Annu. Rev. Public Health. 2000. 21:121–45

# CAUSAL EFFECTS IN CLINICAL AND EPIDEMIOLOGICAL STUDIES VIA POTENTIAL OUTCOMES: Concepts and Analytical Approaches

Roderick J. Little
*University of Michigan, Ann Arbor, Michigan 48109-2029; e-mail: rlittle@umich.edu*

Donald B. Rubin
*Harvard University, Cambridge, Massachusetts 02138; e-mail: rubin@hustat.harvard.edu*

**Key Words** statistical inference, randomization, observational studies, noncompliance, missing data

■ **Abstract** A central problem in public health studies is how to make inferences about the causal effects of treatments or agents. In this article we review an approach to making such inferences via potential outcomes. In this approach, the causal effect is defined as a comparison of results from two or more alternative treatments, with only one of the results actually observed. We discuss the application of this approach to a number of data collection designs and associated problems commonly encountered in clinical research and epidemiology. Topics considered include the fundamental role of the assignment mechanism, in particular the importance of randomization as an unconfounded method of assignment; randomization-based and model-based methods of statistical inference for causal effects; methods for handling noncompliance and missing data; and methods for limiting bias in the analysis of observational data, including propensity score matching and sensitivity analysis.

## INTRODUCTION

A central problem in clinical and epidemiological studies is how to make inferences about the causal effects of treatments or agents; for example, does an unsanitary water supply cause cholera epidemics, and can public health measures to improve the water supply reduce this cause of suffering and mortality? Does exposure to smoking or to high levels of toxic chemicals cause cancer? Does a new treatment for cancer extend survival?

In this article we review an approach to making inferences about causal effects via potential outcomes, and we discuss the application of this approach to

0163-7527/00/0510-0121$14.00

a number of data collection designs and associated problems commonly encountered in clinical research and epidemiology. The formulation is often implicit in other approaches to causal inference, but our focus is on the insights provided by the potential-outcome formulation, rather than a comparison with other approaches. The article is an expansion of the work of Rubin (70), which focuses on the teaching of causal inference in departments of statistics and economics; a textbook on this topic is forthcoming (GW Imbens & DB Rubin, manuscript in preparation).

Below we define the causal effect of a treatment or agent on an outcome, for an individual and for a population, and state the assumptions associated with this definition. Then we emphasize the crucial roles of the mechanisms of treatment assignment and sample selection in estimating causal effects, after which we review three approaches to inference both in the comparatively straightforward setting of a randomized clinical trial (RCT) with full compliance and no missing data and in more complex settings involving covariates or sequential treatment allocation. We consider complications to causal inference in randomized clinical trials arising from noncompliance or missing data. We consider causal inference in observational studies, which can potentially be biased by self-selection of treatments. Analytical approaches such as covariate adjustment, matching methods, propensity matching and subclassification, and sensitivity analysis are reviewed. Finally, we provide some concluding remarks and an appendix with a summary of the notation and a glossary of terms.

## DEFINITION OF CAUSAL EFFECTS THROUGH POTENTIAL OUTCOMES

The definition of "cause" is complex and challenging, but for empirical research, the idea of the causal effect of an agent or treatment seems more straightforward and practically useful. A key concept is the definition of causal effects through potential outcomes. Causal estimands are comparisons of the potential outcomes that would have been observed under different exposures of units to treatments. For example, the causal effect of taking aspirin 2 hours earlier involves a comparison of the state of one's headache now with what it would have been had no aspirin been taken; $Y(1)$ is the outcome (a measure of headache pain) without aspirin, and $Y(2)$ is the outcome with aspirin; the difference, $Y(2) - Y(1)$, is an obvious definition of the causal effect of aspirin on headache. More generally, $Y(\mathrm{W})$ is the outcome under treatment $W = 1$, 2, etc.

This intuitive and transparent definition of causal effects via potential outcomes is known in some circles (24) as Rubin's Causal Model, but the formal notation, in the context of randomization-based inference in randomized experiments, dates back to Neyman (40, 66), and the intuitive idea can be found in the empirical-research literature in several fields, including economics (21, 22, 43, 84), epidemiology (20, 48, 49, 51), the social and behavioral sciences (79, 80, 87), and statistics

(9, 14, 42, 62). Rubin (57, 60, 62) provided a formal framework that extended the idea beyond randomized experiments and randomization-based inference; this extended framework also allows the formal consideration of complications, such as unintended missing data and noncompliance.

The key problem for inference is that we get to observe the outcome $Y$ for only one of the possible treatments, namely the treatment actually assigned—in particular, if we assign aspirin to treat the headache, then we observe $Y(2)$ but do not observe $Y(1)$, the outcome we would have seen if aspirin were not assigned. The outcomes under treatments not assigned can be regarded as missing, and the problem is one of drawing inferences about these missing values with the observed data.

When the individual-level definition of causal effects through potential outcomes is applied to a set of individuals, a complication is that the outcomes for a particular individual may depend on the treatments and outcomes for other individuals in the set. For example, suppose we are in the same room and *your* taking aspirin for your headache affects the state of my headache whether or not I take aspirin (if you don't take it and eliminate your headache, your complaining will be unbearable!). Figure 1 indicates that, for two individuals, there are then four possible outcomes: $Y(11)$, $Y(12)$, $Y(21)$ and $Y(22)$ for each subject, where $Y(jk)$ denotes the outcome when the first individual receives treatment $j$ and the second individual receives treatment $k$. For more than two individuals, the number of potential outcomes increases exponentially. This problem is avoided by the assumption of "no interference between units" (9) or, more generally (62), the "stable unit-treatment value" assumption (SUTVA), which also assumes that there are no hidden versions of treatments (e.g. effective and ineffective aspirin tablets). The main idea is that the causal effect for a particular individual does not depend on assignments of other individuals. Under SUTVA, each unit has just two potential outcomes under treatments 1 and 2. Hence the full set of potential outcomes for $N$ units can be represented by an $N \times 2$ array, as in Figure 2. Without some such exclusion restrictions (to use the language of economists) which limit the range of potential outcomes, causal inference is impossible. Nothing is wrong with making assumptions; on the contrary, they are the strands that join the field of statistics to scientific disciplines. The quality of these assumptions, not their existence, is the issue.

| Subject | $Y(11)$ | $Y(12)$ | $Y(21)$ | $Y(22)$ |
|---|---|---|---|---|
| 1 | a | b | c | d |
| 2 | e | f | g | h |

**Figure 1** The four potential outcomes for two subjects when stable unit-treatment value is not assumed. $Y(jk) =$ outcome given that subject 1 receives treatment $j$ and subject 2 receives treatment $k$.

| Subject | $Y(1)$ | $Y(2)$ |
|---|---|---|
| 1 | a = b | c = d |
| 2 | e = g | f = h |
| … | … | … |
| $N$ | y | z |

**Figure 2** Potential outcomes for $N$ subjects under SUTVA. Entries for first two subjects correspond to Figure 1.

# ASSIGNMENT AND SELECTION MECHANISMS

## The Need for a Posited Assignment Mechanism

Without a model for how treatments are assigned to units, formal causal inference, at least using probabilistic statements, is impossible. This does not mean that we cannot do anything unless we know the assignment mechanism, but it does mean that probabilistic statements about causal effects (such as confidence intervals or claims that estimates are unbiased) cannot be made without assumptions about the nature of the mechanism.

