = \alpha + x\beta$.

When constant intercepts cannot be assumed and nothing else is known about the intercept magnitudes, it may be possible to represent our uncertainty about α_j via the mixed effects model

$$\mu_j(x) = \alpha + x\beta + \varepsilon_j. \tag{8}$$

Here, α_j has been decomposed into $\alpha + \varepsilon_j$, where ε_j has mean zero, and the confounding in $\mu_1(x_1) - \mu_0(x_0)$ has become an unobserved random variable $\varepsilon_1 - \varepsilon_0$. Correlation of the random effects (ε_j) with the treatments (x_j) leads to bias in estimating β. This bias may be attributed to or interpreted as confounding for β in the regression analysis. Confounders are now covariates that "explain" the correlation between ε_j and x_j. In particular, confounders reduce the correlation of x_j and ε_j when entered in the model and so reduce the bias in estimating β.

3. CONTROL OF CONFOUNDING

3.1 Control Via Design

Perhaps the most obvious way to avoid confounding in estimating $\mu_{A1} - \mu_{A0}$ is to obtain a reference population B for which μ_{B0} is known to equal μ_{A0}. Among epidemiologists, such a population is sometimes said to be *comparable to* or *exchangeable with* A with respect to the outcome under the reference treatment. In practice, such a population may be difficult or impossible to find. Thus, an investigator may attempt to construct such a population, or to construct exchangeable index and reference populations. These constructions may be viewed as *design-based* methods for the control of confounding.

Restriction and matching. Perhaps no approach is more effective for preventing confounding by a known factor than *restriction*. For example, gender imbalances cannot confound a study restricted to women. Nonetheless, restriction on many factors can reduce the number of available subjects to unacceptably low levels and may greatly reduce the generalizability of results as well. *Matching* the treatment populations on confounders overcomes these drawbacks and, if successful, can be as effective as restriction. For example, gender imbalances cannot confound a study in which the compared groups have identical proportions of women. Unfortunately, differential losses to observation may undo the initial covariate balances produced by matching. Another problem is that matches may become difficult or impossible to find if one attempts to match on more than a few factors.

Randomization. Neither restriction nor matching prevents (although they may diminish) imbalances on unrestricted, unmatched or unmeasured covariates. In contrast, randomized treatment allocation (randomization) offers a means of dealing with confounding by covariates not explicitly accounted for by the design. It must be emphasized, however, that this solution is only probabilistic and subject to severe practical constraints. For example, protocol violations and loss to follow-up may produce systematic covariate imbalances between the groups (and consequent confounding), and random imbalances may be severe, especially if the study size is small (Fisher, 1935; Rothman, 1977). Blocked randomization can help ensure that random imbalances on the blocking factors will not occur, but it does not guarantee balance of unblocked factors. Thus, even in a perfectly executed randomized trial, the no-confounding condition $\mu_{A0} = \mu_{B0}$ is not a realistic assumption for inferences about causal effects. Successful randomization simply insures that the difference $\mu_{A0} - \mu_{B0}$, and hence the degree of confounding, has expectation zero and converges to zero under the randomization distribution; it also provides a permutation distribution for causal inferences (Fisher, 1935; Cox, 1958, Chapter 5).

Exchangeability. Under randomization, the parameters μ_{A0} and μ_{B0} (and μ_{A1} and μ_{B1} as well) are outcomes of a random process and so can be treated as random variables. Successful randomization renders μ_{A0} and μ_{B0} unconditionally *exchangeable* in the usual probabilistic sense (Cornfield, 1976); in other words, the unconditional joint distribution $F_0(u_A, u_B)$ of μ_{A0}, μ_{B0} is symmetric, so that $F_0(u_A, u_B) = F_0(u_B, u_A)$ for any possible pair of values u_A, u_B for μ_{A0}, μ_{B0}. This exchangeability permits derivation of inferential procedures for (say) $\mu_{A1} - \mu_{A0}$ based on substituting μ_{B0} for μ_{A0} and then allowing for random differences between μ_{A0} and μ_{B0} (Robins, 1988). It applies regardless of what the parameter μ represents; that is, randomization yields exchangeability for all parameters of the outcome distribution. From a Bayesian perspective, μ_{A0} and μ_{B0} can always be treated as random variables. Thus, a practical and sufficient design-based approach to confounding is to find or construct comparison groups such that μ_{A0} and μ_{B0} are exchangeable.

Now consider the regression formulation (8). Here again, it is neither realistic nor necessary to assume absence of confounding to make inferences about the effect parameter β. Rather, it is sufficient to find a population such that the random effects ε_j are exchangeable (so that any correlation of x_j and ε_j is random), as would arise if treatment levels were randomized. This approach is often described in epidemiology as searching for a "natural experiment," that is, a situation in which a compelling argument can be made that the exposure was effectively randomized by natural circumstances. Of course, any inferences may be sensitive to assumptions about the distribution of the random effects (e.g., normality), and the structural form of the model (e.g., linearity); such concerns lead naturally to randomization tests for effects (Fisher, 1935; Cox, 1958; Copas, 1973).

3.2 Control Via Analysis

Design-based methods are often infeasible or insufficient to produce exchangeability. Thus, there has been an enormous amount of work devoted to analytic adjustments for confounding. With a few exceptions, these methods are based on observed covariate distributions in the compared populations. Such methods will successfully control confounding only to the extent that enough confounders are adequately measured and employed in the analysis. Then, too, many methods employ parametric models at some stage, and their success thus depends on the faithfulness of the model to reality. There is a tension between the demands of adjusting for enough covariates and the dependence of the analysis on modeling assumptions. This issue cannot be covered in depth here, but a few basic points are worth noting.

The simplest methods of adjustment begin with stratification on confounders. A covariate cannot be responsible for confounding within a stratum that is internally homogeneous with respect to the covariate. This is so, regardless of whether the covariate was used to define the stratum. For example, gender imbalances cannot confound observations within a stratum composed solely of women. It would seem natural, then, to control confounding due to measured factors by simply stratifying on them all. Unfortunately, one would then confront the well-known *sparse-data* problem: given enough factors, few if any strata would have subjects in both treatment groups, thereby making comparisons inefficient or impossible (Robins and Greenland, 1986).

One solution to this problem begins by noting that within-stratum homogeneity is unnecessary to prevent confounding by a covariate. Within-stratum balance is sufficient, because comparisons within a stratum cannot be confounded by a covariate that is not associated with treatment within the stratum. Hence, a given stratification should be sufficient to control confounding by a set of covariates if the covariates are balanced across the strata, that is, unassociated with treatment within the strata. Rosenbaum and Rubin (1983) showed that, subject to any modeling restrictions used for score estimation, balance in probability for a set of covariates could be achieved by exact stratification on the estimated propensity score, where the propensity score is defined as the probability of treatment given the covariates in the combined (treated and untreated) study population. They further showed that this score was the coarsest score that would produce balance in probability. Stratification on the estimated propensity score thus reduces adjustment for multiple covariates to stratification on a single variable and lowers the risk of sparse-data problems if the model used for propensity scoring is correct. Unfortunately, in sparse data there may be little power to test whether the model is correct.

