Causal Inference Without Counterfactuals

A. P. DAWID

A popular approach to the framing and answering of causal questions relies on the idea of counterfactuals: outcomes that would have been observed had the world developed differently; for example, if the patient had received a different treatment. By definition, one can never observe such quantities, nor assess empirically the validity of any modeling assumptions made about them, even though one's conclusions may be sensitive to these assumptions. Here I argue that for making inference about the likely effects of applied causes, counterfactual arguments are unnecessary and potentially misleading. An alternative approach, based on Bayesian decision analysis, is presented. Properties of counterfactuals *are* relevant to inference about the likely causes of observed effects, but close attention then must be given to the nature and context of the query, as well as to what conclusions can and cannot be supported empirically. In particular, even in the absence of statistical uncertainty, such inferences may be subject to an irreducible degree of ambiguity.

KEY WORDS: Average causal effect; Causes of effects; Causation; Determinism; Effects of causes; Metaphysical model; Potential response; Treatment-unit additivity.

PART I: INTRODUCTION

1. CAUSAL MODELING

Association is not causation. Many have held that statistics, though well suited to investigate the former, strays into treacherous waters when it makes claims to say anything meaningful about the latter. Yet others have proceeded as if inference about the causes of observed phenomena were indeed a valid object of statistical enquiry; and it is certainly a great temptation for statisticians to attempt such "causal inference." Among those who have taken the logic of causal statistical inference seriously, I mention in particular Rubin (1974, 1978), Holland (1986), Robins (1986, 1987), Pearl (1995a), and Shafer (1996). This article represents my own attempt to contribute to the debate as to the appropriate statistical models and methods to use for causal inference, and what causal conclusions can be justified by statistical analysis.

There are many philosophical and statistical approaches to understanding and uncovering causation, and here I do not attempt to attack the problem on a broad front. I continue my attention to a simple decision-based understanding of causation, wherein an external agent can make interventions in, and observe various properties of, some system. Rubin (1978) and Heckerman and Shachter (1995), among others, have emphasized the importance of a clear decision-theoretic description of a causal problem. Understanding of the "causal effects" of intervention will come through the building, testing, and application of *causal models*, relating interventions, responses, and other variables.

In my view, the enterprise of causal statistical modeling is not essentially different from any other kind of statistical modeling, and is most satisfactorily understood from a Popperian hypothetico-deductive viewpoint. A model is not a straightforward reflection of external reality, and to propose a model is not to assert or to believe that nature behaves in a particular way. (Nature is surely utterly indifferent to our attempts to ensnare her in our theories.) Rather, a model is a construct within the mental universe, through which we attempt somehow to describe certain, more or less restricted, aspects of the empirical universe. To do this, we need to have a clear understanding of the semantics of such a description. This involves setting up a clear correspondence between the very different features of these two universes. In particular, we require very clear (if possibly implicit) understandings of:

- what the system modeled is (and so in particular how to distinguish a valid from an invalid instance of the system)
- what real world quantities are represented by variables appearing in the model
- what an intervention involves (for example, "setting" a patient's treatment to "none" by (a) withholding it from him, (b) wiring his jaw shut, or (c) killing him are all very different interventions, with different effects, and must be modeled as such. We must also be clear as to what variables are affected by the intervention, directly or indirectly, and how.)
- what is meant by replication (in time, space, etc.).

Also vital are clearly defined methods for understanding, assessing, and measuring the empirical success of any such attempt at description of the real world by a mathematical model. (One approach to such understanding and assessment in the case of ordinary probability modeling, based on the concept of probability calibration, may be found in Dawid 1985.)

As long as a model appears to describe the relevant aspects of the world satisfactorily, we may continue, cautiously, to use it; when it fails to do so, we need to search for a better one. In particular, any causal understandings that we may feel we have attained must always be treated as tentative and subject to revision should further observation of the world require it.

A. P. Dawid is Professor of Statistics, Department of Statistical Science, University College London, London, WC1E 6BT, U.K. (E-mail: *dawid@stats.ucl.ac.uk*). The ideas finally presented in this article have been festering for many years, in the course of which the author has had valuable discussions (and often heated arguments) with many people. The author particularly wishes to acknowledge the major contributions of Don Rubin, Judea Pearl, Glenn Shafer, Jamie Robins, Ross Shachter, and Volodya Vovk.

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

To be fully general, I should consider models for complex problems, such as those discussed by Robins (1986) and Pearl (1995a), wherein interventions of various kinds are possible at various points in a system, with effects that can cascade through a collection of variables. Although such problems can be modeled and analyzed (using structures such as influence diagrams) within the general philosophical and methodological framework of this article, that would involve additional theoretical development. To keep things simple, I restrict attention here to systems on which it is possible to make a single external intervention, which I refer to as *treatment,* and observe a single eventual *response.* I also suppose, with no further real loss of generality, that just two treatments are available. Another restriction, that could again be relaxed at the cost of further elaboration, is that I do not address the important and challenging problems arising from nonignorable treatment assignment or observational studies (e.g., Rubin 1974, 1978); see, however, Section 8.1 for some related analysis.

2. COUNTERFACTUALS

Much recent analysis of causal inference is grounded in the manipulation of *counterfactuals.* Philosophically, a counterfactual statement is an assertion of the form "if X had been the case, then Y would have happened," made when it is known to be false that X is the case. In a famous historical counterfactual, Pascal (1669, sec. 162), opined:

> Le nez de Cléopâtre: s'il eût été plus court, toute la face de la terre aurait changé.

(If Cleopatra's nose had been shorter, the whole face of the world would have been altered.) More recently, an intriguing, seemingly self-referring, assertion was made by Shafer (1996, p. 108):

> Were counterfactuals to have objective meaning, we might take them as basic, and define probability and causality in terms of them.

One of the aims of this article is to persuade the reader of the genuinely counterfactual nature of this claim.

An archetype of the use of counterfactuals in a causal statistical context is the assertion "if only I had taken aspirin, my headache would have gone by now." It is implicit that I did not take aspirin, and I still have the headache. Such an assertion, if true, could be regarded as justifying an inference that not taking aspirin has "caused" my headache to persist this long; and that if I had taken aspirin, that would have "caused" my headache to disappear by now. The assignment of cause is thus based on a comparison of the real and the counterfactual outcomes.

If Y_{A} denotes the duration of my headache when I take aspirin, and $Y_{\bar{\mathrm{A}}}$ its duration when I don't, then the foregoing assertion is of the form "$Y_{\bar{\mathrm{A}}} > y, Y_{\mathrm{A}} < y$" and relates jointly to the pair of values for $(Y_{\mathrm{A}}, Y_{\bar{\mathrm{A}}})$. An important question, which motivates much of the development in this article, is to what extent such assertions can be validated or refuted by empirical observation. My approach is grounded in a Popperian philosophy, in which the meaningfulness of a purportedly scientific theory, proposition, quantity, or concept is related to the implications it has for what is or could be observed, and, in particular, to the extent to which it is possible to conceive of data that would be affected by the truth of the proposition or the value of the quantity. When this is the case, assertions are empirically refutable and are considered "scientific." When this is not so, they may be branded "metaphysical." I argue that counterfactual theories are essentially metaphysical. This in itself might not be automatic grounds for rejection of such a theory, if the causal inferences that it led to were unaffected by the metaphysical assumptions embodied in it. Unfortunately, this is not so, and the answers that the approach delivers to its inferential questions are seen, on closer analysis, to be dependent on the validity of assumptions that are entirely untestable, even in principle. This can lead to distorted understandings and undesirable practical consequences.

3. TWO PROBLEMS

There are several different problems of causal inference, which are often conflated. In particular, I consider it important to distinguish between causal queries of the two types (Holland, 1986):

I. "I have a headache. Will it help if I take aspirin?"

II. "My headache has gone. Is it because I took aspirin?"

Query I requires inference about the *effects of causes*; that is, comparisons among the expected consequences of various possible interventions in a system. Such queries have long been the focus of the bulk of the standard statistical theory of experimental design (which, it is worth remarking, has in general displayed little eagerness for counterfactual analyses). Query II, in contrast, relates to *causes of effects*; one seeks to understand the causal relationship between an already observed outcome and an earlier intervention. Queries of this second kind might arise in legal inquiries; for example, into whether responsibility for a particular claimant's leukemia can be attributed to the fact that her father worked in a nuclear power station for 23 years. The distinction between queries I and II is closely related to that sometimes made between problems of *general* and of *singular* causation (Hitchcock 1997), although in our formulation both queries relate to singular circumstances.

I consider both types of query valid and important, but they are different, and require different, though related treatments. Evidence, (e.g., findings from epidemiological surveys) that is directly relevant to query I, is often used, inappropriately, to address query II, without careful attention to the difference between the queries.

4. PREVIEW

In Part II I consider the problem of "effects of causes." Section 5 introduces the essential ingredients of the problem and distinguish two varieties of model: a *metaphysical model,* which allows direct formulation of counterfactual quantities and queries, and a *physical model,* which does not. By means of a simple running example, I illustrate how certain inferences based on a metaphysical model are not completely determined by the data, however extensive, but remain sensitive to untestable additional assumptions. I also delimit the extent of the resulting arbitrariness. Section 6 describes an entirely different approach, based on physical

modeling and decision analysis, and shows how it delivers an unambiguous conclusion, avoiding the above problems. Section 7 questions the role of an implicit attitude of "fatalism" in some counterfactual causal models and methods. Section 8 extends the discussion to cases in which additional covariate information is available on individual systems. Section 9 investigates whether certain analyses stemming from a counterfactual approach nevertheless might be acceptable for "physical" purposes; examples are given of both possible answers. Section 10 asks whether it might ever be strictly advantageous to base physical analyses on a metaphysical structure. This appears to be sometimes the case for causal *modeling,* but arguably not so for causal *inference.*

In Part III I address the distinct problem of "causes of effects." For this, purely physical modeling appears inadequate, and the arbitrariness already identified in metaphysical modeling becomes a much more serious problem. Section 11 explains how this arbitrariness can be reduced by taking account of concomitant variables. Section 12 introduces a convention of conditional independence across alternative universes, which helps clarify the counterfactual inference and possibly reduce the intrinsic ambiguity. Section 13 considers the possibility of using underlying deterministic relations to clarify causal questions and inferences. I argue that to be useful, these must involve genuine concomitant variables. A contrast is drawn with "pseudodeterministic models," which are always available in the counterfactual framework. These have a deterministic mathematical structure, but need not involve true concomitants. Such a purely formal structure, I argue, is not enough to support meaningful inferences about the causes of effects. Section 14 discusses in more detail the meaning of concomitance and argues that this is partly a matter of convention, relative to a specific causal inquiry, rather than a property of the physical world.

The general message of this article is that inferences based on counterfactual assumptions and models are generally unhelpful and frequently plain misleading. Alternative approaches can avoid these problems, while continuing to address meaningful causal questions. For inference about the effects of causes, a straightforward "black box" decision-analytic approach, based on models and quantities that are empirically testable and discoverable, is perfectly adequate. For inference about the causes of effects, causal models must be suited to the questions addressed as well as to the empirical world, and understanding of the relationships between observed variables and possibly unobserved, but empirically meaningful, concomitant variables becomes important. The causal inferences justified by empirical findings will still in general retain a degree of arbitrariness and convention, which should be fully admitted.

PART II: EFFECTS OF CAUSES

5. COMPARISON OF TREATMENTS: COUNTERFACTUAL APPROACH

As a simple and familiar setting to discuss and contrast different approaches to inference about the effects of causes, I investigate the problem of making comparisons between two treatments, t and c (e.g., aspirin and placebo control) on the basis of an experiment. In this section I consider counterfactual approaches to this problem and show how they can produce ambiguous answers, unless arbitrary and unverifiable assumptions are imposed.

Consider a large homogeneous population $\mathcal{U}$ of clearly distinguishable individuals, or systems, or (as we shall generally call them) units, u, to each of which one can choose to apply any one treatment, i, out of the treatment set $\mathcal{T} = \{t, c\}$, and observe the resulting response, Y. Once one treatment has been applied, the other treatment can no longer be applied. This property can be ensured by appropriate definition of experimental unit u (e.g., headache episode rather than patient) and treatment (combinations of treatments, if available, being redefined as new treatments).

Experimentation consists in selecting disjoint sets of units $\mathcal{U}_i \subseteq \mathcal{U}$ $(i = t, c)$, applying treatment i to each unit in $\mathcal{U}_i$, and observing the ensuing responses (e.g., time for the headache to disappear). The experimental units might be selected for treatment by some form of randomization, but this is inessential to my argument. For further clarification of the argument, I assume that the treatment groups are sufficiently large so that all inferential problems associated with finite sampling can be ignored.

Homogeneity of the population is an intuitive concept, which can be formalized in a number of ways. From a classical standpoint, the individuals might be regarded as drawn randomly and independently from some large population; a Bayesian might regard them as exchangeable. In this context, homogeneity is also taken to imply that no specific information is available on the units that might serve to distinguish one from another (this constraint is relaxed in Sec. 8). In particular, the experimenter is unable to take any such information into account, either deliberately or inadvertently, in deciding which treatment a particular unit is to receive. To render this scenario more realistic and versatile, suppose that he did in fact have additional measured *covariate* information on each unit, determined by (but not uniquely identifying) that unit. Then one would confine attention to a subpopulation having certain fixed covariate values, and this subpopulation might then be reasonably regarded as homogeneous. That is, this discussion should be understood as applying at the level of the residual variation, after all relevant observed covariates have been allowed for. (One can then also allow treatment assignment to take these observed covariates into account.)

Counterfactual Framework. The counterfactual approach to causal analysis for this problem focuses on the collection of *potential responses* $\mathcal{Y} := (Y_i(u)\colon i \in \mathcal{T}, u \in \mathcal{U})$, where $Y_i(u)$ is intended to denote "the response that would be observed if treatment i were assigned to unit u." One can consider $\mathcal{Y}$ as arranged in a two-way layout of treatments by units, with $Y_i(u)$ occupying the cell for row i and column u. Note that many of the variables in $\mathcal{Y}$ are (to borrow a term from quantum physics) *complementary,* in that they are not simultaneously observable. Specifically, for any unit u, one can observe $Y_i(u)$ for at most one treat-

ment i. Assignment of treatments to units will determine just which (if any) of these complementary variables are to be observed, yielding a collection $\mathcal{X}$ of responses that I call a *physical array*—in contrast to the *metaphysical array* $\mathcal{Y}$. Although the full collection $\mathcal{Y}$ is intrinsically unobservable, counterfactual analyses are based on consideration of all of the $(Y_i(u))$ simultaneously. Current interest in the counterfactual approach was instigated by Rubin (1974, 1978), although it can be traced back at least to Neyman (1935; see also Neyman 1923).

5.1 Metaphysical Model

What kind of models can be reasonably entertained for the metaphysical array $\mathcal{Y}$? The assumption of homogeneity essentially requires us to model the various pairs $(Y_t(u), Y_c(u))$ for $u \in \mathcal{U}$ as iid, given their (typically unknown) bivariate distribution P. I denote the implied marginal distributions for Y_t and Y_c by P_t and P_c. It is important to note that the full bivariate distribution P is not completely specified by these marginals, without further specification of the dependence between Y_t and Y_c.

Although the major points of the discussion apply to a general model of the foregoing form, for definiteness I concentrate on the following specific bivariate normal model.

Example 1. The pairs $\{(Y_t(u), Y_c(u)): u \in \mathcal{U}\}$ are modeled as iid, each with the bivariate normal distribution with means (θ_t, θ_c), common variance ϕ_Y, and correlation ρ.

When $\rho \geq 0$, which seems a reasonable judgment (see section 12), one can also represent this structure by means of the mixed model

$$Y_i(u) = \theta_i + \beta(u) + \gamma_i(u), \tag{1}$$

where all of the $(\beta(u))$ and $(\gamma_i(u))$ are mutually independent normal random variables, with mean 0 and variances $\phi_\beta := \rho\phi_Y$ and $\phi_\gamma := (1 - \rho)\phi_Y$. One can also regard (1) as a (fictitious) representation of the bivariate normal model even when $\rho < 0$, in which case we must have $-\phi_Y \leq \phi_\beta \leq 0$ and $0 \leq \phi_\gamma \leq 2\phi_Y$. Then the calculations below, though based on this fictitious representation, are still valid. Inversely, one could start with (1) as the model, in which case

$$\phi_Y = \phi_\beta + \phi_\gamma \tag{2}$$

and

$$\rho = \frac{\phi_\beta}{\phi_\beta + \phi_\gamma}. \tag{3}$$

In the usual parlance of the analysis of variance, (1) expresses $Y_i(u)$ as composed of a fixed *treatment effect* θ_i associated with the applied treatment i, common to all units; a random *unit effect* $\beta(u)$, unique to unit u, but common to both treatments; and a random *unit–treatment interaction,* $\gamma_i(u)$, varying from one treatment application to another, even on the same unit. [This last term could also be interpreted as incorporating intrinsic random variation, which can not be distinguished from interaction because replicate observations on $Y_i(u)$ are impossible.]

5.2 Causal Effect

The counterfactual approach typically takes as the fundamental object of causal inference the *individual causal effect:* a suitable numerical comparison, for a given unit, between the various potential responses it would exhibit, under the various treatments that might be applied. Note that such a quantity is meaningless unless one regards the several potential responses, complementary though they are, as having simultaneous existence.

Here the individual causal effect (ICE) for unit u is identified with the difference

$$\tau(u) := Y_t(u) - Y_c(u). \tag{4}$$

Alternative possibilities might be $\log Y_t(u) - \log Y_c(u)$ and $Y_t(u)/Y_c(u)$. There seems no obvious theoretical reason, within this framework, to prefer any one such comparison to any other, the choice perhaps being made according to one's understanding of the applied context and the type of inferential conclusion desired. But however defined, an ICE involves direct comparison of complementary quantities and is thus intrinsically unobservable.

In most studies, the specific units used in the experiment are of no special interest in themselves, but merely provide a basis for inference about generic properties of units under the influence of the various treatments. For this purpose, it is helpful to conceive of an entirely new *test unit,* u_0, from the same population, that has not yet been treated, and to regard the purpose of the experiment as to assist in making the decision as to which treatment to apply to it. If one decides on treatment t, then one obtains response $Y_t(u_0)$; if c, one obtains $Y_c(u_0)$. Thus inference needs to be made about these two quantities, and they need to be compared somehow. Note that although $Y_t(u_0)$ and $Y_c(u_0)$ are complementary, neither is (as yet) counterfactual.

The counterfactual approach might focus on the ICE $\tau(u_0) = Y_t(u_0) - Y_c(u_0)$, or a suitable variation thereon. Under (1),

$$\tau(u) = \tau + \lambda(u), \tag{5}$$

with $\tau := \theta_t - \theta_c$, the *average causal effect* (ACE), and $\lambda(u) := \gamma_t(u) - \gamma_c(u)$, the *residual causal effect,* having distribution

$$\lambda(u) \sim \mathrm{N}(0, 2\phi_\gamma). \tag{6}$$

Thus

$$\tau(u) \sim \mathrm{N}(\tau, 2\phi_\gamma). \tag{7}$$

This model holds in particular for the inferential target $\tau(u_0)$. Because $\tau(u_0)$ is probabilistically independent of any data on the units in the experiment, inference about $\tau(u_0)$ essentially reduces to inference about the pair (τ, ϕ_γ).

5.3 Physical Model

Suppose that a particular experimental assignment has been specified. Label, arbitrarily, the units receiving treatment i as $u_{i1}, u_{i2} \ldots, u_{i_{n_i}}$. Then the observed response on unit u_{ij} is $X_{ij} := Y_i(u_{ij})$. The collection $(X_{ij}: i = t, c; j = 1, \ldots, n_i)$ constitutes the physical array $\mathcal{X}$. The

mean response on all units receiving treatment i is $\bar{X}_i := (1/n_i)\sum_{j=1}^{n_i} X_{ij}$.

It follows trivially from the model assumptions of Example 1 that the joint distribution over $\mathcal{X}$ is described by

$$X_{ij} \sim \mathrm{N}(\theta_i, \phi_Y), \tag{8}$$

independently for all (i, j). Equivalently, from (1),

$$X_{ij} = \theta_i + \varepsilon_{ij}, \tag{9}$$

with $\varepsilon_{ij} := \beta(u_{ij}) + \gamma_i(u_{ij}) \sim \mathrm{N}(0, \phi_Y)$ independently for all (i, j).

Now to the extent that the (1) says anything about the empirical world, this must be fully captured in the implied models (8) (one such for each possible physical array). Clearly, from extensive data having the structure (8), one can identify θ_t, θ_c, and ϕ_Y, but the individual components ϕ_β and ϕ_γ in (2)—or, equivalently, the correlation ρ satisfying (3)—are not identifiable; one has *intrinsic aliasing* (McCullagh and Nelder 1989, sec. 3.5) of unit effect and unit–treatment interaction. As far as the desired inference about $\tau(u_0)$ is concerned, one can identify its mean, $\tau = \mathrm{ACE}$, in (7). However, its variance, $2\phi_\gamma$, is not identifiable from the data, beyond the requirement $\phi_\gamma \leq \phi_Y$ (if one restricts to $\rho \geq 0$) or $\phi_\gamma \leq 2\phi_Y$ (for ρ unrestricted).

5.4 A Quandary

This poses an inferential quandary. Consider two statisticians, both of whom believe in (1). However, statistician S1 further assumes that $\phi_\beta = 0$ ($\rho = 0$), and statistician S2 assumes that $\phi_\gamma = 0$ ($\rho = 1$). Both S1 and S2 accept (8) for the physical array, with no further constraints on its parameters. Extensive data, assumed to be fully consistent with (8) for the physical array, lead to essentially exact estimates of θ_t, θ_c, and ϕ_Y. However, S1 infers $\phi_\beta = 0$ and $\phi_\gamma = \phi_Y$, whereas S2 has $\phi_\beta = \phi_Y$ and $\phi_\gamma = 0$. When they come to inference about $\tau(u_0)$, from (7), they will agree on its mean, τ, but differ about its variance, $2\phi_\gamma$. A third statistician, making different assumptions (e.g., $\phi_\beta = \phi_\gamma$, equivalent to $\rho = 1/2$) will come to yet another distinct conclusion. Is it not worrisome that models that are intrinsically indistinguishable, on the basis of any data that could ever be observed, can lead to such different inferences? How can one possibly choose between these inferences?

The aforementioned state of affairs is clearly in violation of what, in another context (Dawid 1984, sec. 5.2), I have called *Jeffreys's law:* the requirement that mathematically distinct models that cannot be distinguished on the basis of empirical observation should lead to indistinguishable inferences. This property can be demonstrated mathematically in cases where those inferences concern future observables, and I consider it to have just as much intuitive force in the present context of causal inference.

There is one important, but very special, case where the foregoing ambiguity vanishes: when ϕ_Y is essentially 0, and hence so are both ϕ_β and ϕ_γ. In this case the units are not merely homogeneous, but *uniform,* in that for each $i, Y_i(u)$ is the same for all units u. The property $\phi_Y \doteq 0$ can, of course, be investigated empirically, and might be regarded as a distinguishing feature of at least some problems in the "hard" sciences. When it holds, one can in effect observe both $Y_t(u)$ and $Y_c(u)$ simultaneously, by using distinct units, thus enabling direct measurement of causal effects. I further consider this case of uniformity, and its extensions, in Section 13.

5.5 Additional Constraints

How should one proceed if one does not have uniformity? It is common in studies based on counterfactual models to impose additional constraints. In the present context, a common additional constraint is that of *treatment–unit additivity* (TUA), which asserts that $\tau(u)$ in (4) is the same for all $u \in \mathcal{U}$. In terms of (1), this is equivalent to $\phi_\gamma = 0$ ($\rho = 1$) and leads to a simple inference: $\tau(u_0) = \tau$, with no further uncertainty (τ having been identified, from a large experiment, as $\bar{X}_t - \bar{X}_c$). However, as pointed out earlier, there is simply no way that TUA can be tested on the basis of any empirically observable data in the context of (1), and it is intuitively clear that the same holds for any other models that might be considered. When for each pair $(Y_t(u), Y_c(u))$, it is never possible to observe both components, how can one ever assess empirically the assertion that $Y_t(u) - Y_c(u)$ (unobservable for each u) is the same for all u? If I had used a more general model in Example 1, whereby I allowed the variance to be different for two responses, say ϕ_t and ϕ_c, then TUA does have the testable implication $\phi_t = \phi_c$, and so could be rejected on the basis of data casting doubt on this property. But such data would still not distinguish between TUA and any of the other models considered earlier, all of which would likewise be rejected. I have assumed throughout that the data are consistent with the physical model (8), so that this issue does not arise.