To illustrate the need to posit an assignment mechanism, consider the situation depicted in Table 1, in which each of four patients is assigned to a control treatment (treatment 1) or a new treatment (treatment 2) by a physician where $Y$ is the number of years lived after the operation. Column 1 refers to the potential outcomes under treatment 1, the vector $Y(1)$, and column 2 to the potential outcomes, $Y(2)$, under treatment 2; this representation makes the SUTVA assumption. The individual causal effects, defined as the difference of years lived under treatments 2 and 1, are given in the third column. The fourth column, labeled $W$, indicates the treatment each patient actually received. The doctor in this example is one each of us would want to use, because the last two columns of Table 1 show that each patient gets assigned the better treatment—not better for the average patient but better for the individual patient. But what general conclusions do the observed data suggest? The average observed length of survival for those given treatment 1 is 1 year more than for those given treatment 2, so the obvious, but incorrect, conclusion is that treatment 1 is superior for the pool of patients for which these four patients are representative. This conclusion is wrong because the typical causal effects (e.g. the average or median of the individual causal effects in Table 1, column 3) clearly favor treatment 2 for these patients, giving an average benefit of 3 years.

The definition of causal effects via potential outcomes and the formal consideration of the assignment mechanism clarify the roles of two strong design features of epidemiological and clinical studies: the inclusion of a control group, and the randomization of treatments to subjects. Specifically, because the causal effect of

**TABLE 1** Artificial example of a confounded assignment mechanism[a]

| | **Years lived after indicated treatment:** | | | |
|---|---|---|---|---|
| **Patient** | $\boldsymbol{Y}(\mathbf{1})$ | $\boldsymbol{Y}(\mathbf{2})$ | $\boldsymbol{Y}(\mathbf{2}) - \boldsymbol{Y}(\mathbf{1})$ | $\boldsymbol{W}$ |
| 1 | (1) | 6 | (5) | 2 |
| 2 | (3) | 12 | (9) | 2 |
| 3 | 9 | (8) | $(-1)$ | 1 |
| 4 | 11 | (10) | $(-1)$ | 1 |
| Mean | 10[b] | 9[b] | (3) | |

[a] $Y(1)$, years lived under control treatment; $Y(2)$, years lived under new treatment; $Y(2) - Y(1)$, difference in years lived under treatments 1 and 2; $W$, treatment each patient actually received. This representation makes the "stable unit-treatment value" assumption; values in parentheses are not observed. For example, subject 1 in the first row is assigned $W = 2$ and survives $Y(2) = 6$ years; if assigned $W = 1$, the subject would have survived $Y(1) = 1$ year, so the causal effect of the new versus control treatment is $Y(2) - Y(1) = 5$ years. The values $Y(1) = 1$ and $Y(2) - Y(1) = 5$ are in parentheses because they are not observed.

[b] Mean of observed values. The difference of these values ($-1$) has the wrong sign as an estimate of the true average causal effect (3).

a particular treatment is defined with reference to a second control treatment, a control group is needed unless the outcome under the control treatment is very well understood, because we get to observe only one of the two outcomes for any given individual. Randomization is an assignment mechanism that allows causal effects for groups to be estimated relatively easily. In particular, in classical completely randomized experiments, $W$ is assigned randomly and hence independently of outcomes. In terms of conditional probability, the probability that an individual with potential outcomes $Y(1)$ and $Y(2)$ is assigned a treatment $W$ is a constant that does not depend on the values of $Y(1)$ and $Y(2)$, that is:

$$\boldsymbol{P}(W \mid Y(1), Y(2)) = \boldsymbol{P}(W) \quad \text{for all} \quad Y(1), Y(2) \qquad 1.$$

where the left side is the conditional probability of assignment $W$ given potential outcomes $Y(1)$ and $Y(2)$. Let $\overline{y}_1$ and $\overline{y}_2$ denote the sample means of the outcomes of individuals assigned to treatments 1 and 2, respectively. Under random assignment, the difference in the sample means $\overline{y}_1 - \overline{y}_2$ is an unbiased estimate of the average causal effect.

Let $X$ denote a set of covariates recorded for an individual, and let $Y_{obs}$ denote the recorded outcome and $Y_{mis}$ the unrecorded outcomes of treatments not assigned. Equation 1 is an example of an "ignorable" mechanism, the essential feature of which is that assignment depends on the data only through $X$ and $Y_{obs}$; that is, an ignorable mechanism has the property that

$$\boldsymbol{P}(W \mid X, Y_{obs}, Y_{mis}) = \boldsymbol{P}(W \mid X, Y_{obs}) \quad \text{for all} \quad Y_{mis} \qquad 2.$$

Important examples of ignorable mechanisms are "unconfounded" experiments, in which assignment depends only on observed covariates *X:*

$$\boldsymbol{P}(W \mid X, Y_{obs}, Y_{mis}) = \boldsymbol{P}(W \mid X) \quad \text{for all} \quad Y_{mis}, Y_{obs} \tag{3.}$$

Other important designs allow assignment to depend not only on covariates but also on observed outcomes, and hence they do not satisfy Equation 3 but do satisfy the weaker ignorability condition, Equation 2. In particular, in studies involving staggered entry of subjects, biased-coin designs (85) allow allocations of subjects to depend on observed outcomes of other subjects assigned earlier in the study. These designs can be used to limit assignments to the treatment that appears to be inferior. In studies involving repeated measurements over time and sequential treatment allocations, the allocation at a particular time may depend on outcomes recorded for the same individual at earlier time points, yielding a form of sequential ignorability (46).

Of course, in many epidemiological studies, assignment by randomization is not possible. In such cases, any analysis of the observed data must take the assignment mechanism into account; because at least half of the potential outcomes (which define causal effects) are missing, the process that makes them missing must be part of the inferential model (58, p. 581).

Mechanisms in which treatment assignments are potentially entangled with the outcomes, as in the example displayed in Table 1, are called "nonignorable." Below we discuss inferential methods for randomized studies, and later we consider observational studies in which the mechanism is potentially nonignorable.

## Selection Mechanisms for Inference about a Population Based on a Sample

Causal inference for a particular sample of individuals depends on the mechanism of assignment of treatments to that sample. Usually we are interested in making inference about a larger population of individuals, which the sample is intended to represent. Such inferences depend not only on the mechanism of treatment allocation but also on the mechanism of selection of the sample from the population. The ideal design for causal inference about a population would involve random selection of subjects and random allocation of treatments to those subjects. However, studies involving randomized selection and treatment allocation are rare in practice. Thus observational studies may involve random sampling of a target population but not random allocation of treatments or conditions to be studied. On the other hand, clinical trials often involve random allocation of treatments but rarely involve random selection of subjects from the population.

The selection mechanism can be formalized as the distribution of the set of binary variables *S* that indicate whether members of the population are selected. Paralleling the definition of Equation 2 for treatment allocation, the main condition for an ignorable-selection mechanism is that selection depends on the data only through observed outcomes; that is

$$\boldsymbol{P}(S \mid X, W, Y_{obs}, Y_{mis}, Y_{exc}) = \boldsymbol{P}(S \mid X, W, Y_{obs}) \quad \text{for all} \quad Y_{mis}, Y_{exc} \qquad 4.$$

where $Y_{exc}$ represents the potential outcomes for units of the population excluded from the sample. If Equation 4 can be justified, then the selection mechanism can be ignored for population causal inference, but otherwise a model for selection needs to be incorporated into the analysis. The extent to which nonignorable selection leads to bias in the estimated population effects depends on the degree of association between treatment effects and selection, after adjusting for measured covariates. This issue is addressed informally when average treatment effects are compared from a variety of studies involving different study samples; relative stability of the estimates is evidence that effects of selection may be minor, but large heterogeneity of effects suggests that the issues of selection need careful attention.

## MAJOR MODES OF INFERENCE FOR RANDOMIZED STUDIES

We do not need any more assumptions to proceed with forms of causal inference that are based solely on the distribution of statistics induced by a randomized assignment mechanism. These forms of inference were developed in the context of classical randomized experiments. It is useful to describe methods for drawing valid causal inferences in this setting before we consider more complicated settings involving nonrandomized data.