The most common method for avoiding sparse-data problems is to impose parametric constraints on the regression of the outcome on the treatment and covariates; such strategies are described in many textbooks (e.g., see Clayton and Hills, 1993; Kelsey et al., 1996; Rothman and Greenland, 1998). Hybrid methods which combine regressions on treatment and outcome have also been developed; see Robins and Greenland (1994) and Rosenbaum (1995) for examples. Nonetheless, theoretical results indicate that no approach can completely solve sparse-data problems, insofar as sample size will always limit the number of degrees of freedom available for covariate adjustment and model testing (Robins and Ritov, 1997).

3.3 Sufficient Control

Without randomization, the evaluation of within-stratum or residual confounding becomes a major concern. For this purpose, we define a stratification as *sufficient for estimation of stratum-specific causal effects* if, within strata, μ_{A0} and μ_{B0} are exchangeable. In a parallel fashion, we define a set of variables as *sufficient for control of confounding* if simultaneous (joint) stratification on all the variables is sufficient in the sense just described. (We note in passing that this is a weaker condition than that of "covariate sufficiency" as used in Stone, 1993.) Randomization ensures sufficiency of the set of measured variables not affected by treatment. In the absence of randomization, however, causal inferences become dependent on and sensitive to the assumption that the set of variables available for analysis is sufficient. It almost always remains logically possible that this set is insufficient because some confounder essential for sufficiency has not been recorded; thus, causal inferences from observational studies almost always hinge on subject-matter priors ("judgements") about what may be missing from the set. Sensitivity of results to possible unmeasured confounders can be assessed via formal sensitivity analysis (Rosenbaum, 1995; Copas and Li, 1997; Robins, Rotnitzky, and Scharfstein, 1999).

There are some systematic ways of deriving the implications of background assumptions. For example, assumptions about the directions and absences of causal relations among variables (measured and unmeasured) can be conveniently encoded in a *causal graph* or *path diagram*, in which arrows (directed arcs) represent cause–effect relations. Conditional on the assumptions underlying the graph, the question of sufficiency of a set of variables (such as the set of measured variables) can be easily answered using a simple graphical algorithm called the "back-door test" (Pearl, 1995). The same algorithm allows one to determine whether subsets of a sufficient set are themselves sufficient. By stepwise deletion and testing, we may thus identify *minimally sufficient* subsets (that is, sufficient subsets with no sufficient proper subsets). The need for such identification arises, for example, in epidemiologic studies in which numerous "lifestyle" covariates (diet, physical activity, smoking and drinking habits, etc.) are measured and are potential confounders of the effect under study. Here, the total set of covariates may be sufficient for control as defined above, but impractical to control in its entirety, even when using propensity score or outcome-regression methods (Greenland, Pearl and Robins, 1999).

Graphical identification of sufficient subsets operates on background assumptions, rather than data. An analogous statistical approach is given by the following result (Robins, 1997).

THEOREM. *A subset S of a sufficient set is itself sufficient if the remainder subset R (those variables in the original set but not in S) can be decomposed into disjoint subsets R_1 and R_2 such that both*

$$R_1 \coprod X|S \quad \textit{and} \quad R_2 \coprod Y|R_1, X, S; \tag{9}$$

that is, R_1 is independent of treatment X given S, and R_2 is independent of the outcome Y given R_1, treatment X, and S. When such a decomposition exists, R_1 can be taken to be the largest subset of R satisfying $R_1 \coprod X|S$.

Any set identified as sufficient by the back-door test must satisfy the conditions of Robins's theorem, although the converse is not true (Robins, 1997). As an example of the theorem, suppose X is measured breast-implant exposure, Y is time to breast cancer from an index time, S contains diet and exercise variables, R contains alcohol-drinking and smoking variables with R_1 = drinking and R_2 = smoking variables, and $S \cup R$ is sufficient. Then the drinking and smoking variables can be omitted if the drinking and implant variables are independent conditional on the diet and exercise variables, and the smoking and breast cancer variables are independent conditional on the drinking, implant, diet and exercise variables. These conditions are testable using observed data; for example, smoking has not been found to be associated with breast cancer upon extensive control of other lifestyle variables, while the association of drinking and implants could be examined in certain existing data bases.

Stone (1993, page 459) defined "no confounding given S" as fulfillment of the above condition (9) with R_1 equal to all unobserved covariates that affect the response (outcome). This "no-confounding" definition neither implies nor is implied by our definition, but Stone showed that it does imply that S is sufficient for control. The above theorem is a generalization of Stone's result, in that it imposes no causal constraints on R_1 or R_2; for example, it does not assume that R_1 contains all (or even any) covariates that affect response.

A set S that is sufficient for estimating stratum-specific effects will also be sufficient for estimating a summary measure of the effect of treatment on the entire target population. Nonetheless, because confounding may "average out" across strata, the converse is not true: a set S may be sufficient for estimating a summary effect even though insufficient for estimating stratum-specific effects (Greenland and Robins, 1986). This notion will be formalized in the next subsection.

3.4 Residual Confounding

Suppose that we subdivide the total study population $(A + B)$ into K strata indexed by k. Let μ_{A1k} be the parameter of interest in stratum k of populations A and B under treatment x_1. The effect of treatment x_1 relative to x_0 in stratum k may be defined as $\mu_{A1k} - \mu_{A0k}$ or μ_{A1k}/μ_{A0k}. The confounding that remains in stratum k is called the *residual confounding* in the stratum, and is measured by $\mu_{A0k} - \mu_{B0k}$ or μ_{A0k}/μ_{B0k}. Note that a sufficient stratification has $E(\mu_{A0k} - \mu_{B0k}) = 0$ for all k by virtue of the exchangeability of μ_{A0k} and μ_{B0k}, yet may have random residual confounding within strata.

Residual confounding may be summarized in a number of ways, for example, by standardization or other weighted-averaging methods (Miettinen, 1972; Rothman and Greenland, 1998). As an illustration, suppose the strata represent age-sex-specific subgroups, and the proportion of the standard population that falls in age-sex stratum k is p_k. Then the effect of x_1 versus x_0 on A standardized to (weighted by) the distribution $p_1, \ldots, p_K$ is

$$(10)\qquad \begin{aligned} D_{AA} &= \sum_k p_k \mu_{A1k} - \sum_k p_k \mu_{A0k} \\ &= \sum_k p_k(\mu_{A1k} - \mu_{A0k}), \end{aligned}$$

whereas the standardized difference comparing A to B is

$$(11)\qquad \begin{aligned} D_{AB} &= \sum_k p_k \mu_{A1k} - \sum_k p_k \mu_{B0k} \\ &= \sum_k p_k(\mu_{A1k} - \mu_{B0k}). \end{aligned}$$

The overall residual confounding in D_{AB} is thus

$$(12)\qquad \begin{aligned} D_{AB} - D_{AA} &= \sum_k p_k \mu_{A0k} - \sum_k p_k \mu_{B0k} \\ &= \sum_k p_k(\mu_{A0k} - \mu_{B0k}), \end{aligned}$$

which may be recognized as the standardized difference comparing A and B when both are given treatment x_0, using $p_1, \ldots, p_K$ as the standard distribution. With this formulation, a stratification can be considered sufficient for inference on D_{AA} if $\sum_k p_k \mu_{A0k}$ and $\sum_k p_k \mu_{B0k}$ are exchangeable.