A similar untestable assumption commonly made in the case of binary responses (Imbens and Angrist 1994) is *monotonicity,* which requires that $P\ (Y_c = 1, Y_t = 0) = 0$ (where the response 1 represents a successful, and 0 an unsuccessful, outcome).

5.6 What Can Be Said?

If inferences are restricted to those that *are* justified by the data, without the imposition of untestable additional constraints, then the most that can be said about $\tau(u_0)$ [assuming (1)] is

$$\tau(u_0) \sim \mathrm{N}(\tau, 2\phi_\gamma), \tag{10}$$

with τ estimated precisely but ϕ_γ subject only to the inequality $0 \leq \phi_\gamma \leq \phi_Y$ (or $0 \leq \phi_\gamma \leq 2\phi_Y$ if one allows $\rho < 0$), whose right side only is estimated precisely. Only if one is fortunate enough to find that ϕ_Y is negligible (the situation of uniformity) can one obtain an unambiguous inference for $\tau(u_0)$.

A very similar analysis can be conducted for other metaphysical models. Although the physical model only allows one to identify the marginal distributions P_t and P_c of the joint distribution P, the distribution of an individual causal effect (however defined) will depend further on the dependence structure of P. (There is a large literature

on properties and inequalities for joint distributions with known marginals; see, e.g., Rüschendorf, Schweizer, and Taylor 1996.) Consequently, even when very large experiments have been conducted, unambiguous inferences about such causal effects cannot be made without making further untestable assumptions, such as TUA or monotonicity.

Two contrasting morals may be drawn from the foregoing analysis, both grounded in the principle that one should be careful not to make "metaphysical inferences" sensitive to assumptions that can not be put to empirical test. Moral 1 is that inference about individual causal effects should be carefully circumscribed, as following (10). Alternatively, one might draw the more revolutionary Moral 2, that if one cannot get a sensible answer to the question, then perhaps the question itself, with its focus on inference for $\tau(u_0)$, is not well posed. In the next section I reformulate the question in an entirely different manner that allows a clear and unambiguous answer.

6. DECISION-ANALYTIC APPROACH

As demonstrated in the foregoing example, the principal difficulty with the counterfactual approach is that the desired inference depends on the joint probability structure of the complementary variables $(Y_t(u), Y_c(u))$, whereas one is only ever able to observe (at most) one of these for each u. One can, however, consistently estimate both marginal distributions P_t, and P_c. Can these separate marginal distributions be put to good use?

I take a straightforward Bayesian decision-analytic approach (see, e.g., Raiffa 1968). One has to decide whether to apply treatment t or treatment c to a new unit u_0. The marginal distributions P_t and P_c of Y_t and of Y_c having been identified, from extensive experimental data on each separate treatment group, these now express the appropriate predictive uncertainty about the response on u_0, conditional on its being given t or c. The consequence (loss) of the decision may be measured by some function $L(\cdot)$ of the eventual yield Y. The decision tree for this problem is given in Figure 1.

At node $\nu_t, Y \sim P_t$, and the (negative) value of being at ν_t is measured by the expected loss $E_{P_t}\{L(Y)\}$. Similarly, ν_c has value $E_{P_c}\{L(Y)\}$. The principles of Bayesian decision analysis now require that at the decision node ν_0, that treatment i leading to the smaller expected loss be chosen.

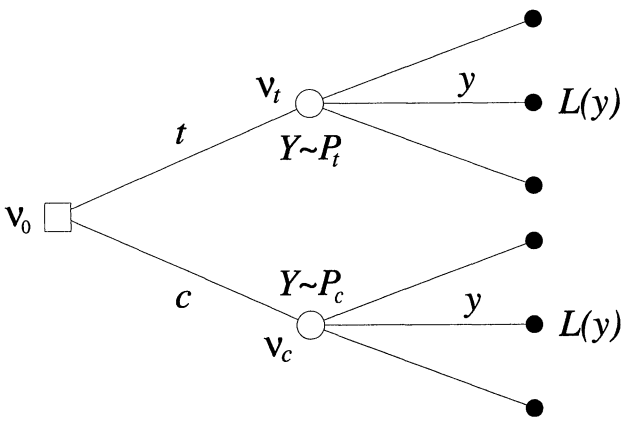

Figure 1. Decision Tree.

Note that whatever loss function is used, this solution involves only the two identifiable marginal distributions, P_c and P_t. In particular, our statisticians S1 and S2 of Section 5.4, who agree on (1) and obtain common estimates of θ_t, θ_c, and ϕ_Y, while disagreeing about ρ, will be led to the identical decision. It simply does not matter that S2 believes that the time for a headache to disappear if aspirin is taken will be *exactly* 10 minutes less than if it is not taken, whereas S1 regards the difference of these times as uncertain, although again with expectation 10 minutes; there is no way in which such differences in beliefs can affect the decision problem.

It is only for simplicity of the argument that I have assumed that the experiment is large enough to allow full identification of P_t and P_c. With a more limited experiment, one could either replace these with suitable estimates or, for a wholeheartedly Bayesian approach, use the appropriate predictive distributions for the response on u_0 (under either hypothetical treatment application, separately), given the experimental data.

My analysis extends readily to the case where one wants to decide how to apply treatments to a number of future units. In a quality control setting, the loss might be a combination of the sample mean and variance of all the responses, for example.

One can also consider models for more complex problems, involving nonhomogeneous populations. For example, in earlier work (Dawid 1988) I used symmetry arguments to justify the construction of certain random-effects–type models for complex experimental layouts, generalizing models such as those of (1) for the metaphysical array or (9) for the physical array. In the general case, one again needs to use the data of the experiment to make appropriate predictive inferences for test units, under varying hypothetical treatment assignments; but these predictive inferences will now be more complex and will also depend on the relationship assumed between the test units and the experimental units. For example, if the experiment involved planting different varieties of cereal on plots (units) nested within blocks nested within fields, and recording their yields, then one might wish to consider predictions for the yield of each variety if planted on a new plot in an old (i.e., experimental) block in an old field, a plot in a new block in an old field, or (more usefully) a plot in a new field. As long as one's models relate the responses of the new and the old units (under arbitrary treatment assignments), and so support the required predictive inferences, one can conduct whatever decision-analytic analysis appears most relevant to one's purpose, eschewing counterfactuals entirely.

7. FATALISM

Many counterfactual analyses are based, explicitly or implicitly, on an attitude that I term *fatalism.* This considers the various potential responses $Y_i(u)$, when treatment i is applied to unit u, as predetermined attributes of unit u, waiting only to be uncovered by suitable experimentation. (It is

implicit that that the unit u and its properties and propensities exist independently of, and are unaffected by, any treatment that may be applied.) Note that because each unit label u is regarded as individual and unrepeatable, there is never any possibility of empirically testing this assumption of fatalism, which thus can be categorized as metaphysical.

The fatalistic worldview runs very much counter to the philosophy underlying statistical modeling and inference in almost every other setting. For example, it leaves no scope for introducing realistic stochastic effects of external influences acting between the times of application of treatment and of the response. Any account of causation that requires one to jettison all of the familiar statistical framework and machinery should be treated with the utmost suspicion, unless and until it has shown itself completely indispensable for its purpose.

7.1 Some Fatalistic Concepts

I do not wish to give the impression that all counterfactual analyses must be fatalistic; there are notable exceptions (e.g., Robins and Greenland 1989). However, it is a very natural bedfellow of counterfactual inference, much of which can not proceed without it. For example, only if one takes a fatalistic attitude does it make sense even to talk of such properties as treatment-unit additivity or monotonicity (Sec. 8).

A fundamental use of fatalism underlies certain counterfactual analyses of treatment non-compliance (see, e.g., Imbens and Rubin 1997), where each patient is supposed categorizable as a *complier* (who would take the treatment if prescribed, and not take it if not prescribed), a *defier* (not take it if prescribed, take it if not prescribed), an *always taker* (take it whether or not prescribed), or a *never taker* (not take it whether or not prescribed). Some causal inferences are based on consideration of the responses to treatment of, say, the group of compliers. However, it is only under the unrealistic assumption of fatalism that this group has any meaningful identity, and thus only in this case could such inferences even begin to have any useful content.

7.1.1 Stable Unit-Treatment Value Assumption. An assumption that has often been considered essential to useful causal inferences is the *stable unit-treatment value assumption* (SUTVA) (Rubin 1980, 1986). To describe this, one has to start from a more general metaphysical model of the effect of experimentation on responses, wherein the response $Y_\xi(u)$ of unit u could in principle depend on the full treatment assignment ξ over all units, not just on the specific treatment i applied to u. Then SUTVA requires that in fact this potential complicating feature be absent, so that one can replace $Y_\xi(u)$ by $Y_i(u)$, thus returning to the situation already considered. But again, without the fatalistic assumption of preexisting values of the $(Y_\xi(u))$, for any assignment ξ, it is not possible to make sense of SUTVA (but see Sec. 10.1.1 for a nonfatalistic reinterpretation of SUTVA).

7.1.2 Decision Analysis and Fatalism. By contrast, the decision-analytic approach requires no commitment to (or, for that matter, against) fatalism. There is no conceptual or mathematical difficulty in regarding the probability distributions of the response (i.e., P_t and P_c in Example 1) as incorporating further uncontrollable influences over and above effects attributable directly to treatment. As far as SUTVA is concerned, the decision analyst has no need of it. In the context of Example 1, SUTVA can be replaced by the much weaker assumption that the application of treatments does not destroy the homogeneity of the units, beyond the obvious difference that some will now have one treatment and some will have another. Then one will still have complete homogeneity of the responses for all units (experimental or future) receiving the same treatment, and can thus use the experimental data to identify the distribution, P_i, of response within treatment group i, which also expresses the uncertainty about the response $Y_i(u_0)$ of a new unit u_0, if it were given treatment i. Hence one is still in a position to set up, and solve, the basic decision problem for u_0.

8. USE OF ADDITIONAL INFORMATION

Now suppose that it is possible to gather, or at least to conceive of gathering, additional information about individual units, which might be used to refine uncertainties about their responses to treatments. Any such information can be described in terms of a generic variable K, determined by a measurement protocol that, when applied to unit u, leads to a measurement $K(u)$. For the analysis of effects of causes I restrict attention to generic variables that are *covariates;* that is, features of units that can be observed prior to experimentation. Nevertheless, before it is observed, each $K(u)$ must be treated as a random variable.

There are several cases to consider, according as whether or not the covariates are observed on the experimental units and/or on test units:

1. Covariates on experimental and test units. Suppose that a covariate K is measured on all experimental units, and also that for a test unit $u_0, K(u_0)$ will be measured before the treatment decision has to be made.

If K takes values in a finite set, then one can simply restrict attention to the subset (assumed large) of the experimental units for which $K(u) = K(u_0)$. Then one essentially recovers the homogeneous population problem that has already been analyzed.

Otherwise, or if the aforementioned restricted subset is not sufficiently large, one can conduct appropriate statistical modeling. A counterfactual treatment would need to model a joint conditional distribution of (Y_c, Y_t) given K; for the decision-analytic treatment, one only needs to use the data to assess and compare the associated predictive distributions of $Y(u_0)$ given $K(u_0)$, for each treatment. Again, the decision-analytic approach, in contrast to the counterfactual approach, is essentially insensitive to any further assumptions about, or modeling of, the joint distribution of potential responses.

2. Covariates on experimental units only. In this case it is appropriate to ignore altogether the covariate information on the experimental units—except that when the experiment

is not large, modeling this more detailed information might enhance the accuracy of estimation of the required marginal predictive distributions of $Y(u_0)$ for each treatment.

3. Covariate on test unit only. This is more problematic, because even for the less demanding decision-analytic approach, the experiment gives no direct information about the required predictive distributions of response given covariate and treatment. Whichever approach one takes, there is no escape from the fact that the solution will be highly dependent on untested (though in principle testable) assumptions about these distributions. One possibility would be to ignore $K(u_0)$ altogether, but this is itself tantamount to an empirically untested assumption of independence between K and Y for each treatment. In any event, however one proceeds, there is no advantage to be gained from the introduction of counterfactuals. Similar comments apply when information of differing extents is available on the experimental and test units.

8.1 Alternatives to Additivity

One argument that can be made for the need for a metaphysical assumption such as treatment-unit additivity (Sec. 5.5) is the following. An experiment (e.g., a clinical trial) will often have very specific *inclusion criteria* that render the experimental units nonrepresentative of the population to which it is intended to generalize the findings. Then, although one may still have homogeneity of units within the experiment, it might no longer be reasonable to regard the test unit u_0 as exchangeable with the experimental units. But if we can assume TUA, so that $Y_t(u) - Y_c(u) \equiv \tau$ for all units, experimental and test, then an estimate of the treatment effect τ from the experiment will still be applicable to u_0. Thus counterfactual analysis based on TUA appears unaffected by this modification to the framework. For the decision-analytic approach, however, the required separate predictive inferences about the response $Y(u_0)$, given either treatment, for a test unit u_0 would be simultaneously more complicated and less reliable when the experimental units cannot be regarded as representative of the test units.

An alternative way of proceeding avoids metaphysical assumptions. For each unit u, let $Q(u)$ be a variable taking values $0, t$, and c, generated by the experimenter as part of the process of designing his experiment. He intends to include u in the experiment and apply treatment t to it if $Q(u) = t$, to include u in the experiment and apply treatment c to it if $Q(u) = c$, and to exclude u from the experiment if $Q(u) = 0$. These intentions do not, however, preclude one from considering other possibilities; one can, for example, meaningfully assess probabilistic uncertainty about $Y(u)$, given that the assignment $Q(u) = t$ has been made, on the hypothesis that u will receive treatment c.

I assume that, for some covariate K, the distribution of $Q(u)$ given $K(u)$ is the same for all units u. Thus K is the information that the experimenter takes into account in generating Q, and so embodies the inclusion and treatment criteria. The distribution of Q given K is assumed unaffected by further conditioning on the applied treatment i and the eventual response Y. Using the notation and properties of conditional independence (Dawid 1979), $Q \perp\!\!\!\perp (i, Y)|K$, whence

$$Y \perp\!\!\!\perp Q|K, i. \tag{11}$$

Consider now the model assumption

$$E(Y|K, i) = \theta_i + \gamma(K) \qquad (i = t, c), \tag{12}$$

for some unknown parameters θ_t and θ_c and parametric function $\gamma(\cdot)$. If this holds, define $\tau = \theta_t - \theta_c$.

Note that by (11), the left side of (12) is unaffected by further conditioning on Q. In particular, (12) implies $E\{Y|K, i, Q = i\} = \theta_i + \gamma(K)$ $(i = t, c)$, so that for any k,

$$E\{Y|K = k, t, Q = t\} - E\{Y|K = k, c, Q = c\} \equiv \tau. \tag{13}$$

Conversely, (13) with (11) implies (12). But $E\{Y|K = k, i, Q = i\}$ can be estimated straightforwardly from the measurements of covariate K and outcome Y on the set of experimental units to which treatment i has been applied. Consequently, property (12) is testable from the experimental data, and, if it can be assumed to hold, the parameter τ is estimable. (A simple unbiased estimator of τ is given by the difference of the mean responses for the two treated groups.)

Also, one can compare hypothetical treatment applications on a test unit u_0, with observed $K(u_0) = k$ and, by construction, $Q(u_0) = 0$, as follows:

$$\begin{aligned} &E\{Y(u_0)|K(u_0) = k, t\} - E\{Y(u_0)|K(u_0) = k, c\} \\ &\quad = E\{Y|K = k, t, Q = 0\} - E\{Y|K = k, c, Q = 0\} \\ &\quad = E\{Y|K = k, t\} - E\{Y|K = k, c\}, \end{aligned}$$

once again using (12). But this is just τ, as identified from the experiment. (If $K(u_0)$ is not observed, then one must take a further expectation over K, but this clearly has no effect.)

The foregoing approach, based on the testable assumption (12) rather than the metaphysical assumption of TUA, thus allows one to generalize readily from the experiment to the target population, even in the face of differential selection and treatment criteria.

It has been assumed in the foregoing that it is appropriate to focus directly on the expected response. In the general framework of Section 6, with a loss function L, one could replace $E(Y)$ by $E\{L(Y)\}$ throughout. (A counterfactual analysis would similarly require that TUA be modified to $L\{Y_t(u)\} - L\{Y_c(u)\} \equiv \tau$, all u.)

9. SHEEP AND GOATS

I have argued that any elements of a theory that have no observable or testable consequences (e.g., TUA) are to be regarded as metaphysical, and, in accordance with Jeffreys's law, should not be permitted to have any inferential consequences either. Causal analyses can be classified into *sheep* (those obeying this dictum) and *goats* (the rest). I have shown that the decision-analytic approach is a sheep.

What of the counterfactual approach? It certainly has the potential to generate goats. In particular, any inference dependent on assumptions requiring the acceptance of fatalism (e.g., TUA, or monotonicity, or assertions about the

group of compliers in clinical trial) must be a goat. However, specific inferential uses of counterfactual models may turn out to be sheep. The following section describes one such use.

9.1 Average Causal Effect

Suppose that in the counterfactual approach, one were to define the ICE for unit u as $f\{Y_t(u)\} - f\{Y_c(u)\}$, for some function f. For example, one might use the linear form $Y_t(u) - Y_c(u)$, or the logarithmic form $\log\{Y_t(u)/Y_c(u)\}$. If $\mathcal{U}$ is effectively infinite, then the ACE [population average of ICE(u)] is $E_P\{f(Y_t) - f(Y_c)\}$. But this is just $E_{P_t}\{f(Y)\} - E_{P_c}\{f(Y)\}$ and thus depends only on the marginal distributions P_c and P_t (and is exactly the criterion determining the solution of the decision problem having $L \equiv f$). Hence this particular use of counterfactual analysis, focusing on an infinite-population ACE, is consistent with the decision-analytic approach and involves only terms subject to empirical scrutiny. It is fortunate that many of the superficially counterfactual analyses in the literature, from Rubin (1978) onward, have in fact confined attention to ACE and thus lead to acceptable conclusions.

However, seemingly minor variations of the foregoing form for ICE, such as $Y_t(u)/Y_c(u)$, can not be handled in this way. $E_P(Y_t/Y_c)$ is not determined by the marginals P_t and P_c alone, although these can be used to set bounds (Rachev 1985). So any form of inference focusing on such causal effects, at either the individual or the population average level, would be a metaphysical goat, dependent on untestable ingredients of the metaphysical model and hence likely to be misleading.

9.2 Neyman and Fisher

Here is a variation on ACE, using even the simple definition (4), that is nevertheless a goat. It is the basis of the approach introduced by Neyman (1935) and followed through by Wilk and Kempthorne (1955, 1956, 1957).

Let $\mathcal{U}^* := \mathcal{U}_t \cup \mathcal{U}_c$ be the set of experimental units, say N in total. (In the literature, the units are not completely homogeneous, but are classified in an experimental layout; e.g., a row-column structure with treatments imposed to form a latin square. However, this does not affect the essential logic.) Neyman expressed the null hypothesis of "no treatment effect" as asserting that $Y_t^* = Y_c^*$, where $Y_i^* := N^{-1} \sum_{u \in \mathcal{U}^*} Y_i(u)$ is the average response that would have been observed in the experiment had all units been given treatment i (thus both Y_t^* and Y_c^* are genuinely counterfactual quantities). Wilk and Kempthorne (1955) considered averages over a larger, but still finite, population $\mathcal{U}$ from which $\mathcal{U}^*$ was drawn. In these approaches, inference is based on the distribution generated by random treatment assignment (and, where appropriate, random sampling of the levels used for the experiment), under assumed values for the metaphysical array of all potential responses $(Y_i(u))$, these values playing the role of parameters in the randomization model. Such an approach (even when extended by introducing random errors of observation) is clearly based on a fatalistic worldview.

Neyman showed that for the latin square, the usual t test was an unbiased test of his null hypothesis only if TUA could be assumed; similarly, the analyses of Wilk and Kempthorne give different answers, according to whether or not one assumes TUA. These workers concluded that one needs to think very carefully, in each particular context, about the validity of the TUA assumption, and tailor one's inferences accordingly. However, because there are no conceivable data that could shed any light on this validity, it is not clear how to act on this advice. Two statisticians with observationally equivalent models could arrive at discrepant conclusions. This suggests very strongly that Neyman's approach is not a helpful one, and that his metaphysical null hypothesis is misguided.

Fisher, in the rapporteur's account of his comments on Neyman (1935), rejected this approach, arguing instead that the appropriate null hypothesis was

$$H_0\colon\ \tau = 0,$$

for which the standard t test *is* valid.

Fisher's null hypothesis is often taken to have been

$$H_0^*\colon\ \tau(u) \equiv 0;$$

that is, $\tau = 0$ and $\phi_\gamma = 0$, implying $Y_t(u) = Y_c(u)$ for all u. This, too, is a metaphysical hypothesis. However, it is not certain that this was Fisher's intention. In any case, as far as the observable structure (8) is concerned, these two hypotheses are indistinguishable, as are the resulting tests. This identity extends to more complex layouts; in earlier work (Dawid 1988), I showed how the standard tests may be justified purely on the basis of a hypothesis of invariance of the joint distribution of responses under suitable relabelling of units, which is very much weaker than H_0^* (see also Cox 1958). The broader hypothesis H_0 is equivalent to $P_t = P_c$, which is all that is needed for indifference in the decision problem—and is, of course, a sheep, being testable from the data.

10. INSTRUMENTAL USE OF COUNTERFACTUALS

Even if one accepts that the output of a causal analysis should not involve any direct assertions about counterfactuals, the example of Section 9.1 demonstrates that it is at least possible to use counterfactual models for acceptable purposes. However, that example also shows no obvious advantage to doing so, and the use of counterfactual models always lays one open to the danger of producing "goat-like" inferences, without signalling when that is the case (as for the variant forms of ACE considered at the end of Sec. 9.1).

It nevertheless remains conceivable that purely mathematical use of the richer structure inherent in the modeling of the metaphysical array might actually simplify some derivations and analyses of acceptable "sheep-like" inferences. An analogy might be the fruitfulness of coupling arguments in probability theory, or of complex analysis in number theory.

In my view, there may be a limited place for such instrumental use of counterfactuals in the context of causal *model-building*. However, I remain to be persuaded of the

usefulness of counterfactuals, even in a purely instrumental role, for causal *inference.*

10.1 Counterfactuals for Modeling

The model (9) for the physical array was derived by marginalizing the metaphysical model of Example 1, so as to focus on the subcollection of variables picked out by the experimental design. This may be regarded as an instrumental use of counterfactuals for the purposes of modeling. However, in this simple example this looks like overkill; (9) is itself a very natural structure to impose on the physical array directly.

In more complicated problems, there may be some genuine advantage to modeling at the metaphysical level. Thus, suppose that the experimental units are laid out in a row-column structure. One way to build appropriate models for outcomes is to apply the ideas of symmetry modeling (Dawid 1988). If one associates with each plot the full vector of (complementary) potential responses it would exhibit under the various different possible treatment applications, then it might be reasonable to regard the joint distribution for all of these vectors as invariant under separate relabellings of rows and columns. If (less compellingly, and purely for simplicity of exposition) we also impose invariance under relabellings of the treatments, symmetry arguments imply that we can represent the probability structure of the metaphysical array $\mathcal{Y} = (Y_{irc})$ (where i labels treatments, r labels rows, and c labels columns) by the random-effects model

$$Y_{irc} = \mu + \alpha_i + \beta_r + \gamma_c + (\alpha\beta)_{ir} + (\alpha\gamma)_{ic} + (\beta\gamma)_{rc} + (\alpha\beta\gamma)_{irc}, \quad (14)$$

with all the terms uncorrelated, $\mathrm{var}(\alpha_i) = \sigma_\alpha^2$, and so on.

