Causal Inference Using Potential Outcomes: Design, Modeling, Decisions

Donald B. RUBIN

Causal effects are defined as comparisons of potential outcomes under different treatments on a common set of units. Observed values of the potential outcomes are revealed by the assignment mechanism—a probabilistic model for the treatment each unit receives as a function of covariates and potential outcomes. Fisher made tremendous contributions to causal inference through his work on the design of randomized experiments, but the potential outcomes perspective applies to other complex experiments and nonrandomized studies as well. As noted by Kempthorne in his 1976 discussion of Savage's Fisher lecture, Fisher never bridged his work on experimental design and his work on parametric modeling, a bridge that appears nearly automatic with an appropriate view of the potential outcomes framework, where the potential outcomes and covariates are given a Bayesian distribution to complete the model specification. Also, this framework crisply separates scientific inference for causal effects and decisions based on such inference, a distinction evident in Fisher's discussion of tests of significance versus tests in an accept/reject framework. But Fisher never used the potential outcomes framework, originally proposed by Neyman in the context of randomized experiments, and as a result he provided generally flawed advice concerning the use of the analysis of covariance to adjust for posttreatment concomitants in randomized trials.

KEY WORDS: Analysis of covariance; Assignment-based causal inference; Assignment mechanism; Bayesian inference; Direct causal effects; Fieller–Creasy; Fisher; Neyman; Observational studies; Principal stratification; Randomized experiments; Rubin causal model.

1. PROLOGUE

I greatly appreciate the invitation of the COPSS selection committee to contribute this year's R. A. Fisher Memorial Lecture. It certainly is humbling to consider the massive contributions of this giant of twentieth century statistics, as well as the published versions of the previous Fisher lectures. I will not attempt to compete with the incredibly encompassing lecture by Jimmie Savage (1976), with an assist from John Pratt, who helped complete it posthumously, but rather focus on one part of Fisher's work that has influenced me greatly, the design of experiments for causal inference, and attempt to relate some aspects of his contributions to current developments concerning inference for causal effects in more general settings. This presentation, however, will be more idiosyncratic than Cox's (1989) Fisher lecture on a somewhat similar topic, in that I will make no systematic attempt to refer to the many outstanding contributions made by others to this area, but rather will concentrate on how Fisher's work connects to the perspective that I advocate.

I never met Fisher in person; he died in 1962, a time when I was still doing physics as an undergraduate at Princeton University. Most of my knowledge of him, besides that obtained through reading his contributions, was gained from my Ph.D. advisor at Harvard University, Bill Cochran. Bill was a wonderful man with a charming and warm sense of humor.

Bill noted that Fisher, as everyone familiar with him knew, was a man of seemingly unbounded brilliance and arrogance. Bill had a variety of stories that he used to illustrate both of these characteristics, often with great humor with Bill as the butt of the story. One story, which illustrates the arrogance more than the brilliance, is relevant to the topic of this presentation, a connection made in the final section. It concerned the Fieller–Creasy controversy as recorded in the Royal Statistical Society (RSS) Symposium on Interval Estimation in 1954. Fieller (1954) and Creasy (1954) proposed two distinct "fiducial" solutions to the problem, in essence, of obtaining an interval estimate for the ratio of two means of independent normal distributions with known variances. Mr. Fieller, an established researcher, had proposed a solution years earlier that had Fisher's endorsement as *the* fiducial solution. Moreover, Fieller (1944) showed that it satisfied Neyman's (1934) criterion for a confidence interval.

Miss Creasy, in contrast, was a young researcher who had proposed a fiducial interval based on the same framework that Fisher had used to obtain the fiducial distribution for the difference between the means of two independent normal distributions with unknown variances, the Behrens–Fisher problem. Fisher was fairly brutal to the young Miss Creasy in his published discussion and, apparently, according to Bill, was even more disparaging of her efforts at the meeting.

At the time of the meeting, however, Cochran could not understand why the Creasy derivation was faulty, based as it was on Fisher's endorsed fiducial solution to the Behrens–Fisher problem. Cochran found Fisher in his office a few days after the RSS meeting, and Fisher immediately went to the blackboard, muttering words to the effect that only an idiot could not understand something so simple. Fisher began to write the assumptions with accompanying condescending comments, and Cochran could see after a few lines that Fisher was heading toward the Creasy solution! Fisher abruptly stopped writing, paused, and then quickly rubbed out all of his "derivation" and concluded his "proof" with something like, "From here it's obvious, even to you!" He proceeded to dismiss Cochran, having wasted enough time on this junior Scottish fool.

Cochran, who had daughters, told me that he felt that Fisher was undoubtedly especially dismissive of Creasy because she was Miss Creasy, and such people had little place in such scientific debates. Bill clearly thought otherwise.

Savage's (1976, p. 446) conclusion on the merits of Fisher's argumentation on this topic is consistent with Cochran's:

Donald B. Rubin is John L. Loeb Professor of Statistics, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *rubin@stat.harvard.edu*). This article is the written version of the 2004 Fisher Lecture, presented August 11, 2004 at the Joint Statistical Meetings in Toronto. The author thanks Constantine Frangakis, Andrew Gelman, and Roderick Little for extremely penetrating and helpful comments on earlier drafts of this article.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Review Article
DOI 10.1198/016214504000001880

On one occasion, Fisher (1954) struck out blindly against a young lady who had been anything but offensive or incompetent. His conclusion was that had the lady known what she was about she would have solved a certain problem in a certain fashion; he was right about that but failed to notice that she had solved it in just that fashion.

We return to this story after our brief journey through causal inference.

2. THE CAUSAL ESTIMAND = "THE SCIENCE"

When facing any problem of statistical inference, it is most important to begin by understanding the quantities that we are trying to estimate—the estimands. Doing so is particularly critical when dealing with causal inference, where mistakes can easily be made by describing the technique (e.g., computer program) used to do the estimation without any description of the object of the estimation.

In standard problems of causal inference, the causal estimand is the array of values depicted in Figure 1. Here there are N units, which are physical objects at particular points in time (e.g., plots of land, individual people, one person at repeated points in time). Each unit can be exposed or not to a treatment, here called "active treatment" if exposed and "control treatment" if not exposed; for example, the taking of aspirin or not. We generally denote these by "treatment" and "control" without ambiguity. The column labeled "Covariates," X, represents variables that take their values before the treatment assignment or, more generally, simply cannot be affected by the treatment, such as preaspirin headache pain or sex of the unit.

The columns labeled "Potential Outcomes" present the values of the outcome variable Y for each unit at a particular point in time after the action (e.g., headache pain 2 hours after taking aspirin or not), $Y(1)$ under the active treatment, and $Y(0)$ under the control treatment. Any information that is to be analyzed is included in X, $Y(1)$, and $Y(0)$, and thus the indexing of the units is completely random, that is, simply a random permutation of $1, \ldots, N$.