4. COLLAPSIBILITY

4.1 Collapsibility in Contingency Tables

Consider the $I \times J \times K$ contingency table representing the joint distribution of three discrete variables X, Y, Z, the $I \times J$ marginal table representing the joint distribution of X and Y, and the set of conditional $I \times J$ subtables (strata) representing the joint distributions of X and Y within levels of Z. Generalizing Whittemore (1978) (who considered log-linear model parameters), we say a measure of association of X and Y is *strictly collapsible* across Z if it is constant across the strata (subtables) and this constant value equals the value obtained from the marginal table.

Noncollapsibility (violation of collapsibility) is sometimes referred to as *Simpson's paradox*, after a celebrated article by Simpson (1951). This phenomenon had been discussed by earlier authors, including Yule (1903); see also Cohen and

Nagel (1934). Some statisticians reserve the term *Simpson's paradox* to refer to the special case of noncollapsibility in which the conditional and marginal associations are in opposite directions, as in Simpson's numerical examples. Simpson's algebra and discussion, however, dealt with the general case of inequality. The term "*collapsibility*" seems to have arisen in later work; see Bishop, Fienberg and Holland (1975).

Table 1 provides some simple examples. The difference of probabilities that $Y = 1$ (the risk difference) is strictly collapsible. Nonetheless, the ratio of probabilities that $Y = 1$ (the risk ratio) is not collapsible because the risk ratio varies across the Z strata, and the odds ratio is not collapsible because its marginal value does not equal the constant conditional (stratum-specific) value. Thus, collapsibility depends on the chosen measure of association.

Now suppose that a measure is not constant across the strata, but that a particular summary of the conditional measures does equal the marginal measure. This summary is then said to be *collapsible* across Z. As an example, in Table 1 the ratio of risks standardized to the marginal distribution of Z is

$$\begin{aligned}&\big[P(Z=1)P(Y=1|X=1,\ Z=1)\\ &\quad + P(Z=0)P(Y=1|X=1,\ Z=0)\big]\\ &\quad\cdot\big[P(Z=1)P(Y=1|X=0,\ Z=1)\\ &\qquad + P(Z=0)P(Y=1|X=0,\ Z=0)\big]^{-1}\\ &\quad = \frac{0.50(0.80)+0.50(0.40)}{0.50(0.60)+0.50(0.20)} = 1.50,\end{aligned}$$

equal to the marginal (crude) risk ratio. Thus, this measure is collapsible in Table 1. Various tests of collapsibility and strict collapsibility have been developed; see Whittemore (1978); Asmussen and Edwards (1983); Ducharme and LePage (1986); Greenland and Mickey (1988) and Geng (1989) for examples.

TABLE 1
Examples of collapsibility and noncollapsibility in a three-way distribution

	Z = 1		**Z = 0**		**Marginal**	
	X = 1	**X = 0**	**X = 1**	**X = 0**	**X = 1**	**X = 0**
$Y = 1$	0.20	0.15	0.10	0.05	0.30	0.20
$Y = 0$	0.05	0.10	0.15	0.20	0.20	0.30
Risks[a]	0.80	0.60	0.40	0.20	0.60	0.40
Risk differences	0.20		0.20		0.20	
Risk ratios	1.33		2.00		1.50	
Odds ratios	2.67		2.67		2.25	

[a] Probabilities of $Y = 1$.

4.2 Regression Formulation

The above definition of strict collapsibility extends to regression contexts. Consider a generalized linear model for the regression of Y on three regression vectors $\underline{W}$, $\underline{X}$, $\underline{Z}$:

$$\begin{aligned}&g[E(Y|\underline{W}=\underline{w},\ \underline{X}=\underline{x},\ \underline{Z}=\underline{z})]\\ &\qquad = \alpha + \underline{w\beta} + \underline{x\gamma} + \underline{z\delta}.\end{aligned} \tag{13}$$

The regression is said to be *collapsible* for $\underline{\beta}$ over $\underline{Z}$ if $\underline{\beta} = \underline{\beta}^*$ in the regression omitting $\underline{Z}$,

$$g[E(Y|\underline{W}=\underline{w},\ \underline{X}=\underline{x})] = \alpha^* + \underline{w\beta}^* + \underline{x\gamma}^*, \tag{14}$$

and is *noncollapsible* if $\underline{\beta} \neq \underline{\beta}^*$ (Clogg, Petkova and Shihadeh, 1992). Thus, if the regression is collapsible for $\underline{\beta}$ over $\underline{Z}$ and $\underline{\beta}$ is the parameter of interest, $\underline{Z}$ need not be measured to estimate $\underline{\beta}$. If $\underline{Z}$ is measured, however, tests of $\underline{\beta} = \underline{\beta}^*$ can be constructed (Hausman, 1978; Clogg, Petkova and Shihadeh, 1992; Clogg, Petkova and Haritou, 1995).

The preceding definition generalizes the original contingency table definition to arbitrary variables. There is a technical problem with the above regression definition, however: If the first (full) model is correct, it is unlikely that the second (reduced) regression will follow the given form; that is, most families of regression models are not closed under deletion of $\underline{Z}$. If, for example, Y is Bernoulli and g is the logit link function, so that the full regression is first-order logistic, the reduced regression will not follow a first-order logistic model except in special cases. One way around this dilemma (and the fact that neither of the models is likely to be exactly correct) is to define the model parameters as the asymptotic means of the maximum-likelihood estimators. These means are well defined and interpretable even if the models are not correct (White, 1994).

It may be obvious that, if the full model is correct, $\underline{\delta} = 0$ implies collapsibility for $\underline{\beta}$ and $\underline{\gamma}$ over $\underline{Z}$. Suppose, however, that neither $\underline{\beta}$ nor $\underline{\delta}$ is zero. In that case, marginal independence of the regressors does not ensure collapsibility for $\underline{\beta}$ over $\underline{Z}$ except when g is the identity or log link (Gail, Wieand and Piantadosi, 1984; Gail, 1986); conversely, collapsibility can occur even if the regressors are associated (Whittemore, 1978); see the example below. Thus, it is not generally correct to equate collapsibility over $\underline{Z}$ with simple independence conditions, although useful results are available for the important special cases of linear, log-linear, and logistic models (e.g., see Gail, 1986; Wermuth, 1987, 1989; Robinson and Jewell, 1991; Geng, 1992; Guo and Geng, 1995).

4.3 Other Collapsibility Concepts

The literature on graphical probability models distinguishes a number of properties that have been referred to as types of collapsibility; see Frydenberg (1990), Whittaker (1990, Section 12.5) and Lauritzen (1996, Section 46.1) for examples. Both definitions given above are special cases of *parametric collapsibility* (Whittaker, 1990).