If one considers the implications of this model for the marginal joint distribution of some physical array $\mathcal{X} = (X_{rc})$, in which a specified treatment $i = i(r, c)$ is applied to the unit in row r and column c, then one finds a similar representation, but with the last two terms intrinsically confounded, just as the separate terms $\beta(u)$ and $\gamma_i(u)$ in (1) are confounded in the term ε_{ij} of (9). If one further confines attention to latin square designs, so that no treatment appears more than once in any row or column, then there is additional (extrinsic) confounding, resulting in the model

$$X_{rc} = \mu + \alpha_i + \beta_r + \gamma_c + \varepsilon_{rc}, \quad (15)$$

where, with $i = i(r, c)$,

$$\varepsilon_{rc} = (\alpha\beta)_{ir} + (\alpha\gamma)_{ic} + (\beta\gamma)_{rc} + (\alpha\beta\gamma)_{irc}. \quad (16)$$

This is of course the (random-effects version of) the usual model for the observables in the latin square design. The extrinsic confounding between the $(\alpha\beta), (\alpha\gamma)$, and $(\beta\gamma) + (\alpha\beta\gamma)$ terms in (16) will, however, make predictive inferences, which depend on these terms individually, especially sensitive to assumptions that cannot be tested with such a design.

On the other hand, one could initially restrict attention to the physical array $\mathcal{X}$ and consider the group of symmetries that preserve its structure. Such a symmetry is represented by the combination of a row permutation and a column permutation having the additional property that any two units receiving identical treatments before permutation also receive identical treatments after permutation. This group will depend very specifically on the way in which treatments are assigned to units, and can have highly variable structure for different latin square layouts (Bailey 1991, ex. 4). Because of these additional restrictions on the symmetry of the physical array $\mathcal{X}$, the implied symmetry model constructed directly for $\mathcal{X}$ can be considerably more complex than that expressed by (15). In such a case, modeling the metaphysical array directly, for the purely instrumental use of deriving an appropriate model for the physical array, appears to be the more fruitful approach.

Another example of the usefulness (or at least convenience), for constructing models of the physical domain, of direct modeling of the metaphysical domain (using "pseudostructural nested distribution models") was given by Robins and Wasserman (1997).

10.1.1 Compatibility. Taking the approach of modeling each possible physical array by marginalising from a single joint model for the metaphysical array, the resulting collection of physical models will have a property that I term *compatibility:* For two different experimental layouts that both result in unit u receiving treatment i, the marginal models for the associated response on that unit are identical. This identity extends to the joint model for the responses of a collection of units that happen to be treated in the same way in both experiments. This property can be regarded as a noncounterfactual counterpart of the counterfactual SUTVA (see Sec. 7).

I further distinguish two forms, *strong* and *weak,* of compatibility for a collection of physical models under varying treatment assignments. Weak compatibility (which seems the more natural, and makes no reference whatsoever to counterfactuals) simply requires the earlier stated property of identity of common marginal models. Strong compatibility requires the existence of a single joint model for the metaphysical array that can be used to generate, by appropriate marginalization, the various different physical models. To extend the analogy with quantum theory, strong compatibility requires the existence of "hidden variables," underlying all observations that might be made. Although strong compatibility always implies weak compatibility, in full generality the converse need not hold. Consider, for example, variables (Y_1, Y_2, Y_3), where Y_i is either 1 or -1 and where one can observe any of the pairs $(Y_1, Y_2), (Y_2, Y_3)$, and (Y_3, Y_1) but cannot observe all three variables simultaneously. The corresponding bivariate distributions are specified by $Y_1 = Y_2, Y_2 = Y_3$, and $Y_3 = -Y_1$, with Y_i either 1 or -1, each with probability $1/2$. Then these distributions are weakly, but not strongly, compatible. (I am grateful to Steffen Lauritzen for this example.) Although the structure of this example is not quite the same as that of the current problem, it is conceivable that causal models also could have weak compatibility without strong compatibility. This opens up the possibility of a still deeper analogy with quantum theory, where observable behavior cannot be explained by means of a "hidden variable" theory.

In the decision-analytic approach, the property of compatibility, although possibly very useful in streamlining the modeling, has no fundamental role to play. All that is needed is to construct appropriate models relating the outcomes on the experimental units, according to the treatment assignments actually made, with those on as-yet untreated units, under various assumptions about how those new units might be treated. Then these can be used to make predictive inferences under the varying assumptions, and so assess the relative value of future interventions.

10.2 Counterfactuals for Inference?

There are many problems where workers who have grown familiar and comfortable with counterfactual modeling and analysis evidently consider that it forms the only satisfactory basis for *causal inference.* However, I have not as yet encountered any use of counterfactual models for inference about the effects of causes that is not either (a) a goat, delivering misleading inferences of no empirical content, or (b) interpretable, or readily reinterpretable, in noncounterfactual terms. I have already given examples of (a) and also, in Section 9.1, of (b). Here are some more cases of (b).

Robins (1986) initially developed causal inferential methods on the basis of a counterfactual model. However, in recent work (Robins and Wasserman 1997), both the underlying model and the associated methods are reexpressed in noncounterfactual terms.

Conversely, Pearl (1993), in introducing a semantics for graphical models of causal structures, did so in a way that avoided counterfactuals. Later (Pearl 1995a), he translated this into a counterfactual language, based on functional models, but to no obvious advantage; his specific analyses (e.g., in Pearl 1995a, app.) make no necessary use of this additional structure.

An interesting problem that did initially appear to require a counterfactual model is the development of inequalities for (sheep-like) causal effects in clinical trials with imperfect treatment compliance (Balke and Pearl 1994b). However, I have been able to derive the identical inequalities without the additional baggage of functional models or counterfactuals (indeed, an example of just such a derivation was given in Pearl 1995b).

Another interesting recent example of (b) given by Greenland, Robins, and Pearl (1999) purports to define confounding in terms of counterfactuals, but explicitly introduces an alternative interpretation based on exchangeability. Most of its analyses make no essential use of counterfactuals. Two appendixes, considering carefully the interpretation of counterfactual assertions in a number of cases, represent to me convincing demonstrations of their meaninglessness and pointlessness (although the authors themselves stop short of this conclusion).

PART III: CAUSES OF EFFECTS

11. INFERENCE ABOUT CAUSES OF EFFECTS

I now address the problem of inference about the causes of effects. As I demonstrate, this is still more problematic than inference about the effects of causes, and it may be impossible to avoid a degree of ambiguity in the resulting inferences.

The major new ingredient is that, along with having the experimental data, one now has a further unit u_0, of individual interest, to which treatment t has already been applied and the response $Y_t(u_0) = y_0$ observed. (One may also have further relevant information about u_0 or its environment, perhaps even gathered between the application of treatment and observation of response. I consider this possibility later but for the moment assume that this is not so.) Interest centers on whether, for the specific unit u_0, the application of t "caused" the observed response. It appears that, to address this question, there is no alternative but to somehow compare the observed valued y_0 with the counterfactual quantity $Y_c(u_0)$, the response that would have resulted from application of c to u_0. Equivalently, inference about the individual causal effect $\tau(u_0) = y_0 - Y_c(u_0)$ is required. However, the fact that such an inference may be desirable does not, in itself, render it possible. I now explore what can be justified scientifically from data.

Example 2. Consider again the bivariate normal counterfactual model of Example 1. Suppose that there is no possibility of ever measuring any other relevant information on any unit, beyond its response to treatment.

The conditional distribution of $\tau(u_0) \equiv Y_t(u_0) - Y_c(u_0)$, given $Y_t(u_0) = y_0$, is normal, with mean and variance

$$\lambda := E\{\tau(u_0)|Y_t(u_0) = y_0\} = y_0 - \theta_c - \rho(y_0 - \theta_t) \quad (17)$$

and

$$\delta^2 := \mathrm{var}\{\tau(u_0)|Y_t(u_0) = y_0\} = (1 - \rho^2)\phi_Y. \quad (18)$$

Now, as already emphasised, from the extensive experimental data [even when extended with the additional observation $Y_t(u_0) = y_0$], only θ_t, θ_c, and ϕ_Y can be learned. The correlation ρ cannot be identified. Hence, even with extensive data, residual arbitrariness remains. When $\rho = 0$ ($\phi_\beta = 0$, or independence of Y_t and Y_c), $\lambda = y_0 - \theta_c$ and $\delta^2 = \phi_Y$. The value $\rho = 1$ ($\phi_\gamma = 0$, or TUA) yields $\lambda = \theta_t - \theta_c$ and $\delta^2 = 0$ (or, at the other extreme, if $\rho = -1$, then $\lambda = 2y_0 - \theta_t - \theta_c$, and $\delta^2 = 0$ again). Assuming $\rho \geq 0$, only the inequalities

$$\lambda \text{ lies between } \theta_t - \theta_c \text{ and } y_0 - \theta_c$$

and

$$\delta^2 \leq \phi_Y$$

can be inferred. Thus only when y_0 is sufficiently close to θ_t will one get an unambiguous conclusion about λ, insensitive to empirically untestable assumptions about ρ; and only when ϕ_Y is sufficiently small will one be able to say anything empirically supportable and unambiguous about δ^2. If one takes $\rho = 1$, equivalent to TUA, then one obtains a seemingly deterministic inference, $\tau(u_0) = \theta_t - \theta_c$, but this is of little real value when the data give no reason to choose any particular value of ρ over any other. (The

inequalities developed here rely on the assumption, itself untestable, of joint normality. Even though the data may support marginal normality for each of Y_t and Y_c, any further aspects of the joint distribution must remain unknowable, and, in principle, the distribution of Y_c, given the observed value $Y_t = y$, could be anything so long as $\phi_Y > 0$. Thus a complete skeptic could hold that inference about the causes of effects, on the basis of empirical evidence, is impossible.)

Note that, if one does assume TUA, but not otherwise, then the retrospective inference about $\tau(u_0)$ is not affected by the additional information $Y_t(u_0) = y_0$ on the new unit, and thus is the same as for the case of arguing about effects of causes. Because the TUA assumption is so prevalent in the literature, the essential distinction between inference about the effects of causes and inference about the causes of effects has not usually been noted.

The aforementioned sensitivity to assumptions extends to, for example, Bayesian inference, which would require integration of the distribution defined by (17) and (18) over the posterior distribution of all the parameters. In this posterior, θ_t, θ_c, and ϕ_Y will be essentially degenerate at their sample estimates, so that one can substitute these in (17) and (18), and just integrate over the conditional distribution of the nonidentified parameter ρ, given $(\theta_t, \theta_c, \phi_Y)$. However, this will be exactly the same in the posterior as in the prior, and thus the inference will remain sensitive to the assumed form of the prior.

No amount of wishful thinking, clever analysis, or arbitrary untestable assumptions can license unambiguous inference about causes of effects, even when the model is simple and the data are extensive (unless one is lucky enough to discover uniformity among units).

11.1 Concomitants

It appears from the foregoing that there is an inherent ambiguity in inference about the causes of effects. However, some progress toward reducing this may be possible if one can probe more deeply into the hidden workings of the units, by observing suitable additional variables. This is the basis and purpose of scientific investigation. As demonstrated in Sections 6 and 8, such deeper scientific understanding is not essential for assessing "effects of causes," which can proceed by essentially a "black box" approach, simply modeling dependence of the response on whatever covariate information happens to be observed for the test unit. However, it is vital for any study of inference about "causes of effects," which must take into account what has been learned from experiments about the inner workings of the black box.

Thus suppose that it is possible to measure *concomitant variables* associated with a unit. These might be covariates, as already considered. However, other quantities can also be allowed, as long as they can be assumed to be unaffected by the treatment applied (although use of the term "unaffected" itself begs many causal and counterfactual questions; see sec. 14). An example might be the weather between the times of planting and of harvesting a crop. Typically the variation in the response conditional on concomitants will be smaller than that unconditionally.

Example 3. Suppose that, in the context of Example 1, detailed experiments have measured a concomitant K and have found that, conditional on $K(u) = k$ and the application of treatment i, the response $Y(u)$ is normally distributed with residual variance ψ_K, say, and mean $\theta_i + k$. From these experiments, the values of ψ_K and the θ's have been calculated.

Define $\phi_K := \mathrm{var}(K)$ and $\psi_0 := \phi_Y = \phi_K + \psi_K$. Then $\mathrm{cov}(K, Y_c) = \mathrm{cov}(K, Y_t) = \phi_K$. Combining these with the covariance structure for the complementary pair (Y_c, Y_t) implied by (1), the full dispersion matrix of (K, Y_c, Y_t) is seen to be

$$\begin{pmatrix} \phi_K & \phi_K & \phi_K \\ \phi_K & \phi_Y & \rho\phi_Y \\ \phi_K & \rho\phi_Y & \phi_Y \end{pmatrix}.$$

Thus the conditional correlation between Y_c and Y_t, given K, is

$$\rho_{ct\cdot K} := \frac{\rho\phi_Y - \phi_K}{\phi_Y - \phi_K} = 1 - (1-\rho)\frac{\psi_0}{\psi_K}. \tag{19}$$

In parallel to Example 2, the arbitrary parameter $\rho_{ct\cdot K} \in [-1, 1]$ cannot be identified from these more refined experiments (although it might be reasonable to take $\rho_{ct\cdot K} \geq 0$).

Now consider inference about "causes of effects" on a test unit u_0. I again distinguish between the cases where concomitant information is, or is not, available for u_0:

1. If one observed $K(u_0) = k$, say, then one could conduct an analysis very similar to that of Example 2. In particular, (17) would be replaced by $E\{\tau(u_0)|Y_t(u_0) = y, K(u_0) = k\} = (y - \theta_c - k) - \rho_{ct\cdot K}(y - \theta_t - k)$, which, because the final term in parentheses is now of order $\sqrt{\psi_K}$, rather than $\sqrt{\psi_0}$ as before, should be less sensitive to the arbitrariness in the correlation, now $\rho_{ct\cdot K}$. Similarly, (18) would be replaced by $\mathrm{var}\{\tau(u_0)|Y_t(u_0) = y, K(u_0) = k\} = (1-\rho^2_{ct\cdot K})\psi_K$, now bounded above by $\psi_K < \psi_0$, rather than by $\phi_Y = \psi_0$. Clearly these improvements are more substantial with smaller residual variance ψ_K of Y given K.

2. Now suppose that one does not observe $K(u_0)$, or any other concomitant variable, on u_0. In this case—in contrast to case 2 of in Section 8 for effects of causes—the analysis *is* affected by the more detailed findings in the experiments performed.

Define $\gamma_K := \phi_K/\phi_Y = 1 - \psi_K/\psi_0$. By (19), one has (assuming that $\rho_{ct\cdot K} \geq 0$)

$$\gamma_K \leq \rho \leq 1 \tag{20}$$

(or, for $\rho_{ct\cdot K}$ unrestricted, $2\gamma_K - 1 \leq \rho \leq 1$). Consequently, the experimental identification of K, even though it can not be observed on u_0, has reduced the "interval of ambiguity" for ρ from [0, 1] to $[\gamma_K, 1]$ (or, for $\rho_{ct\cdot K}$ unrestricted, from $[-1, 1]$ to $[2\gamma_K - 1, 1]$), and thus yields tighter limits on λ and δ^2 in (17) and (18).

From this perspective, the ultimate aim of scientific research may be seen as discovery of a concomitant variable, K^* say, that yields the smallest achievable residual variance $\psi^* := \psi_{K^*}$, and thus, with $\gamma^* := \gamma_{K^*} = 1 - \psi^*/\psi_0$, the shortest possible interval of ambiguity, $[\gamma^*, 1]$, for ρ. (I am here assuming, for simplicity, that the model of Example 3 applies for any concomitant K that might be considered. Although the mathematics are more complicated if this assumption is dropped, the essential logic continues to apply.) I term such a variable a *sufficient concomitant.* (The collection of all concomitants is always sufficient in this sense, but one would hope to be able to reduce it without explanatory loss.) However, unless $\psi^* = 0$, and rarely even then, it will not usually be possible to know whether this goal has been attained.

Nonetheless, using (20) with (17) and (18), one can still make scientifically sound (though imprecise) inferences on the basis of whatever current level of understanding, in terms of discovered explanatory concomitant variables K, has been attained. This will take into account that there is a nonstatistical component of uncertainty or arbitrariness in the inferences, expressed by interval bounds on the quantitative causal conclusions.

I have assumed that the experiments performed have been sufficiently large that purely statistical uncertainty can be ignored. In practice this will rarely be the case. However, an appropriate methodology for combining such statistical uncertainty with the intrinsic ambiguity that still remains in the limit is not yet available. Techniques for dealing with this problem are urgently needed.

12. CONDITIONAL INDEPENDENCE

Suppose that K^* is a sufficient concomitant. Assuming that $\rho_{ct \cdot K^*} \geq 0$, one has, from (19), the ultimate residual variance $\psi^* \geq (1 - \rho)\psi_0$. In particular, $\rho < 1$ implies that $\psi^* > 0$. If $\psi^* = 0$ (and thus $\rho = 1$), then the value of K^* determines both potential responses Y_t and Y_c, without error, and so, once K^* is identified, the ambiguity in the inferences entirely disappears. I call such a situation *deterministic,* and consider it further in Section 13.

However, for reasons discussed in Section 14, I regard determinism as exceptional, rather than routine. In this section I consider further the nondeterministic case, having $\psi^* > 0$, and, by (19), ρ constrained only to the interval of ambiguity $[\gamma^*, 1]$ (as $\rho_{ct \cdot K^*}$ ranges from 0 to 1), with $\rho^* = 1 - \psi^*/\psi_0$.

As far as any empirical evidence is concerned, there is no constraint whatsoever on $\rho_{ct \cdot K^*}$. However, it would seem odd to hypothesize, for example, $\rho_{ct \cdot K^*} = 1$, because this would imply $\rho = 1$, complete dependence between real and counterfactual responses, at the same time as asserting nondeterminism, in the sense that there is no concomitant information one could gather that would allow one to predict the response perfectly. Likewise, to hypothesize any other value of $\rho_{ct \cdot K^*} > 0$ would appear to leave open the possibility of finding a more powerful set of predictors that would explain away this residual dependence, thus further reducing the residual variance.

To limit the arbitrariness in the value of ρ, one could attempt to give ρ further meaning by requiring that $\rho_{ct \cdot K^*} = 0$; the totally inexplicable components of variation of the response, in the real and in the counterfactual universes, should be independent. Extending this, one might require that all variables be treated as conditionally independent across complementary universes, given all the concomitants (which are, of course, constant across universes). Under this assumption, the interval of ambiguity for ρ shrinks to the point $\gamma^* = 1 - \psi^*/\psi_0$.

The foregoing conditional independence assumption is best regarded as a *convention,* providing an interpretation of just what one intends by a counterfactual query. It leads to a factor-analysis–type decomposition of the joint probabilistic structure of complementary variables, into (a) a part fully explained by the concomitants, and common to all the complementary universes, and (b) residual "purely random" errors, modeled as independent (for any given unit) across universes. In this way, one can at last give a clear structure and meaning (albeit partly conventional) to a metaphysical probability model for the collection of all potential responses. Note that if one accepts this conditional independence convention, then one obtains, on using (19), $\rho = \gamma_{K^*} \geq 0$—providing some justification for imposing this condition. (Without the convention, and with no constraints on $\rho_{ct \cdot K^*}$, one can only assert $\rho \geq 2\gamma_{K^*} - 1$.)

Once a sufficient concomitant K^* is identified, leaving aside for the moment the question of how one could know this, the conditional independence convention renders counterfactual inference in principle straightforward and unambiguous. In the context of Example 3, one can take $\rho = \gamma^* = \psi^*/\psi_0$, thus eliminating the ambiguity. More generally, from detailed experiments on treated and untreated units, we can discover the joint distribution of K^* and Y_t, and of K^* and Y_c. For a new unit u_0 on which no concomitants are observed, on observing $Y_t(u_0) = y$ one can condition (using, e.g., Bayes's theorem) in the joint distribution of (K^*, Y_t) to find the revised distribution of K^*, and then combine this with the conditional distribution of Y_c given K^* to obtain the appropriate distribution of the counterfactual Y_c. This two-stage procedure is valid if and only if one accepts the conditional independence property. Alternatively (and equivalently), one can use this property to combine the two experimentally determined distributions into a single joint distribution for (K^*, Y_t, Y_c) and marginalize to obtain that of (Y_t, Y_c), then finally condition on $Y_t(u_0) = y$ in this bivariate distribution. Minor variations will handle the case where one has also observed the value of some concomitant variables on u_0.

Example 4 (with acknowledgment to V. G. Vovk). A certain company regularly needs to send some of its workers into the jungle. It knows that the probability that a typical worker will die (D) if sent to the jungle (J) is $\text{pr}(D|J) = 3/4$, compared with $\text{pr}(D|\bar{J}) = 1/4$ if the worker is retained at the head office. Joe is sent to the jungle, and dies. What is the probability that Joe would have died if he had been kept at the head office?

1. Suppose first that all workers are equally robust, and that the risk of dying is governed purely by the unspecified dangers of the two locations. One might then regard the complementary outcomes as independent, so that the answer to the question is 1/4.

2. Now suppose that, in addition to external dangers, the fate of a worker depends in part on his natural strength. With probability 1/2 each, a worker is either strong (S) or weak ($\bar{S}$). A strong worker has probability of dying in the jungle $\text{pr}(D|J,S) = 1/2$, and at the head office $\text{pr}(D|\bar{J},S) = 0$. A weak worker has respective probabilities $\text{pr}(D|J,\bar{S}) = 1$ and $\text{pr}(D|\bar{J},\bar{S}) = 1/2$. [These values are consistent with the earlier probabilities assigned to $\text{pr}(D|J)$ and $\text{pr}(D|\bar{J})$.] Given that Joe died in the jungle, the posterior probability that he was strong is 1/3. If one assumes conditional independence, given strength, between the complementary outcomes, the updated probability that he would have died if kept at the head office now becomes $1/3 \times 0 + 2/3 \times 1/2 = 1/3$.

3. In fact, Joe was replaced at the head office by Jim, who took his desk. Jim died when his filing cabinet fell on him. This gives additional information about the dangers Joe might have faced had he stayed behind. How should one take it into account? There is no right answer. If one regards the toppling of the filing cabinet, killing whoever is at the desk, as unaffected by who that occupant may be, and include it as a concomitant, then the answer becomes 1. Or one could elaborate, allowing the probability that the occupant is killed by the falling cabinet to depend on whether he is strong or weak. But it would be equally reasonable to consider that had Joe stayed behind, the dangers he would have met would have been different from those facing Jim. In this case the previous arguments and answers (according as whether or not one accounts for strength) could still be reasonable.

As should be clear from the foregoing example, even with the conditional independence convention the answer to a query about "causes of effects" must depend in part on what variables it is considered reasonable to regard as concomitants. I consider this issue further in Section 14.

12.1 Undiscovered Sufficient Concomitants

What if, as will usually be the case, one has measured concomitants K in experiments, but has not yet identified a sufficient concomitant K^*? In Example 3, one could then only assert $\psi^* \leq \psi_K$ and thus, using the conditional independence property $\rho = \gamma^*, \rho \geq \gamma_K$. Hence the convention of conditional independence at the level of the sufficient concomitant has not, in this case, resulted in any reduction in the interval of ambiguity for ρ.

Nevertheless, one can think, in the light of current knowledge and having regard to the potentially available concomitants (see sec. 14 below), about plausible values of the ultimate residual variance ψ_{K^*}, and use this in setting reasonable limits, or distributions, for $\rho = 1 - \psi_{K^*}/\psi_0$. This still leaves the inference dependent on (as yet) experimentally unverified assumptions, but it might at least be possible to present reasoned arguments for the assumptions made. This approach based on conditional independence also obviates the need for new methods of statistical inference, combining ambiguity and uncertainty.

13. DETERMINISM

In certain problems of the 'hard' sciences, it can happen that, by taking account of enough concomitant variables, the residual variation in the response for any treatment can be made to disappear completely (at least for all practical purposes), thus inducing at this more refined level the situation of uniformity considered in Section 5.4 when all problems of causal inference and prediction disappear. In Example 3, this would occur if one found $\psi_K = 0$, which would imply $\rho = 1$ and so eliminate all ambiguity. Such problems may be termed *deterministic,* because the response is then given as a function $Y = f(i, D)$ of the appropriate *determining concomitant* D (which is then necessarily sufficient) and the treatment i, without any further variability. This property is in principle testable when D is given. (If it is rejected, it may be possible to reinstate it, at a deeper level, by refining the definition of D.) However, even when such underlying determinism does exist, discovering that this is the case and identifying the determining concomitant D and the form of f may be practically difficult or impossible, requiring a large-scale, detailed, and expensive scientific investigation and sophisticated statistical analyses.