The column labeled "Unit-Level Causal Effects" provides the collection of individual unit-level causal effects, which for the ith unit are, the comparison of $Y_i(1)$ and $Y_i(0)$, typically the difference, $Y_i(1) - Y_i(0)$, but not always; the ratio or any such comparison could be used to define unit-level causal effects. Of course, we can never observe both $Y_i(1)$ and $Y_i(0)$ for any unit i, because we cannot unwind time and go back and expose the ith unit to the other treatment. This is called the "fundamental problem of causal inference" (Holland 1986). Each potential outcome is observable, but we can never observe all of them.

As indicated by the last column, "summary" causal effects can also be defined at the level of collections of units, such as the mean unit-level causal effect for all units, the median unit-level causal effects for all males, the difference for females between the median $Y_i(1)$ and the median $Y_i(0)$, and so on. Many summary causal effects are typical unit-level causal effects, such as the mean or median unit-level causal effect for a collection of units. Other summary causal effects are marginal in that they compare some aspect of the marginal distributions of $Y(1)$ and $Y(0)$ for a collection of units, such as for females, median $Y_i(1)$ − median $Y_i(0)$. Mean causal effects have a simple interpretation because they are both typical unit-level and marginal causal effects.

		Potential outcomes			
Units	Covariates X	Treatment $Y(1)$	Control $Y(0)$	Unit-level Causal effects	Summary Causal effects
1	X_1	$Y_1(1)$	$Y_1(0)$	$Y_1(1)$ v. $Y_1(0)$	Comparison of $Y_i(1)$ v. $Y_i(0)$ for a common set of units
⋮	⋮	⋮	⋮	⋮	
i	X_i	$Y_i(1)$	$Y_i(0)$	$Y_i(1)$ v. $Y_i(0)$	
⋮	⋮	⋮	⋮	⋮	
N	X_N	$Y_N(1)$	$Y_N(0)$	$Y_N(1)$ v. $Y_N(0)$	

Figure 1. "Science"—The Causal Estimand.

The critical requirement, however, is that to be a causal effect, the comparison must be a comparison of $Y_i(1)$ and $Y_i(0)$ for a common set of units, such as females. More formally, a causal effect must be a comparison of the ordered sets $\{Y_i(1), i \in S\}$ and $\{Y_i(0), i \in S\}$, not $\{Y_i(1), i \in S_1\}$ and $\{Y_i(0), i \in S_0\}$, $S_1 \neq S_0$. One cannot compare medical costs for male smokers with medical costs for female nonsmokers and claim that this comparison is a causal effect of smoking without making heroic and unwarranted assumptions. This requirement is more subtle than it might first appear. In fact, as I discuss later, Fisher himself did make such an error throughout his life. I see this error as possibly attributable to Fisher's unwillingness to utilize the marvelous contribution of formal notation for potential outcomes, originally due to Jerzy Neyman. As Savage (1976, p. 446) stated:

I am surely not alone in having suspected that some of Fisher's major views were adopted simply to avoid agreeing with his opponents (Neyman 1961, pp. 148–149).

In any case, the depiction in Figure 1 requires assumptions for it to be adequate—in particular, SUTVA (stable unit treatment value assumption) (Rubin 1980), which comprises two subassumptions. First, it assumes that there is no interference between units (Cox 1958); that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that there are no hidden versions of treatments; no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0.

An assumption that is also implicit in the representation in Figure 1 is that the science—the covariates and the potential outcomes—is not affected by how or whether we try to learn about it, whether by completely randomized experiments, randomized blocks designs, observational studies, or another method. That is, whether I took an aspirin because I tossed a fair coin to decide or because I happened to have an aspirin nearby does not affect the pain that would be observed under either treatment. Or, more topically, whether I try to learn about the effects of hormone replacement therapy on postmenopausal women through a randomized experiment (e.g., the Woman's Health Initiative Study Group 1998) or through an observational study (e.g., the Nurses Heath Study, Grodstein, Clarkson, and Manson 2003), the science does not change—the causal effects of taking versus not taking hormones for the units are not affected. Without these assumptions, causal inference using potential outcomes is far more complicated, although still possible in principle by, for example, allowing only some units to interfere with each other, as with models of additive carryover effects of drugs, or by allowing the act of observation to

be an additional intervention, in accordance with quantum mechanics. As a result, SUTVA is commonly made, although often implicitly and sometimes without much thought.

Nothing is wrong with making assumptions; causal inference is impossible without making assumptions, and they are the strands that link statistics to science. It is the scientific quality of those assumptions, not their existence, that is critical. There is always a trade-off between assumptions and data—both bring information. With better data, fewer assumptions are needed. But in the causal inference setting, assumptions are always needed, and it is imperative that they be explicated and justified. One reason for providing this detail is so that readers can understand the basis of conclusions. A related reason is that such understanding should lead to scrutiny of the assumptions, investigation of them, and, ideally, improvements. Sadly, this stating of assumptions is typically absent in many analyses purporting to be causal and replaced by a statement of what computer programs were run, which I regard as entirely inadequate scientifically.

Because at least half of the potential outcomes are missing and the underlying assumptions about them are so critical, this notation explicitly representing both potential outcomes is an exceptional contribution to causal inference. However, despite its apparent simplicity, it did not arise until 1923, and then only in the context of completely randomized experiments.

3. FISHER AND NEYMAN ON THE POTENTIAL OUTCOME NOTATION IN RANDOMIZED EXPERIMENTS AND BEYOND

Neyman (1923), in his Ph.D. thesis, appears to have been the first writer to use this potential outcome notation, and he did so only in the context of a completely randomized experiment, in particular, a hypothetical agricultural experiment, where the units were distinct plots of land, and the potential outcomes were called potential yields (of crops). In addition to introducing the potential outcome notation, Neyman also proved two important facts about the completely randomized experiment. First, the difference of observed sample means between the n_1 treated units and the n_0 control units, $\bar{y}_1 - \bar{y}_0$, was an unbiased estimator of the average causal effect over all of the units,

$$\bar{Y}_1 - \bar{Y}_0 = \sum_{i=1}^{N} \frac{Y_i(1) - Y_i(0)}{N}.$$

Here unbiased means averaging over all possible randomizations, with the potential outcomes treated as fixed values. Neyman also showed that the usual estimate of the variance of the difference between two sample means, $s_1^2/n_1 + s_0^2/n_0$, where s^2 refers to within-group sample variances, is generally a positively biased estimate of the true variance of $\bar{y}_1 - \bar{y}_0$ over all possible randomizations unless $Y_i(1) - Y_i(0)$ is constant for all i, in which case it is unbiased.

Notice that Neyman was writing about randomized experiments a couple of years before Fisher (1925) explicitly proposed them:

> Ex. 44. Accuracy attained by random arrangement. The direct way of overcoming this difficulty is to arrange the plots wholly at random.

Yet Fisher is generally credited with the "invention" or "discovery" of randomized experiments. This attribution of randomization to Fisher rather than to Neyman reflects the dominant position of the "English school" in the statistics of that time, and more important historically, I believe, because Neyman himself endorsed this attribution (Reid 1982, p. 45):

> On one occasion, when someone perceived him as anticipating the English statistician R. A. Fisher in the use of randomization, he objected strenuously: "... I treated *theoretically* an unrestrictedly randomized agricultural experiment and the randomization was considered a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization, an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fisher's achievements."