5. CONFOUNDING AND NONCOLLAPSIBILITY

5.1 The Divergence

Much of the statistics literature does not distinguish between the concept of confounding as a bias in effect estimation and the concept of noncollapsibility; for example, Becher (1992) defines confounding as $\underline{\beta} \neq \underline{\beta}*$ in models (13) and (14), in which case the elements of Z are called confounders; similarly, Guo and Geng (1995) define Z to be a nonconfounder if $\beta = \beta^*$. Nonetheless, the two concepts are distinct: confounding may occur with or without noncollapsibility and noncollapsibility may occur with or without confounding (Miettinen and Cook, 1981; Greenland and Robins, 1986; Wickramaratne and Holford, 1987). Mathematically identical conclusions have been reached by other authors, albeit with different terminology in which noncollapsibility corresponds to "bias" and confounding corresponds to "covariate imbalance" (Gail, 1986; Hauck, Neuhaus, Kalbfleisch and Anderson, 1991).

Noncollapsibility without confounding. Table 2 gives the response distributions under treatments x_1 and x_0 for a hypothetical target population A and the response distribution under treatment x_0 for a hypothetical reference population B. Suppose A receives treatment x_1, B receives x_0, and we wish to estimate the effect that receiving x_1 rather than x_0 had on A. If we take the odds of response as the outcome parameter μ, we get $\mu_{A1} = 0.6/(1-0.6) = 1.50$, and $\mu_{A0} = \mu_{B0} = 0.4/(1-0.4) = 0.67$. Hence, there is no confounding of the odds ratio: $\mu_{A1}/\mu_{A0} = \mu_{A1}/\mu_{B0} = 1.50/0.67 = 2.25$. Nonetheless, the covariate Z is associated with response in A and B. Furthermore, the odds ratio is not collapsible: within levels of Z, the odds ratios comparing A under treatment x_1 to either A or B under x_0 are $(0.8/0.2)/(0.6/0.4) = (0.4/0.6)/(0.2/0.8) = 2.67$, higher than the unconditional (crude) odds ratio of 2.25 obtained when Z is ignored.

The preceding example illustrates a peculiar property of the odds ratio as an effect measure: treatment x_1 (relative to x_0) elevates the odds of response by 125% in population A, yet within each stratum of Z it raises the odds by 167%. If Z is associated with response conditional on treatment but unconditionally unassociated with treatment, the stratum-specific odds ratios must be farther from 1 than the unconditional odds ratio if the latter is not 1 (Gail, 1986; Hauck et al., 1991). This phenomenon is often interpreted as a "bias" in the unconditional odds ratio, but in fact there is no bias if one takes care to not misinterpret the unconditional effect as an estimate of the stratum-specific or individual effects (Miettinen and Cook, 1981; Greenland, 1987).

TABLE 2
Distribution of responses for hypothetical index population A under treatments x_1 and x_0, and for reference population B under treatment x_0: Example of noncollapsibility without confounding of the odds ratio

	Population A		
	Response probability if		
Stratum	$X = x_1$	$X = x_0$	**Stratum size**
$Z = 1$	0.8	0.6	1,000
$Z = 0$	0.4	0.2	1,000
Unconditional	0.6	0.4	
	Population B		
	Response probability if		
Stratum	$X = x_1$	$X = x_0$	**Stratum size**
$Z = 1$	$*^a$	0.6	1,000
$Z = 0$	$*^a$	0.2	1,000
Unconditional	$*^a$	0.4	

[a] Not used in example.

Confounding without noncollapsibility. To create a numerical example in which the odds ratio is collapsible and yet is confounded for the overall effect, we need only modify Table 2 slightly, by changing the stratum size for $Z = 0$ in population B to 1,500. With this change, the proportion with $Z = 1$ in population B drops from 0.5 to 0.4, the unconditional response probability in population B under treatment x_0 becomes $0.4(0.6)+0.6(0.2) = 0.36$, and the unconditional response odds μ_{B0} in population B under x_0 becomes $0.36/(1-0.36) = 0.5625$. Thus, $\mu_{B0} = 0.5625 < 0.67 = \mu_{A0}$, with consequent confounding of the odds ratio: μ_{A1}/μ_{A0}, the true effect that x_1 had on the odds in population, equals 2.25 (as before), which is less than the unconditional odds ratio $\mu_{A1}/\mu_{B0} = 1.50/0.5625 = 2.67$. Nonetheless, this unconditional odds ratio equals the stratum-specific odds ratios, which are unchanged from the previous example.

5.2 Conditions for Equivalence

The example in Table 2 shows that, when μ is the odds of the outcome, μ_{A0} may equal μ_{B0} (no confounding) even when the odds ratio is not collapsible over the confounders. Conversely, the modified example shows that we may have $\mu_{A0} \neq \mu_{B0}$ even when the odds ratio is collapsible. A probabilistic explanation of the discrepancy between nonconfounding and collapsibility is that μ_{A0} will equal μ_{B0} whenever Z is sufficient for control and is unconditionally unassociated with treatment, as in Table 2, whereas collapsibility of the odds ratio will occur whenever Z is unassociated with treatment conditional on response, as in the modified example (Bishop, Fienberg and Holland, 1975). Thus, the discrepancy is just a consequence of the nonequivalence of unconditional and conditional associations.

If the effect measure is the difference or ratio of response proportions, results of Gail (1986) imply that this measure will be collapsible over Z if Z has the same distribution in A and B (that is, if Z and treatment are unconditionally unassociated). It follows that, when examining such measures, the above phenomena (noncollapsibility without confounding and confounding without noncollapsibility) cannot occur if Z is sufficient for control. More generally, when the effect measure can be expressed as the average effect on population members [e.g., under the linear causal model (4)], the conditions for noncollapsibility and confounding will be identical, provided the covariates in question form a sufficient set for control. In such cases, noncollapsibility and confounding become equivalent, which may explain why the two concepts are often not distinguished. The nonequivalence of the two concepts for odds ratios simply reflects the fact that the unconditional effect of a treatment on the odds is not the average treatment effect on population members (Greenland, 1987).

5.3 Regression Formulations

The preceding conclusions correspond to well-known results for correlated-outcome regression. For example, the difference between the stratum-specific and crude odds ratios in Table 2 corresponds to the differences between cluster-specific and population-averaged (marginal) effects in binary regression (Neuhaus, Kalbfleisch and Hauck, 1991): the clusters of correlated outcomes correspond to the strata, the cluster effects correspond to the covariate effects, the cluster-specific treatment effects correspond to the stratum-specific log odds ratios, and the population-averaged treatment effect corresponds to the crude log odds ratio.

More generally, consider a situation in which the full regression model [model (13)] is intended to represent causal effects of the regressors on Y. Noncollapsibility over $\underline{Z}$ (that is, $\underline{\beta} \neq \underline{\beta}^*$) does not correspond to confounding of effects unless g is the identity or log link. That is, it is possible for $\underline{\beta}$ to unbiasedly represent the effect of manipulating $\underline{W}$ within levels of $\underline{X}$ and $\underline{Z}$, and, at the same time, for $\underline{\beta}^*$ to unbiasedly represent the effect of manipulating $\underline{W}$ within levels of $\underline{X}$, even though $\underline{\beta}^* \neq \underline{\beta}$. Table 2 demonstrates this point for logistic models, and shows that noncollapsibility in a logistic model does not always signal a bias. The divergence between $\underline{\beta}$ and $\underline{\beta}^*$ corresponds to the distinction between cluster-specific and population-averaged effects: the cluster-specific model corresponds to the full model (13) in which $\underline{Z}$ is an unobserved univariate cluster-specific random variable independent of $\underline{W}$ and $\underline{X}$, with mean zero and unit variance; $\underline{\delta}^2$ is then the vector of random-effects variances.