Fundamentally, there are three formal statistical modes of causal inference, one Bayesian and two randomization based (67). Of the two distinct forms of randomization-based inference, one is due to Neyman (40) and the other is due to Fisher (15).

### Fisherian Randomization-Based Inference

Fisher's approach is the more direct conceptually, and it is closely related to the mathematical idea of proof by contradiction. It is basically a stochastic proof by contradiction, giving the *P* value, a measure of the plausibility of the null hypothesis of absolutely no effect whatsoever on the subjects included in the study.

The first element in Fisher's mode of inference is the null hypothesis, which is nearly always $Y(1) = Y(2)$ for all units; both treatments have exactly the same outcomes. Under this null hypothesis, all potential outcomes are known from the observed outcome $Y_{obs}$ because $Y(1) = Y(2) = Y_{obs}$. It follows that, under this null hypothesis, the value of any statistic, $T$, such as the difference of the observed averages for units exposed to treatment 2 and units exposed to treatment 1, $\overline{y}_2 - \overline{y}_1$, is known not only for the observed assignment, but for all possible assignments $W$.

Suppose we choose a statistic *T*, such as $\bar{y}_2 - \bar{y}_1$, and calculate its value under each possible assignment (assuming that the null hypothesis is true) and also calculate the probability of that assignment under the randomized assignment mechanism. In most classical experiments, these probabilities are either zero or a common value for all possible assignments. For example, in a completely randomized experiment with $N = 2n$ units, *n* are randomly chosen to receive treatment 2 and *n* to receive treatment 1. Then any assignment *W* that has *n* units assigned to each treatment has probability $1/C_n^N$, where $C_n^N = (N) \times (N-1) \times \cdots \times (n+1)/(1 \times 2 \times \cdots \times (N-n))$ is the binomial coefficient, and all other *W*'s have zero probability. Knowing the value of *T* and its probability for each *W*, we can then calculate the probability (under the assignment mechanism and the null hypothesis) that we would observe a value of *T* that is as "unusual" as or more unusual than the observed value of *T*, where "unusual" is defined a priori, typically by how discrepant *T* is from zero. This probability is the *P* value of the observed value of the statistic *T* under the null hypothesis; small values are interpreted as evidence against the null hypothesis.

This form of inference is very elegant, but limited; the null model is very restricted, the test statistic is somewhat arbitrary, and a small *P* value does not necessarily imply that deviations from the null hypothesis are substantively important. Nevertheless, there is merit in calculating a *P* value; if the null hypothesis is not rejected for an appropriate choice of statistic *T*, it is hard to claim evidence for treatment differences.

## Neyman's Randomization-Based Inference

Neyman's form of randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics over the distribution induced by the assignment mechanism to calculate a confidence interval for the typical causal effect. The essential idea is the same as in Neyman's (39) classic article on randomization-based (now often called design-based) inference in surveys. Typically, an unbiased estimator of the causal estimand (the typical causal effect, e.g. the average) is created, and an unbiased or upwardly biased estimator of the variance of that unbiased estimator is found (bias and variance both defined over the randomization distribution). Then an appeal is made to the central limit theorem for the normality of the estimator over its randomization distribution, from which a confidence interval for the causal estimand is obtained.

To be more explicit, the causal estimand is typically the average causal effect $\bar{Y}_2 - \bar{Y}_1$, where the averages are over all units in the population being studied and the traditional statistic for estimating this effect is the difference in sample averages for the two groups, $\bar{y}_2 - \bar{y}_1$, which can be shown to be unbiased for $\bar{Y}_2 - \bar{Y}_1$ under simple random assignment of treatments. In completely randomized experiments with $n_1$ units assigned the first treatment and $n_2$ units assigned the second treatment, the estimated variance is $se^2 = s_1^2/n_1 + s_2^2/n_2$, where $s_1^2$ and $s_2^2$ are the sample variances in the two treatment groups; $se^2$ overestimates the actual variance of $\overline{y}_2 - \overline{y}_1$ unless all individual causal effects are the same (that is, the causal effect is additive). The standard 95% confidence interval for $\overline{Y}_2 - \overline{Y}_1$ is $\overline{y}_2 - \overline{y}_1 \pm 1.96se$, which, in large enough samples, includes $\overline{Y}_2 - \overline{Y}_1$ in $\geq$95% of the possible random assignments. If we view the $N$ units as being a random sample from an essentially infinite population of units, then $se^2$ is exactly unbiased for the variance of the $\overline{y}_2 - \overline{y}_1$ over random sampling of units and randomized treatment assignment.

Neyman's form of inference is not prescriptive in telling us what to do, but rather it is aimed at evaluations of procedures for causal inference; in repeated applications, how often does the interval $\overline{y}_2 - \overline{y}_1 \pm 1.96se$ include $\overline{Y}_2 - \overline{Y}_1$? Nevertheless, it forms the basis for most of what is done in important areas of application (e.g. the worlds of pharmaceutical development and approval and randomized clinical trials in medicine). This form of inference is also the basis for most of what is done for causal inference in epidemiology when a randomized study has been done or assumed. That is, the analysis of nonrandomized epidemiological data is nearly always based on Neymanian inference under an implicit assumption that at some level, discussed later, randomization took place.

## Model-Based Inference

The third form of statistical inference for causal effects is model-based inference, in which the model for the assignment mechanism, $\boldsymbol{P}(W \mid X, Y(1), Y(2))$, is supplemented with a model for the data, $\boldsymbol{P}(Y(1), Y(2) \mid X)$, which is nearly always indexed by unknown parameters $\theta$. In our view, the Bayesian version of model-based inference (62) is the most general and conceptually satisfying. [In this approach, $\theta$ is given a prior distribution and causal inference is obtained from the conditional distribution of the causal estimand given observed data, which follows by Bayes' theorem from the observed data and the models for the assignment mechanism and the data.]

More explicitly, suppose that the causal estimand is $\overline{Y}_2 - \overline{Y}_1$, as before; then its posterior distribution, that is, its conditional distribution given the model specifications and the observed values of $W$, $X$, and $Y$, follows from the posterior predictive distribution of the missing values $Y_{mis}$, evaluated at the observed values of $W$, $X$, and $Y_{obs}$. With large samples, the Bayesian analysis for a simple independent normal model agrees very closely with Neyman's confidence approach. The posterior distribution of $\overline{Y}_2 - \overline{Y}_1$ is normal with mean $\overline{y}_2 - \overline{y}_1$ and variance $s_1^2/n_1 + s_2^2/n_2 - c$, where $c$ is zero for inference about means in an infinite population and in a finite population $c$ is proportional to the variance of the individual causal effects and thus is zero when the causal effect is additive (66). For more complex problems that lack simple analytical posterior distributions, posterior distributions can be computed by Markov chain Monte Carlo methods (82) that effectively multiply-impute (65) the values of the missing potential outcomes $Y_{mis}$. The approach of simulating the missing potential outcomes is very intuitive, even for nonstatisticians.

It should be emphasized here that parameters $\theta$ of the models for the data (for example, regression coefficients in normal linear regression models) are *not* causal effects. For example, suppose that $Y$ is a binary variable indicating 5-year survival. It may be reasonable to model the log-odds of the probability of survival as an additive function of treatment group and covariates. On the other hand, the causal effects of interest may be direct comparisons of survival rates under new and control treatments. The investigator defines the causal estimand, but nature chooses the distributional forms for the variables. Bayesian inference via simulation is ideally suited to address such questions.