Neyman evidently recognized the enormous difference between doing mathematical calculations in statistics and understanding their implications for the actual conduct of statistical practice.

The year after the publication of the first edition of *Statistical Methods for Research Workers*, Fisher (1926) presented an extended discussion of randomization and introduced the "Fisher sharp null hypothesis," $Y_i(1) = Y_i(0)$ for all i. This null hypothesis is sharp in the sense that under it, all potential outcomes are known for the units exposed to either treatment. Therefore under this hypothesis, from the one actual randomization, we know the hypothetical observed value of any statistic (i.e., any function of the observed data, such as $\bar{y}_1 - \bar{y}_0$) under all possible randomizations of the units in this study. Thus we can calculate a significance level (or p value) stating how unusual the actual observed statistic is relative to all possible values of that statistic that might have been observed with these units. I view this as Fisher's formulation of "proof by stochastic contradiction."

But at the time Fisher "discovered" or "invented" physical randomization, he did not use notation like Neyman's and could not have been aware of his thesis, which was written in Polish. On the other hand, Fisher certainly seemed to have the idea of potential outcomes lurking in his consciousness before 1923; as he wrote (Fisher 1918, p. 214):

> If we say, "This boy has grown tall because he has been well fed," we are not merely tracing out the cause and effect in an individual instance; we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter.

Clearly, Fisher is contemplating an alternative hypothesis and the associated potential outcome under that hypothesis. But having the idea is very different from both having it and formulating explicit mathematical notation for it. As far as I can tell, Fisher never used the potential outcome notation despite its common use in the context of randomized experiments after about 1937 (e.g., Welch 1937; Pitman 1938; McCarthy 1939; Anscombe 1948; Kempthorne 1952; Cox 1958; Brillinger, Jones, and Tukey 1978).

Although there are allusions to the intuitive idea of using potential outcomes to define causal effects generally in the preceding Fisher (1918) quotation and in many other places, such as the economics literature (e.g., Haavelmo 1943; Tinbergen 1930; Hurwicz 1962), I can find no explicit use of that notation to describe causal effects in nonrandomized studies until a half century later (in Rubin 1974). The extension of Neyman's potential outcome notation to define causal effects in both nonrandomized and randomized studies is sometimes called the "Neyman–Rubin" model (e.g., Pearl 1996). In the economics literature,

the use of the potential outcomes notation to define causal effects has recently (e.g., Heckman 1996) been attributed to Roy (1951) or Quandt (1958), which is puzzling because neither of these articles addresses causal inference, and the former has no mathematical notation at all. For seeds of potential outcomes in economics, the earlier references cited at the start of this paragraph are much more relevant; see the rejoinder by Angrist, Imbens, and Rubin (1996) for more on this topic.

Some authors (e.g., Greenland, Pearl, and Robins 1999; Dawid 2000) call the potential outcomes "counterfactuals," borrowing the term from philosophy (e.g., Lewis 1973). I much prefer Neyman's implied term "potential outcomes," because these values are not counterfactual until after treatments are assigned, and calling all potential outcomes "counterfactuals" certainly confuses quantities that can never be observed (e.g., your height at age 3 if you were born yesterday in the Arctic) and so are truly a priori counterfactual, with unobserved potential outcomes that are not a priori counterfactual (see Frangakis and Rubin 2002; Rubin 2004; and the discussion and reply for more on this point).

In any case, whatever name we wish to use for potential outcomes or for their use to define causal effects, they have rapidly become standard in many branches of social, economic, and biomedical sciences over the last two decades (e.g., Frangakis and Baker 2001; Gelman and King 1990; Heckman 1989; Imbens and Angrist 1994; Pratt and Schlaifer 1988; Sobel 1996; Winship and Morgan 1999). This recent use of potential outcomes in the literature of nonrandomized studies is in stark contrast to the previous literature, which used a deficient "observed value" notation defined in Section 4 (e.g., Heckman 1979; Pratt and Schlaifer 1984, to use two examples with identical authors writing before and after the transition).

4. THE ASSIGNMENT MECHANISM

To see why randomized experiments are so special, we need to embed them in a larger class of designs without their special properties. This larger class that encompasses randomized experiments is now generally called "assignment mechanisms"—methods that assign treatments to units. An assignment mechanism can be thought of as a special type of missing-data mechanism that creates missing potential outcomes (Rubin 1976, 1978).

Let W_i indicate the assignment for unit i, where 1 implies the active treatment and 0 implies the control, and $*$ implies neither, as with a future unit, $W = (W_1, \ldots, W_i, \ldots, W_N)^T$. Then the assignment mechanism gives the probability of the vector W given fixed values of the science, X, $Y(1)$, and $Y(0)$, where this notation refers to the full array of values for all units. Thus the assignment mechanism can be written as

$$\Pr\big(W|X, Y(1), Y(0)\big). \tag{1}$$

In (1), the vector W is the only random variable; the science is regarded as fixed and waiting to be partially revealed by the assignment mechanism. Thus Neyman's results in Section 3 can be stated as

$$E\big(\bar{y}_1 - \bar{y}_0|X, Y(1), Y(0)\big) = \bar{Y}(1) - \bar{Y}(0)$$

and

$$V\big(\bar{y}_1 - \bar{y}_0|X, Y(1), Y(0)\big) \geq E\big(s_1^2/n_1 + s_0^2/n_0|X, Y(1), Y(0)\big),$$

where $E(\cdot|\cdot)$ and $V(\cdot|\cdot)$ refer to expectation and variance over the distribution of W given X, $Y(1)$, $Y(0)$, that is, over the randomization distribution. Also, Fisher's p value for $\bar{y}_1 - \bar{y}_0$ can be stated, in an imprecise but hopefully clear notation, as

p value

$$= \Pr\big(\bar{y}_1 - \bar{y}_0 \geq \bar{y}_{1,\text{obs}} - \bar{y}_{0,\text{obs}}|X, Y(1) \equiv Y(0), Y(0) \equiv Y(1)\big),$$

where, again, $\bar{y}_1 - \bar{y}_0$ is a random variable because W is a random variable, and $\bar{y}_{1,\text{obs}} - \bar{y}_{0,\text{obs}}$ is the actual observed value of $\bar{y}_1 - \bar{y}_0$; this can formally be viewed as a posterior predictive p value (Rubin 1984b; Meng 1994; Gelman, Meng, and Stern 1996) and so has a Bayesian justification, which I find intellectually pleasing.

An assignment mechanism must be posited for probabilistic causal inference, and in certain circumstances, such as randomized experiments, it is the only model that needs to be posited to make inferential progress. That is, no model on the science may be needed beyond SUTVA, which is what both Fisher and Neyman showed in their seminal 1926 and 1923 publications.