6. CONFOUNDING IN INTERVENTION STUDIES: FURTHER ISSUES

In this section we briefly discuss some special issues of confounding that arise in studies of interventions, such as clinical trials and natural experiments.

6.1 Adjustment in Randomized Trials

Some controversy has existed about adjustment for random covariate imbalances in randomized trials. Although Fisher asserted that randomized comparisons were "unbiased," he also pointed out that they could be confounded in the sense used here (e.g., see Fisher, 1935, page 49). Fisher's use of the word "unbiased" was unconditional on allocation, and therefore of little guidance for analysis of a given trial. Some arguments for conditioning on allocation are given in Greenland and Robins (1986) and Robins and Morgenstern (1987). Other arguments for adjustment in randomized trials have been given by Rothman (1977); Miettinen and Cook (1981) and Senn (1989).

6.2 Intent-to-Treat Analysis

In a randomized trial, noncompliance can easily lead to confounding in comparisons of the groups actually receiving treatments x_1 and x_0. One somewhat controversial solution to noncompliance problems is intent-to-treat analysis, which defines the comparison groups A and B by treatment assigned rather than treatment received. Detractors of intent-to-treat analysis consider it an attempt to define away a serious problem, especially when treatment received is the treatment of scientific interest. Supporters of intent-to-treat analysis emphasize that intent-to-treat tests (tests of assigned-treatment effects) remain valid tests of received-treatment effects under broader conditions than conventional tests of received-treatment effects. See the volume edited by Goetghebeur and van Houwelingen (1998) for discussions of these and related issues, including alternatives to intent-to-treat analysis.

A crucial point is that confounding can affect even intent-to-treat analyses. For example, apparently random assignments may not be random, as when blinding is insufficient to prevent the treatment providers from protocol violations or when there is differential loss to follow-up. Even when these problems do not occur, random imbalances remain possible. A more subtle problem is that noncompliance can produce bias away from the null in an intent-to-treat analysis of equivalence trials. With noncompliance, the sharp null hypothesis (of equivalence of the two treatments) does *not* by itself imply that the distribution of outcomes will be the same in both treatment arms, because noncompliance represents movement into a third untreated state that does not correspond to any assigned treatment (Robins, 1998). To illustrate, suppose treatments A and B are both 100% effective and thus completely equivalent with respect to their effect on the outcome, so that the equivalence null is satisfied. Suppose, however, that treatment A causes a harmless but unpleasant flushing sensation, whereas treatment B does not, and as a consequence compliance is 70% for A but 100% for treatment B. Then the intent-to-treat test will reject the null hypothesis of equivalence solely because of the lower compliance with treatment A. Thus, in this example, noncompliance confounds the intent-to-treat analysis away from the correct null hypothesis of equivalence.

6.3 Choice of Target

In observational epidemiologic studies, the usual goal is to estimate the effect that treatment had on the treated group, $\mu_{A1} - \mu_{A0}$, or would have had on the untreated group, $\mu_{B1} - \mu_{B0}$, depending on the ultimate policy objectives; for example, $\mu_{A1} - \mu_{A0}$ may be of interest as a measure of harm caused by the treatment. However, in randomized trials there are several reasons for orienting the estimation goal toward comparison of the expected outcome of the entire (treated + untreated) study group if everyone had been treated, $\mu_{+1} = (\mu_{A1} + \mu_{B1})/2$, and the expected outcome of this group if no one had been treated, $\mu_{+0} = (\mu_{A0} + \mu_{B0})/2$ (Robins, 1988). Both these outcomes are counterfactual, and so analysis requires substituting the estimable pair (μ_{A1}, μ_{B0}) for (μ_{+1}, μ_{+0}). Consequently, the no-confounding condition becomes $\mu_{A1} = \mu_{+1}$ *and* $\mu_{B0} = \mu_{+0}$ or, equivalently, $\mu_{A1} = \mu_{B1}$ *and* $\mu_{A0} = \mu_{B0}$. This condition is not realistic, but it is also not necessary. Inferential procedures for (say) $\mu_{+1} - \mu_{+0}$ can be derived from the randomization-induced unconditional exchangeability of μ_{A1} with μ_{B1} and of μ_{A0} with μ_{B0}.

One advantage of focusing on $\mu_{+1} - \mu_{+0}$ rather than $\mu_{A1} - \mu_{B1}$ is that standard inferential procedures for treatment effects on risks yield conservatively valid inferences for $\mu_{+1} - \mu_{+0}$ (Copas, 1973; Robins, 1988; see also Neyman, 1935). Another argument for focusing on μ_{+1} and μ_{+0} applies if the entire cohort is a random sample from a specified target population: in that situation $\mu_{+1} - \mu_{+0}$ will in expectation be closer to the treatment effect in the target than $\mu_{A1} - \mu_{B0}$, because the former will deviate from the target effect only because of sampling variability, whereas the latter will incorporate randomization variability as well (Robins, 1988).

6.4 Ignorability and Confounding

It is sometimes possible to evaluate a treatment-assignment mechanism even though the mechanism is not under control; examples include the way utility companies assign water and power sources to homes. In these settings, the mechanism may clearly not be random but may nonetheless satisfy weaker conditions that allow inferences about the effect of interest.

As an example, suppose one is interested in just one particular outcome variable, such as time to death. A treatment-assignment mechanism is *strongly ignorable* for the outcome variable if (1) the treatment-assignment variable X it defines is independent of the vector $\mathbf{y} = (y_0, y_1, \ldots, y_K)$ of potential outcomes, and (2) each unit has nonzero probability of assignment to each treatment level (Rosenbaum and Rubin, 1983). Standard randomization methods are strongly ignorable for all outcomes when noncompliance and censoring are absent or purely random.

Like randomization, strongly ignorable treatment assignment insures that parameters of the population-specific distributions of potential outcomes will be exchangeable across populations. For example, suppose we have two treatment levels (x_0, x_1), a strongly ignorable assignment mechanism, and (as above) we label the x_1- and x_0-treated groups by A and B; then, absent any other information related to μ_{A0} or μ_{B1}, the pair (μ_{A1}, μ_{A0}) should be exchangeable with the pair (μ_{B1}, μ_{B0}).

For most purposes, however, only component-specific exchangeability of μ_{A1} with μ_{B1} or μ_{A0} with μ_{B0} is needed (Greenland and Robins, 1986; Robins, 1987b); this condition is implied by but does not imply "weak ignorability," in which condition (1) above is replaced by the condition that X is independent of each component of $\mathbf{y}$ considered separately, rather than jointly (Stone, 1993). Ignorability (strong or weak) is thus stronger than needed for identification of the causal parameter $\mu_{A1} - \mu_{A0}$.