If one had a deterministic model, one could use it to *define* potential responses: $Y_i(u) = f(i, D(u))$. (Necessary here is the property that D, being a concomitant, is unaffected by treatment. But because D need not be a covariate, this model is not necessarily fatalistic.) One could determine the value of any potential response on unit u by measuring $D(u)$. Thus in this special case one can indeed consider the complementary variables $(Y_i(u)) \equiv (f(i, D(u)))$, for fixed unit u but varying treatment i, as having real, rather than merely metaphysical, simultaneous existence.

Note in particular that even in this rare case where one can give empirical meaning to counterfactuals, the causal modeling is not based on a primitive notion of counterfactual; rather, the counterfactuals are grounded in, and take their meaning from, the model. [In the same way, I consider that Lewis's (1973) interpretation of counterfactuals in terms of "closest possible worlds" is question-begging, because closeness cannot be sensibly defined except in terms of an assumed causal model.]

A deterministic model, when available, can also be used to make sense of nonmanipulative accounts of causation. Given D, the potential responses, for various real or hypothetical values of the variable "treatment," are determined and can be compared directly, however the specification of treatment may be effected.

For inference about the causes of effects, assume that one has observed $Y_t(u_0) = y_0$, but not $D(u_0)$, and wishes to assess uncertainty about $Y_c(u_0)$. In the context of Example 3, $\rho = 1$, eliminating all ambiguity and (in this rare case) justifying TUA and the inference $\tau(u_0) = \theta_t - \theta_c$. More generally, suppose that detailed experimentation has identified a deterministic model $Y_i(u) = f(i, D(u))$. Although

one has not observed $D(u_0)$, one can assess a distribution for it. This should reflect both typical natural variation of D across units (as discovered from experiments) and any additional concomitant information one may have on u_0. From this distribution, one can derive the induced joint distribution over the collection $(f(i, D(u_0)))$ of complementary potential responses. Then one can condition the distribution of $D(u_0)$ on the observation $f(t, D(u_0)) = y_0$ and thus arrive at appropriate posterior uncertainty about a genuine counterfactual such as $Y_c(u_0) \equiv f(c, D(u_0))$. In this way, a fully deterministic model (if known) allows an unambiguous solution to the problem of assessing the "causes of effects." The essential step is generation of the joint distribution over the set of complementary responses (together with any observed concomitants), this being fully grounded in an understanding of their dependence on determining concomitants, and a realistic probabilistic assessment of the uncertainty about those determining concomitants.

The foregoing procedure is merely a special case of that described in Section 12, but not now dependent on the convention of conditional independence of residual variation across parallel universes—because in this case there is no residual variation.

Example 5. Suppose that a major scientific investigation has demonstrated the validity of the model (1), but now reinterpreted as a deterministic model, with all of the β's and γ's identified as concomitant variables that can, with suitable instruments, be measured for any unit and have been so measured in the experimental studies. Further, from these studies, the previously specified independent normal distributions for these quantities have been verified, and all of the parameters $(\theta_t, \theta_c, \phi_\beta, \phi_\gamma)$ have been identified.

One now examines a new unit u_0, which has been given treatment t, and observes the associated response $Y_t(u_0) = y$. The individual causal effect $\tau(u_0)$ is $\gamma_t(u_0) - \gamma_c(u_0)$, which is now in principle measurable. In practice, measurement of the β's and γ's for unit u_0 may not be possible. Then (in the absence of any further relevant information) one might describe the uncertainty about their values using their known joint population distribution. The appropriate uncertainty about $\tau(u_0)$ is then expressed by the normal distribution with mean λ and variance δ^2 given by (17) and (18); however, because the value of $\rho = \phi_\beta/(\phi_\beta + \phi_\gamma)$ is now available from the scientific study, the ambiguity in this inference has been eliminated.

Note that it is vital for the foregoing analysis that the quantities $\gamma_t(u)$ and $\gamma_c(u)$ be *simultaneously* measurable, with the specified independent distributions. It is not enough only to identify $\beta(u)$ and define the γ's as error terms, $\gamma_i(u) = Y_i(u) - \theta_i - \beta(u)$; in that case, because one cannot simultaneously observe both $Y_t(u)$ and $Y_c(u)$, one cannot verify the required assumption of independence between $\gamma_t(u)$ and $\gamma_c(u)$.

13.1 Undiscovered Determinism

If one believes that the problem is deterministic, but has not yet completely identified the determining concomitant D or the function f, then one can propose parametric forms for f and the distribution of D, and attempt to estimate these (or integrate out over the posterior distribution of their parameters) using the available data. In principle, sufficiently detailed experimentation would render such assumptions empirically testable and identify the parameters. In practice, however, this may be far from the case. Thus consider Example 2, in which no concomitants have been measured. One could propose an underlying deterministic model of the form

$$Y = \theta_i + D_i, \quad (i = t, c),$$

with D_t and D_c determining concomitants, supposedly measurable on any unit by further, more refined, experiments. In the current state of knowledge, however, one can say no more than $D_i \sim \mathrm{N}(0, \phi_Y)$. Further, one has no information on the correlation ρ between D_t and D_c. It is clear that, until one is able to conduct the more detailed experiments, merely positing such an underlying deterministic structure makes no progress toward removing current ambiguities, and our inferences remain highly sensitive to our assumptions. In such a case there seems to be no obvious advantage in assuming determinism; one might just as well conduct analyses such as that of Example 3, basing them only on experimentally observed quantities and deriving suitably qualified inferences encompassing the remaining ambiguity—which should not be artificially eliminated by imposing unverified constraints on the model. (Nevertheless, it may be, as suggested in sec. 12.1, that thinking about the possibilities for what one might discover in further experiments could aid a reasonable and defensible resolution—subject to later empirical confirmation or refutation—of some of the ambiguities.)

13.2 Pseudodeterminism

It seems to me that behind the popularity of counterfactual models lies an implicit view that all problems of causal inference can be cast in the deterministic paradigm (which in my view is only rarely appropriate), for a suitable (generally unobserved) determining concomitant D. If so, this would serve to justify the assumption of simultaneous existence of complementary potential responses. Heckerman and Shachter (1995), for example, take a lead in this from Savage (1954), who based his axiomatic account of Bayesian decision theory on the supposed existence of a "state of nature," entirely unaffected by any decisions taken, which, together with those decisions, determines all variables. Shafer (1986) has pointed up some of the weaknesses of this conception.

The functional graphical model framework of Pearl (1995a) posits that underlying observed distributional stabilities of observed variables are functional relationships, involving the treatments and further latent variables. When such a deterministic structure can be taken seriously, with all its variables in principle observable, it leads to the possibility (at least) of well-defined counterfactual inferences, as described earlier. These will again, quite reasonably, be sensitive to the exact form of the functional relationships in-

volved, over and above any purely distributional properties of the manifest variables; but these functional relationships are in principle discoverable. Balke (1995) and Balke and Pearl (1994a) investigated the dependence of causal inferences on the functional assumptions.

However, often the "latent variables" involved in such models are not genuine concomitants (measurable variables, unaffected by treatment). Then there is no way, even in principle, of verifying the assumptions made—which will nevertheless affect the ensuing inferences, in defiance of Jeffreys's law. I term such functional models *pseudodeterministic* and regard it as misleading to base analyses on them. In particular, I regard it as unscientific to impose intrinsically unverifiable assumed forms for functional relationships, in a misguided attempt to eliminate the essential ambiguity in our inferences.

Within the counterfactual framework, it is always possible to construct, mathematically, a pseudodeterministic model: Simply define $D(u)$ to be the complementary collection of all potential outcomes on unit u. In Example 1 one would thus take $D = (Y_t, Y_c)$. One then has the trivial deterministic functional relationship $Y = f(i, D)$, where f has the *canonical form* $f(i, (y_t, y_c)) = y_i$ $(i = t, c)$. If a joint distribution were now assigned to (Y_t, Y_c), then the analysis presented earlier for inferring "causes of effects" in deterministic models could be formally applied.

This is not a true deterministic model: D is not a true concomitant, because it is not, even in principle, observable. Construction of such a pseudodeterministic model makes absolutely no headway toward addressing the nonuniqueness problems exposed in Sections 5.4 and 11; it remains the case that no amount of scientific investigation will suffice to justify any assumed dependence structure for (Y_t, Y_c), or eliminate the sensitivity to this of the inferences about causes of effects. This can be done only by taking into account genuine concomitants.

14. CONTEXT

In basing inference about the causes of effects on concomitant variables (as in Sec. 11.1), it appears that I am departing from my insistence that metaphysical assumptions should not be allowed to affect inferences. This is because to say that a variable is a concomitant involves an assertion that it is unaffected by treatment, and hence would take the same value, both in the real universe and in parallel counterfactual universes in which different treatments were applied. Such an assumption is clearly not empirically testable. Nevertheless, one's causal inferences will depend on the assumptions made as to which variables are to be treated as concomitants. This arbitrariness is over and above the essential inferential ambiguity that I have already identified, which remains even after the specification of concomitants has been made.

My attitude is that there is indeed an arbitrariness in the models that one can reasonably use to make inferences about causes of effects, and hence in the conclusions that are justified. But I would regard this as relating, at least in part, to differences in the nature of the questions being addressed. The essence of a specific causal inquiry is captured in the largely conventional specification of what may be termed the context of the inference—namely, the collection of variables one considers it appropriate to regard as concomitants; see Example 4. Appropriate specification of context, relevant to the specific purpose at hand, is vital to render causal questions and answers meaningful. It may be regarded as providing necessary clarification of the ceteris paribus ("other things being equal") clause often invoked in attempts to explicate the idea of cause. Differing purposes will demand differing specifications, requiring differing scientific and statistical approaches and yielding differing answers. In particular, whether it is reasonable to use a deterministic model must depend on the context of the problem at hand, as this will determine whether it is appropriate to regard a putative determining variable D as a genuine concomitant, unaffected by treatment. For varying contexts one might have varying models, some deterministic (involving varying definitions of D) and some nondeterministic.

Example 6. Consider an experiment in which the treatments are varieties of corn and the units are field plots. Suppose that variety 1 has been planted on a particular field plot, and its yield measured. One might ask "What would the yield have been on this plot if variety 2 had been planted?." Before this question can be addressed, it must be made more precise; and this can be done in various ways, depending on one's meaning and purpose.

First, the answer must depend in part on the treatment protocol. For example, this might lay down the weight, or alternatively the number, of seeds to be planted. In the former case, the counterfactual universe would be one in which the weight of variety 2 to be planted would the same as the weight of variety 1 actually planted; in the latter case, "weight" would need to be changed to "number," so specifying different counterfactual conditions and leading one to expect a different answer. (In either case the actual and counterfactual responses will depend in part on the particular seeds chosen, introducing an irreducibly random element into each universe.) One might choose to link the treatments in the two universes in further ways; for example, if one had happened to choose larger than average seeds of variety 1, then one might want to consider a counterfactual universe in which we also chose larger than average seeds of variety 2. This would correspond to a fictitious protocol in which the treatment conditions were still more closely defined.

The same counterfactual question might be asked by a farmer who had planted variety 1 in nonexperimental conditions. In this case there was no treatment protocol specified, and there is correspondingly still more freedom to specify the fictitious protocol linking the real and the counterfactual universe. But only when one has clearly specified one's hypothetical protocol can one begin to address the counterfactual query.

This done, one must decide what further variables one will regard as concomitants, unaffected by treatment. It might well be reasonable to include among these certain physical properties of the field plot at the time of planting,

and perhaps also the weather in its neighbourhood, subsequent to planting.

One might also want to take into account the effect of insect infestation on yield. It would probably not be reasonable to treat this as a concomitant, because different crops are differentially attractive to insects. Instead, one might use some specification of the abundance and whereabouts of the insects prior to planting. However, it would be simply unreasonable to expect this specification to be in any sense complete. Would one really want to consider the exact initial whereabouts and physical and mental states of all insects as identical in both the real and the counterfactual universe, and so link (though still far from perfectly) the insect infestations suffered in the two universes? If one did, then one would need a practically unattainable understanding of insect behaviour before one could formulate and interpret, let alone answer, the counterfactual query. Furthermore, to insist (perhaps in an attempt to justify a deterministic model) on fixing the common properties of the two universes at an extremely fine level of detail risks becoming embroiled in unfathomable arguments about determinism and free will. Would one really have been at liberty to apply a different treatment in such a closely determined alternative universe? To go down such a path seems to me to embark on a quest entirely inappropriate to any realistic interpretation of the query. Instead, one could imagine a counterfactual universe agreeing with the real one at a much less refined level of detail (in which initial insect positions are perhaps left unspecified). This corresponds to a broader view of the relevant context, with fewer variables considered constant across universes. It is up to the person asking the counterfactual query, or attempting causal inference, to be clear about the appropriate specification, explicit or implicit, of the relevant context.

The conditional independence convention further allows one to tailor counterfactual inferences to the appropriate context, as in Example 4, without embarking on fruitless searches for "ultimate causes." In Example 6, one may wish to omit from specification of context any information about, or relevant to, the population and behavior of the insects. One could then take the amounts of insect infestation, in the real and the counterfactual universes, as independent, conditionally on whatever concomitants are regarded as determining context. This choice may be regarded as making explicit one's decision to exclude insect information from the context, rather than as saying anything meaningful about the behavior of the world. With this understanding, the very meaning (and hence the unknown value) of the correlation ρ between Y_t and Y_c (or of any other measure of the dependence between such complementary quantities) will involve, in part, one's own specification of the context considered appropriate to the counterfactual questions.

The relation between the partly conventional specification of context and general scientific understanding is a subtle one. Certainly the latter should inform the former, even when it does not determine it; general scientific or intuitive understandings of meteorological processes must underlie any identification of the weather as a concomitant, unaffected by treatment. Moreover, it is always possible that further scientific understanding might lead to a refinement of what is regarded as the appropriate context; thus the discovery of genetics has enabled identification of previously unrecognized invariant features of an individual and thus discarding of previously adequate, but now superseded, causal theories. Causal inference is, even more than other forms of inductive inference, only tentative; causal models and inferences need to be revised, not only when theories and assumptions on which they are based cease to be tenable in the light of empirical data, but also when the specification of the relevant context has to be reformulated—be this due to changing scientific understanding or to changing requirements of the problem at hand.

15. CONCLUSION

I have argued that the counterfactual approach to causal inference is essentially metaphysical, and full of temptations to make "inferences" that cannot be justified on the basis of empirical data and are thus unscientific. An alternative approach based on decision analysis, naturally appealling and fully scientific, has been presented. This approach is completely satisfactory for addressing the problem of inference about the effects of causes, and the familiar "black box" approach of experimental statistics is perfectly adequate for this purpose.

However, inference about the causes of effects poses greater difficulties. A completely unambiguous solution can be obtained only in those rare cases where it is possible to reach a sufficient scientific understanding of the system under investigation as to allow the identification of essentially deterministic causal mechanisms (relating responses to interventions and concomitants, appropriately defined). When this is not achievable (whether the difficulties in doing so be fundamental or merely pragmatic), the inferences justified even by extensive data are not uniquely determined, and one must be satisfied with inequalities. However, these may be refined by modeling the relevant context and conducting experiments in which concomitants are measured. A major and detailed scientific study may be required to reduce the residual ambiguity to its minimal level (and, even then, there can be no prior guarantee that it will do so).

Thus, if one wants to make meaningful and useful assertions about the causes of effects, then one must be very clear about the meaning and context of one's queries. And then there is no magical statistical route that can bypass the need to do real science to attain the clearest possible understanding of the operation of relevant (typically nondeterministic) causal mechanisms.

[Received October 1997. Revised July 1999.]

REFERENCES

Bailey, R. A. (1991), "Strata for Randomized Experiments" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 53, 27–78.

Balke, A. A. (1995), "Probabilistic Counterfactuals: Semantics, Computation, and Applications," Ph.D. dissertation, University of California, Los Angeles, Dept. of Computer Science.

Balke, A. A., and Pearl, J. (1994a), "Probabilistic Evaluation of Counterfactual Queries," in *Proceedings of the Twelfth National Conference on*

Artificial Intelligence (AAAI-94), Seattle, Vol. I, pp. 230–237.
——— (1994b), "Counterfactual Probabilities: Computational Methods, Bounds and Applications," in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, eds. R. Lopez de Mantaras and D. Poole, San Mateo, CA: Morgan Kaufmann, pp. 46–54.
Cox, D. R. (1958), "The Interpretation of the Effects of Nonadditivity in the Latin Square," *Biometrika*, 45, 69–73.
Dawid, A. P. (1979), "Conditional Independence in Statistical Theory" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 41, 1–31.
——— (1984), "Present Position and Potential Developments: Some Personal Views. Statistical Theory. The Prequential Approach" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 147, 278–292.
——— (1985), "Calibration-Based Empirical Probability" (with discussion), *The Annals of Statistics*, 13, 1251–1285.
——— (1988), "Symmetry Models and Hypotheses for Structured Data Layouts" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 50, 1–34.
Greenland, S., Robins, J. M., and Pearl, J. (1999), "Confounding and Collapsibility in Causal Inference," *Statistical Science*, 14, 29–46.
Heckerman, D., and Shachter, R. (1995), "Decision-Theoretic Foundations for Causal Reasoning," *Journal of Artificial Intelligence Research*, 3, 405–430.
Hitchcock, C. (1997), "Causation, Probabilistic," in *Stanford Encyclopedia of Philosophy*, online at http://plato.stanford.edu/entries/causation-probabilistic/.
Holland, P. W. (1986), "Statistics and Causal Inference" (with discussion), *Journal of the American Statistical Association*, 81, 945–970.
Imbens, G. W., and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–476.
Imbens, G. W., and Rubin, D. B. (1997), "Bayesian Inference for Causal Effects in Randomized Experiments With Noncompliance," *The Annals of Statistics*, 25, 305–327.
Lewis, D. K. (1973), *Counterfactuals*, Oxford, U.K.: Blackwell.
McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.
Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles," *Roczniki Nauk Rolniczych*, X, 1–51 (in Polish), English translation of Section 9 by D. M. Dabrowska and T. P. Speed, (1990), *Statistical Science*, 9, 465–480.
——— (1935), "Statistical Problems in Agricultural Experimentation" (with discussion), *Supplement to Journal of the Royal Statistical Society*, 2, 107–180.
Pascal, B. (1669), *Pensées sur la Réligion, et sur Quelques Autres Sujets*, Paris: Guillaume Desprez. (Edition Garnier Frères 1964).
Pearl, J. (1993), "Aspects of Graphical Models Connected With Causality," *Proceedings of the 49th Session of the International Statistical Institute*, 391–401.
——— (1995a), "Causal Diagrams for Empirical Research" (with discussion), *Biometrika*, 82, 669–710.
——— (1995b), "Causal Inference From Indirect Experiments," *Artificial Intelligence in Medicine*, 7, 561–582.
Rachev, S. T. (1985), "The Monge–Kantorovich Mass Transference Problem and Its Stochastic Applications," *Theoretical Probability and its Applications*, 29, 647–671.
Raiffa, H. (1968), *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Reading, MA: Addison-Wesley.
Robins, J. M. (1986), "A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect," *Mathematical Modelling*, 7, 1393–1512.
——— (1987), Addendum to "A New Approach to Causal Inference in Mortality Studies With Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect," *Computers and Mathematics with Applications*, 14, 923–945.
Robins, J. M., and Greenland, S. (1989), "The Probability of Causation Under a Stochastic Model for Individual Risk," *Biometrics*, 45, 1125–1138.
Robins, J. M., and Wasserman, L. A. (1997), "Estimation of Effects of Sequential Treatments by Reparameterizing Directed Acyclic Graphs," Technical Report 654, Carnegie Mellon University, Dept. of Statistics.
Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.
——— (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–68.
——— (1980), Comment on "Randomization Analysis of Experimental Data: The Fisher Randomization Test" by D. Basu, *Journal of the American Statistical Association*, 81, 961–962.
——— (1986), "Which Ifs Have Causal Answers?" (comment on "Statistics and Causal Inference" by P. W. Holland), *Journal of the American Statistical Association*, 81, 961–962.
Rüschendorf, L., Schweizer, B., and Taylor, M. D. (Eds.) (1996), *Distributions with Fixed Marginals and Related Topics*, Institute of Mathematical Statistics Lecture Notes Monograph Series, Vol. 28, Hayward, CA: Institute of Mathematical Statistics.
Savage, L. J. (1954), *The Foundations of Statistics*, New York: Wiley.
Shafer, G. (1996), *The Art of Causal Conjecture*, Cambridge, MA: MIT Press.
——— (1986), "Savage Revisited" (with discussion), *Statistical Science*, 4, 463–501.
Wilk, M. B., and Kempthorne, O. (1955), "Fixed, Mixed, and Random Models," *Journal of the American Statistical Association*, 50, 1144–1167.
——— (1956), "Some Aspects of the Analysis of Factorial Experiments in a Completely Randomized Design," *Annals of Mathematical Statistics*, 27, 950–985.
——— (1957), "Nonadditivities in a Latin Square," *Journal of the American Statistical Association*, 52, 218–236.

Comment

D. R. Cox

I very much admire Professor Dawid's original, lucid, and penetrating discussion of causality. And yet: has the philosophical coherence, if not thrown the baby out with the bathwater, at least left the baby seriously bruised in some vital organs? Dawid's formulation of the purpose of causal discussion involves a decision about treatment allocation to a new individual. Most experiments with which I have been involved have as their purpose the gaining of some understanding of a phenomenon. This may lead eventually to recommendations on specific decisions but that comes later. The noun "understanding" is probably too vague for merciless philosophical discussion, and I realize that the decision making does not have to be taken too literally, but has something been lost in the decision-oriented formulation?

D. R. Cox, Department of Statistics and Nuffield College, Oxford, U.K.

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

Causality is thus from one viewpoint bound up with understanding, of course approximate and tentative, of an underlying generating process (Cox 1992; Cox and Wermuth 1996, pp. 219–227). (For a cogent discussion in a sociological context, see Goldthorpe 1999.) Dawid himself seems to edge toward this view in his final sentence. Causality is also connected with the notions of generalizability and specificity; that is, of the extent to which conclusions can be applied in new situations and of the extent to which an effect applies to a potential new individual and not just in some average sense. The latter point clearly connects with Dawid's discussion. Here the two crucial features are understanding of underlying process and the demonstration of absence of interaction with baseline features—the latter playing an interesting and important role also in Dawid's discussion.

The distinction between these two versions of causality (and there are others) can be seen as follows. Imagine that a careful experiment, preferably randomized, or even a whole series of such experiments shows that T produces a higher response than C but that there is no understanding, no matter how tentative, of why this is. Has causality been established? In one sense it has, and yet I believe that many working scientists would be uneasy using the term in such situations.

Dawid is rightly critical of naive interpretations of the assumption of unit-treatment additivity. But surely it is clear that average effects are all that can be estimated and the deterministic formulation is just a convenient simplification? At least some very applied accounts (Cox 1958, pp. 15–19) are absolutely explicit on this point.

It is hard to disagree with Dawid's distaste for assumptions that can never be tested even in principle, and his distinction between sheep and goats is valuable. Yet at a work-a-day level, the point is more that any assumptions should not be pressed too far beyond the limits to which they can be tested and, importantly, that assumptions can be tested indirectly via their consequences as well as directly. Thus unit–treatment additivity also implies that the distributions in different treatment groups differ only by translation. In particular, they have equal variance—a much more important reason for being interested in equality of variance than possible effects on tests of significance or other factors. Further, it can be regarded as a major limitation of the Popperian viewpoint that it gives no account of how the ideas to be subject to test are to be obtained. In this, not directly testable assumptions may play a vital role and goats and sheep may interact fruitfully; think of, for example, the role of genes in classical genetics and of atoms in eighteenth or nineteenth century physics.