All randomized experiments are assignment mechanisms with two critically important properties. First, they are "ignorable" (Rubin 1976, 1978),

$$\Pr\big(W|X, Y(0), Y(1)\big) = \Pr(W|X, Y_{\text{obs}}), \tag{2}$$

where Y_{obs} is the collection of observed potential outcomes, $Y_{\text{obs}} = (Y_{\text{obs},1}, \ldots, Y_{\text{obs},i}, \ldots, Y_{\text{obs},N})$; $Y_{\text{obs},i}$ is the observed value of Y for unit i,

$$Y_{\text{obs},i} = W_i Y_i(1) + (1 - W_i) Y_i(0). \tag{3}$$

Analogously, we have that Y_{mis} is the collection of missing or unobserved potential outcomes, $Y_{\text{mis}} = (Y_{\text{mis},1}, \ldots, Y_{\text{mis},N})$,

$$Y_{\text{mis},i} = (1 - W_i) Y_i(1) + W_i Y_i(0). \tag{4}$$

Also, using a somewhat imprecise but unambiguous notation,

$$Y = (Y_{\text{obs}}, Y_{\text{mis}}).$$

Second, in a randomized experiment, the unit-level probabilities of treatment assignment, the "propensities" (Rosenbaum and Rubin 1983), are between 0 and 1,

$$0 < \Pr(W_i = 1|X, Y_{\text{obs}}) < 1. \tag{5}$$

Explicit mathematical notation for a general assignment mechanism did not appear in the literature until the 1970s, and it now appears to be widely used and accepted. The literature on randomized experiments gave expressions for ignorable assignment mechanisms. Sequential experiments (e.g., Chernoff 1972; play-the-winner rules, Ware 1989) were described with explicit dependence of assignment on Y_{obs}, whereas classical randomized experiments (e.g., Cochran and Cox 1950; Kempthorne 1952) were described with no dependence of assignment on Y_{obs} but only on blocking factors, incorporated into X, and can be called "unconfounded,"

$$\Pr\big(W|X, Y(0), Y(1)\big) = \Pr(W|X).$$

But the possible dependence on Y_{mis} does not seem to have been formalized until Rubin (1975, 1976, 1978), as noted by Pratt and Schlaifer (1988, pp. 23–24) (where, throughout, bracketed expressions in direct quotations are added by me an attempt to increase clarity):

A [causal] law can be observed in data if and only if the process that generated the [observed] data [the assignment mechanism] satisfied a condition first stated precisely by Rubin (1978) but implicit in R. A. Fisher's examples of the proper analysis of experiments. . . . We are in the position unusual for critics of insisting that Rubin's contribution is much more important than he or his followers have even suggested it is.

The literature on observational studies (briefly alluded to at the end of Sec. 3) used models relating Y_{obs} to both (W, X) and hypothetical parameters θ (e.g., via least squares regression equations), completely eschewing the use of potential outcomes in favor of the collapsed Y_{obs}. But Y_{obs} mixes up science—the potential outcomes, and what we do to try to learn about science—and the assignment mechanism, as is clear from (2) for Y_{obs}. Consequently, this notation is inherently deficient and can lead the brilliant astray, even Fisher, as we discuss later. Using only the collapsed notation, Y_{obs}, and the treatment indicator W in place of the potential outcomes, one cannot even directly state the critical benefit of randomization in achieving ignorability. Even those who wrote with great clarity in the context of randomized experiments used this deficient notation in the context of observational studies and made mistakes. For example, in my discussion of Cochran's contributions to observational studies (Rubin 1984a), I pointed out that Bill's posthumous book, *Planning and Analysis of Observational Studies* (Cochran 1983), faltered when discussing matching with nonparallel response surfaces, that is, situations where the regression of $Y(1)$ on X was linear with a different slope than the regression of $Y(0)$ on X. I pointed out other similar errors later (Rubin 1990).

Relying only on a model for the assignment mechanism, we can make tremendous progress on statistical inference for causal effects, even in observational studies, using, for example, propensity scores (see, e.g., Rosenbaum and Rubin 1983, 1984, 1985), and the explosion of recent literature on propensity scores in applied journals (Google 2004 returned more than 13,000 hits for "propensity score"; see also Rosenbaum 2002). But models on the science still have a critical role to play in causal inference, as I now discuss using Bayesian statistics, something Fisher eschewed just as he eschewed Neyman's potential outcomes.

5. MODELS ON THE SCIENCE—BAYESIAN INFERENCE

Thus far, we have not presented any models on the science—no concepts such as regression modeling, relative risks, odds ratios, or hazards models. I feel that such models arise most naturally in causal inference within a Bayesian framework. The assignment-based perspectives of Fisher and Neyman do not rely on models of science for their validity and in this sense are inherently more robust, because models on the assignment mechanism can be essentially correct, as in a randomized experiment. However, do not forget the assumption of SUTVA, which is on the science and an ingredient of Neyman's assignment-based approach (although automatically satisfied under Fisher's null hypothesis). Essentially all models on the science are wrong, because we are trying to understand what nature is doing. Nevertheless, even though all such models may be wrong, some are very useful—a comment attributable to George Box. The benefits of modeling the science in causal inference include the ability to deal with more complex situations and to summarize results more logically.

Within the model-based Bayesian framework for causal inference (Rubin 1975, 1978), we directly confront the fact that at least half of the potential outcomes are missing by creating a posterior predictive distribution for them. More precisely, from a model on the science,

$$\Pr\big(X, Y(1), Y(0)\big), \tag{6}$$

and the model for the assignment mechanism, (1), we can find the posterior predictive distribution of Y_{mis}—the missing potential outcomes, given the observed values of W, X, and Y_{obs}. This posterior predictive distribution can be written as

$$\Pr(Y_{\text{mis}}|X, Y_{\text{obs}}, W) \propto \Pr\big(X, Y(1), Y(0)\big)\Pr\big(W|X, Y(1), Y(0)\big), \tag{7}$$

where Y_{mis} is the only unobserved random variable; X, Y_{obs}, and W are all are assumed to be observed. From (7), we can calculate the posterior distribution of any causal estimand, because all causal estimands are functions of the observed X, Y_{obs}, and the missing Y_{mis}. Essentially, by multiply imputing Y_{mis} according to (7), we can simulate the posterior distribution of any causal estimand that we want: Draw a value of Y_{mis}, impute it, calculate the causal estimand, redraw Y_{mis}, and so on.

Two critical facts simplify this approach and create a bridge to standard likelihood theory. The first fact is that the modeling of the science is not as daunting as it first may appear. Recall that the indexing of the units is, by definition, a random permutation of $1, \ldots, N$, and thus any distribution on the science must be row-exchangeable, that is, constant under permutation of the row indices. Thus, by de Finetti's theorem, with essentially no loss of generality we can write

$$\Pr\big(X, Y(0), Y(1)\big) = \int \prod_{i=1}^{N} f\big(X_i, Y_i(0), Y_i(1)|\theta\big)p(\theta)\,d\theta, \tag{8}$$

where $f(\cdot|\theta)$ is an iid model for each unit's science given a hypothetical parameter θ with prior (or marginal) distribution $p(\theta)$. Because of the central role of (8) in Bayesian inference for causal effects and Savage's enormous influence on Bayesian statistics, it is interesting to note that neither de Finetti's comments following Savage's Fisher lecture (1976) nor the lecture itself reflects awareness of this role of de Finetti's theorem or the role of randomization in Bayesian causal inference, discussed next.