In our view, the Bayesian model-based approach is by far the most direct and flexible mode of inference for causal effects. However, model-based approaches achieve these benefits by postulating a distribution for the data $\boldsymbol{P}(Y(1), Y(2) \mid X)$, which the randomization-based approaches avoid. Such a distribution can be very useful, but it can be like a loaded gun in the hands of a child, fraught with danger for the naive data analyst. Models need to be carefully chosen to reflect the key features of the problem and should be checked for important violations of key assumptions and modified where necessary.

In summary, one should be willing to use the best features of all of these inferential approaches. In simple classical randomized experiments with normal-like data, the three approaches give similar practical answers, but they do not in more difficult cases in which each perspective provides different strengths. Ideally, in any practical setting, the answers from the perspectives are not in conflict, at least with appropriate models, proper conditioning, and large samples.

## The Role of Covariates in Randomized Studies

Covariates are variables whose values are not affected by the treatment assignment and subsequent outcomes, such as variables that are recorded before randomization into treatment groups (e.g. year of birth or pretest scores in an educational evaluation). Covariates are typically critical to a sensible analysis of an epidemiological study. But it is important to understand first their role in randomized experiments. If a covariate is used in the assignment mechanism, as with a blocking variable in a randomized block design, that covariate must be used in analysis for valid inference, regardless of inferential perspective. In classical completely randomized experiments in which covariates are not used for treatment assignment, covariates can still be used to increase efficiency of estimation.

The role of covariates in increasing efficiency is most transparent with Bayesian inference; for example, in a clinical trial to test a new medical treatment, observed pretreatment disease severity scores are typically predictive of missing post-treatment outcomes, for example, post-treatment outcomes under the treatment not received. Therefore, using pretreatment scores improves prediction of $Y_{mis}$ and reduces the variance of the imputed missing potential outcomes. For instance, when using a linear regression model, the residual variance of post-treatment scores after adjusting for pretreatment scores is less than the variance of raw post-treatment scores.

From the Fisherian and Neymanian perspectives, we can use these covariates to define a new statistic to estimate causal estimands. For example, one can use the difference in average observed gain scores $(\overline{y}_2 - \overline{x}_2) - (\overline{y}_1 - \overline{x}_1)$, where $\overline{x}_2$ and $\overline{x}_1$ are average observed pretreatment scores and $\overline{y}_2$ and $\overline{y}_1$ are average observed post-treatment scores, rather than the difference in average post-treatment scores $\overline{y}_2 - \overline{y}_1$, to estimate $\overline{Y}_2 - \overline{Y}_1$. If $X$ and $Y$ are correlated from the Neymanian perspective, the variance of the difference in average gain scores will be less than the variance of the difference in average post-test scores, which translates into smaller estimated variances and shorter confidence intervals. From the Fisherian perspective, this reduced variance translates into more powerful tests of the null hypothesis; under an alternative hypothesis, smaller real deviations from the null hypothesis are more likely to lead to more significant $P$ values.

Suppose as an extreme case that the new treatment adds essentially 10 points to everyone's pretreatment score, whereas the old treatment does nothing. The observed gain scores have essentially zero variance in each treatment group, whereas the post-treatment scores have the same variances as the pretreatment scores. This means that the Neymanian confidence interval for the difference in gain scores is much shorter than that for the post-treatment scores. Also, the observed value of the difference of gain scores is the most extreme value that can be observed under the Fisherian null hypothesis, and so the observed result with gain scores is as significant as possible, which is not true for the difference in observed post-treatment scores.

## More Complex Randomized Experiments

The potential outcome formulation of causal effects extends in a natural way to more complex randomized designs. For example, if $K > 2$ treatments are randomized, then each subject has $K$ potential outcomes, one for each treatment. If the treatments are combinations in a factorial design with $J$ factors, with factor $j$ having $k_j$ levels, then there are $K = k_1 \times k_2 \times \cdots \times k_J$ treatment combinations, and analysis methods of classical experimental design (6, 9) can be applied to study the causal effects of factors and their interactions.

In a longitudinal study with a sequence of treatment decisions and sequential ignorability, suppose that there are $J$ time points at which treatments are assigned, and at time point $j$ there are $k_j$ possible treatment choices. Again, there are $K = k_1 \times k_2 \times \cdots \times k_J$ possible treatment combinations and $K$ potential outcomes for each individual in the study. The assignment of a treatment option at a particular time point may depend on $Y_{obs}$ through outcomes for treatments recorded before that time point. In a series of papers (45–50), Robins and colleagues develop models for assessing the causal effects of particular treatment combinations in this setting, typically the combination that assigns the active treatment at every time point compared with the combination that assigns the control treatment at every time point. See Wasserman (86) for an accessible account of the basic ideas underlying these methods. Because some treatment combinations may not be observed, particularly when $K$ is large, these methods by necessity need to make relatively strong modeling assumptions to make inferences feasible. Nevertheless, if the assumptions are plausible, answers are likely to be superior to those obtained by methods that fail to take into account the sequential nature of the treatment allocation process.

When the probability of treatment assignment is allowed to depend on covariates or recorded outcomes and hence varies from unit to unit, the individual assignment probabilities $p_i \equiv \boldsymbol{P}(W_i \mid X_i, Y_{obs})$ are called propensity scores (54, 74). For Neymanian inference, they are required to be strictly between 0 and 1, and they are key quantities in the analysis. When the propensity scores are known, the assignment mechanism is essentially known, and simple generalizations of Fisherian and Neymanian modes of inference can be applied. In particular, Horvitz-Thompson estimation (27), in which observations are weighted by the inverse probabilities of their being observed, plays a key role for both modes of inference. An important point is that when there is little or no overlap in the propensity scores in the treatment groups, no causal inference is possible without strong external assumptions (60).

With Bayesian inference, an unconfounded assignment mechanism is ignorable (62), so that, after including the covariates that determine the assignment probabilities or an adequate summary (such as, possibly, the vector of propensity scores in the model), analysis in principle proceeds as in a classical randomized experiment. In this situation, however, there can be much greater sensitivity to model specification than in a classical randomized experiment. This sensitivity is the Bayesians' analog of the increased variance of the Horvitz-Thompson estimator in unbalanced designs and the decreasing power in such designs as they become more unbalanced.

## THREATS TO UNCONFOUNDED TREATMENT ASSIGNMENT

### Introduction

Randomized studies with full compliance and no missing data are relatively easily analyzed by the methods described above. However, in practice, inference for causal effects is often complicated by lack of compliance and incomplete data. The potential outcome formulation provides conceptual clarity and a set of tools for negotiating these complications. In the next section we consider the issue of noncompliance, and later we consider the issue of missing data.

### Noncompliance

Analysis of a randomized study is complicated when subjects do not comply with their assigned treatment. An intention-to-treat (ITT) analysis compares the outcomes of subjects by randomization groups (an as-randomized analysis), ignoring the compliance information. The ITT effect measures the effect of treatment randomization rather than the effect of treatment for those who actually received it. The ITT estimator is protected from selection bias by randomized treatment assignment, but in general is a distorted measure of the effect of the treatment itself. Alternatively, subjects can be classified by the treatments actually received (an "as-treated" analysis). The problem then is that randomization is violated, and confounding factors associated with switching from the assigned treatments potentially corrupt the causal interpretation of treatment effects.

The potential outcome definition of causal effects provides a useful basis for understanding noncompliance problems and assumptions implied by various estimation strategies (1, 28, 29). Participants are classified as one of four types, compliers, defiers, never-takers, and always-takers, in the context of this experiment. Compliers are people who would adopt whatever treatment is assigned, never-takers are people who would take the control treatment regardless of what they are assigned, always-takers would take the active treatment regardless of what they are assigned, and defiers are those who would adopt the opposite treatment to their assignment. Let $C$ denote the compliance indicator of the participant, where $C = 1$ for compliers and $C = 0$ for all noncompliers: defiers, never-takers, and always-takers.