I am a bit puzzled by the sharp distinction drawn in the discussion of effects of causes and causes of effects, although I see why it is needed in Dawid's formulation. It would be widely agreed that the interpretation of retrospective studies tends to be more hazardous than that of corresponding prospective studies. But this is partly because of the greater possibility of measurement biases in recording past events and partly because there is often a lack of clarity about the definition of an appropriate control group. These do not seem to figure in the present discussion. Is the second point related in some way to Dawid's account?

Finally, I repeat that I learned much from the article, which is an important contribution to an important topic.

ADDITIONAL REFERENCES

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

——— (1992), "Causality: Some Statistical Aspects," *Journal of the Royal Statistical Society*, Ser. A, 155, 291–301.

Cox, D. R., and Wermuth, N. (1996), *Multivariate Dependencies*, London: Chapman and Hall.

Goldthorpe, J. (1999), *Causation, Statistics and Sociology*, Dublin: Economic and Social Research Institute.

Comment

George CASELLA and Stephen P. SCHWARTZ

1. INTRODUCTION

Professor Dawid has presented a thought-provoking analysis of causal inference, and has certainly caused us to think hard about these matters. We comment on three main topics: the structure of models, the object of inference, and the philosophy of inference.

The desire to make a causal inference leads one to a particular class of models. From the model (and the data), an inference need be made. The model, and an associated parameter of interest, directs the possible type of inference. We then must decide on a reference set (or population) to which the inference will be made. All of these pieces work together in an inferential philosophy. There are many choices to be made at each stage of the process (model, parameters, inference). Dawid insists that such choices, and

George Casella is Liberty Hyde Bailey Professor of Biological Statistics, Department of Biometrics, Cornell University, Ithaca, NY 0000 (E-mail: *qc15@cornell.edu*). Stephen P. Schwartz is Professor, Department of Philosophy and Religion, Ithaca College, Ithaca, NY 0000. This is technical report BU-1452-M in the Department of Biometrics, Cornell University. The research was supported by National Science Foundation grant DMS-9971586.

inferences, must be based on strict principles that can be verified empirically. We believe that such a program is so overly rigid that, in the end, science is not served.

2. INFERRING FROM ...

An individual counterfactual model necessarily involves unobservable quantities. These quantities can lead to *unidentifiable* models such as (1). When faced with such a model, we would normally think that the statistician would try to refine the model to make valid *individual causal effects* inferences possible, if such inferences are the desire of the experimenter.

One way of doing this is to shift the target of inference to *average causal effects*. Then (8), which is free of nonidentifiability baggage, can be used. To us, this is a way out of the problems inherent in (1). Let us look at this switch of inferential target a bit more closely.

To switch, one is forced to place assumptions on the structure of the parameters, thus bringing in a "metaphysical component" that Dawid finds so distasteful. But, in a sense, this is a reality of inference. When faced with an unwieldy model, we must make assumptions to obtain usable inferences.

One assumption that results in average causal effects becoming the inferential target is that of *treatment unit additivity* (TUA), which Dawid does not like. But there is another road to average causal effect, based on the thinking of the eighteenth century philosopher David Hume (1748):

> It appears, then, that this idea of a necessary connection among events arises from a number of similar instances which occur, of the constant conjunction of these events; nor can that idea ever be suggested by any one of these instances surveyed in all possible lights and positions.

Following Hume, causal inference is necessarily shifted from the individual to the group. This eliminates any counterfactual problems because, at the group level, the counterfactual is observable (one group did not get aspirin).

Dawid's insistence on empirical verification would reject the foregoing line of reasoning. Such an insistence not only severely restricts the range of possible models, but also may disregard the scientific input of the subject matter expert (who may insist that TUA is entirely plausible for the experiment at hand).

A crucial point is that if we can reduce the inferential target to one based only on marginal distributions, then we can provide a reasonable inference (to us, this means that we are working with an identifiable model). As Dawid rejects TUA (and presumably the argument based on Hume) as metaphysical, he applies Bayesian decision theory to reduce the inference to a marginal one. However, in doing so, he has substantially changed the inferential target. The primary target of inference is $Y_t(u) - Y_c(u)$, the individual difference, which is unobservable. Using either TUA or Hume, this target becomes $\theta_t - \theta_u$, the average causal effect. Dawid's decision theory argument leads to the inferential target being $u_0|\text{treatment} = t$, the distribution of the response given that the treatment was t. Although this may be a reasonable target of inference, it may not be the one that the experimenter cares about.

3. INFERRING TO ...

Whatever the chosen target of inference, an inference must be drawn. Some of us think in terms of populations or reference sets, often described by the experimenter. For example, in Dawid's Example 6, we can specify a number of reference sets. We have usually left the choice of such to the experimenter, whose greater subject matter knowledge can be used to choose the appropriate frame of inference.

3.1 Empirical Verification

Relying strictly on empirical verification, Dawid deals with the shortcomings of a model like (1) by invoking a principle known as *Jeffreys's law* to decree what types of inferences are allowed from nonidentifiable models.

Jeffreys's law is the *likelihood principle* in another guise. The likelihood principle states that if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that the likelihood $L(\theta|\mathbf{x})$ is proportional to $L(\theta|\mathbf{y})$ for all θ, then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical.

Since the landmark work of Birnbaum (1962), the likelihood principle has been the focus of much debate. It is probably fair to say that with the exception of the strictest Bayesian, most statistical practice violates the likelihood principle. Why this is so is perhaps best explained by Berger and Wolpert (1984):

> We emphatically believe that the LP (likelihood principle) is always valid, in the sense that the experimental evidence concerning θ *is* contained in $l_X(\theta)$ (the likelihood function). Because of limited time and resources, however, interpreting or making use of this evidence *may* involve use of measures violating the LP.

This sentiment may be closest to what most statisticians feel. There are compelling arguments for embracing the likelihood principle, but in reality, we need to go beyond it. We must use, among other things, metaphysical assumptions to thoroughly evaluate an inference.

To adhere to empirical verification and the limitations imposed on inferences by Jeffreys's law leads inexorably to a Popperian view, as Dawid explains:

> My approach is grounded in a Popperian philosophy, in which meaningfulness of a purportedly scientific theory, proposition, quantity, or concept is related to the implications it has for what is or could be observed and, in particular, to the extent to which it is possible to conceive of data that would be affected by the truth of the proposition or the value of the quantity. When this is the case, assertions are empirically refutable and considered "scientific." When not so, they may be branded "metaphysical."

However, this view is based on a philosophical orientation that is outmoded and has been rejected by virtually all mainstream philosophers of science.

3.2 Popper is Out

The "Popperian" philosophy that grounds Dawid's approach was part of the much larger *logical positivist* philosophical movement that had great currency up to perhaps 40 years ago. Logical positivism's main tenet is that meaningful propositions must be either analytic (mathematical)

or empirically falsifiable or verifiable by possible sensory observations. Karl Popper emphasized falsifiability and, for example, famously directed an attack against Marxism, arguing that it was unscientific and just a matter of faith. To that extent, positivism served a useful purpose, helping rid the intellectual arena of much philosophical and pseudoscientific dross. It was like a breath of fresh air. Logical positivism has also been influential in science. For example, behaviorism is based on the idea that we can observe behavior but cannot directly observe other people's minds. Therefore, behavior, but not the mind, is a fit subject of scientific study.

Starting in about 1950, logical positivism was subjected to a withering series of criticisms and has now entirely lost favor among philosophers. The attack was based primarily on the logical work of W. V. Quine (1961) and the historical work of Thomas Kuhn (1970), with much help from many other thinkers and researchers. The criticisms demonstrated that the logical positivist program was too rigid and technically unworkable and that logical positivism did not represent the actual practice of scientists. If held to the rigid standard of Popperian philosophy, then little or no actual science would get done. The demise of logical positivism has had the beneficial effect of expanding the horizons of scientific pioneers. For example, cognitive science has now replaced behaviorism as the leading orientation in psychology.

3.3 Counterfactuals are In

Among the many technical problems facing logical positivists was what to do about counterfactuals. Certainly, many counterfactuals are unverifiable and do not seem to be scientifically meaningful. For example, "If I had been born in China, I would now be able to speak Chinese." On the other hand, many other counterfactuals clearly seem to be meaningful and indeed true; for example, "If Nixon had not resigned, he would have been impeached." The fact is that counterfactuals are indispensable in many areas, but attempts to analyze them in terms of direct observation foundered. The problem of how to understand them is still a matter of philosophical controversy. Probably the most widely accepted view today is that of David Lewis cited by Dawid. Lewis analyzes counterfactuals in terms of other possible worlds, ways that things could have been but are not—anathema to the logical positivists and Dawid.

3.4 A Fatal Flaw?

Dawid's use of tendentious vocabulary clouds his argument and obscures the motivation for his views. For example, besides the questionable empirical versus metaphysical distinction, Dawid rejects a view he terms "fatalism":

> Many counterfactual analyses are based, explicitly or implicitly, on an attitude that I term fatalism. This conceives of the various potential responses $Y_i(u)$, when treatment i is applied to unit u, as predetermined attributes of unit u, waiting only to be uncovered by suitable experimentation. (It is implicit that the unit u, and its properties and propensities, exist independently of, and are unaffected by, any treatment that may be applied.)

If by this Dawid means that the world and its objects exist independently of our attempts to know them, then this view is quite respectable and usually goes under the rubric "realism." And it seems that even Dawid sometimes embraces a "fatalist" view, as he says, "Nature is surely utterly indifferent to our attempts to ensnare her in our theories." "Fatalism" seems to be a highly misleading name for a rather commonplace and obvious idea. If Dawid means something else by his use of "fatalism," then we fear that he is attacking a "straw man" view that no one holds.

4. AND FINALLY ...

Clearly, there is something right about the positivist approach in general. Certainly we want our scientific theories to be verifiable or falsifiable in some sense, but it turns out that verifiability and falsifiability are much more flexible, elastic, and looser notions than the logical positivists supposed. The upshot is that we need to take a more tolerant approach to verification and falsification and abandon the kind of tendentious and rigid distinctions that the logical positivists, and following them Dawid, use. Scientific theories are not verified or falsified by direct observation or crucial experiment, except in very rare instances. Rather, theories are accepted or rejected by scientists on the basis of how well they explain selected sets of data, how elegant, simple, and useful they are, how well they do against competing theories, and so on. In fact, in his discussion of Barndorff-Nielsen's paper, Dawid (1976) expressed a similar sentiment when he said (our italics):

> A constant theme in the development of statistics has been the search for justification for what statisticians do. To read the textbooks, one might easily get the distorted idea that "Student" proposed his t test because it was the uniformly most powerful test of a normal mean, but it would be more accurate to say that the concept of UMPU gains much of its appeal because it produces the t test, and *everyone knows the t test is a good thing.*

Everyone knows that the simple, elegant, and useful t test is a good thing because it has performed admirably for almost 100 years. In the interest of science, performance counts for more than rigid adherence to philosophical principles.

ADDITIONAL REFERENCES

Berger, J. O., and Wolpert, R. W. (1988), *The Likelihood Principle* (2nd ed.), Hayward, CA: Institute of Mathematical Statistics.

Birnbaum, A. (1962), "On the Foundations of Statistical Inference" (with discussion), *Journal of the American Statistical Association*, 57, 269–306.

Dawid, A. P. (1976), "Discussion of the Paper by Barndorff-Nielsen," *Journal of the Royal Statistical Society*, Ser. B, 38, 123–125.

Hume, D. (1748), *An Inquiry Concerning Human Understanding*. C. W. Hendel (Ed.) (1955), Indianapolis: Bobbs-Merrill.

Kuhn, T. S. (1970), *The Structure of Scientific Revolutions* (2nd ed.), Chicago: University of Chicago Press.

Quine, W. V. (1961), *Two Dogmas of Empiricism*, reprinted in *From a Logical Point of View*, New York: Harper Torchbooks.

Comment

Judea PEARL

1. BACKGROUND

The field of statistics has seen many well-meaning crusades against threats from metaphysics and other heresy. In its founding prospectus of 1834, the Royal Statistical Society resolved "to exclude carefully all Opinions from its transactions and publications—to confine its attention rigorously to facts." This clause was officially struck out in 1858, when it became obvious that facts void of theory could not take statistics very far (*Annals of the Royal Statistical Society* 1934, p. 16).

Karl Pearson launched his own metaphysics "red scare" about causality in 1911: "Beyond such discarded fundamentals as 'matter' and 'force' lies still another fetish amidst the inscrutable arcana of modern science, namely, the category of cause and effect" (Pearson 1911, p. iv). Pearson's objection to theoretical concepts such as "matter" and "force" was so fierce and his rejection of determinism so absolute that he consigned statistics to almost a century of neglect within the study of causal inference. Philip Dawid was one of a handful of statisticians who boldly protested the stalemate over causality: "Causal inference is one of the most important, most subtle, and most neglected of all the problems of statistics" (Dawid 1979).

In the past two decades, owing largely to progress in counterfactual, graphical, and structural analyses, causality has been transformed into a mathematical theory with well-defined semantics and well-founded logic, and many practical problems that were long regarded as either metaphysical or unmanageable can now be solved using elementary mathematics. (See Pearl 2000 for a gentle introduction to the counterfactual, graphical, and structural equation approaches to causality.) In the article, Professor Dawid welcomes the new progress in causal analysis but expresses mistrust of the quasi-deterministic methods by which this progress has been achieved.

Attitudes of suspicion toward counterfactuals and structural equation models are currently pervasive among statisticians, and Dawid should be commended for bringing such concerns into the open. By helping to dispel misconceptions about counterfactuals, Dawid's article may well have rescued statistics from another century of stagnation over causality.

2. THE EMPIRICAL CONTENT OF COUNTERFACTUALS

The word "counterfactual" is a misnomer. Counterfactuals carry as clear an empirical message as any scientific laws, and indeed are fundamental to them. The essence of any scientific law lies in the claim that certain relationships among observable variables remain invariant when the values of those variables change relative to our immediate observations. For example, Ohm's law ($V = IR$) asserts that the ratio between the current (I) and the voltage (V) across a resistor remains constant for all values of I, including yet-unobserved values of I. We usually express this claim in a function or a hypothetical sentence: "Had the current in the resistor been I (instead of the observed value I_0) the voltage would have been $V = I(V_0/I_0)$," knowing perfectly well that there is no way to simultaneously measure I and I_0. (Every mathematical function is interpreted hypothetically, and the study of counterfactuals is merely a study of standard mathematical functions.) Such sentences appear to be *counterfactual,* because they deal with unobserved quantities that differ from (and hence seem to contradict) those actually observed. Nonetheless, this circumstantial nonobservability and apparent contradiction do not diminish whatsoever the ability to submit physical laws to empirical tests. Scientific methods thrive on attempts to confirm or falsify the predictions of such laws.

The same applies to stochastic processes (or data-generation models), usually written in the form of functional relations $y = f(x, u)$, where X and U stand for two sets of random variables, with joint distribution $P(x, u)$, and f is a function (usually of unknown form) that determines the value of the outcome $Y = y$ in terms of observed and unobserved quantities, $X = x$ and $U = u$. To see how counterfactuals and joint probabilities of counterfactuals emerge from such a stochastic model, I consider a simple case where Y and X are binary variables (e.g., treatment and response) and U is an arbitrary complex set of all other variables that may influence Y. For any given condition $U = u$, the relationship between X and Y must be one of the (only) four binary functions

$$\begin{aligned} f_0&: y = 0 \text{ or } \{Y_0 = 0, Y_1 = 0\},\\ f_1&: y = x \text{ or } \{Y_0 = 0, Y_1 = 1\},\\ f_2&: y \neq x \text{ or } \{Y_0 = 1, Y_1 = 0\}, \end{aligned}$$

and

$$f_3: y = 1 \text{ or } \{Y_0 = 1, Y_1 = 1\}. \quad (1)$$

As u varies along its domain, the only effect it can have on the model is to switch the relationship between X and Y among these four functions. This partitions the domain of U into four equivalence classes, where each class contains those points u that correspond to the same function. The probability $P(u)$ thus induces a probability function over the potential response pairs $\{Y_0, Y_1\}$ shown in (1). This construction is the inverse of the one discussed in Dawid's Section 13; one starts with genuine concomitants U, and

Judea Pearl is Professor of Computer Science and Statistics, University of California, Los Angeles, CA 90024 (E-mail: *judea@cs.ucla.edu*, Web: *www.cs.ucla.edu/~judea/*).

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

they turn into jointly distributed counterfactual concomitants $\{Y_0, Y_1\}$ that Dawid calls metaphysical and fatalistic.

Admittedly, when u stands as the identity of a person, the mapping of u into the pair $\{Y_0, Y_1\}$ appears horridly fatalistic, as if that person is somehow doomed to react in a predetermined way to treatment $(X = 1)$ and no treatment $(X = 0)$. However, if one views u as the sum total of all experimental conditions that might possibly affect that individual's reaction (including biological, psychological, and spiritual factors, operating both before and after the application of the treatment), then the mapping is seen to evolve reasonably and naturally from the functional model $y = f(x, u)$. This quasi-deterministic functional model mirrors Laplace's conception of nature (Laplace 1814), according to which of nature's laws are deterministic, and randomness surfaces merely due to our ignorance of the underlying boundary conditions. (The structural equation models used in economics, biology, and stochastic control are typical examples of Laplacian models.) Dawid detests this conception. This is not because it ever failed to match macroscopic empirical data (only quantum mechanical phenomena exhibit associations that might conflict with the Laplacian model), but rather because it appears to stand contrary to the "familiar statistical framework and machinery" (Sec. 7). I fail to see why a framework and machinery that did not exactly excel in the causal arena should be deprived of enhancement and retooling.

3. EMPIRICISM VERSUS IDENTIFIABILITY

Dawid's empiricism is summarized in his abstract:

> By definition, one can never observe such [counterfactual] quantities, nor assess empirically the validity of any modeling assumption made about them, even though one's conclusions may be sensitive to these assumptions.

This warning is not entirely accurate. Many counterfactual modeling assumptions do have testable implications; for example, exogeneity (or ignorability) $(Y_1 \perp\!\!\!\perp X)$ and monotonicity $(Y_1(u) \geq Y_0(u))$ each can be falsified by comparing experimental and nonexperimental data (Pearl 2000, p. 294). More important, the warning is either empty or self-contradictory. If one's conclusions have no practical consequences, then their sensitivity to invalid assumptions is totally harmless, and Dawid's warning is empty. If, on the other hand, one's conclusions do have practical consequences, then their sensitivity to assumptions automatically makes those assumptions testable, and Dawid's warning turns contradictory.

The two queries about aspirin and headache, which Dawid uses to distinguish effects of causes from causes of effects ("sheep" from "goats"), may serve well to illustrate the inconsistency in Dawid's philosophy. The two queries are

I. I have a headache. Will it help if I take aspirin?

II. My headache has gone. Is it because I took aspirin?

Letting $X = 1$ stand for "taking aspirin" and and $Y = 1$ stand for "having a headache" (after 1/2 hour, say), the counterfactual expressions for the probabilities of these two queries read:

$$Q_{\mathrm{I}} = P\ (Y_1 = 0) - P\ (Y_0 = 0)$$

and

$$Q_{\mathrm{II}} = P\ (Y_0 = 1 | X = 1, Y = 0). \tag{2}$$

In words, Q_{II} stands for the probability that my headache would have stayed had I not taken aspirin $(Y_0 = 1)$, given that I did in fact take aspirin $(X = 1)$ and the headache has gone $(Y = 0)$. (I restrict the population to persons who have headaches prior to considering aspirin.) Dawid is correct in stating that the two queries are of different types, and the language of counterfactuals displays this difference and its ramifications in vivid mathematical form. By examining their respective formulas, one can immediately detect that Q_{II} is conditioned on the outcome $Y = 0$, whereas Q_{I} is unconditioned. This implies that some knowledge of the functional relationship (between X and Y) must be invoked in estimating Q_{II} (Balke and Pearl 1994). I challenge Dawid to express Q_{II}, let alone formulate conditions for its estimation in a counterfactual-free language. For background information, the identification of Q_{I} requires exogeneity (i.e., randomized treatment), whereas that of Q_{II} requires both exogeneity and monotonicity; both assumptions have testable implications (Pearl 2000, p. 294). Epidemiologists are well aware of the difference between Q_{I} and Q_{II} [they usually write $Q_{\mathrm{II}} = Q_{\mathrm{I}}/P\ (Y = 0|X = 1)$], though the corresponding identification conditions for Q_{II} are often not spelled out as clearly as they could (Greenland and Robins 1988).

What is puzzling in Dawid's article is that he considers Q_{II} to be, on one hand, valid and important (Sec. 3) and, on the other hand, untestable (Sec. 11); the two are irreconcilable. If Q_{II} is valid and important, then one should expect the magnitude of Q_{II} to affect some future decisions, and can then use the correctness of those decisions as a test (hence interpretation) of the empirical claims made by Q_{II}. What are those claims, and how can they be tested?

According to the interpretation given in the previous section, counterfactual claims are merely conversational shorthand for scientific predictions. Hence Q_{II} stands for the probability that a person will benefit from taking aspirin in the *next* headache episode, given that aspirin proved effective for that person in the past (i.e., $X = 1, Y = 0$). Therefore, Q_{II} is testable in sequential experiments where subjects' reactions to aspirin are monitored repeatedly over time. (One needs to assume that a person's characteristics do not change over time, an assumption that is testable in principle.) In such tests one can easily verify whether subjects who have had one positive experience with aspirin $(X = 1, Y = 0)$ have a higher than average probability of benefiting from aspirin in the future.

I have argued elsewhere (Pearl 2000, p. 217) that counterfactual queries of type II are the norm in practical decision making, whereas causal effect queries (type I) are the exception. The reason is that decision-related queries are usually brought into focus by observations that could be modified by the decision (e.g., a patient suffering from a set of symp-

toms). The case-specific information provided by those observations is essential for properly assessing the effect of the decision, and conditioning on these observations leads to queries of type II, as in Q_{II}. The Bayesian approach proposed by Dawid cannot properly handle conditioning on factors that are affected by the treatment, and thus precludes answering the most common type of decision-related queries. (Detailed dynamic models or temporally indexed data for every conceivable set of observations would be needed for specifying the probabilities in the decision trees of such analyses.)

I agree with Dawid that certain assumptions needed for identifying causal quantities are not easily understood (let alone ascertained) when phrased in counterfactual terms. Typical examples are assumptions of ignorability (Rosenbaum and Rubin 1983), which involve conditional independencies among counterfactual variables. However, this cognitive difficulty comes not because counterfactuals are untestable, but rather because dependencies among counterfactuals are derived quantities that are a few steps removed from the way we conceptualize cause-effect relationships. To overcome this difficulty, a hybrid form of analysis can be used, in which assumptions are expressed in the friendly form of functional relationships (or diagrams), and causal queries (e.g., Q_{II}) are posed and evaluated in counterfactual vocabulary (Pearl 2000, p. 215–7, 231–4). Functional models, in the form of nonparametric structural equations, thus provide both formal semantics and conceptual basis for a complete mathematical theory of counterfactuals.

In Section 5.4, Dawid restates his empiricist philosophy in the form of a requirement which he calls *Jeffreys's law:*

> ... mathematically distinct models that cannot be distinguished on the basis of empirical observation should lead to indistinguishable inference.

This requirement reads like a tautology: If two models entail two distinguishable inferences, and if the difference between the two inferences matters at all, then the two models can easily be distinguished by whatever (empirical) criterion used to distinguish the two inferences. Dawid may have meant the following:

> ... mathematically distinct models that cannot be distinguished on the basis of past empirical observation should lead to indistinguishable inference regarding future observation (which may be obtained under new experimental conditions).

This is none other but the requirement of identifiability (see, e.g., Pearl 1995). It requires, for example, that if our data are nonexperimental, then two models that are indistinguishable on the basis of those data entail the same value of the average causal effect (ACE)—a quantity discernible in experimental studies. It likewise requires that if one's data come from static experiments, then two models that are indistinguishable on the basis of those data entail the same value of Q_{II}—a quantity discernible in sequential experiments.

If the aim of Dawid's empiricism is to safeguard identifiability, his proposal would be welcome by all causal analysts, including adventurous counterfactualists. Unfortunately, careful reading of his article shows that David aims to impose an overly restrictive and unworkable type of safeguard, a type rejected in almost every branch of science.