The second critical fact is that if the treatment assignment mechanism is ignorable (e.g., randomized), then when the expression for the assignment mechanism (2) is evaluated at the observed data, it is free of dependence on Y_{mis}. Thus the rightmost factor in (7) is a constant, and so the explicit conditioning on W in (7) can be ignored (hence the term "ignorable assignment mechanism"):

$$\Pr(Y_{\text{mis}}|X, Y_{\text{obs}}, W) \propto \Pr(Y_{\text{mis}}|X, Y_{\text{obs}}). \tag{9}$$

Introducing the parameter θ from (8), we then have

$$\Pr(Y_{\text{mis}}|X, Y_{\text{obs}}) = \int \Pr(Y_{\text{mis}}|X, Y_{\text{obs}}, \theta)\Pr(\theta|X, Y_{\text{obs}})\,d\theta,$$

where $\Pr(\theta|X, Y_{\text{obs}})$ is, by definition, the posterior distribution of θ, equal to the prior distribution $p(\theta)$ times the likelihood of θ,

$$L(\theta|X, Y_{\text{obs}}) \propto \prod_{i=1}^{N} f(Y_{i,\text{obs}}, X_i|\theta)$$
$$= \prod_{W_i=1} f(Y_i(1), X_i|\theta) \prod_{W_i=0} f(Y_i(0), X_i|\theta),$$

using an imprecise but hopefully clear notation.

Fisher made tremendous contributions to likelihood theory itself (starting, essentially, with Fisher 1922) and to specific models for $f(\cdot|\theta)$. But Fisher never related his vast work on likelihoods and models to his vast work on experimental design. As Kempthorne (1976) noted when discussing Savage's (1976) Fisher lecture:

> The work of Fisher abounds in curiosities. One which has struck me forcibly is the absence of any discussion of the relationship of Fisher's ideas on experimentation (DOE) to his general ideas on inference (SI). The latter book contains no discussion of ideas of randomization... which made DOE so interesting and compelling to investigators in noisy experimental sciences. Can the ideas on randomization and on parametric likelihood theory be fused into a coherent whole? I think not.

Yet the bridge between these two classes of Fisher's contributions is almost immediate if we are willing to use Neyman's potential outcomes and the Bayesian formulation sketched here, because the resulting combination is an embedding of both the assignment-based perspective and the modeling perspective in one coherent framework. This framework, which (1) extends Neyman's potential outcomes to define causal effects in all situations, (2) includes the formulation of an assignment mechanism with explicit possible dependence on all potential outcomes including the missing ones, and (3) embeds both assignment-based and Bayesian-likelihood inference in a common framework, is the framework that I advocate, and it is now sometimes called "Rubin's causal model" (Holland 1986; also see Angrist et al. 1996 for discussion). For a review of this basic perspective, see the article by Little and Rubin (2000), and for an expanded treatment, including aspects such as traditional econometric instrumental variables analyses, see the forthcoming text by Imbens and Rubin (2005).

6. DECISIONS: BASED ON CURRENT KNOWLEDGE OF SCIENCE AND ON COSTS OF DECISIONS

One consequence of the Bayesian approach to causal inference is that the posterior distribution of causal estimands, obtained from the posterior predictive distribution of the missing potential outcomes (7), is viewed as a summary of all that is currently known about the science from the current data and the prior science. Assuming that this summary of the current state of knowledge is accurate, this can be combined with various assessments of the costs and benefits of various decisions to choose which decisions to make. This perspective is something of an idealization; for example, decisions concerning which experiments to conduct, which treatments to include, what sample sizes to use, and so on, can greatly influence the scientific summary. But the perspective that (1) the science exists independently of how we try to learn about it and that (2) if the model used for analysis of the resulting data is approximately correct, then the resulting posterior distribution will give a fair summary of the current state of knowledge of that science seems, at least to me, consistent with common views of the scientific enterprise. For a statement of this attitude in the context of noncompliance in pharmaceutical randomized experiments, see the article by Sheiner and Rubin (1994) and the earlier related article with discussion by Efron and Feldman (1991).

Fisher might have agreed with this general claim if he had been more sympathetic to the Bayesian perspective (Fisher 1956, pp. 102–103):

> It is important that the scientific worker introduces no cost functions for faulty decisions.... To do so would imply that the purposes to which new knowledge was to be put were known and capable of evaluation. If, however, scientific findings are communicated for the enlightenment of other free minds, they may be put sooner or later to the service of a number of purposes, of which we can know nothing.

Such statements by Fisher were often in the context of debates over the use of "significance tests," which summarized evidence against a null hypothesis by a p value, versus the use of "accept/reject" tests, which Fisher viewed as appropriate for accepting or rejecting products, but not for scientific summarization. But a claim that the scientific evidence for a causal inference could be summarized, in any generality, by a single number between 0 and 1 seems almost as far-fetched to me as the idea that it could be summarized by a 0 (accept) or a 1 (reject).

Perhaps Fisher was thinking of the use of fiducial distributions as a way to summarize science, but I do not find this view articulated in these debates; also, few have found the fiducial perspective to be satisfactory in any generality (see, e.g., Savage 1976, sec. 4.6, and the ensuing discussions). The logic underlying the fiducial perspective seems to work best when the argument for it is presented very quickly and rubbed out before one can think too hard about it, as in Cochran's Fisher–Creasy story!

An argument can be made that Fisher viewed the likelihood function as such a summary, but I find the inability in principle to integrate the likelihood over nuisance parameters to obtain a marginal summary of an estimand to be major problem, especially in high-dimensional situations. To many, like myself, who are sympathetic with the Bayesian argument, a posterior distribution with clearly stated prior distributions is the most natural way to summarize evidence for a scientific question. Combining this summary with the costs of decisions then also becomes natural and supports an expanded view of Fisher's preceding statement.

I see here Fisher's general resistance to acknowledge or use the contributions of others as interfering with a clear delineation between science and decisions, which he appears to have supported in principle. But this limitation in perspective is certainly not an outright error on Fisher's part. I conclude with an example where Fisher's failure to use Neyman's potential outcomes for causal inference in the context of "complex" randomized experiments did lead to flawed advice.

7. COMPLEX EXPERIMENTS: "DIRECT" AND "INDIRECT" CAUSAL EFFECTS

Consider the problem of adjusting for a "concomitant" variable—an outcome variable that is not the outcome of primary interest, but may be "on the causal pathway" of the treatment affecting the primary outcome variable, Y. That is, the

concomitant variable, C, is not a covariate but rather is something like a covariate in that we wish to "adjust" for it. For example, we may wish to estimate the "direct" effect of treatment versus control on Y for the set of units for which the treatment does not effect C.

Fisher wrote about this problem in *Design of Experiments* (DOE) from the first edition in 1935 to the last (8th) edition in 1966. His views remained unchanged, even though he was a discussant of Neyman's RSS lecture on randomized experiments (Neyman 1935), which used the potential outcome notation, and in my view would have helped to reveal the general flaws in the following advice.