In practice, we observe compliance status only for the assigned treatment, so the full compliance status of subjects is incompletely observed. Specifically, if a subject is assigned to the new treatment and complies, then that subject may be a complier or an always-taker. If a subject is assigned to the new treatment and fails to comply, then that subject may be a never-taker or a defier. If a subject is assigned to the control and complies, then that subject may be a complier or a never-taker. If a subject is assigned to the control and obtains the new treatment, then that subject may be an always-taker or a defier. Imbens & Rubin (28) treat this partial information about compliance as a missing-data problem.

Let us make the SUTVA assumption, so that individual-level causal effects can be defined without reference to other individuals in the study. For a respondent assigned to group $W$, let $A(W)$ be the adopted treatment and let $Y(W)$ be the potential outcome. The unit-level causal effect of treatment assignment ($W$) on treatment adopted ($A$) is $A(2) - A(1)$, and the unit-level causal effect of treatment assignment ($W$) on outcome ($Y$) is $Y(2) - Y(1)$. These unit level effects of $W$ on $A$ and $Y$ are generally not observable, because individuals cannot be assigned to both the experimental and control treatments. However, the average ITT effects on both $A$ and $Y$ can be readily estimated by the averages over groups of respondents in a randomized trial.

Another causal effect of interest in many studies of treatments is the Complier-Average Causal Effect (CACE), which is the average causal effect for the subpopulation of compliers, denoted by $C = 1$. For compliers, $A(1) = 1$ and $A(2) = 2$, and

$$\text{CACE} = E(Y(2) - Y(1) \mid C = 1) \qquad 5.$$

where the conditioning on $C = 1$ is identical to conditioning on $A_{(1)} = 1$ and $A(2) = 2$. The CACE is a valid causal effect because it is a summary measure of individual-level effects in a subpopulation of interest, namely compliers. The CACE and the ITT effect capture two treatment features of potential interest—the effect of the treatment on those who can be induced to take it in this experiment and the overall impact of a treatment on the whole population, including noncompliers.

Two common methods of analysis, "as-treated" and "per-protocol" analysis, are generally flawed because they do not estimate the CACE or any other useful summary of individual-level causal effects (77). The as-treated analysis classifies subjects by the adopted treatment $A(W)$ and estimates

$$E(Y(W) \mid A(W) = 2) - E(Y(W) \mid A(W) = 1)$$

This is not an average of individual-level causal effects, because it compares averages of $Y$ for groups with different characteristics. Specifically, it compares the average outcome of those adopting treatment 2 (namely, compliers assigned treatment 2, defiers assigned treatment 1, and always-takers assigned either treatment), with the average outcome of those adopting treatment 1 (namely, compliers assigned treatment 1, defiers assigned treatment 2, and never-takers assigned either treatment). A "per-protocol" analysis compares subjects who actually adopted their assigned treatments and estimates

$$E(Y(2) \mid A(2) = 2) - E(Y(1) \mid A(1) = 1)$$

which is also not an average of individual-level causal effects. It compares the outcomes of those who were assigned and adopted treatment 2 (i.e. compliers and always-takers assigned to 2) with the outcomes of those who were assigned and adopted treatment 1 (that is, compliers and never-takers assigned to 1). The potential outcome formulation of causal effects helps to clarify the deficiencies of these common analysis methods. The crux of the CACE formulation is the distinction between the definition of a "true" complier (compliance under *both* treatments) and an "observed" complier (compliance under the treatment actually assigned).

The CACE in Equation 5 is a valid causal estimand, but it is not immediately obvious how to estimate it because the compliance status $C$ of individuals is generally unknown. However, the CACE can be estimated from the data under SUTVA and random assignment of $W$, if we make the following additional assumptions (1):

1. Exclusion restriction of treatment assignment given treatment received: for never-takers and always-takers, whose adopted treatment is the same regardless of which treatment is assigned, the outcome $Y$ is the same regardless of which treatment is assigned; that is, $Y(1) = Y(2)$ if $A(1) = A(2)$;
2. Monotonicity of treatment assignment and treatment adopted (there are no defiers);
3. Nonzero denominator (the population includes some compliers).

To derive an estimate of the CACE under these assumptions, note that:

$$\begin{aligned} E(Y(2) - Y(1)) &= E(Y(2) - Y(1) \,|\, C = 1)\, \boldsymbol{P}(C = 1) \\ &\quad + E(Y(2) - Y(1) \,|\, C \neq 1)\, \boldsymbol{P}(C \neq 1) \end{aligned} \qquad 6.$$

Under assumption 2, the treatment effect for never-takers and for always-takers is identically zero. If, in addition, assumption 3 holds (there are no defiers), then the second term on the right side of Equation 6 is zero, and

$$E(Y(2) - Y(1) \,|\, C = 1) = E(Y(2) - Y(1))/\boldsymbol{P}(C = 1) \qquad 7.$$

that is, the CACE is the ITT effect divided by the proportion of compliers. Now under randomized treatment allocation, the standard ITT estimator for the effect of $W$ on $Y$, the difference in sample means $\bar{y}_2 - \bar{y}_1$ is an unbiased estimate of the numerator of Equation 7. Also, let $p_2$ be the proportion of subjects in the treatment group who adopt the new treatment, and let $p_1$ be the proportion of subjects in the control group who adopt the new treatment. Then $p_2$ is an unbiased estimate of the proportion of compliers or always-takers (those who adopt the treatment when assigned it), and $p_1$ is an unbiased estimate of the proportion of always-takers (those who adopt the treatment when assigned the control). Hence $p_2 - p_1$ is an unbiased estimate of the proportion of compliers, $\boldsymbol{P}(C = 1)$. Thus an approximately unbiased estimate of the CACE is

$$IVE = \frac{\bar{y}_2 - \bar{y}_1}{p_2 - p_1}, \qquad 8.$$

which is the estimated ITT effect divided by difference in the proportions that adopt the new treatment in the new treatment and control groups. Assumption 3 assures that the denominator in Equation 8 has a nonzero expectation. Equation 8 is sometimes called the instrumental variable (IV) estimator. See Sommer & Zeger (81) and Ettner (13) for applications of this method. Our derivation of this estimator makes explicit assumptions that are hidden in other formulations (5). In particular, the exclusion restriction, assumption 1, plays a key role, and it is not a consequence of randomization of treatments.

Another advantage of the potential-outcome formulation is that it leads to more efficient estimators of the CACE than the IV estimator (Equation 8). Bayesian formulations for inference about the CACE treat the compliance indicator $C$ as missing data (23, 28), and they make use of iterative simulation techniques such as data augmentation (83). The general topic of causal inference when there is noncompliance is currently an active area of research (2–4, 12, 17, 18, 23, 37).

## Missing Data

Missing data (32–34, 65, 76) is a pervasive problem in epidemiological and clinical studies, particularly when they involve repeated measurements over time. Statistical analyses when there are missing values are too often confined to simple and relatively ad hoc fixes, such as discarding the incomplete cases or imputing unconditional means or the last recorded value in a longitudinal study. These analyses make strong assumptions about the nature of the missing data mechanism and often yield biased inferences.

Methods that impute or fill in the missing values have the advantage that observed values in the incomplete cases are retained. Thus, in randomized trials, the balancing for confounders afforded by the randomization is preserved. Better methods of imputation impute a draw from a predictive distribution of the missing values, given the observed data. For example, in stochastic regression imputation (33, section 4.5), each missing value is replaced by its regression prediction plus a random error, with variance equal to the estimated residual variance. A serious defect with all such imputation methods is that they "make up" data that do not exist. More precisely, a single imputed value cannot represent all of the uncertainty about which value to impute, so analyses that treat imputed values just like observed values generally underestimate uncertainty, even if nonresponse is modeled correctly.