4. PRAGMATIC VERSUS DOGMATIC EMPIRICISM

The requirement of identifiability, as just stated, is a restriction on the type of queries one may ask (or inferences one may make) and not on the type of models one may use. This brings up the difference between pragmatic and dogmatic empiricism. A pragmatic empiricist insists on asking empirically testable queries, but leaves the choice of theories to convenience and imagination; the dogmatic empiricist insists on positing only theories that are expressible in empirically testable vocabulary. As an extreme example, a strictly dogmatic empiricist would shun the use of negative numbers, because negative quantities are not observable in isolation. For a less extreme example, a pragmatic empiricist would welcome the counterfactual model of individual causal effects (ICE) (see Sec. 5.2) as long as it leads to valid and empirically testable estimation of the quantity of interest (e.g., ACE). Dawid rejects this model a priori because it starts with unobservable unit-based counterfactual terms, $Y_1(u)$ and $Y_0(u)$, and thus fails the dogmatic requirement that the entire analysis, including all auxiliary symbols and all intermediate steps, "involve only terms subject to empirical scrutiny." What is gained by this prohibition, according to Dawid, is protection from asking nonidentifiable queries. His proposal, in the form of Bayesian decision trees, indeed ensures that one does not ask certain forbidden questions, but unfortunately, it also ensures that one never asks or answers important questions (such as Q_{II}) that cannot be expressed in his restricted language. It is a stifling insurance policy, analogous to banning division from arithmetics to protect one from dividing by 0. (Overprotection may also tempt the counterfactual camp; see Imbens and Rubin 1995.)

Science rejected this kind of insurance long ago. The Babylonians astronomers were masters of black box prediction, far surpassing their Greek rivals in accuracy and consistency (Toulmin 1961, pp. 27–30). Yet science favored the creative-speculative strategy of the Greek astronomers, which was wild with metaphysical imagery: circular tubes full of fire, small holes through which the fire was visible as stars, and hemispherical earth riding on turtle backs. It was this wild modeling strategy, not Babylonian rigidity, that jolted Eratosthenes (276–194 B.C.) to perform one of the most creative experiments in the ancient world and measure the radius of the earth.

This creative speculate-test-reject strategy (which is my understanding of Popperian empiricism) is practiced throughout science because it aims at understanding the mechanisms behind the observations and thus gives rise to new questions and new experiments, which eventually yield predictions under novel sets of conditions. Quantum mechanics was invented precisely because J. J. Thomson and others dared take deterministic classical mechanics very seriously, and boldly asked "metaphysical" questions about physical properties of electrons when electrons were un-

observable. The language of counterfactuals likewise enables the statistician to pose and reject a much richer set of "what if" questions than does the language of Bayesian decision theory. Giving up this richness is the price to pay for Dawid's insurance.

5. COUNTERFACTUALS AS INSTRUMENTS

Dawid reports (at the end of Sec. 10.2) that the bounds for causal effects in clinical trials with imperfect compliance (Balke and Pearl 1997) are "sheep-like"—namely valid, meaningful, and safe even for counterfactually averse statisticians. Ironically, when we examine the conditional probabilities that achieve those bounds, we find that they represent subjects with deterministic behavior, *compliers, never-takers,* and *defiers,* precisely the kind of behavior that Dawid rejects as "fatalistic" (Sec. 7.1). The lesson is illuminating: Even starting with the best sheep-like intentions, there is no escape from counterfactuals and goat-like determinism in causal analysis.

This lesson leads to a new way of legitimizing counterfactual analysis in the conservative circles of statistics. Researchers who mistrust the quasi-deterministic models of Laplace (i.e., $y = f(x,u)$) can now view these models as limit points of a space of nondeterministic models $P(y|x)$ constrained to agree with the observed data. Accordingly, the mistrustful analysis of counterfactuals can now be viewed as a benign analysis of limit points of ordinary probability spaces, in much the same way that irrational numbers can be viewed as limit points (or Dedekind cuts) of benign sets of rational numbers.

Dawid is correct in noting that many problems about the effects of causes can be reinterpreted and solved in noncounterfactual terms. Analogously, some of my colleagues can derive De-Moivre's theorem, $\cos n\theta = \mathrm{Re}[(\cos\theta + i\sin\theta)^n]$, without the use of those mistrustful imaginary numbers. So, should we strike complex analysis from our math books? Examining the major tangible results in causal inference in the past two decades (e.g., propensity scores, identification conditions, covariate selection, asymptotic bounds) reveals that, although these results *could* have been derived without counterfactuals, they simply *were not.* This may not be taken as a coincidence if one asks why it was Eratosthenes that measured the size of the earth and not some Babylonian astronomer, master in black box prediction. The success of the counterfactual language stems from two ingredients necessary for scientific progress in general: (a) the use of modeling languages that are somewhat richer than the ones needed for routine predictions, and (b) the use of powerful mathematics to filter, rather than muzzle, the untestable queries that such languages tempt us to ask.

Dawid is inviting causality to submit to the Babylonian safeguard of black box mentality. I dare predict that causality will reject his offer.

ADDITIONAL REFERENCES

Annals of the Royal Statistical Society 1834–1934 (1934), The Royal Statistical Society, London, p. 16.

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1172–1176.

Greenland, S., and Robins, J. (1988), "Conceptual Problems in the Definition and Interpretation of Attributable Fractions," *American Journal of Epidemiology*, 128, 1185–1197.

Imbens, G. W., and Rubin, D. R. (1995), Discussion of "Causal Diagrams for Empirical Research" by J. Pearl, *Biometrika*, 82, 694–695.

Laplace, P. S. (1814), *Essai Philosophique sure les Probabilites*, New York: Courcier, English translation by F. W. Truscott and F. L. Emory, New York: Wiley, 1902.

Pearl, J. (2000), *Causality*, New York: Cambridge University Press.

Pearson, K. (1911), *Grammar of Science*, (3rd ed.), London: A. and C. Black.

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of Propensity Score in Observational Studies for Causal Effects," *Biometrica*, 70, 41–55.

Toulmin, S. (1961), *Forecast and Understanding*, Bloomington, IN: Indiana University Press.

Comment

James M. ROBINS and Sander GREENLAND

By narrowly concentrating on randomized experiments with complete compliance, Dawid, in our opinion, incorrectly concludes that an approach to causal inference based on "decision analysis" and free of counterfactuals is completely satisfactory for addressing the problem of inference about the effects of causes. We argue that when attempting to estimate the effects of causes in observational studies or in randomized experiments with noncompliance (termed broken experiments by Barnard et al. (1998), reliance on counterfactuals or their logical equivalents cannot be avoided.

Causal inference from observational data and broken experiments historically has been viewed as problematic, and even illegitimate, by most statisticians. Thus we regard it as a serious oversight for Dawid to deny the usefulness of a counterfactuals without a more careful consideration of observational studies and broken experiments. The purpose of this discussion is to redress that oversight, by reviewing the considerations that have led

James M. Robins is Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. Sander Greenland is Professor of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095.

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

so many to adopt a counterfactual approach to causal inference.

1. THE PROBLEM OF CONFOUNDING

Suppose that we have discrete pretreatment covariates A and B in an observational study. At each level of A, suppose that treatment T taking values in $\{t, c\}$ is positively correlated with a disease outcome Y, but at each joint level AB of A and B, treatment is independent of outcome. In the language of the school of probabilistic causality (PC), AB screens off T from Y (Suppes 1970). Some PC texts would then say that treatment does not probabilistically cause Y relative to the causal field determined by A and B. However, this statement does not reflect the common language and appropriate policy meaning of a cause, which is that manipulating T would change Y. Indeed, there is a potential for an infinite regress wherein the association of T and Y varies among positive, negative, and null as one adjusts for additional covariates.

In epidemiology, it has been common to view the association adjusted for all measured pretreatment covariates as most likely to be causal. But Greenland and Robins (1986) noted that additional adjustment can increase confounding, in that the more adjusted association could be further from the true average causal effect than the less adjusted association. This problem has also been noted in the PC literature. As a result, the most sophisticated PC texts state that an adjusted effect is guaranteed to have a causal interpretation only when one has succeeded in adjusting for all nontreatment causes X of the outcome. It then follows as a theorem that the association of treatment and the outcome within levels of the measured covariates, say W, has a causal interpretation if either (a) the other elements $X \setminus W$ of X are independent of T given W or (b) $X \setminus W$ is independent of Y given W and T (Robins and Morgenstern 1987).

Unfortunately, these sufficient "conditions for no confounding" are never empirically testable from observational data, because by definition X contains all nontreatment causes, including those unmeasured and those not even known to exist. Hence the question of the existence and magnitude of confounding by the unmeasured factors $X \setminus W$ in an observational study is metaphysical in Dawid's sense, even under his preferred PC theory. It follows that causal inference from observational data is a Dawidian goat. In more standard statistical parlance, the average causal effect of a treatment is not identified from observational data without making nonidentifiable assumptions about the magnitude and direction of confounding. We are confident that Dawid does not wish to join R. A. Fisher (1959) in thereby concluding that causal inferences from observational data are illegitimate, including the inference that cigarette smoking is a cause of lung cancer (Stolley 1991). If we are correct, then Dawid has no choice but to recognize the need for untestable assumptions.

In an attempt to stave off the need for untestable assumptions, some commentators have argued that one should consider as potential confounders only those (often few) variables for which one can make a plausible case that they are common causes of treatment and the outcome. We find this reasoning unacceptable. Not only does it make confounding a property of the mind rather than of the physical world, but it rewards ignorance. The less one knows about possible causes, the freer one is to make definitive causal statements. The price of this freedom is that more, if not most, of these statements will be false.

1.1 Counterfactuals

As Dawid recognizes, in a deterministic (i.e., Newtonian or Laplacian) world with a single time-independent treatment T, the "all causes" approach to causal inference implies the existence of counterfactuals: If the world is deterministic (i.e., fatalistic) and $X = X(u)$ includes all nontreatment causes for subject u, then the outcome must be a deterministic function $f(i, X(u))$ of $X(u)$ and the treatment $i \in \{t, c\}$. We can then define the counterfactual $Y_i = Y_i(u)$ to be $f(i, X(u))$. In the general case with time-varying treatments, covariates, and outcomes, Robins (1995a, 1997) proved that Pearl's "all causes" nonparametric structural equation model is mathematically equivalent to a special case of the general counterfactual causal model of Robins (1986, 1987) (see also Galles and Pearl 1997). Indeed, the counterfactuals $Y_i(u), i \in \{t, c\}$ are exactly the ultimate covariates needed for adjustment. Because $Y = Y(u)$ is a deterministic function of $(T(u), Y_t(u), Y_c(u))$, all other variables are independent of $Y(u)$ given treatment $T = T(u)$ and $(Y_t(u), Y_c(u))$.

Allowing stochastic counterfactuals as done by Robins (1986, 1988), Greenland (1987), and Robins and Greenland (1989, 1991), we can show that even in nondeterministic settings, the "all causes" approach implies the existence of counterfactuals. Hence we reject Dawid's argument that the "all causes" approach is less metaphysical than the counterfactual approach because of the latter's reliance on "complementary" variables. To be specific, suppose that $Y(u)$ is Bernoulli. Consider a stochastic counterfactual model with the following properties: (a) There exist counterfactual probabilities $(\theta_t, \theta_c) = (\theta_t(u), \theta_c(u))$ that are deterministic functions of the individual u; (b) the function $Y_i(u)$ is the outcome of a Bernoulli experiment with success probability $\theta_i(u)$ when $T(u) = i$; $Y_i(u)$ is undefined when $T(u) \neq i$; and (c) $Y(u) = Y_{T(u)}(u)$. This model implies that $\theta_t(u)$ and $\theta_c(u)$ have a joint distribution but $Y_t(u)$ and $Y_c(u)$ do not. If we take the "all causes" approach as a primitive, we can, in complete parallel with our argument in the deterministic case, define the counterfactuals $\theta_i(u)$ to be the deterministic function $f(i, X(u))$ for which (a)–(c) hold.

1.2 Stochastic Versus Deterministic Worlds

The deterministic counterfactual model is the limiting special case of the stochastic in which $\theta_t(u)$ and $\theta_c(u)$ are always either 1 or 0. As it is impossible to use observational data to empirically decide whether the world is deterministic versus stochastic, we now investigate the inferential consequences of this inability.

1.2.1 No Unmeasured Confounders. Let W denote the measured pretreatment covariates. In a counterfactual

model, we say there are no unmeasured confounders if $\theta_i \amalg T|W$ for $i \in \{t, c\}$. This assumption will always hold in a randomized experiment with complete compliance. Given no unmeasured confounders, the marginal distributions P_c and P_t of Y_c and Y_t given W are identified and equal to the distributions of Y given W and $T = c$ and $T = t$. Thus, as discussed by Dawid and by Robins (1986), if the goal is to determine treatment for a subject u_0 exchangeable with the study subjects by comparing P_t to P_c, then it does not not matter whether the world is stochastic or deterministic.

We agree with Dawid's concern that an analyst may obtain inconsistent estimates of P_t and P_c by specifying a parametric model for nonidentifiable features of the joint distribution of (Y_t, Y_c). Our conclusion is not to reject counterfactuals models, however, but rather to criticize models and measures of effect that depend on nonidentifiable features (Greenland 1987) and to develop semiparametric counterfactual models (i.e., structural nested models, marginal structural models, and models based on the g-computation algorithm) that place no restrictions on those features (Robins 1997, 1999). Our approach completely obviates Dawid's concern.

1.2.2 Unmeasured Confounders. Because of the potential for confounding by unmeasured factors, causal effects are not identified by observational data, and the distribution of those data only implies bounds on the causal effect. For deterministic counterfactual models, the bounds always include the causal null hypothesis (Robins 1989). For the stochastic model in which for each $i \in \{t, c\}$, the $\theta_i(u)$ have the same value for all subjects u within a stratum of the measured covariates, there is no possibility of confounding, association is causation, and the upper and lower bounds coincide. Other assumptions concerning the joint distribution of (T, θ_t, θ_c) will result in bounds intermediate in length. Because whether the world is deterministic is not testable, any value lying within the deterministic bounds can never be rejected by the data. When bounds are too wide to be useful, other approaches to incorporating uncertainty due to unmeasured confounding include sensitivity analysis and formal Bayesian inference (Robins, Scharfstein, and Rotnitzky 1999). As with bounds, the resulting inferences may depend on whether one specifies a deterministic versus a stochastic counterfactual model.

1.2.3 Counterfactual Analyses That Make a Fundamental Use of Determinism. Dawid notes that in certain counterfactual analyses, the causal contrasts of interest may have no meaning if the world is stochastic. Dawid cites Imbens and Rubin (1997) for one example. The counterfactual analysis of death as a competing risk by Robins (1986, remark 12.2; 1995b) is a second. We describe a simplified single-occasion discrete-time version of Robins's analysis and provide a new approach that yields meaningful causal contrasts in both deterministic and stochastic worlds.

Example: Competing risks in a deterministic world. We observe data (ZY, Y, T), where $T = T(u)$ is a randomized treatment, $Y = Y(u) = 1$ if subject u is alive at 6 months and $Y(u) = 0$ otherwise, and $Z = Z(u)$ is blood pressure measured at 6 months, which is observed only if $Y(u) = 1$. We refer to death as a "competing risk" for the ability to observe $Z(u)$. In the literature, the counterfactuals $(Z_i(u), Y_i(u)), i \in \{t, c\}$, are often assumed to exist, in which case average causal effect of treatment on blood pressure is $E[Z_t(u) - Z_c(u)]$. This assumption implies that blood pressure $Z_i(u)$ at 6 months under treatment i is defined (although never observable) even though the subject u would be dead; that is, $Y_i(u) = 0$. Odd though it may seem, this may sometimes be a useful assumption; for example, if we were studying young children in a developing country. It would be much less reasonable if we were studying adults for whom hypertension is an important cause of death. Even when assumed to be well defined, the measure $E[Z_t(u) - Z_c(u)]$, like the other measures of the effect of treatment on blood pressure considered later, is not nonparametrically identified from the data (ZY, Y, T); the distribution of the data only imply bounds for the measure. In contrast, an effect measure relevant for choosing the optimal treatment under a particular utility function for the joint outcome (ZY, Y) *will* be identifiable. Nonetheless, a basic scientist's interest may lie in the unidentified effect of treatment on blood pressure.

Kalbfleisch and Prentice (1980) argued that it was never sensible to view $Z_i(u)$ as well-defined function of u if $Y_i(u) = 0$, in which case $E[Z_t(u) - Z_c(u)]$ is undefined as well. In that case, Robins (1986) noted that a meaningful measure of the effect of treatment on blood pressure would be its effect $\Delta_{tc} = E[Z_t(u) - Z_c(u)|Y_c(u) = Y_t(u) = 1]$ on subjects who would survive to 6 months under either treatment. This definition has two drawbacks. First, as noted by Robins (1986), it can result in nontransitivity of treatment comparisons when the treatment has three or more levels. For example, if T has support $\{t, c, r\}$, then it is possible that Δ_{tc}, Δ_{cr}, and Δ_{rt} are all positive, so that t is "preferred" to c, c is preferred to r, and r is preferred to t. Transitivity can be restored by replacing the measure Δ_{tc} by $\Delta^*_{tc} = E[Z_t(u) - Z_c(u)|\{Y_i(u) = 1; i \in \text{support}(T)\}]$, but then the probability of being in the conditioning set may be small or even 0 if the support of T is big.

A Stochastic World Generalization. The world may be stochastic. Under the stochastic counterfactual model of Section 1.1, $Y_c(u)$ and $Y_t(u)$ do not have a joint distribution, but unless they do, Δ_{tc} is without meaning. One solution is to add to our stochastic counterfactual model the assumption that Y_c and Y_t have a joint distribution. For the model to continue to satisfy properties analogous to (a)–(c), we assume that, with (θ_c, θ_t) as in Section 1.1, the conditional density $f(Y_c, Y_t|\theta_c, \theta_t)$ factors as $f(Y_c|\theta_c)f(Y_t|\theta_t)$, so that Y_c and Y_t are independent given (θ_c, θ_t). Further, we impose the restriction that $Z_i \amalg Y_j | Y_i = 1, \theta_c, \theta_t$ for $i, j \in \{t, c\}$, reflecting the fact that Y_j is purely random given (θ_c, θ_t). This model is similar to that in Dawid's Section 12. Under this model, it is an elementary calculation to show that $\Delta_{tc} = E^*[\phi_{tc}(u)]$, where $\phi_{tc}(u) \equiv \phi_{tc}(\theta_c(u), \theta_t(u))$ is the random variable

$$\phi_{tc}(u) \equiv E[Z_t(u)|\theta_c(u), \theta_t(u), Y_t(u) = 1] - E[Z_c(u)|\theta_c(u), \theta_t(u), Y_c(u) = 1]$$

and $E^*[\cdot]$ denotes an expectation taken with respect to the weighted density $f^*(\theta_c, \theta_t) \propto \theta_c \theta_t f(\theta_c, \theta_t)$.

This approach has two deficiencies when the world is truly stochastic. First, it is no longer logically necessary to define the effect of treatment only for the (possibly quite small) subset with $Y_c(u) = Y_t(u) = 1$. Second, in assuming a joint distribution for $Y_c(u)$ and $Y_t(u)$, the approach fails to satisfy Dawid's desire to keep metaphysical (i.e., untestable) assumptions to a minimum. The following alternative solution overcomes both deficiencies. We take $\phi_{tc}(u)$ as the definition of the causal effect of treatment on subject u's blood pressure whenever $\phi_{tc}(u)$ is defined; that is, $\theta_t(u)\theta_c(u) \neq 0$. For subjects for whom $\theta_t(u)\theta_c(u) = 0$, we leave the causal effect undefined. Then $\Phi_{tc} = E[\phi_{tc}(u) I\{\theta_t(u)\theta_c(u) \neq 0\}]/\mathrm{pr}[\theta_t(u)\theta_c(u) \neq 0]$ is the average causal effect of treatment on blood pressure among all subjects u for whom the causal effect is defined, where $I(\cdot)$ is the indicator function. On the one hand, suppose the world is deterministic. Then, as required, $\Phi_{tc} = \Delta_{tc}$. On the other hand, suppose the world is "fully stochastic" in the sense that $\theta_t(u)\theta_c(u) \neq 0$ for all u, and it makes sense to regard $Z_i(u)$ as defined even if $Y_i(u) = 0$. Then $\Phi_{tc} = E[Z_t(u) - Z_c(u)]$, when we assume that $Z_i \amalg Y_i | \theta_c, \theta_t$ for $i \in \{t, c\}$ so as to reflect the Y_i's being purely random given (θ_c, θ_t). Thus the approach to the problem of competing risks based on our alternative solution yields all previously proposed measures for the effect of treatment on blood pressure as special cases.

2. COUNTERFACTUALS, VAGUENESS, AND OBSERVATIONAL STUDIES

Historically, the main criticism of counterfactuals has not been the statistical objection to positing a joint distribution for complementary variables, but rather the incontrovertible fact that most counterfactuals are inherently vague or ill-defined. We argue, however, that, to misquote the Bard, "the vagueness is not in our counterfactuals but in our attempt to make causal inferences from observational data." To forswear vagueness is to join with Fisher and foreswear causal inference from nonexperimental data.

The following proposition of Quine's (1950) effectively ended counterfactual analysis among philosophers until the 1960s: If Bizet and Verdi had been of the same nationality, they both would have been French. Quine argued that because Bizet was French and Verdi was Italian, by symmetry considerations, this counterfactual could not have a truth value and thus was an ill-defined proposition. David Lewis (1973) rejoined that even though some counterfactual propositions may be ill-defined and nearly all are somewhat vague, many are useful. We agree. In fact, we believe that counterfactuals are "vague" precisely to the degree to which one fails to make precise the hypothetical interventions and the causal contrasts under consideration. For example, suppose that one collects observational data to examine the hypothesis that drinking alcohol protects against heart disease. Alcohol may protect against heart disease via a variety of pathways: It may have a direct effect on blood lipid composition; it may relax type A personalities, thereby decreasing stress-induced hypertension; it may stimulate liver enzymes that detoxify cardiac toxins such as cigarette smoke; it may displace in the diet other items, such as rich desserts, that themselves cause heart disease. If the causal contrast of interest is the direct biological effect of alcohol not mediated through its effect on diet, then one might compare an intervention wherein the daily consumption of alcohol is set to 200 kilocalories (about 2 ounces) and the diet is fixed at a prespecified menu to one in which alcohol consumption is prevented and diet is again held to the same menu.

If, however, alcohol delivered in spirits could have a different effect from alcohol delivered in wine, then these interventions must also specify the source of alcohol. Like attempts to specify all potential common causes (confounders), any attempt to eliminate all vagueness from our interventions leads to an infinite regress wherein we need to specify the type of wine, the vineyard, the year, and other factors. On the way, we eliminate the relevance of any empirical data to our causal query. For example, we might have available disaggregated data on wine and spirit consumption, but similar data on vineyard and year are out of the question.

Only in a randomized experiment in which the interest lies in the causal effect of the entire protocol (so that problems of noncompliance and lack of double-blindings are irrelevant) can we succeed in eliminating all vagueness. In observational studies, the source of the vagueness is the fundamental unavoidable difficulty in formulating just what it is we mean by the causal effect of alcohol on heart disease; the vagueness of counterfactuals is a symptom, not the cause of this difficulty. Dawid appears to express closely related sentiments in his Section 14. Thus we were surprised by Dawid's comment in Section 10 that the two appendixes of Greenland et al. (1999) were convincing illustrations of the meaninglessness and pointlessness of counterfactuals, for we can only interpret his comment as saying that causal inference from nonexperimental data is meaningless and pointless.

3. TESTABILITY AND POPPER

Contrary to Dawid's comments, the fact that counterfactual models have untestable elements does not make them "unscientific" according to either the philosophy of Popper or more modern philosophies of science. Popper made clear that falsifiability means a theory must have *some* observable predictions that would lead to its rejection were those predictions to fail, not that *every* feature of the theory be testable (Popper 1974). Counterfactual causal theories meet this requirement by having testable (observable) consequences for the marginal outcome distributions in randomized trials with complete compliance. That observational data do not always provide such critical tests is an inherent difficulty with the data source, not with the theory. Popper also made clear that "metaphysical" (i.e., apparently untestable) elements of theories could be scientifically important in providing guidance for the further development of both theory and method (Popper 1982). From this perspective, as Dawid recognizes, the counterfactual approach

has already shown itself to be an invaluable metaphysical research program for causal inference from observational studies. Similarly, counterfactuals play a key role in several speculative interpretations of quantum phenomena (e.g., Penrose 1994, pp. 237–306; Price 1996, pp. 132–194).