In DOE, Fisher (1935, 1954, 1956, 1966, chap. IX, sec. 55) wrote:

> In agricultural experiments involving the yield following different kinds of treatments, it may be apparent that the yields of the different plots have been much disturbed by variations in the number of plants which have established themselves. If we are satisfied that this variation in plant number is not itself an effect of the treatments being investigated [in which case plant number is a true covariate], or if we are willing to confine our investigation to the effects on yield, excluding such as flow directly or indirectly from effects brought about by variations in plant number, then it will appear desirable to introduce into our comparisons a correction which makes allowance, at least approximately, for the variations in yield directly due to variation in plant number itself.

He also wrote about this in *Statistical Methods for Research Workers* (1970, sec. 49.1, pp. 283–284), in a way that was consistent with this previous quotation:

> Thus, if we were concerned to study the effects of agricultural treatments upon the purity index of the sugar extracted from sugar-beet, a variate which might be much affected by concomitant variations in (a) sugar-percentage, and (b) root weight, an analysis of covariance applied to the three variates, purity, sugar percentage, and root weight, for the different plots of the experiment, would enable us to make a study of the effects of experimental treatments on purity alone; i.e., after allowance for any effect they may have on root weight or concentration, without our needing to have observed in fact any two plots agreeing exactly in both root weight and sugar percentage.

Fisher's recommendation is to conduct an analysis of covariance (ANCOVA) of $Y_{\text{obs},i}$ on W_i and $C_{\text{obs},i}$, where $C_{\text{obs},i} = W_i C_i(1) + (1 - W_i)C_i(0)$. This analysis is equivalent to a regression analysis of observed outcome on treatment indicator and observed predictor, which was criticized in Section 4 for its naivete in the context of observational studies. Essentially, an ANCOVA compares the average observed $Y_i(1)$ with the average observed $Y_i(0)$ for units with a common value of $C_{\text{obs},i}$, which generally does not estimate a causal effect of any kind, because generally the resultant estimand does not satisfy the definition for a causal estimand given in Section 2. The problem with this approach is illustrated in Figures 2 and 3.

Suppose that Figures 2 and 3 represent very large randomized experiments of N units, for concreteness, say N plots; the concomitant C is the number of plants established in each plot, the primary outcome Y is the yield in each plot, the treatment is a new fertilizer, and the control is the standard fertilizer. In each experiment, half of the units are randomly assigned to the active treatment and half of the units are assigned to the control treatment. That is, the assignment mechanism is ignorable with

$$\Pr\big(W_i = 1 | X_i, Y_i(1), Y_i(0)\big) = \Pr\big(W_i = 0 | X_i, Y_i(1), Y_i(0)\big) = 1/2 \quad \text{for } i = 1, \ldots, N. \tag{10}$$

The left set of columns in Figure 2 give the potential outcome in the first experiment. The first two rows represent those $N/2$ units with common values of the potential outcomes, and so randomly assigning them would result in half being assigned to control, represented by the first row, and half being assigned to treatment, represented by the second row, and analogously for the second pair of rows. The resultant observed data are represented in the right half of Figure 2. Each pair of rows corresponds to a "principal stratum" using the terminology of Frangakis and Rubin (2002), where each principal stratum is defined by common values of $C_i(1)$ and $C_i(0)$. The left columns reveal that for all units, there is a causal effect of treatment on the concomitant variable, C, of size 1, but there is no effect on the primary outcome variable Y for any unit. Because all units experience a treatment effect on C and no effect on Y, the answer to the question of the "direct" effect of treatment on Y, after adjusting for C, seems to be a matter of definition, most simply chosen to be 0.

An examination of the right set of columns in Figure 2, however, reveals that conditioning on the observed value of the concomitant, C_{obs}, which is equivalent to Fisher's life-long recommendation, leads to a contradictory conclusion: When $C_{\text{obs}} = 3$, those plots that received the new treatment do worse, a treatment effect of -2 (compare the second and third rows in the observed data). Also in this case, the regressions of $Y_{\text{obs},i}$ on $C_{\text{obs},i}$ in the $W_i = 0$ and $W_i = 1$ groups are linear and parallel, with constant treatment minus control difference equal to -2. So the conclusion of Fisher's recommended ANCOVA is that the treatment's effect on Y, after making allowance for any effect on C, is negative! This clearly seems incorrect, as is revealed by an examination of the potential outcomes in Figure 2.

The example in Figure 3 illustrates a flaw in Fisher's proposed solution even when there does appear to be a well-defined "direct" treatment effect on Y after controlling for C. The example is analogous to the one in Figure 2 except that first, there is a constant treatment effect on Y of size 1 for all units, and second, for one-third of the units, represented by the middle two rows, there is no treatment effect on the concomitant, C. For the other units, the treatment effect on the concomitant is 1; for the

Fraction of population	Potential outcomes				Observed data		
	$C(1)$	$C(0)$	$Y(1)$	$Y(0)$	W	C_{obs}	Y_{obs}
1/4	3	2	10	10	0	2	10
1/4	3	2	10	10	1	3	10
1/4	4	3	12	12	0	3	12
1/4	4	3	12	12	1	4	12

Figure 2. An Example With a Treatment Effect on the Concomitant, C, But No Treatment Effect on the Primary Outcome, Y.

Fraction of population	Potential outcomes				Observed data		
	$C(1)$	$C(0)$	$Y(1)$	$Y(0)$	W	C_{obs}	Y_{obs}
1/6	3	2	11	10	0	2	10
1/6	3	2	11	10	1	3	11
1/6	3	3	13	12	0	3	12
1/6	3	3	13	12	1	3	13
1/6	4	3	15	14	0	3	14
1/6	4	3	15	14	1	4	15

Figure 3. An Example With a Constant Treatment Effect on the Outcome, Y, and a "Direct" Effect for Units With No Treatment Effect on the Concomitant, C.

principal stratum where there is no treatment effect on the concomitant, the answer to the question about the direct effect of treatment seems to be clear: It is size 1. Yet once again, Fisher's advice provides an incorrect answer, despite the parallel linear regression lines in the $W_i = 0$ and $W_i = 1$ groups; the ANCOVA of $Y_{\text{obs},i}$ on W_i and $C_{\text{obs},i}$ implies that the "direct" causal effect of treatment on the primary outcome, after accounting for the effect of treatment on concomitant, is of size -1, which is the average $Y_{\text{obs},i}$ for the treated with $C_{\text{obs},i} = 3$ (the average $Y_{\text{obs},i}$ in rows 2 and 4, i.e., 12) minus the average $Y_{\text{obs},i}$ for the controls with $C_{\text{obs},i} = 3$ (the average $Y_{\text{obs},i}$ in rows 3 and 5, i.e., 13).