Rubin (1991) has referred to an assignment mechanism that satisfies condition (1) (i.e., assigns treatment independently of $\mathbf{y}$) as "unconfounded." Following Fisher (1935), we prefer to call such a mechanism "unbiased," because traditional usage of "unconfounded" refers to the actual allocation produced by the mechanism. Rubin's usage, like Stone's (1993) definition of "no confounding," does not allow for random allocation errors; as discussed earlier for randomization, however, random variation in unbiased mechanisms can by chance produce confounded allocations.

7. CONCLUSION

Concepts of confounding have been discussed by philosophers and scientists for centuries. It is only in more recent decades, however, that precise formal definitions of these concepts have emerged within statistical theory. These developments have revealed the distinction between counterfactual and collapsibility-based concepts of confounding. This distinction deserves mention in basic statistics education, because the counterfactual definition of confounding is nonparametric and specific to causal inference, whereas collapsibility depends on the choice of association parameter and requires no reference to causality or effects.

Our discussion has assumed that both the treatment variable and the confounders can be fully characterized by fixed covariates. Further subtleties can arise when these variables are time-dependent; see Robins (1986, 1987a, b, 1997) and Pearl and Robins (1995). We also have not considered issues of confounding in separating direct and indirect effects; for discussions, see Robins (1986, 1997), Robins and Greenland (1992, 1994), Pearl and Robins (1995) and Pearl (1997).

We wish to end on the cautionary note that confounding is but one of many problems that plague studies of cause and effect. Biases of comparable or even greater magnitude can arise from measurement errors, selection (sampling) biases, and systematically missing data, as well as from model-specification errors. Even when confounding and other systematic errors are absent, individual causal effects will remain unidentified by statistical observations (Greenland and Robins, 1988; Robins and Greenland, 1989). It remains a serious challenge to create a statistical theory that can encompass all these problems coherently and also yield practical methods for data analysis.

APPENDIX 1: RESTRICTIONS IMPOSED BY THE POTENTIAL-OUTCOMES APPROACH

First, causal effects are defined only for *comparisons* of treatment levels. To state that "drinking two glasses of wine a day lengthened Smith's life by four years" is meaningless by itself. A reference level (e.g., no wine at all) must be at least implicit to make sense of the statement. Smith might have lived even longer had she consumed one rather than two glasses per day, in which case the statement would be false relative to one glass a day. As given, the statement could refer to no wine or four glasses per day or any other possibility.

Second, the definition assumes that y_{ik}, the outcome of unit i under treatment k, remains conceptually meaningful even if unit i is *not* given treatment k. That is, the analyst must be prepared to treat $y_{i0}, \dots, y_{iK}$ as parameters unaffected by treatment assignment. Treatment assignment only determines which one of these $K + 1$ parameters we observe [that is, the realization of Y_i (Rubin, 1974, 1978, 1991)]; the other K parameters remain latent traits of individual i. In the philosophy literature, analysis of this assumption is one of the core tasks of counterfactual logic (Simon and Rescher, 1966; Lewis, 1973a; Galles and Pearl, 1998; see also the discussion of Holland, 1986). The statement "if x_k had been administered, the response Y_i of unit i would have been y_{ik}" is called a *counterfactual conditional* (Lewis, 1973b; Stalnaker, 1968); it asserts that Y_i would have equaled y_{ik} if x_k had been administered to unit i, *even if x_k had not in fact been administered to unit i.*

Third, the effects captured by the above counterfactual definition are *net effects*, in that they include all indirect effects and interactions not specifically excluded by treatment definitions. For example, Smith's consumption of two glasses of wine per day rather than none may have given her four extra years of life solely because one night at a formal dinner it made her feel unsteady and she had a friend drive her home; had she not drunk, she would have driven herself, skidded on a patch of ice, hit a tree and been killed. This sort of indirect effect is not one we would wish to capture when studying biologic effects of wine use. It is nonetheless included in our measure of effect (as well as any estimate) unless we amend our treatment definition to include holding constant all "risky" activities that take place during Smith's life. Such an amendment is sometimes (simplistically) subsumed under the clause of "all other things being equal (apart from treatment)," but can be a serious source of ambiguity when the intervention that enforces the amendment is not well defined.

Consider next an illustration in which redefinition of the treatment is not an option. Suppose the objective is to estimate the effect of coffee use on risk of myocardial infarction (MI). If coffee had not been used by members of a cohort of current users, some of these members might have instead taken up or increased smoking to achieve their desired level of stimulation. Thus, coffee use may well have prevented or reduced smoking and so indirectly reduced MI risk. Would we want this indirect effect of coffee included in our target effect parameter? The answer to this question depends on the research goals, rather than statistics. If the answer is no—we wish only to estimate direct physiologic effect of coffee use, apart from its influence on smoking habits—we would have to redefine our reference level to one of no increased smoking, as well as no coffee use.

A fourth restriction, which may be considered an aspect of the third, is that the definition assumes that treatments not applied to a unit could have been applied. Suppose Smith would not and could not stop daily wine consumption unless forced physically to do so. The effect of her actual two-glass-a-day consumption versus the counterfactual "no wine" would now be undefined without amending the treatment definition to include forcing Smith to drink no wine, for example, by removing all alcohol from public availability.

Some authors account for the preceding restriction by requiring that the counterfactual definition of "effect" applies only to "treatment variables." The latter are defined informally as variables subject to intervention or to manipulation of their levels (e.g., see Holland, 1986). One may sense an echo of the circularity (as in ordinary definitions of cause and effect), for the notion of manipulation embodies having an effect on treatment levels (i.e., on x_k) and is itself somewhat ambiguous. Further ambiguity arises because ordinary and useful notions of cause do not impose such restrictions; for example, the statement "trisomy 21 causes Down's Syndrome" is meaningful, even though no method for intervention on trisomy 21 is known. Nonetheless, it has been argued that one strength of the counterfactual approach is its explication of the ambiguities inherent in defining cause and effect (Lewis, 1973a; Holland, 1986; Rubin, 1990). Although some authors impose no restrictions on the definition of cause (e.g., Lewis, 1973a), in our presentation we assume that we are dealing only with causes that can be manipulated, such as drug treatments.

Implicit in most discussions of potential outcomes, including the present one, is that the outcome y_{ik} of unit i under treatment x_k does not depend on the treatment given to any other unit. This postulate is often called the assumption of no interference among units, or the stability assumption (Cox, 1958; Rubin, 1978, 1990). Chapter 2 of Cox (1958) gives a careful discussion of conditions that lead to interference in experimental trials. In epidemiologic studies, interference arises readily when the outcome is contagious; here, phenomena such as herd immunity may lead to complex dependencies of subject-specific outcomes on the entire population distribution of treatment. In such situations, the potential outcomes of a single unit must at least be written $y_{i\omega}$, where ω ranges over all $(K+1)^N$ possible allocations of the $K+1$ treatments among the N population units; further complexities arise if dependencies among the unit specific outcomes must be directly modeled (Halloran and Struchiner, 1995).