Multiple imputation (34, 65, 68, 71, 72, 76) is a modification of imputation that fixes many of the defects of that method. Instead of imputing a single set of draws for the missing values, a set of $M$ (e.g. $M = 5$) datasets is created, each containing different sets of draws of the missing values from their predictive distribution. We then apply the analysis to each of the $M$ datasets and combine the results in a simple way. In particular, for scalar estimands, the multiple imputation estimate is the average of the estimates from the $M$ datasets, and the variance of the estimate is the average of the variances from the five datasets plus $1 + 1/M$ times the sample variance of the estimates over the $M$ datasets (the factor $1 + 1/M$ is a small-$M$ correction). The second component of the variance estimates the contribution to the variance from imputation uncertainty, omitted by single imputation methods.

Often multiple imputation is not much more difficult than doing a single imputation—the additional computing from repeating an analysis $M$ times is not a major burden, and methods for combining inferences are straightforward. Most of the work is in generating good predictive distributions for the missing values, not in drawing repeatedly once a good distribution is created. The underlying theory of multiple imputation is Bayesian, and the method is closely related to Bayes' simulation methods (83). However, the method also has good frequentist properties if the imputation model is carefully specified (68, 71).

Another powerful and principled approach to handling the missing data, especially in large samples, is maximum likelihood based on a statistical model for the data and missing-data mechanism (33, 76). This method is related to multiple imputation, uses all of the information in the complete and incomplete cases, and yields inferences that account for the loss of information arising from the fact that values are missing. The methods are principled in that they are based on explicit assumptions about the data and missing-data mechanism, which can be assessed and modified as necessary.

Until recently, most methods for handling missing data have assumed the missing data are missing at random (MAR), in the sense formalized by Rubin (58).

Loosely speaking, this assumption holds if missingness does not depend on the missing values after conditioning on the values observed in the data set. This assumption is realistic in some settings but not in others. For example, in longitudinal smoking cessation trials (7), dropout is non-MAR if it is associated with treatment outcome, as is the case if subjects fail to show for a clinic visit because they have failed and do not want to admit it or, on the other hand, if subjects fail to show because they have quit smoking and are not motivated to participate further in the study. Other forms of dropout, such as that arising from relocation, may be plausibly unrelated to the outcome and hence potentially MAR, particularly if characteristics associated with relocation and the outcome are included in the analysis.

When data are not MAR, a model for the missing-data mechanism needs to be incorporated into the analysis to yield appropriate inferences. Little (31) reviews models for longitudinal data with dropouts that include a model for the missing-data mechanism. A useful feature of these models is that the data analyst is forced to consider the nature of the missing-data process and the reasons for nonresponse. For example, in a longitudinal study of smoking interventions, different reasons for dropout, such as treatment failure or relocation, can be distinguished in the analysis and modeled in different and appropriate ways. Sensitivity to nonignorable nonresponse can be reduced by assuming that certain kinds of nonresponse are ignorable, thus confining the nonignorable component of the model to particular forms of dropout, and by including in the analysis variables measured before treatment dropout that are predictive of treatment outcome.

A problem with such models is that the data often do not provide reliable information for estimating the parameters without tenuous and untestable assumptions (30). Thus the analyst is often led to sensitivity analyses, in which inferences are displayed under a variety of alternative assumptions about the mechanism (35, 61). Little & Yau (36) describe a sensitivity analysis for an ITT analysis involving longitudinal data, in which missing outcomes after dropout are multiply imputed by an as-treated model. Frangakis & Rubin (16) discuss estimation of the CACE when there is both noncompliance and missing data, and they show that, even for valid inference about the ITT effect, it helps to model the noncompliance.

## CAUSAL INFERENCE IN OBSERVATIONAL STUDIES

### Observational Studies Conceptualized in Terms of Unconfounded Designs

In observational studies, in which the assignment of treatments is not under the control of the investigator, many epidemiological studies assume that treatment assignment is ignorable, perhaps after controlling for recorded confounders. That is, there is an assumption, which may or may not be made explicit, that the treatment assignment corresponds to that of an unconfounded randomized experiment. This is a critical yet unverifiable assumption, generally dependent on the extent and nature of the recorded covariates. The latter play a far more critical role than in randomized experiments, because adjustment for them is essential to reduce or avoid bias.

Good observational studies, like good experiments, are designed, not "found." William Cochran's work on this topic before the mid-1970s is particularly important (64). When designing an experiment, we do not have any outcome data, but we plan the collection, organization, and analysis of the data to improve our chances of obtaining valid, reliable, and precise causal answers. The same principle should be applied in an observational study. In an observational study comparing two treatments, the key step is to assemble data such that, within blocks, the covariate distribution for those in the control group is approximately the same as the treatment group; when the blocks have just one control and one treated unit, this corresponds to a matched-pairs design. The use of estimated propensity scores can be particularly helpful here, as discussed next.

## Methods for Adjusting for Measured and Unmeasured Confounders

When the assignment mechanism can be assumed to be unconfounded after conditioning on observed confounders, as in Equation 3, valid causal inferences can be obtained by adjusting estimated treatment effects for the covariates $X$. Methods for trying to achieve this include the following:

1. Regression of the outcome on dummy variables for treatments and on the covariates, including interactions if these are regarded as important. This method is very common in epidemiological studies. It can be very unreliable when the treatment and control groups differ markedly in their covariate distributions, because the extrapolation implied by the regression model relies strongly on assumptions such as linearity.
2. Methods that adjust for an estimate of the propensity score $p(W = 1 \mid X)$, typically obtained by discriminant analysis or logistic regression of treatment indicators on covariates. These methods include adding the estimated propensity score as a covariate in a regression model, which is also subject to problems of extrapolation, or better matching (56, 73, 78) or subclassification (55, 69) of treated units and controls based on the estimated propensity score, or weighting cases by estimates of the propensity score. For example, treated-control pairs might be formed with similar values of the propensity score (and perhaps the covariates) (56). Alternatively, subjects might be cross-classified by treatment group and by quintiles of the estimated propensity score; separate estimates of treatment effects are made within each propensity score subclass and then combined across the five subclasses, for example, using direct standardization (55, 69). An important advantage of these methods over regression

adjustment is that the investigator may discover that there is essentially no overlap in the distributions of the covariates in the treated and control groups. In that case, there is no hope of drawing valid causal inferences from these data without making strong external assumptions involving extrapolation (60). In such situations, it is important to allow that the data set cannot support a strong causal inference.

3. Combinations of propensity score methods and regression modeling (10, 44, 60, 64, 75). Examples are to fit a regression model within each propensity score subclass and then combine these estimates (44) or to fit a regression model to matched-pair differences. Simulations (59, 63, 75) and actual applications (10, 44, 69) indicate that combined strategies are generally superior to either strategy alone, for the goal of achieving essentially the same answer from an observational study as the parallel randomized experiment.
4. Methods based on the assumption of sequential ignorability (45–51) can also been applied to longitudinal observational studies when appropriate.

### Sensitivity Analyses and Bounds

Typically in epidemiological studies, we do not know that we have available a set of covariates that is adequate to support the claim of an unconfounded assignment mechanism. Thus, even if we have successfully adjusted for all covariates at hand, we cannot be sure that there is not some hidden unmeasured covariate that, if included, would alter the results.

A sensitivity analysis posits the existence of such a covariate and how it relates to both treatment assignment and outcome, and it examines how the results change. A classic example is the analysis by Cornfield et al (8) of the relationship between lung cancer and smoking. More recent work includes methods based on maximum likelihood (53), methods based on the randomization distribution (52), and bounding methods (26, 38).