Another way to see what is wrong with these analyses is to realize, despite treatment being ignorable in these examples, that forcing the conditioning on C_{obs} leads to a nonignorable treatment assignment mechanism. For instance, from Figure 2, we see that in this case, in contrast to (10), we have

$$\Pr\big(W_i = 1 | C_{\text{obs},i}, Y_i(1), Y_i(0)\big) = \begin{cases} 1 & \text{if } C_{\text{obs},i} = 3 \text{ and } Y_i(1) = 10 \\ 1 & \text{if } C_{\text{obs},i} = 4 \\ 0 & \text{otherwise.} \end{cases} \quad (11)$$

Similarly, for the example depicted in Figure 3,

$$\Pr\big(W_i = 1 | C_{\text{obs},i}, Y_i(1), Y_i(0)\big) = \begin{cases} 1 & \text{if } C_{\text{obs},i} = 3 \text{ and } Y_i(1) = 11 \text{ or } 13 \\ 1 & \text{if } C_{\text{obs},i} = 4 \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

Fisher's ANCOVA is predicated on an ignorable assignment mechanism because it implicitly assumes that the set of subjects with any fixed value of $C_{\text{obs},i}$ are randomized into treatment and control, which is not true from (11) and (12). A correct analysis conditional on $C_{\text{obs},i}$ has to account for the nonignorable assignment mechanism created by this forced conditioning.

There are other valid ways to analyze data like these. One way is to combine C and Y into one outcome variable, such as the ratio Y/C. Another way is to consider (C, Y) as a bivariate outcome variable, which is related to, but definitely not the same as, Fisher's suggestion. A third way, and the one that attempts to penetrate the treatment effect on Y after "adjusting" for C, is to use the principal stratification approach of Frangakis and Rubin (2002). For more on this approach, see Rubin (2004) on "direct" and "indirect" causal effects, which advocates a Bayesian analysis where the latent principal strata are essentially multiply imputed, as well as the accompanying discussion and rejoinder. Also see Frangakis, Rubin, and Zhou (2002), where some of the invited discussion of the article appears to repeat Fisher's flawed reasoning.

8. CONCLUSIONS REGARDING CAUSAL INFERENCE USING POTENTIAL OUTCOMES

We have taken a short and idiosyncratic, but I hope interesting, trip through the world of inference for causal effects—idiosyncratic in that it clearly represents my own views and the influence of various giants of statistics on these views. Despite other approaches advocated by people whom I greatly respect (e.g., Dawid 2000; Lauritzen 2004; Pearl 2000), the potential outcomes formulation of causal effects, whether in randomized experiments or in observational studies, has achieved widespread acceptance. The potential outcomes, together with covariates, define the science in the sense that all causal estimands are functions of these values. Assumptions are required for causal inference, and stability (i.e., SUTVA) is the most straightforward of these in many circumstances.

It is necessary to posit an assignment mechanism when drawing causal inferences using probabilistic statements because at least half of the potential outcomes are missing in any investigation, even after making the simplifying SUTVA assumption. Randomized experiments have two special features that make such inferences particularly easy to draw: ignorability and propensities between 0 and 1. Fisher and Neyman each proposed a method of inference for causal effects based solely on the assignment mechanism; Fisher's method involved significance tests of sharp null hypotheses, whereas Neyman's method involved the repeated sampling expectations of statistics.

We can supplement the assignment mechanism with a model on the science and thus adopt, in essence, a Bayesian framework to inference for causal effects, which makes immediate ties to Fisher's contributions to likelihood theory based on models, even though Fisher himself never made this connection. The Bayesian perspective is extremely flexible and is especially convenient for summarizing the current state of knowledge about the science in complex situations. This summarization of knowledge can be, and I believe generally should be, viewed as a enterprise distinct from making decisions, which involve all sorts of trade-offs with various losses and gains, as Fisher pointed out, often unfathomable at the time that the science is being summarized.

When conducting causal inference, maintaining the distinction between observed values (e.g., concomitants such as $C_{i,\text{obs}}$) and which potential outcomes they reflect [e.g., $C_i(1)$ vs. $C_i(0)$] is critical for clear thinking. It is all too easy to slip into ignoring the difference and reaching invalid conclusions. Data are not only measurements, but also reflections of what they measure. A statement like "the value of C_{obs} is 3, but I'm not going to tell you whether that's a measurement of $C(1)$ or $C(0)$" is not generally wise or helpful. Scientific inferences change, in general, when $C_{\text{obs}} = 3$ implies that $C(1) = 3$ versus when it implies that $C(0) = 3$.

Even the brilliant Fisher was trapped by this to some extent, because he failed to recognize important contributions from Neyman and others he seemed to view as intellectually inferior. As Fisher himself stated:

> The example is a useful reminder of the truth that in general, a change in the data [e.g., whether C_{obs} is a realization of $C(1)$ or $C(0)$] may be expected to lead to a change in the inferences.

This was Fisher's concluding line (1954, p. 213) in his criticism of Monica Creasy's discussion paper (1954), which started and now completes our journey.

[Received October 2004. Revised October 2004.]

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables" (with discussion), *Journal of the American Statistical Association*, 91, 444–472.

Anscombe, F. J. (1948), "The Validity of Comparative Experiments," *Journal of the Royal Statistical Society*, Ser. A, 61, 181–211.

Brillinger, D. R., Jones, L. V., and Tukey, J. W. (1978), "Report of the Statistical Task Force for the Weather Modification Advisory Board," in *The Management of Western Resources, Vol. II: The Role of Statistics on Weather Resources Management*, Stock No. 003-018-00091-1, Washington, DC: U.S. Government Printing Office.