APPENDIX 2: DEFINING POTENTIAL OUTCOMES WHEN COMPETING RISKS ARE PRESENT

The definition of potential outcomes can be especially difficult in survival analysis when competing risks are present. Consider again Smith's drinking. Suppose she contracted cancer at age 70 and drank two glasses of wine a day, but would have instead died of a myocardial infarction at age 68 if she had drunk no wine. How could we define her counterfactual age-at-cancer given no wine? Without such a definition, the effect of two glasses of wine versus none would be undefined.

One school of thought maintains that, to offer a definition, one must have an unambiguous concept of the manner in which competing risks would be removed, as well as the counterfactual time of the outcome event. These requirements are *not* met by conditioning on "absence of competing risks": Such hypothetical absence is itself not a treatment or other well-defined counterfactual state, even though standard probability calculations (as used in product-limit estimates) make it appear otherwise (Kalbfleisch and Prentice, 1980, page 166; Prentice and Kalbfleisch, 1988). To deal with this problem, one must specify a treatment that would prevent the competing risks, as well as the treatment x_k of primary interest.

An opposing view maintains that it can be scientifically useful to assume that a potential outcome y_{ik} remains a well-defined quantity even if a competing risk would occur and prevent its observation under treatment x_k (Robins, 1986, 1987a; Slud, Byar and Schatzkin, 1988). In such a case, y_{ik} is the time the outcome would have occurred *if* no competing risks occurred before y_{ik}. That is, we imagine that the causal mechanism leading to the outcome would have taken its course to a particular value y_{ik} had it not been interrupted by the competing risk. Returning to our example, we could imagine that, if Smith had drunk no wine, her pathophysiologic state at age 68 (when she died from a myocardial infarction) would have been such as to produce a malignant tumor after just one more year. Under this admittedly very hypothetical scenario, we could say that her drinking extended her age-at-cancer by one year because she in fact contracted cancer at age 70. Note that this latent-outcome model does *not* imply that the competing risks are independent.

A third approach to the competing-risk problem is based on extending the ordinary definitions of effect in absence of competing risks to encompass instances in which the outcome is undefined under one or more treatments. For details, see Robins (1995b).

ACKNOWLEDGMENTS

The authors thank Leon Gleser and Jay Kaufman for their helpful comments and suggestions for the revisions of this paper.

REFERENCES

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statist. Sci.* **10** 364–376.

Asmussen, S. and Edwards, D. (1983). Collapsibility and response variables in contingency tables. *Biometrika* **70** 567–578.

Balke, A. and Pearl, J. (1994). Counterfactual probabilities: computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence 10* (R. Mantaras and D. Poole, eds.) 46–54. Morgan Kaufmann, San Francisco.

Becher, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine* **11** 1747–1758.

Berkane, M., ed. (1997). *Latent Variable Modeling and Applications to Causality.* Springer, New York.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Bross, I. D. J. (1967). Pertinency of an extraneous variable. *J. Chronic Disease* **20** 487–495.

Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford Univ. Press.

Clogg, C. C., Petkova, E. and Shihadeh, E. S. (1992). Statistical methods for analyzing collapsibility in regression models. *J. Educ. Statist.* **17** 51–74.

Clogg, C. C., Petkova, E. and Haritou, A. (1995). Statistical methods for comparing regression coefficients between models (with discussion). *Amer. J. Sociololgy* **100** 1261–1305.

Cohen, M. R. and Nagel, E. (1934). *An Introduction to Logic and the Scientific Method.* Harcourt Brace, New York.

Copas, J. B. (1973). Randomization models for matched and unmatched 2 × 2 tables. *Biometrika* **60** 467–476.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 55–95.

Cornfield, J. (1976). Recent methodological contributions to clinical trials. *Amer. J. Epidemiol.* **104** 408–421.

Cornfield, J., Haenszel, W., Hammond, W. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22** 173–203.

Cox, D. R. (1958). *The Planning of Experiments.* Wiley, New York.

Dawid, A. P. (2000). Causal inference without counterfactuals. *J. Amer. Statist. Assoc.* To appear.

Ducharme, G. R. and LePage, Y. (1986). Testing collapsibility in contingency tables. *J. Roy. Statist. Soc. Ser. B* **48** 197–205.

Feynman, R. P. (1963). *Lectures on Physics*. Addison-Wesley, Reading, MA.

Fisher, R. A. (1918). The causes of human variability. *Eugenics Rev.* **10** 213–220.

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

Frydenberg, M. (1990). Marginalization and collapsibility in graphical statistical models. *Ann. Statist.* **18** 790–805.

Gail, M. H. (1986). Adjusting for covariates that have the same distribution in exposed and unexposed cohorts. In *Modern Statistical Methods in Chronic Disease Epidemiology* (S. H. Moolgavkar, and R. L. Prentice, eds.) 3–18. Wiley, New York.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71** 431–444.

Galles, D. and Pearl, J. (1998). An axiomatic characterization of causal counterfactuals. *Found. Sci.* **3** 151–182.

Geng, Z. (1989). Algorithm AS 299. Decomposability and collapsibility for log-linear models. *J. Roy. Stat. Soc. Ser. C* **38** 189–197.

Geng, Z. (1992). Collapsibility of relative risk in contingency tables with a response variable. *J. Roy Statist. Soc. Ser. B* **54** 585–593.

Goetghebeur, E. and van Houwelingen, H., eds. (1998). Analyzing noncompliance in clinical trials. *Statististic in Medicine* **17** 247–389.

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *Amer. J. Epidemiol.* **125** 761–768.

Greenland, S. (1990). Randomization, statistics, and causal inference. *Epidemiology* **1** 421–429.

Greenland, S. and Mickey, R. M. (1988). Closed-form and dually consistent methods for inference on collapsibility in $2 \times 2 \times K$ and $2 \times J \times K$ tables. *J. Roy. Statist. Soc. Ser. C* **37** 335–343.

Greenland, S. and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *Internat. J. Epidemiol.* **15** 413–419.

GREENLAND, S. and ROBINS, J. M. (1988). Conceptual problems in the definition and interpretation of attributable fractions. *Amer. J. Epidemiol.* **128** 1185–1197.

GREENLAND, S., PEARL J. and ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

GROVES, E. R. and OGBURN, W. F. (1928). *American Marriage and Family Relationships* 160–164. Holt, New York.

GUO, J. AND GENG, Z. (1995). Collapsibility of logistic regression coefficients. *J. Roy Statist. Soc. Ser. B* **57** 263–267.

HALLORAN, M. E. and STRUCHINER, C. J. (1995). Causal inference for infectious diseases. *Epidemiol.* **6** 142–151.

HAMILTON, M. A. (1979). Choosing a parameter for 2×2 table or $2 \times 2 \times 2$ table analysis. *Amer. J. Epidemiol.* **109** 362–375.

HAUCK, W. W., NEUHAS, J. M., KALBFLEISCH, J. D. and ANDERSON, S. (1991). A consequence of omitted covariates when estimating odds ratios. *J. Clin. Epidemiol.* **44** 77–81.

HAUSMAN, J. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271.

HECKMAN, J. J. and HOTZ, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training (with discussion). *J. Amer. Statist. Assoc.* **84** 862–874.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.

HUME, D. (1739). *A Treatise of Human Nature.* Oxford Univ. Press. (Reprinted 1888.)