## CONCLUSION

We have attempted to convey the power of the potential-outcomes formulation of causal effects in a variety of settings. We by no means provide a comprehensive review of the vast literature on the analysis of causal effects, which cannot be covered in a single article. For example, we have not reviewed the extensive literature on structural equation modeling (11) and modeling through graphs (19, 41, 88), although these subjects are closely related. The potential-outcomes formulation of causal effects is implicit in much of the work from other perspectives, and we believe that more examination of these other methods from the potential outcome perspective might be useful for clarifying assumptions and spawning new methodological advances (25).

Current research in causal inference is exceedingly lively, involving extensive interactions between behavioral scientists, computer scientists, economists, epidemiologists, philosophers, statisticians, and others. Armed with simple but powerful ideas such as the potential-outcomes definition of causal effects, we look forward to continued methodological developments in this crucial and fascinating area of empirical research, of critical importance to public health.

# APPENDIX

## Summary of Notation and Glossary of Selected Terms

$A(1)$: actual treatment when assigned treatment 1 (with the possibility of noncompliance).

$A(2)$: actual treatment when assigned treatment 2 (with the possibility of noncompliance).

$C$: compliance indicator; equals 1 if subject would comply with all potential treatment assignments, and zero otherwise.

$\boldsymbol{P}(A \mid B)$: the conditional probability of $A$ given $B$.

$S$: set of binary variables indicating whether members of the population are selected or not.

$W$: treatment assignment.

$X$: covariates characterizing subjects, unaffected by treatment assignment.

$Y(1)$: outcome when assigned treatment 1.

$Y(2)$: outcome when assigned treatment 2.

$Y_{exc}$: excluded data, outcomes of units in the population not included in the study sample.

$Y_{mis}$: missing data, including outcomes of treatments not assigned.

$Y_{obs}$: observed data, recorded outcomes of treatments assigned.

$\bar{y}_1, \bar{y}_2$: observed sample means of the outcomes of individuals assigned to treatments 1 and 2, respectively.

$\bar{Y}_1, \bar{Y}_2$: population means of the outcomes of individuals when all are assigned to treatments 1 or 2, respectively.

Additive causal effect: a causal effect that is assumed constant for all units in the population.

As-treated analysis: an analysis of a clinical trial where subjects are classified by their adopted treatment, irrespective of the treatment assigned.

Biased-coin design: a method of randomized treatment allocation where the probability of assigning a treatment is allowed to vary depending on recorded information from earlier subjects in the study.

Causal estimand: a feature of the population that measures the causal effect of a treatment.

Complier-Average Causal Effect (CACE): average causal effect of a treatment in the subpopulation of compliers—see Eq. (5).

Exclusion restriction: a modeling assumption that places a restriction on potential outcomes.

Ignorable treatment assignment: a treatment assignment that allows causal effects to be estimated without explicitly modeling the assignment mechanism—see Eq. (2).

Instrumental Variable (IV) estimator: an estimator of the CACE, defined by Eq. (8).

Intention-to-Treat (ITT) analysis: an analysis where subjects are classified by their assigned treatment, regardless of the treatment actually adopted.

Missing At Random (MAR): with missing values, an assumption that missingness depends only on the values of variables recorded in the data set, and given these, not on missing values.

Per-protocol analysis: an analysis restricted to subjects who adopted their assigned treatments.

Propensity score: the probability of being assigned a treatment as a function of observed covariates.

Randomized block design: a design where subjects are organized into blocks with similar values of covariates, and then randomized to treatments within these blocks.

Randomized Clinical Trial (RCT): a clinical trial where treatments are assigned to subjects using a randomization device, such as random numbers.

Sequential ignorability: a form of ignorable treatment assignment in the context of studies involving a sequence of treatments administered over time.

Stable Unit-Treatment Value Assumption (SUTVA): the assumption that an outcome for a particular individual is not affected by the treatment assignments or outcomes of other individuals in the study, and there are no hidden versions of treatment.

Unconfounded treatment assignment: a particular form of ignorable treatment assignment where assignment depends only on the values of observed covariates X—see Eq. (3).

Weighting class adjustment: a method of statistical analysis where subjects are given a weight based on the fraction of respondents within a weighting class defined by covariate information.

## LITERATURE CITED

1. Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables (with discussion and rejoinder). *J. Am. Stat. Assoc.* 91:444–72
2. Baker SG. 1998. Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *J. Am. Stat. Assoc.* 93:929–34
3. Baker SG, Lindeman KS. 1994. The paired availability design: a proposal for evaluating epidural analgesia during labor. *Stat. Med.* 13:2269–78
4. Barnard J, Frangakis C, Hill J, Rubin DB. 2000. School choice in NY City: a Bayesian analysis of an imperfect randomized experiment. In *Case Studies in Bayesian Statistics*, Vol. 5, ed. R Kass, B Carlin, A Carriquiry, A Gelman, I Verdinelli, et al. New York: Springer-Verlag. In press
5. Bloom HS. 1984. Accounting for no-shows in experimental evaluation designs. *Eval. Rev.* 8:225–46
6. Cochran WG, Cox GM. 1992. *Experimental Designs.* New York: Wiley & Sons. 2nd ed. 640 pp.
7. The COMMIT Research Group. 1995. Community intervention trial for smoking cessation (COMMIT). I. Cohort results from a four-year community intervention. II. Changes in adult cigarette smoking prevalence. *Am. J. Public Health* 85:183–200
8. Cornfield J, Haenszel W, Hammond EC, Lilienfeld A, Shimkin M, Wynder EL. 1959. Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* 22:173–203
9. Cox DR. 1992. *Planning of Experiments.* New York: Wiley & Sons, reprint ed. 320 pp.
10. Dehejia RH, Wahba S. 1999. Causal effects in non-experimental studies: re-evaluating the evaluation of training programs. *J. Am. Stat. Assoc.* 94:1053–62
11. Duncan OD. 1974. *Introduction to Structural Equation Models.* New York: Academic. 180 pp.
12. Efron B, Feldman D. 1991. Compliance as an explanatory variable in clinical trials. *J. Am. Stat. Assoc.* 86:9–17
13. Ettner SL. 1996. The timing of preventive services for women and children: the effect of having a usual source of care. *Am. J. Public Health* 86:1748–54
14. Fisher RA. 1918. The causes of human variability. *Eugenics Rev.* 10:213–20
15. Fisher RA. 1973. *Statistical Methods for Research Workers.* New York: Hafner. 14th ed. 382 pp.
16. Frangakis C, Rubin DB. 1999. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* 86:366–79
17. Frangakis C, Rubin DB, Zhou X-H. 1999. The clustered encouragement design. *Proc. Biomet. Sec. Am. Stat. Assoc. 1999,* pp. 71–79. Alexandria, VA: American Statistical Association
18. Goetghebeur E, Molenberghs G. 1996. Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *J. Am. Stat. Assoc.* 91:928–34
19. Greenland S, Pearl J, Robins JM. 1999. Causal diagrams for epidemiologic research. *Epidemiology* 10:37–48
20. Greenland S, Robins J. 1986. Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* 15:413–19
21. Haavelmo T. 1944. The probability approach in econometrics. *Econometrica* 15:413–19