Chernoff, H. (1972), *Sequential Analysis and Optimal Design*, Philadelphia: Society for Industrial and Applied Mathematics.
Cochran, W. G. (1983), *Planning and Analysis of Observational Studies*, New York: Wiley.
Cochran, W. G., and Cox, G. M. (1950), *Experimental Designs*, New York: Wiley.
Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.
——— (1989), "Role of Models in Statistical Analysis," *Statistical Science*, 5, 169–174.
Creasy, M. A. (1954), "Limits for the Ratio of Means," *Journal of the Royal Statistical Society*, Ser. B, 16, 186–194.
Dawid, A. P. (2000), "Causal Inference Without Counterfactuals" (with discussion), *Journal of the American Statistical Association*, 95, 407–448.
Efron, B., and Feldman, D. (1991), "Compliance as an Explanatory Variable in Clinical Trials" (with discussion), *Journal of the American Statistical Association*, 86, 9–26.
Fieller, E. C. (1944), "A Fundamental Formula in the Statistics of Biological Assay, and Some Applications," *Quarterly Journal of Pharmacology*, 17, 117–123.
——— (1954), "Some Problems in Interval Estimation," *Journal of the Royal Statistical Society*, Ser. B, 16, 175–185.
Fisher, R. A. (1918), "The Causes of Human Variability," *Eugenics Review*, 10, 213–220.
——— (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society*, Ser. A, 222, 309–368.
——— (1925), *Statistical Methods for Research Workers*, London: Oliver & Boyd.
——— (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513.
——— (1935), *Design of Experiments*, Edinburgh: Oliver & Boyd.
——— (1954), Comment on "Limits of the Ratio of Means," by M. A. Creasy, *Journal of the Royal Statistical Society*, Ser. B, 16, 212–213.
——— (1956), *Statistical Methods and Scientific Inference*, New York: Hafner.
——— (1966), *Design of Experiments* (8th ed.), Edinburgh: Oliver & Boyd.
——— (1970), *Statistical Methods for Research Workers* (14th ed), Edinburgh: Oliver & Boyd.
Frangakis, C. E., and Baker, S. G. (2001), "Compliance Adjusted Double-Sampling Designs for Comparative Research: Estimation and Optimal Planning," *Biometrics*, 57, 899–908.
Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.
Frangakis, C. E., Rubin, D. B., and Zhou, X.-H. (2002), "Clustered Encouragement Designs With Individual Noncompliance: Bayesian Inference With Randomization, and Application to Advance Directive Forms" (with discussion), *Biostatistics*, 3, 147–177.
Gelman, A., and King, G. (1990), "Estimating Incumbency Advantage Without Bias," *American Journal of Political Science*, 34, 1142–1164.
Gelman, A., Meng, X. L., and Stern, H. S. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.
Greenland, S., Pearl, J., and Robins, J. (1999), "Causal Diagrams for Epidemiologic Research," *Epidemiology*, 10, 37–48.
Grodstein, F., Clarkson, T. M., and Manson, J. E. (2003), "Understanding the Divergent Data on Postmenopausal Hormone Therapy," *New England Journal of Medicine*, 348, 645–650.
Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11, 1–12.
Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 125–129.
——— (1989), "Causal Inference and Nonrandom Samples," *Journal of Educational Statistics*, 14, 159–168.
——— (1996), Discussion of "Identification of Causal Effects Using Instrumental Variables," by J. D. Angrist, G. W. Imbens, and D. B. Rubin, *Journal of the American Statistical Association*, 91, 459–462.
Holland, P. W. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 81, 945–970.
Hurwicz, L. (1962), "On the Structural Form of Interdependent Systems," in *Logic, Methodology, and Philosophy of Science* (Proceedings of the 1960 International Congress), eds. E. Nagel, P. Suppes, and A. Tarski, Stanford, CA: Stanford University Press.
Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.
Imbens, G. W., and Rubin, D. B. (2005), "Causal Inference in Statistics and the Medical and Social Sciences," unpublished manuscript.
Kempthorne, O. (1952), *The Design and Analysis of Experiments*, New York: Wiley.
——— (1976), Comment on "On Rereading R. A. Fisher," by L. A. Savage, *The Annals of Statistics*, 4, 495–497.
Lauritzen, S. (2004), Discussion of "Direct and Indirect Causal Effects via Potential Outcomes," by D. B. Rubin, *Scandinavian Journal of Statistics*, 31, 189–192.
Lewis, D. (1973), *Counterfactuals*, Oxford U.K.: Blackwell.
Little, R. J. A., and Rubin, D. B. (2000), "Causal Effects in Clinical and Epidemiological Studies via Potential Outcomes: Concepts and Analytical Approaches," *Annual Review of Public Health*, 21, 121–145.
McCarthy, M. D. (1939), "On the Application of the z-Test to Randomized Blocks," *The Annals of Mathematical Statistics*, 10, 337.
Meng, X. L. (1994), "Posterior Predictive p Values," *The Annals of Statistics*, 22, 1142–1160.
Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Roczniki Nauk Rolniczych Tom X* [in Polish]; translated in *Statistical Science*, 5, 465–480.
——— (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, Ser. A, 97, 558–606.
——— (1935), "Statistical Problems in Agricultural Experimentation" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 2 (suppl.), 107–108.
——— (1961), "Silver Jubilee of My Dispute With Fisher," *Journal of the Operations Research Society of Japan*, 3, 145–154.
Pearl, J. (1996), "Causation, Action and Counterfactuals," in *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge*, ed. Y. Shoham, San Francisco, CA: Morgan Kaufman, pp. 57–73.
——— (2000), *Causality: Models, Reasoning, and Inference*, Cambridge, U.K.: Cambridge University Press.
Pitman, E. J. G. (1938), "Significance Tests Which Can Be Applied to Samples From Any Populations. III. The Analysis of Variance Test," *Biometrika*, 29, 322–335.
Pratt, J. W., and Schlaifer, R. (1984), "On the Nature and Discovery of Structure" (with discussion), *Journal of the American Statistical Association*, 79, 3–33.
——— (1988), "On the Interpretation of Observational Laws," *Econometrics*, 39, 23–52.
Quandt, R. E. (1958), "The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes," *Journal of the American Statistical Association*, 53, 873–880.
Reid, C. (1982), *Neyman From Life*, New York: Springer-Verlag.
Rosenbaum, P. R. (2002), *Observational Studies* (2nd ed.), New York: Springer-Verlag.
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.
——— (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.
——— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Incorporating the Propensity Score," *The American Statistician*, 39, 33–38.
Roy, A. (1951), "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.
Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.
——— (1975), "Bayesian Inference for Causality: The Importance of Randomization," in *Proceedings of the Social Statistics Section*, American Statistical Association, pp. 233–239.
——— (1976), "Inference and Missing Data" (with discussion), *Biometrika*, 63, 581–592.
——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.
——— (1980), Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test," by D. Basu, *Journal of the American Statistical Association*, 75, 591–593.
——— (1984a), "William G. Cochran's Contributions to the Design, Analysis, and Evaluation of Observational Studies," in *W. G. Cochran's Impact on Statistics*, eds. P. S. R. S. Rao and J. Sedransk, New York: Wiley, pp. 37–69.
——— (1984b), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.
——— (1990), "Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, 5, 472–480.
——— (2004), "Direct and Indirect Causal Effects via Potential Outcomes" (with discussion), *Scandinavian Journal of Statistics*, 31, 161–170, 195–198.
Savage, L. A. (1976), "On Rereading R. A. Fisher," *The Annals of Statistics*, 4, 441–500.
Sheiner, L. B., and Rubin, D. B. (1994), "Intention-to-Treat Analysis and the Goals of Clinical Trials," *Clinical Pharmacology and Therapeutics*, 87, 6–15.

Sobel, M. E. (1996), "An Introduction to Causal Inference," *Sociological Methods & Research*, 76, 353–379.

Tinbergen, J. (1930), "Determination and Interpretation of Supply Curves: An Example," *Zeitschrift fur Nationalokonomie*; reprinted in Hendry, D. F., and Morgan, M. S. (eds.) (1995), *The Foundations of Econometrics*, Cambridge, U.K.: Cambridge University Press, p. 233.

Ware, J. H. (1989), "Investigating Therapies of Potentially Great Benefit: ECMO" (with discussion), *Statistical Science*, 4, 298–340.

Welch, B. L. (1937), "On the z Test in Randomized Blocks and Latin Squares," *Biometrika*, 29, 21–52.

Winship, C., and Morgan, S. L. (1999), "The Estimation of Causal Effects From Observational Data," *Annual Review of Sociology*, 25, 659–707.

Women's Health Initiative Study Group (1998), "Design of the Women's Health Initiative Clinical Trial and Observational Study," *Controlled Clinical Trials*, 19, 61–109.