HUME, D. (1748). *An Enquiry Concerning Human Understanding.* Open Court Press, LaSalle. (Reprinted 1888.)

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure-Time Data*. Wiley, New York.

KELSEY, J. L., WHITTEMORE, A. S., EVANS, A. S. and THOMPSON, W. D. (1996). *Methods in Observational Epidemiology*, 2nd ed. Oxford Univ. Press.

KITAGAWA, E. M. (1955). Components of a difference between two rates. *J. Amer. Statist. Assoc.* **50** 1168–1194.

LAURITZEN, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

LEWIS, D. (1973a). Causation. *J. Philos.* **70** 556–567.

LEWIS, D. (1973b). *Counterfactuals*. Blackwell, Oxford.

MACMAHON, B. and PUGH, T. F. (1967). Causes and entities of disease. In *Preventive Medicine* (D. W. Clark and B. MacMahon, eds.) 11–18. Little Brown, Boston.

MCKECHNIE, J. L. (ed.) (1979). *Webster's New Twentieth Century Dictionary*. Simon and Schuster, New York.

MIETTINEN, O. S. (1972). Components of the crude risk ratio. *Amer. J. Epidemiol.* **96** 168–172.

MIETTINEN, O. S. and COOK, E. F. (1981). Confounding: essence and detection. *Amer. J. Epidemiol.* **114** 593–603.

MILL, J. S. (1843). *A System of Logic, Ratiocinative and Inductive*. (Reprinted 1956 by Longmans, Green, London.)

MILL, J. S. (1862). *A System of Logic, Ratiocinative and Inductive*, 5th ed. Parker, Bowin, London.

NEUHAUS, J. M., KALBFLEISCH, J. D. and HAUCK, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *Internat. Statist. Rev.* **59** 25–35.

NEYMAN, J. (1923). Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. [English translation of excerpts (1990) by D. Dabrowska and T. Speed, it Statist. Sci. **5** 463–472.]

NEYMAN, J. (1935). Statistical problems in agricultural experimentation (with discussion). *J. Roy. Statist. Soc. Suppl.* **2** 107–180.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

PEARL, J. (1997). On the identification of nonparametric structural models. In *Latent Variable Modeling with Application to Causality* (M. Berkane, ed.) 29–68. Springer, New York.

PEARL, J. and ROBINS, J. M. (1995). Probabilitic evaluation of sequential plans from causal model with hidden variables. In *Uncertainty in Artificial Intelligence* (P. Besnard and S. Hanks, eds.) **11** 444–453. Morgan-Kaufman, San Francisco.

PRENTICE, R. L. and KALBFLEISCH, J. D. (1988). Author's reply. *Biometrics* **44** 1205.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modeling* **7** 1393–1512.

ROBINS, J. M. (1987a). Addendum to "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect." *Computers Math. Appl.* **14** 923–945.

ROBINS, J. M. (1987b). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J. Chronic Dis.* **40** (suppl. 2) 139s–161s.

ROBINS, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine* **7** 773–785.

ROBINS, J. M. (1995a). Discussion of "Causal diagrams for empirical research" by J. Pearl. *Biometrika* **82** 695–698.

ROBINS, J. M. (1995b). An analytic method for randomized trials with informative censoring. *Lifetime Data Analysis* **1** 241–254.

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling with Applications to Causality* (M. Berkane, ed.) Springer, New York, 69–117.

ROBINS, J. M. (1998). Correction for non-compliance in equivalence trials. *Statistics in Medicine* **17** 269–302.

ROBINS, J. M. and GREENLAND, S. (1986). The role of model selection in causal inference from nonexperimental data. *Amer. J. Epidemiol.* **123** 393–402.

ROBINS, J. M. and GREENLAND, S. (1989). The probability of causation under a stochastic model for individual risks. *Biometrics* **46** 1125–1138.

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

ROBINS, J. M. and GREENLAND, S. (1994). Adjusting for differential rates of prophylaxis therapy for PCP in high versus low dose AZT treatment arms in an AIDS randomized trial. *J. Amer. Statist. Assoc.* **89** 737–749.

ROBINS, J. M. and MORGENSTERN, H. (1987). The mathematical foundations of confounding in epidemiology. *Computers Math. Appl.* **14** 869–916.

ROBINS, J. M. and RITOV, Y. (1997). Toward a curse-of-dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Statistics in Medicine* **16** 285–319.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (1999). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology* (E. Halloran, ed.) Springer, New York.

ROBINSON, L. D. and JEWELL, N. P. (1991). Some surprising results about covariate adjustment in logistic regression. *Int. Statist. Rev.* **59** 227–240.

ROSENBAUM, P. R. (1995). *Observational Studies*. Springer, New York.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

ROTHMAN, K. J. (1977). Epidemiologic methods in clinical trials. *Cancer* **39** 1771–1775.

ROTHMAN, K. J. and GREENLAND, S. (1998). *Modern Epidemiology*, 2nd ed. Lippincott-Raven, Philadelphia.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.* **66** 688–701.

RUBIN, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6** 34–58.

RUBIN, D. B. (1990). Comment on "Neyman (1923) and causal inference in experiments and observational studies." *Statist. Sci.* **5** 472–480.

RUBIN, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* **47** 1213–1234.

SENN, S. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine* **8** 467–475.

SIMON, H. A. and RESCHER, N. (1966). Cause and counterfactual. *Philos. Sci.* **33** 323–340.

SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statis. Soc. Ser. B* **13** 238–241. [Reprinted in (1987) *The Evolution of Epidemiologic Ideas* (S. Greenland, ed.) 103–107. ERI Press, Chestnut Hill, MA.]

SLUD, E. V., BYAR, D. P. and SCHATZKIN, D. P. (1988). Dependent competing risks and the latent-failure model. *Biometrics* **44** 1203–1204.

SOBEL, M. E. (1995). Causal inference in the social and behavioral sciences. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (G. Arminger, C. C. Clogg, and M. E. Sobel, eds.) Plenum Press, New York.

STALNAKER, R. C. (1968). A theory of conditionals. In *Studies in Logical Theory* (N. Rescher, ed.) Blackwell, Oxford.

STONE, R. (1993). The assumptions on which causal inference rest. *J. Roy. Statist. Soc. Ser. B* **55** 455–466.

WERMUTH, N. (1987). Parametric collapsibility and lack of moderating effects in contingency tables with a dichotomous response variable. *J. Roy. Statist. Soc. Ser. B* **49** 353–364.

WERMUTH, N. (1989). Moderating effects of subgroups in linear models. *Biometrika* **76** 81–92.

WHITE, H. A. (1994). *Estimation, Inference, and Specification Analysis*. Cambridge Univ. Press.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, New York.

WHITTEMORE, A. S. (1978). Collapsing multidimensional contingency tables. *J. Roy. Statist. Soc. Ser. B* **40** 328–340.

WICKRAMARATNE, P. and HOLFORD, T. (1987). Confounding in epidemiologic studies: the adequacy of the control group as a measure of confounding. *Biometrics* **43** 751–765.

YULE, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika* **2** 121–134.