22. Heckman JJ. 1996. Comment on "Identification of Causal Effects Using Instrumental Variables." *J. Am. Stat. Assoc.* 91:459–62
23. Hirano K, Imbens G, Rubin DB, Zhou X-H. 2000. Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics* 1:1–20
24. Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81:945–70
25. Holland PW. 1988. Causal inference, path analysis, and recursive structural equation models. *Soc. Methodol.* 18:449–84
26. Horowitz JL, Manski CF. 2000. Nonparametric analysis of randomized experiments with missing covariates and outcome data (with discussion and rejoinder). *J. Am. Stat. Assoc.* In press
27. Horvitz DG, Thompson DJ. 1952. A generalization of sampling without replacement from a finite population. *J. Am. Stat. Assoc.* 47:663–85
28. Imbens GW, Rubin DB. 1997. Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Stat.* 25:305–27
29. Imbens GW, Rubin DB. 1997. Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* 64:555–74
30. Little RJ. 1985. A note about models for selectivity bias. *Econometrica* 53:1469–74
31. Little RJ. 1995. Modeling the drop-out mechanism in longitudinal studies. *J. Am. Stat. Assoc.* 90:1112–21
32. Little RJ. 1997. Missing data. In *Encyclopedia of Biostatistics*, ed. P Armitage, T Colton, pp. 2622–35. London: Wiley
33. Little RJ, Rubin DB. 1987. *Statistical Analysis with Missing Data*. New York: Wiley & Sons
34. Little RJ, Schenker N. 1994. Missing data. In *Handbook for Statistical Modeling in the Social and Behavioral Sciences*, ed. G Arminger, CC Clogg, ME Sobel, pp. 39–75. New York: Plenum
35. Little RJ, Wang Y-X. 1996. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* 52:98–111
36. Little RJ, Yau L. 1996. Intent-to-treat analysis in longitudinal studies with drop-outs. *Biometrics* 52:1324–33
37. Little RJ, Yau L. 1998. Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychol. Methods.* 3:147–59
38. Manski CF, Nagin DS. 1998. Bounding disagreements about treatment effects: a study of sentencing and recidivism. *Soc. Methodol.* 28:99–137
39. Neyman J. 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. R. Stat. Soc. Ser. A* 97:558–606
40. Neyman J. 1990. On the application of probability theory to agricultural experiments: essay on principles, section 9. *Translated in Stat. Sci.* 5:465–72
41. Pearl J. 1995. Causal diagrams for empirical research (including discussion and rejoinder). *Biometrika* 82:669–710
42. Pratt JW, Schlaifer R. 1984. On the nature and discovery of structure. *J. Am. Stat. Assoc.* 79:9–33
43. Pratt JW, Schlaifer R. 1988. On the interpretation and observation of laws. *J. Economet.* 39:23–52
44. Reinisch J, Sanders S, Mortensen E, Rubin DB. 1995. In-utero exposure to phenobarbital and intelligence deficits in adult men. *J. Am. Med. Assoc.* 274:1518–25
45. Robins JM. 1987. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J. Chron. Dis.* 40(Suppl.):139S–161S
46. Robins JM. 1989. The control of confounding by intermediate variables. *Stat. Med.* 8:679–701
47. Robins JM. 1997. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to*

*Causality: Lecture Notes in Statistics*, ed. M Berkane, pp. 69–117. New York: Springer-Verlag

48. Robins JM. 1999. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, ed. ME Halloran, D Berry, pp. 95–134. New York: Springer-Verlag
49. Robins JM, Blevins D, Ritter G, Wulfsohn M. 1992. G-estimation of the effect of prophylaxis therapy for *Pneumocystis carinii* pneumonia on the survival of AIDS patients. *Epidemiology* 3:319–36
50. Robins JM, Greenland S, Hu F-C. 1999. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome (with discussion and rejoinder). *J. Am. Stat. Assoc.* 94:687–712
51. Robins JM, Hernan M, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology.* In press
52. Rosenbaum PR. 1995. *Observational Studies.* New York: Springer-Verlag
53. Rosenbaum PR, Rubin DB. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B* 45: 212–18
54. Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
55. Rosenbaum PR, Rubin DB. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 79:516–24
56. Rosenbaum PR, Rubin DB. 1985. Constructing a control group using multivariate matched sampling incorporating the propensity score. *Am. Stat.* 39:33–38
57. Rubin DB. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66:688–701
58. Rubin DB. 1976. Inference and missing data. *Biometrika* 63:581–92
59. Rubin DB. 1976. Multivariate matching methods that are equal percent bias reducing. I. Some examples; II, maximums on bias reduction for fixed sample sizes. *Biometrics* 32:109–32. Printer's erratum. 1976. *Biometrics* 32:955
60. Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *J. Educ. Stat.* 2:1–26. Printer's erratum. 1978. *J. Educ. Stat.* 3:384
61. Rubin DB. 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Stat. Assoc.* 72:538–43
62. Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 7:34–58
63. Rubin DB. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Am. Stat. Assoc.* 74:318–28
64. Rubin DB. 1984. William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In *W. G. Cochran's Impact on Statistics*, ed. PSRS Rao, J Sedransk, pp. 37–69. New York: Wiley
65. Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley & Sons
66. Rubin DB. 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Stat. Sci.* 5:472–80
67. Rubin DB. 1990. Formal modes of statistical inference for causal effects. *J. Stat. Plan. Inf.* 25:279–92
68. Rubin DB. 1996. Multiple imputation after 18+ years. *J. Am. Stat. Assoc.* 91:473–89
69. Rubin DB. 1997. Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* 127:757–63
70. Rubin DB. 1999. Teaching causal inference in experiments and observational

studies. *Proc. Stat. Educ. Sec. Am. Stat. Assoc. 1999.* Alexandria, VA: American Statistical Association

71. Rubin DB, Schenker N. 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Stat. Assoc.* 81:366–74
72. Rubin DB, Schenker N. 1991. Multiple imputation in health care databases: an overview and some applications. *Stat. Med.* 10:585–98
73. Rubin DB, Thomas N. 1992. Affinely invariant matching methods with ellipsoidal distributions. *Ann. Stat.* 20:1079–93
74. Rubin DB, Thomas N. 1992. Characterizing the effect of matching using linear propensity score methods with normal covariates. *Biometrika* 79:797–809
75. Rubin DB, Thomas N. 2000. Combining propensity score matching with additional adjustments for prognostic covariates. *J. Am. Stat. Assoc.* In press
76. Schafer JL. 1996. *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall. 430 pp.
77. Sheiner LB, Rubin DB. 1994. Intention-to-treat analysis and the goals of clinical trials. *Clin. Pharm. Ther.* 1:6–15
78. Smith HL. 1997. Matching with multiple controls to estimate treatment effects in observational studies. *Soc. Methodol.* 27:325–53
79. Sobel ME. 1990. Effect analysis and causation in linear structural equation models. *Psychometrika* 55:495–515
80. Sobel ME. 1995. Causal inference in the social and behavioral sciences. In *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, ed. A Arminger, CC Clogg, ME Sobel, pp. 1–38. New York: Plenum
81. Sommer A, Zeger S. 1991. On estimating efficacy from clinical trials. *Stat. Med.* 10:45–52
82. Tanner MA. 1996. *Tools for Statistical Inference.* New York: Springer-Verlag. 3rd ed. 207 pp.
83. Tanner MA, Wong WH. 1987. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82:528–50
84. Tinbergen J. 1995. Determination and interpretation of supply curves: an example. In *The Foundations of Econometric Analysis*, ed. DF Henry, MS Morgan, pp. 232–45. Cambridge: Cambridge University Press
85. Ware JH. 1989. Investigating therapies of potentially great benefit: ECMO (with discussion and rejoinder). *Stat. Sci.* 4:298–340
86. Wasserman L. 1999. Comment. *J. Am. Stat. Assoc.* 94:704–6
87. Wilkinson L. 1999. Statistical methods in psychology journals: guidelines and expectations. *Am. Psychol.* 54:594–604
88. Wright S. 1921. Correlation and causation. *J. Agric. Res.* 20:557–85

Annual Review of Public Health
Volume 21, 2000

# CONTENTS