In summary, we regard counterfactuals as a powerful tool for eliminating, to the extent possible, vaguessness as to the causal contrasts and hypothetical interventions under consideration. They do so by requiring interested parties to explicate the scientifically important features of the "closest possible worlds" in which all subjects receive or do not receive treatment. Although presented in the context of explicating the difficulty of estimating the causes of effects rather than the effects of causes in observational studies and broken experiments, we agree that many of the points made by Dawid in Sections 11–14 are genuine and difficult problems, and we have considered them in this discussion as well as in our other writings. We believe that these problems are fundamental problems of causal inference that can either be revealed or concealed by a causal theory but never eliminated. Because counterfactuals force these problems into the open, we regard Dawid's essay as a "shoot the messenger" response to counterfactual theory.

ADDITIONAL REFERENCES

Barnard, J., Du, J., Hill, J. L., and Rubin, D. B. (1998), "A Broader Template for Analyzing Broken Randomized Experiments," *Sociological Methods and Research*, 27, 285–317.

Fisher, R. A. (1959), *Smoking—The Cancer Controversy: Some Attempts to Assess The Evidence*, Edinburgh: Oliver and Boyd.

Galles, D., and Pearl, J. (1998), "An Axiomatic Characterization of Causal Counterfactuals," *Foundations of Science*, 3, 151–182.

Greenland, S. (1987), "Interpretation and Choice of Effect Measures in Epidemiologic Analysis," *American Journal of Epidemiology*, 125, 761–768.

Greenland, S., and Robins, J. M. (1986), "Identifiability, Exchangeability, and Epidemiologic Confounding," *International Journal of Epidemiology*, 15, 413–419.

Kalbfleisch, J. D., and Prentice, R. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Penrose, R. (1994), *Shadows of the Mind*, New York: Oxford University Press.

Popper, K. R. (1974), "The Problem of Demarcation," in *Popper Selections*, ed. D. M. Miller, Princeton, NJ: Princeton University Press, pp. 101–117.

——— (1982), "A Metaphysical Epilogue," in *Quantum Theory and The Schism in Physics*, (ed. K. R. Popper, Totowa, NJ: Rowman and Littlefield, chap. 4.

Price, H. (1996), *Time's Arrow and Archimedes' Point*, New York: Oxford University Press.

Quine, W. V. (1950), *Methods of Logic*, New York: Holt, Reinhardt, and Winston.

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

——— (1989), "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, DC: U.S. Public Health Service, National Center for Health Services Research, pp. 113–159.

——— (1995a), Discussion of "Causal Diagrams for Empirical Research" by J. Pearl, *Biometrika*, 82, 695–698.

——— (1995b), "An Analytic Method for Randomized Trials With Informative Censoring: Part I," *Lifetime Data Analysis*, 1, 241–254.

——— (1997), "Causal Inference From Complex Longitudinal Data," in *Latent Variable Modeling and Applications to Causality*, ed. M. Berkane, New York: Springer-Verlag, pp. 69–117.

——— (1999), "Marginal Structural Models Versus Structural Nested Models as Tools for Causal Inference," in *Statistical Models in Epidemiology: The Environment and Clinical Trials*, eds. M. E. Halloran and D. Berry, New York: Springer-Verlag, pp. 95–134.

Robins, J. M., and Greenland, S. (1991), "Estimability and Estimation of Years of Life Lost Due to a Hazardous Exposure," *Statistics in Medicine*, 10, 79–93.

Robins, J. M., and Morgenstern, H. (1987), "The Foundations of Confounding in Epidemiology," *Computers and Mathematics With Applications*, 14, 869–916.

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

Stolley, P. D. (1991), "When Genius Errs," *American Journal of Epidemiology*, 133, 416–425.

Suppes, P. (1970), *A Probabilistic Theory of Causation*, Amsterdam: North-Holland.

Comment

Donald B. RUBIN

Once again, Professor Dawid has provided a stimulating article on a subject of great interest to statistics, the use of potential outcomes to define causal effects; I prefer the more general expression "potential outcomes" to "counterfactuals" to describe the perspective, because as Dawid himself points out before his equation (5), it is only after treatment assignments are known that some potential outcomes become known, whereas others become counterfactual.

The first time that we crossed formal discussions on this was two decades ago, and I suspect that as we have aged, our respective positions have become more entrenched, albeit not necessarily more convincing to each other.

> I feel that a source of the problem with Professor Dawid's formulation of causal inference is his choice to let the outcome variable, Y, have a joint distribution with treatment, X,

Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: *rubin@stat.harvard.edu*).

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

> rather than be t-variate where t is the number of levels of X. Letting Y be t-variate as in Rubin (1978) allows us to define causal effects phenomenologically; that is, as comparisons among t observable quantities rather than as comparisons among t hypothetical conditional distributions of Y given X for fixed values of a sufficient set of covariates. With phenomenological definitions of causal effects, valid inferences for causal effects are predictions of unobserved observables conditional on recorded data, and thus generally change as more covariates are recorded, just as valid predictions change as more predictors are recorded. With distributional definitions, inferences for causal effects are apparently viewed as incorrect unless they arrive at the correct distribution, that is, unless they are conditional on a sufficient set of covariates, an unachievable goal in real world experiments (Rubin 1979, p. 28).
>
> Professor Rubin is one of the small brave band who are beginning to chart the murky depths of causal inference. I differ from him on some matters of personal taste, the most important being his willingness to assign a *joint* distribution to all the conceptual responses of an individual under all applicable treatments, when in fact only one such response can ever be observed. I dislike this because I consider it "nonphenomenological" (there must be a shorter word!) and can only register surprise that this does not bother him too. As Quantum Theory discovered long ago, it is meaningless to assign probabilities to the joint occurrence of events which cannot occur jointly (Dawid 1979, p. 30).

Although it is interesting theoretically to classify problems of causal inference into ones where a joint distribution of potential outcomes is useful and ones where it is superfluous, I find the benefits of formulating all problems of causal inference using potential outcomes so substantial for practice that I have no desire to avoid this perspective, despite Dawid's repeated assertions that we should do so.

I could make a great variety of remarks in reply to specific comments of Dawid, but here I will simply try to convey why I find the potential outcomes perspective so appealing for causal inference. There are two main reasons: for teaching and for addressing real problems.

Teaching

For nearly a decade, I have been teaching a relatively small advanced undergraduate/graduate course on causal inference, sometimes in the Economics Department jointly with Guido Imbens (now at UCLA) and sometimes in the Statistics Department. The students are from various parts of the university, including the Faculty of Arts and Sciences, the School of Public Health and the Medical School. The course is generally very successful, I believe, to some extent because of the focus on the potential outcomes perspective, which is easily taught without resorting to Bayesian decision theory or ad hoc frequentist arguments, easily comprehended and internalized, and easily applied to real problems. The success stories include undergraduate and graduate theses in economics or statistics applying the ideas, which eventually appear in statistics and economics journals, and even win prizes, as well as anecdotes concerning how the course clarified essentially all participants' views of causal inference and nonstatisticians' views of the relevance of statistics as a field. I have recently given a very brief summary outline of this course (Rubin 1999).

A major benefit is the immediate separation of the assignment mechanism—a model for treatment assignments given potential outcomes and covariates, which we can control at times, and a theory for nature—a model for the potential outcomes given the covariates, which we cannot control. This is a point vividly made by Dawid: "Nature is surely utterly indifferent to our attempts to ensnare her in our theories," and is surely a valuable one for students to learn.

Moreover, the perspective seems to have taken over may fields, including economics (e.g., compare Heckman 1979 with Heckman 1989, after discussion in Holland 1989); epidemiology (e.g., Greenland and Poole 1988; Greenland and Robins 1986), and social science (e.g., Gelman and King 1991; Sobel 1995), as well as statistics (e.g., Holland 1986; Rosenbaum 1995).

Addressing Problems

With respect to addressing and clarifying real problems, the potential outcomes framework is extremely helpful. For a pedagogical example, it immediately resolves "Lord's paradox" (Holland and Rubin 1983).

The perspective also has been exceedingly helpful in bridging the gap between traditional econometric instrumental variables (IV) ideas and traditional statistical ideas on causal inference, from both the randomization-based perspective (Angrist, Imbens, and Rubin 1996; Rubin 1998) and the Bayesian perspective (Imbens and Rubin 1997; Hirano, Imbens, Rubin, and Zhou 2000). The key idea is to deal with noncompliance to an assigned treatment as a potential outcome itself, and classify units by their joint values of these compliance potential outcomes. This formulation also leads to progress on even more complex problems involving noncompliance and either missing outcomes (called "broken experiments" in Barnard, Du, Hill, and Rubin 1998; also see Frangakis and Rubin 1999) or censored outcomes (Frangakis and Rubin 2000).

Despite Dawid's aversion to this type of application in Section 7.1, further expansion of the potential outcomes perspective is even more revealing of its utility, so I close with a general example, with the implied challenge to Dawid to formulate a practically more appealing solution without the use of potential outcomes.

"Censored" Outcomes Due to Death

Consider a randomized experiment with two drug treatment conditions and two outcomes at one year after randomization: "alive/dead" ($D = 0, 1$) and "quality-of-life health indicator" ($Y > O$); also available are covariates $X =$ prerandomization health indicators. There is full compliance and no unintended missing data. The patients in the study are fairly ill, and some will die before completion of the study, with the result that the outcome Y is undefined in some sense or "censored" due to death. (More generally, D is an indicator for a condition that makes Y undefined.)

Drawing inferences about the effect of the treatment on D is standard. Drawing inferences about the effect of treatment on Y is more problematic. Some have proposed treat-

ing these Y values as missing or censored; I regard this as inappropriate in most situations (and certainly always truly counterfactual), although often done (e.g., Diggle and Kenward 1994, secs. 5.1 and 5.2; Rotnitzky, Robins, and Scharfstein 1998, sec. 2.2). The approach used here applies potential outcomes as did Rubin (1998, sec. 6), and differs from the traditional competing risk perspective in a way analogous to the way the IV perspective of Angrist et al. (1996) differs from the traditional econometric perspective.

Formally, for each patient there are potential outcomes corresponding to control and new treatment assignment, $(D(0), Y(0))$ and $(D(1), Y(1))$. As with noncompliance, I use the potential outcomes on D to classify the patients into four groups:

- those who would live under either treatment assignment, $LL = \{i | D_i(0) = D_i(1) = 0\}$
- those who would die under either treatment assignment, $DD = \{i | D_i(0) = D_i(1) = 1\}$
- those who would live under control but die under treatment, $LD = \{i | D_i(0) = 0, D_i(1) = 1\}$
- those who would die under control but live under treatment, $DL = \{i | D_i(0) = 1, D_i(1) = 0\}$.

For the LL patients, who comprise π_{LL} proportion of the group, there is a joint distribution of individual potential outcomes of Y under treatment and control, F_{LL}, which implies two marginal distributions of Y. For the DD patients, there is no information on Y. For the LD patients, who comprise π_{LD} proportion of the group, there is a distribution of Y under the control condition, F_{LD}. For the DL patients, who comprise π_{DL} proportion of the population, there is a distribution of Y under the new-treatment condition, F_{DL}.

Thus, in addition to the causal effects on D, which are functions of $(\pi_{LL}, \pi_{LD}, \pi_{DL})$, there are two marginal distributions of Y, F_{LD} and F_{DL}, and one joint distribution of the Y potential outcomes, F_{LL}. The causal estimands for Y follow from F_{LL}; in fact, the LL group is the only group for which causal estimands for Y involve only well-defined values of Y. Thus there are two population-level causal estimands that can be validly assessed; the effect of treatment on D for all patients and the effect of treatment on Y for those patients who would live under both assignments. Of course, the covariate X can be used to estimate causal estimands in more refined subpopulations defined by components of X.

It is true that posterior inference will be sensitive to the prior specification of the parameters of the conditional association (given X) between $(D^{(0)}, Y^{(0)})$ and $(D^{(1)}, Y^{(1)})$, but less so as more covariates are collected to predict (D, Y). With real data, I regard the usual asymptotics letting the sample size go to infinity as somewhat more relevant than letting the number of covariates go to infinity, but I am willing to think about both "asymptotic" settings.

This approach to "censored outcomes due to death" was applied several years ago, using a Bayesian model, to collaborative Amgen, Inc. data in a randomized trial on use of a neurotrophic factor to treat amyotropic lateral sclerosis (ALS), where Y was "forced vital (lung) capacity." Not only were the Bayesian answers reasonable, but moreover, simulations showed that the implied tests (posterior predictive, Rubin 1984, sec. 5; Meng 1994) of the two null hypotheses corresponding to the two causal questions (difference in death rates for all; difference in Y means for those who would live under both treatments), each had essentially the correct frequentist level, both were reasonably powerful, and they were effectively orthogonal in operating characteristics, as they should be because they addressed two distinct scientific questions.

Such a perspective seems critically important in studies of very ill patients where "quality of life" must be considered an outcome distinct from "death." Defining Y when the patient dies to be the worst possible value of Y simply misses the scientific/medical/ethical point to distinguish between these outcomes. There appear to be other situations, as well, as in the examples of Diggle and Kenward (1994) cited earlier involving cows and milk production, where this perspective would be valuable.

It may be that situations with "censoring due to death" can be addressed in a straightforward way without the use of potential outcomes, but to me, the essence of the causal inference situation is immediately conveyed, both intuitively and technically, by the use of potential outcomes with a joint distribution (as proposed in Rubin 1978).

ADDITIONAL 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.

Barnard, J., Du, J., Hill, J., and Rubin, D. B. (1998), "A Broader Template for Analyzing Broken Randomized Experiments," *Sociological Methods and Research*, 27, 285–318.

Diggle, P., and Kenward, M. G. (1994), "Informative Drop-Out in Longitudinal Data Analysis," *Journal of the Royal Statistical Society*, Ser. C, 43, 49–73.

Frangakis, C., and Rubin, D. B. (1999), "Addressing Complications of Intention-To-Treat Analysis in the Combined Presence of All-or-None Treatment Noncompliance and Subsequent Missing Outcomes," *Biometrika*, 86, 366–379.

——— (2000), "A Note on Addressing An Idiosyncracy in Estimating Survival Curves Using Double Sampling in the Presence of Self-Selected Right Censoring." Submitted for publication.

Gelman, A., and King, G. (1991), "Estimating Incumbency Advantage Without Bias," *American Journal of Political Science*, 34, 1142–1164.

Greenland, S., and Poole, C. (1988), "Invariants and Noninvariants in the Concept of Interdependent Effects," *Scandinavian Journal of Work and Environmental Health*, 14, 125–129.

Greenland, S., and Robins, J. (1986), "Identifiability, Exchangeability and Epidemiological Confounding," *International Journal of Epidemiology*, 15, pp. 413–419.

Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

——— (1989), "Causal Inference and Nonrandom Samples," *Journal of Educational Statistics*, 14, 159–168.

Hirano, K., Imbens, G., Rubin, D. B., and Zhou, X. H. (2000), "Estimating the Effect of an Influenza Vaccine in an Encouragement Design," *Biostatistics*, 1, 69–88.

Holland, P. W. (1989), "It's Very Clear," comment on "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training" by J. Heckman and V. Hotz, *Journal of the American Statistical Association*, 84, 875– 877.

Holland, P. W., and Rubin, D. B. (1983), "On Lord's Paradox," in *Principles of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, eds. Wainer and Messick, Hillsdale, NJ: Earlbaum, pp. 3–25.

Meng, X. L. (1994), "Posterior Predictive p Values," *Annals of Mathematical Statistics*, 22, 1142–1160.

Rosenbaum, P. R. (1995), *Observational Studies*. New York: Springer-Verlag.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1999), "Adjusting for Nonignorable Dropout Using Semiparametric Models," *Journal of the American Statistical Association*, 94, 1321–1339.

Rubin, D. B. (1979), Discussion of "Conditional Independence in Statistical Theory" by A. P. Dawid, *Journal of the Royal Statistical Society*, Ser. B, 41, 27–28.

——— (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

——— (1998), "More Powerful Randomization-Based p Values in Double-Blind Trials With Noncompliance" (with discussion), *Statistics in Medicine*, 17, 371–389.

——— (1999), "Teaching Causal Inference in Experiments and Observational Studies," *Proceedings of the Statistical Education Section, American Statistical Association*, pp. 126–131.

Sobel, M. E. (1995), "Causal Inference in the Social and Behavioral Sciences," in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, eds. Arminger, Clogg, and Sobel, New York: Plenum, pp. 1–38.

Comment

Glenn SHAFER

In recent years, a number of statisticians and computer scientists have suggested that casual reasoning requires that questions with hypotheses counter to fact have well-defined answers. Phil Dawid's elegant and insightful article is the first critical examination of this suggestion. As such, it is an essential contribution to the philosophy of probability and causality. It moves the discussion of causality in statistics to a new level of sophistication.

The article should prove an effective exercise in persuasion, because Dawid meets the proponents of counterfactuals on their own ground. He begins with the *counterfactual variables* Y_t and Y_c that appear in the models formulated by Neyman (1923), Rubin (1974, 1978), and Holland (1986), and he makes every effort to understand how much sense and how much use can be made of these variables.

Dawid's central theme is that counterfactuals should be held up to de Finetti's observability criterion. This criterion says that it is legitimate to assess a probability distribution for a quantity Y only if Y is observable at least in principle. On this criterion, it is legitimate to assess probabilities for $Y_t(u)$, because we can apply the treatment t to the unit u and then observe $Y_t(u)$. It is also legitimate to assess probabilities for $Y_c(u)$, because we can apply the control c to u and then observe $Y_c(u)$. But if we cannot do both, then it is not legitimate to assess probabilities for the pair $(Y_t(u), Y_c(u))$. Dawid vindicates de Finetti's criterion by showing that persuasive examples of causal inference that seem to require a joint distribution for $Y_t(u)$ and $Y_c(u)$ can be reformulated so that they clearly do not involve any such joint distribution.

Dawid's discussion of the instrumentalist use of counterfactual variables, one of the article's highlights, demonstrates the effectiveness of his conciliatory approach. As Dawid makes clear, reservations about the empirical meaningfulness of counterfactual variables need not prevent one from using them for mathematical convenience.

My main reservation about the article is that it does not take advantage of Dawid's own path-breaking work on predictive probability (see, e.g., Dawid 1985, 1992; Dawid and Vovk 1997). In his effort to find common ground with those who tout counterfactual variables, Dawid emphasizes the case of a finite homogeneous population, where optimal predictions are merely population averages, and he hints that other cases can be reduced to this case by restricting attention to a subpopulation with a specific value of a covariate. This downplays the link between causality and prediction and obscures the potential richness of that link. It makes it difficult, for example, to recognize that the predictions authorized by casual regularities may often fall short of the full panoply of predictions that would be authorized by a probability distribution (Shafer 1996, 1998).

In the end, Dawid concedes too much, especially on the topic of causes of effects. These concessions can be avoided if one remembers that counterfactual variables do not provide the only framework for discussing causality. Frameworks that make a direct place for probabilistic prediction also have their uses, and they are needed to help distinguish causal statements that have empirical content from those that are irremediably arbitrary and subjective.

Glenn Shafer is Professor, Department of Accounting and Information Systems, Faculty of Management, Rutgers University, Newark, New Jersey (E-mail: *gshafer@andromeda.rutgers.edu)*. These comments benefited from research supported by National Science Foundation grant SES-98199116, and also from discussions of causality with Phil Dawid over many years. These discussions were most recently pursued in the context of a workshop generously supported by the Fields Institute for Research in Mathematical Sciences, which brought together a number of students of causality, including Vanessa Didelez, Mervé Eerola, Michael Eichler, Paul Holland, Steffen Lauritzen, Wayne Oldford, James Robins, Don Rubin, Richard Scheines, and Ross Shachter, in addition to Dawid and myself. Dawid discussed his article at this workshop, and I am grateful to all of the participants for their discussion of the issues it raises.

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

1. CONDITIONAL EXPECTED VALUES SUFFICE FOR DELIBERATION

John and his physician deliberate on whether John should undergo an operation. They decide to go ahead, and John dies on the operating table. How long would John have lived had the operation not been undertaken? Before the decision is made, it is surely legitimate for the physician to talk about her expectations for how long John will live with and without the operation. If the physician is a mathematician, she may write Y(John) for John's longevity and assess the two expected values $E(Y(\text{John})|\text{operation})$ and $E(Y(\text{John})|\text{no operation})$, where

$$E(Y(\text{John})|\text{operation}) := \text{John's expected longevity if the operation is undertaken}$$

and

$$E(Y(\text{John})|\text{no operation}) := \text{John's expected longevity if the operation is not undertaken.} \quad (1)$$

At this point, before the decision of whether to operate, there is nothing counterfactual about either of these quantities. After John's death on the operating table, $E(Y(\text{John})|\text{no operation})$ can be called counterfactual, because it involves a hypothesis that is now contrary to fact. It is a counterfactual expected value. But this term is not particularly enlightening; a better one might be *past conditional.*

There is, I believe, no disagreement about the meaningfulness or usefulness of past conditionals such as $E(Y(\text{John})|\text{no operation})$, nor is there disagreement about the meaningfulness or usefulness of analogous predictions in cases where we can predict the consequences of treatments on a unit u for certain. In such cases, the conditional expected values $E(Y(u)|\text{treatment})$ and $E(Y(u)|\text{control})$, reduce to the conditional categorical predictions

$$Y_t(u) := \text{the value } Y(u) \text{ will take if } u \text{ is given the treatment } t \quad (2)$$

and

$$Y_c(u) := \text{the value } Y(u) \text{ will take if } u \text{ is given the control } c. \quad (3)$$

Again, there is nothing counterfactual about $Y_t(u)$ and $Y_c(u)$ before the decision is made whether to give u the treatment or the control, although if t is given, then it is acceptable afterward to call $Y_c(u)$ a counterfactual prediction.

There is also no disagreement about the importance of quantities like (1)–(4) in causal assertion and hence in causal inference. The essence of causality lies in the fact that different actions will have different consequences or at least different expected consequences, and these differences can still be discussed after the passage of time changes "will have" into "would have had."

Controversy arises only when we ask whether causal questions should always be answered in terms of categorical predictions such as (3) and (4) or whether probabilistic predictions such as (1) and (2) can also be used. The *counterfactual approach* that Dawid criticizes bases causality on quantities of the form (3) and (4) in all cases, even when no such predictions can be made, even in principle. In this approach, even probabilistic predictions are interpreted not as conditional expected values, as I have done in (1) and (2), but rather as expected values of counterfactual variables. The conditional expected values $E(Y(u)|\text{treatment})$ and $E(Y(u)|\text{control})$ are recast as unconditional expected values $E(Y_t(u))$ and $E(Y_c(u))$.

Dawid concedes too much when he assents to this notational trick. The conditional expected values really are conditional. Yes, $E(Y(u)|\text{treatment})$ becomes $E(Y_t(u))$ when the decision is made to apply t to u, but $E(Y(u)|\text{control})$ remains a conditional expected value, now with respect to past rather than current probabilities. There is no need for us to imagine an alternative universe in which it has been promoted to an unconditional expected value.

2. WHY SHOULD THE BLACK BOX CONTAIN DETERMINATE PREDICTIONS?

As we have just seen, a thorough understanding of causal structure is not needed for deliberation. As Dawid explains, the assessment of the effects of possible actions "can proceed by an essentially 'black box' approach, simply modeling dependence of the response on whatever covariate information happens to be observed for the test unit." To understand the "causes of effects," we need to probe inside the black box.

An autopsy may reveal facts about John that the physician could not have suspected or learned beforehand but that made the operation's failure likely. Any such facts obviously need to be taken into account in a discussion of the causes of John's death. For the conditional expected values in (1) and (2) to have causal meaning, they must take all such facts into account. When they do so, will they still be merely probabilities and expected values, or will they necessarily become determinate predictions, of the form (3) and (4)?

There are three powerful arguments against expecting determinate predictions:

1. Twentieth century physics has repeatedly refuted efforts to eliminate probability from the predictions of quantum mechanics.
2. Our mundane experience also provides no support for the proposition that the effects of our actions are always determinate.
3. The very formulation of the question rests on an assumption that John and his physician can choose freely between having the operation and not having it. So how can we coherently deny that there may be later free choices, such as those suggested by Figure 1, that may also effect John's longevity?

The proponents of counterfactuals often respond to the mention of quantum mechanics by suggesting that it is too

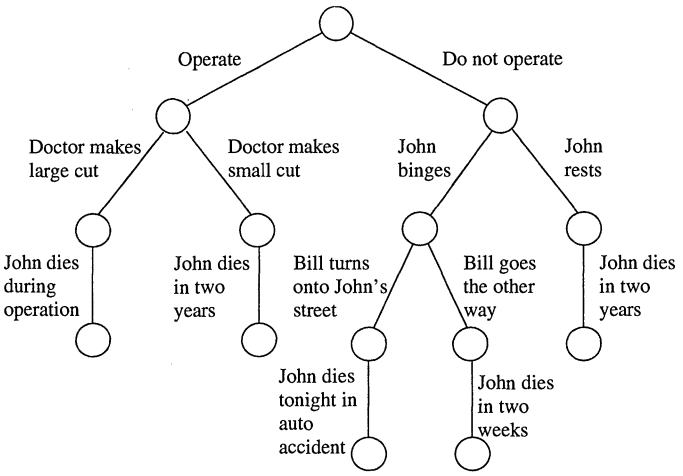

Figure 1. No One Can Predict Exactly the Results of Treatment and Control.

esoteric to be relevant to everyday concerns. In medicine, business, and law, they argue, we can make do with Newtonian mechanics, where actions have predictable consequences. Unfortunately, as the second argument reminds us, Newtonian laws do not get us very far in understanding the choices that we face in medicine, business, and law. Even Laplace's vision of determinism, in which a superior but human-like intelligence can predict the future states of the world from knowledge of the present state and a small number of laws, demands only the possibility of prediction for states in which the world is actually found. If causal laws predict everything, they predict that the physician will undertake the operation. Thus the Laplacean vision does not require that the superior intelligence should be able to make a prediction about what would happen if the operation is not undertaken.

The force of the third argument was already acknowledged in the work by Don Rubin that launched the revival of counterfactuals in statistics in the 1980s and 1990s. Rubin (1978, pp. 39–40) provided three conditions that should be met before one assumes that treatments have determinate results whether or not they are applied:

1. Each treatment should be defined by a series of actions that can be applied to the individual. For example, if one insists on studying the causal effect of being female, then one must specify the particular actions to be taken to make the individual female.
2. Any pretreatment manipulations should be included in this series of actions. For example, if different medical treatments are preceded by different physical examinations, then the examination should be considered part of the treatment.
3. "We cannot attribute cause to one particular action in the series of actions that define a treatment. Thus treatments that appear similar because of a common salient action are not the same treatment and may not have similar causal effects."

The third condition can be elaborated by saying that the series of actions defining a treatment must include all human actions that can affect the outcome. Thus we need to include in the definition of John's treatment not only the doctor's actions, but also John's and Bill's actions in Figure 1.

3. CAUSAL STRUCTURE WITH OBJECTIVE PROBABILITIES

We have been led to the conclusion that probabilities with causal meaning—objective probabilities, if you will—are those based on all the information humanly possible to have and use in a given situation. As Dawid might put it, these are the probabilities based on a *sufficient covariate*.

This conception of objective probability hardly new. In the mid-nineteenth century, Antoine Augustin Cournot, adapting Laplace's formulation of determinism, proposed that objective probabilities are those that a superior but human-like intelligence would obtain using the current facts about the world and knowledge of causal regularities. The idea echoed well into the twentieth century, in the work of authors such as Henri Poincaré (1908) and Émile Borel (1924).

What can one say about causality when one has only probabilistic predictions such as (1) and (2) instead of categorical predictions such as (3) and (4)? As I have argued in *The Art of Causal Conjecture* (1996), one can say a great deal. One may assert that a particular action (by a person, an animal, or some inanimate actor, such as a storm or a comet) changes the expected value of some variable or the probability of some particular outcome. The action is, in this sense, a cause. In some cases, one may conjecture broader causal regularities—for example, that a given type of action always raises the expected value of a given variable, or that all the causes of one variable are causes of another variable. For example, one may conjecture that most actions that increase the expected value of a person's smoking decrease the expected length of that person's life.

One of the attractions of counterfactual variables for statisticians over the past several decades has been precisely the fact that they obviate the need for objective probability. At least they allow reducing objective probability to the simpler concept of frequency in a finite population. This is attractive to some because of their Bayesian persuasions and to others because of their weariness with long-running debates about the meaning of probability. It now seems clear, however, that nothing has been gained by replacing objective probabilities with categorical counterfactuals. Objective probabilities have an empirical meaning at least in principle—they represent what one might obtain in the limit as predictions are improved through additional experience and knowledge. Categorical counterfactuals, everyone agrees, are often unknowable even in principle.

4. ASKING THE WRONG QUESTION

In my view, Dawid concedes too much when he allows that categorical counterfactuals may be needed for inferences about the causes of effects. He concedes too much by agreeing to pose a question that has no meaning.

Something has happened, and we are being asked whether one particular step in the course of events has caused it:

- My headache is gone. Is it because I took aspirin?
- John died on the operating table. Is it because the physician operated?
- Our corn crop failed. Is it because of the variety of seed we planted?

Dawid poses the general question in these words: "We are interested in whether, for the specific unit u_0, the application of t 'caused' the observed response." He lets us know, with the quotation marks around *caused,* that he is asking a silly question. Unfortunately, the quotation marks do not save him from becoming entangled in silly answers.

Imagine that there was a categorical rule about the effect of aspirin on headaches:

- At least two aspirin with at least a cup of water: the headache goes away.
- Less aspirin or less water: the headache persists.

I take the requisite aspirin with the requisite water. My headache goes away. Is it because I took aspirin? I understand the causal structure perfectly, but cannot answer the question with a simple yes or no.

It is equally silly to isolate a single action and ask whether it is *the* cause when the action's effect depends on something that is settled later—what Dawid calls a "determining concomitant." Figure 2 shows a very simple example. I am required to bet \$1 on the outcome of a toss of a coin. I decide to bet on heads, the coin lands tails, and so I loose my \$1. Did my choice of heads "cause" my handing over \$1 instead of receiving \$1?

Here again, the causal structure is perfectly understood. The coin is fair; the chance of its landing heads is 50% regardless of how I bet. The outcome Y (which will be either $+1$ or -1) is completely determined by the treatment T (which will be either "bet on heads" or "bet on tails") together with the determining concomitant D (which will be either "coin lands heads" or "coin lands tails"). I understand exactly what happened. But this does not enable to me to give a yes or no answer to the question whether T "caused" Y.

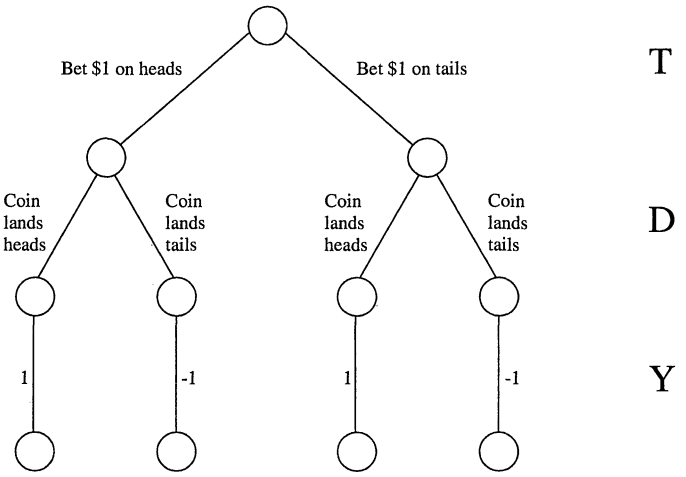

Figure 2. A Determining Concomitant.

Had I instead bet on tails, would the toss have come out the same way? What is gained by asking this question or by making up an answer to it? One can make up whatever answer one wants. If one assumes that this particular determining concomitant D comes out the same in the counterfactual world where I bet on tails, then I win in that world. If I choose a different determining concomitant D', whose possible values are "coin lands the way I bet" and "coin lands opposite the way I bet," and assume that it comes out the same in the counterfactual world, then I lose in that world. What is the point or content of either assumption?

Dawid sees the arbitrariness clearly and so arrives at this formulation: "The essence of a specific causal inquiry is captured in the largely conventional specification of what we may term the context of the inference, namely, the collection of variables that is is considered appropriate to regard as concomitants..." Specification of the context, he concludes, "is vital to render causal questions and answers meaningful."

In practice, I am willing to trust Phil Dawid to take the air out of silly questions by showing how they depend on the arbitrariness of the choice of a context. But I distrust his formulation, for it seems to say that all singular causal questions partake in this silliness—that all causal answers depend on the arbitrary specification of concomitants.

5. ASKING THE RIGHT QUESTION

My father quit smoking in the early 1960s, after using cigarettes heavily for more than 20 years. He died in 1997, at age 75. How long would he have lived had he continued smoking? How much did his quitting smoking change his life expectancy?

These are both questions about the causes of an effect. They are both questions about singular causation, and they are both questions about my father's quitting smoking as a cause of his observed longevity. The first is a wrong question; it will remain a silly question no matter how many different concomitants Phil Dawid tries out on us. The second is a right question. It is a scientific question, which comes with its own context and requires no arbitrary specification of concomitants by Dawid or anyone else.

In the United States, litigation continues between the tobacco industry and the federal government, which seeks compensation for the costs of caring for people whose health was damaged by smoking. The tobacco companies are held liable on the grounds that they took actions to increase cigarette consumption despite their own knowledge of smoking's ill effects. To measure the effect of these actions, we can ask a scientific question:

> How much was the expected government expense on caring for the ill increased by the actions of the tobacco companies?

Alternatively, we can ask a counterfactual question:

> How much less would the government have spent on caring for the ill had the companies had not taken these actions?

The scientific question is very difficult. Its answer is subject to great uncertainty. The counterfactual question adds arbitrariness to the uncertainty. An insistence on the coun-

terfactual question will lead in the end to denial of responsibility. How much blame to place on the tobacco industry becomes not a scientific question but a purely political one: What counterfactual world does one want to imagine?

We all indulge, in anger and regret, in counterfactual talk: "If they had not operated, John would be alive today"; "If I had not said that, she would not have left me"; "If I had chosen a different publisher, my book on causality without counterfactuals would have sold 10,000 copies." The more fortunate among us have someone to remind us that we are talking nonsense. Calmer heads will remind John's son and widow that his length of life had the physician not operated does not have a determinate value.

The physician's responsibility is to compare (1) and (2) based on the best evidence she can reasonably gather, and to perform the operation, if she does perform it, with expertise and care. One can ask for no more. As Jacob Bernoulli, the inventor of mathematical probability, wrote, De Actionum humanarum pretio non statuendum ex eventu (Bernoulli 1713): Do not judge human action by what happens.

ADDITIONAL REFERENCES

Bernoulli, J. (1713), *Ars Conjectandi*. Basel: Thurnisorum.

Borel, E. (1924), "A Propos d'un Certain Traité," *Revue Philosophique*, 98, 321–326, 1924. Reprinted *Oeuvres de Émile Borel*, Vol. 4, Paris: Centre National de la Recherche Scientifique, 1972.

Cournot, A. A. (1851), *Essai sur les Fondements de nos Connaissances et sur les Charactères de la Critique Philosophique*. Paris: Hachette, Paris, 1851. Reprinted in (J. C. Pariente, editor) *Oeuvres Complètes*, Vol. II, ed. J. C. Pariente, Paris: J. Vrin, 1975.

Dawid, A. P. (1985), "Calibration-Based Empirical Probability" (with discussion), *The Annals of Statistics*, 13, 1251–1285.

——— (1992), "Prequential Data Analysis," *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, eds. M. Ghosh and P. K. Pathak, Hayward, CA: Institute of Mathematical Statistics, pp. 113–126.

Dawid, A. P., and Vovk, V. G. (1997), "Prequential Probability: Principles and Properties," *Bernoulli*, 3, 1–38.

Kenes, J. M. (1921), *A Treatise on Probability*, London, Macmillan.

Kyburg, H. E. Jr., and Smokler, H. E., editors (1964), *Studies in Subjective Probability*. Wiley, New York.

Poincaré, H. (1908), *Science et Méthode*, Paris: Flammarion.

Shafer, G. (1996), *The Art of Causal Conjecture*, Cambridge, MA: MIT Press.

——— (1998), "Mathematical Foundations for Probability and Causality," in *Mathematical Aspects of Artificial Intelligence*, ed. F. Hoffman, Providence: American Mathematical Society, pp. 207–270.

Comment

Larry WASSERMAN

In his essay, Phil Dawid takes a stand against the use of counterfactuals in causal inference, arguing that they are unhelpful and potentially misleading. Instead, he favors a decision-theoretic approach. These are important issues, and yet they are rarely discussed in the statistics literature. I appreciate Dawid's thoughtful article and the opportunity to comment on it.

My view on counterfactuals is very different. I believe that counterfactuals give the simplest and clearest explanation of causality. I agree that there are potential dangers in performing inference with models based on counterfactuals. A careless user could end up trying to estimate nonidentifiable parameters. But nonidentifiability lurks in many statistical models. Overall, I think that the advantages of counterfactuals far outweigh their disadvantages.

1. WHY COUNTERFACTUALS ARE USEFUL

Phil Dawid concedes that counterfactuals are useful for "... causal model building." I wish to go further and claim that counterfactuals are useful for developing a clear understanding of causal inference. I usually find that it is easy to clear up most confusion about causal issues using explanations based on counterfactuals. I believe this is simpler than the proposed decision-theoretic approach.

Let us review the basics of counterfactuals in causal inference. Suppose that the treatment variable T is binary and let Y be the outcome. The counterfactuals in this case are (Y_0, Y_1), where Y_0 is the outcome if not treated and Y_1 is the outcome if treated. (Note that my notation departs slightly from that in the article.) The set of random variables is $V = (Y_0, Y_1, T, Y)$, where the observed outcome Y is related to the other random variables by the *consistency relation* $Y = TY_1 + (1 - T)Y_0$. Evidently, if $T = 1$, we see Y_1 but not Y_0. Similarly, if $T = 0$, we see Y_0 but not Y_1. To me, both Y_0 and Y_1 are nonetheless well defined, a point I return to in Section 3.

The average causal effect is defined by $\theta = E(Y_1) - E(Y_0)$, which does not equal the association $\alpha = E(Y|T = 1) - E(Y|T = 0)$. The adage "association is not causation" is thus easily rendered mathematically precise. Moreover, one can show that in an observational study, θ is not identifiable. If we randomize the assignment of treatment, then T is independent of (Y_0, Y_1) and we have that $\theta = E(Y_1) - E(Y_0) = E(Y_1|T = 1) - E(Y_0|T = 0) = E(Y|T = 1) - E(Y|T = 0) = \alpha$. Hence with counterfactuals, it is easy to show that randomization makes the causal effect equal to an identifiable parameter α.

As a further example, consider how easily Simpson's paradox can be explained using counterfactuals. In addi-

Larry Wasserman is Professor, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213. This research supported by National Institutes of Health grant R01-CA54852-01 and National Science Foundation grant DMS-9803433. The author thanks Isabella Verdinelli, Sander Greenland, and Jamie Robins for helpful comments.

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

tion to the random variables (Y_0, Y_1, T, Y), consider also a covariate Z, such as sex. It is possible to construct a joint distribution P such that

$$P(Y = 1|T = 1) > P(Y = 1|T = 0) \quad (1)$$

and yet

$$P(Y = 1|T = 1, Z = z) < P(Y = 1|T = 0, Z = z) \quad \forall\, z. \quad (2)$$

So far, there is no paradox. But if we interpret (1) to mean "the treatment is better than the control" and interpret (2) to mean "the control is better than the treatment for each sex," then we do have a paradox. The problem is the translation of the equations into English statements. In fact, using counterfactuals, "the control is better than the treatment for each sex" corresponds to $P(Y_1 = 1|Z = z) < P(Y_0 = 1|Z = z)$ for all z. It then follows that $P(Y_1 = 1) = \sum_z P(Y_1 = 1|Z = z)P(Z = z) < \sum_z P(Y_0 = 1|Z = z)P(Z = z) = P(Y_0 = 1)$ and hence that "the control is better than the treatment."

As further evidence of the pedagogical value of counterfactuals, consider the two problems, labeled I and II in the article, which are called the *effects of causes* and *causes of effects*. The difference between the two is again transparent using counterfactuals. The first question refers to $\Pr(Y_1 = 1) - \Pr(Y_0 = 1)$ and the second refers to $\Pr(Y_0 = 0|T = 1, Y = 1)$. In a randomized study, the first is identifiable, and the second is not.

2. IMPLICIT USE OF COUNTERFACTUALS

As I discuss in Section 4, some parameters of interest are not identifiable and so cannot be estimated from data alone. But in some cases, we can bound the parameter of interest. An excellent example is the bounds found by Manski (1990), Robins (1989), and Balke and Pearl (1997) for causal effects in clinical trials with imperfect compliance. Balke and Pearl prefer directed acyclic graphs to counterfactuals, but a close examination of their proof shows that their calculations are essentially based on a counterfactual representation. Generally, it is easier to bound causal parameters using counterfactuals. In Section 10.2, Dawid notes that he is able to derive these inequalities without counterfactuals. Maybe so—but he already knew the answer! The more important question is whether most people could have derived the bounds without counterfactual reasoning.

3. THE PHYSICAL MEANING OF COUNTERFACTUALS

Dawid claims that counterfactuals have no real physical meaning, and he makes an analogy with incompatible variables in quantum mechanics. I am no expert in quantum mechanics, but I think that such an analogy is potentially misleading. Our inability to measure two incompatible variables simultaneously is built into the mathematics of quantum mechanics. There is no idealized experiment in which we could observe both variables. This is a deep, mathematical fact about quantum mechanics. The limitations on observing both counterfactuals in, say, a medical study, are much different. These limitations are mainly practical limitations (carryover effects of the aspirin, etc.) and, at least in principle, I can imagine an idealized experiment where I can almost observe both Y_0 and Y_1. I do not see this as being mathematically precluded in the same way as it is in quantum mechanics.

4. IDENTIFIABILITY

Recall that the set of random variables is $V = (Y_0, Y_1, T, Y)$, where $Y = Y_1T + Y_0(1 - T)$, the causal effect is $\theta = E(Y_1) - E(Y_0)$, and the association is $\alpha = E(Y|T = 1) - E(Y|T = 0)$. Let P denote the joint law of V. It can be shown that α is identifiable assuming that $P(T = 1)P(T = 0) > 0$. In a randomized study, $\theta = \alpha$, and hence θ is also identifiable. In an observational study, one can show that θ is not identifiable (without further assumptions), but identifiable bounds can be placed on θ. On the other hand, even in a randomized study, there are other functionals of P that are not identifiable, such as $\Pr(Y_0 = 1|Y_1 = 1)$. I think it is the lack of identifiability that leads Dawid to claim that counterfactuals are potentially misleading. But lack of identifiability is possible in any statistical problem. It is always important to ensure that the quantities being estimated are identifiable and, if not, to state identifiable bounds on the quantity of interest. In this sense, counterfactuals are no different than any statistical model. In fact, counterfactuals actually help, because they allow one to rigorously prove what is and is not identifiable.

5. CONCLUSION

Dawid has raised a host of interesting issues. Statisticians often shy away from causation. I hope that this article will encourage statisticians to delve further into causation.

As I have stated, my opinion about counterfactuals is that they are useful, even crucial, for obtaining a clear understanding of causation. As long as we remain vigilant about nonidentifiability, counterfactuals are not dangerous. In his conclusion, Phil states that "the counterfactual approach to causal inference is essentially metaphysical, and full of temptations to make 'inferences' that cannot be justified on the basis of empirical data and are thus unscientific." I suggest that we continue to use counterfactuals but educate users to resist the temptation to indiscriminantly make inferences for nonidentified parameters in all models, not just causal models.

ADDITIONAL REFERENCES

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Manski, C. (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, 80, 319–323.

Robins, J. (1989), "The Analysis of Randomized and Nonrandomized AIDS Treatment Trials Using a New Approach to Causal Inference in Longitudinal Studies," in *Health Service Research Methodology: A Focus on AIDS*, eds. L. Sechrest, H. Freeman, and A. Mulley, Washington, DC: U.S. Public Health Service, pp. 113–159.

Rejoinder

A. P. DAWID

I am happy that my article has received serious attention from such an impressive range of deep-thinking individuals. Rather than deal with each discussant individually, I have organized this rejoinder around central and recurring themes.

1. MORE GENERAL CAUSAL MODELING

I largely restricted my analysis to the simple case of homogeneous populations. Shafer would like me to have explored more complex predictive structures; he would also like "acts of nature," as well as of external agents, to count as causes. Robins and Greenland consider that I have avoided the main point by excluding observational studies and related complications. Cox complains that the search for "some understanding of a phenomenon" hardly figures in my account. Casella and Schwartz emphasise the importance of the reference set in providing an appropriate frame of inference.

I am grateful to these discussants for prompting me to think more deeply about these valid and important concerns. To address them, I have to accept that a vital aspect of causal modeling and inference is the identification of modular subprocesses, persisting across and linking a variety of changing circumstances. This may require complex scientific investigations and understandings.

Consider an observational study. One is not interested in making predictions about the behavior of a future unit exchangeable with those in the study, but rather for a new unit subjected to one or a number of possible interventions. However, unless one can somehow link the behaviors of units under the observational and the interventional regimes, no useful inference is possible. Most of the discussants make this link using "potential responses," defined in terms of a hypothetical intervention experiment and implicitly assumed to continue to have meaning under observational conditions. Pearl extends this idea using functional models. I would rather make the link by means of a *stable probabilistic mechanism,* assumed to be the same in both the observational and the interventional settings, describing the distribution of a response Y given some suitable covariate K and treatment i (see my Sec. 8.1). Notwithstanding the preferences of Wasserman, Robins and Greenland, and Rubin, I consider that the benefits of randomization (for example) can be explained more meaningfully and convincingly in terms of the independence of T and K, rather than the independence of T and the metaphysical collection of all potential responses. Similarly, fully adequate treatment can be given for all other problems of observational studies, including Simpson's paradox (Dawid 1979) and the assumption of "no unobserved confounders." I am currently developing these ideas and analyses, and hope to present my case for them in due course.

To clarify the aforementioned notion of "stable probabilistic mechanism," I need to say more about my conception of probability. All of the discussants except Shafer and Robins and Greenland seem to be out-and-out Laplacian deterministes, for whom nothing short of a functional model relating outputs to inputs will do as a description of nature. And all discussants seem to believe that the relevant relationships between variables, be they deterministic or stochastic, are genuine features of the external world, rather than of the models we choose to describe the world, and that only models incorporating those features make sense.

My own attitude is very different. Although at heart I too am a Laplacian determinist, I do not conclude that one must *model* nature as behaving deterministically; the level of detail required to make sense of such a model is, typically, simply not appropriate to the kind of question that we aim to address. I see "probability" as an inescapably theoretical term, with only an indirect connection to the empirical world. I would judge the empirical success of a probabilistic model by means of the *calibration criterion* (Dawid 1982), which requires that certain averages of probabilities calculated from the model coincide, in the long run, with corresponding relative frequencies observed in the world. An essential ingredient qualifying this criterion, however, is the level of detail that the model is intended to express, and that in turn determines just which collections of probabilities it is appropriate to average. I made this idea (there termed the *information base*) precise in an earlier article (Dawid 1985), where I essentially showed that for any given information base, there will be just one "correct" assignment of probability values to the observed events—but this will typically vary with the information base used. In particular, it is possible that at some sufficiently refined level of detail, those "correct" probabilities are all 0 or 1, corresponding to a fully deterministic description, whereas at a more interesting or realistic level of detail, the "correct" probabilities are noncategorical. So, unlike Robins and Greenland, and Shafer, I cannot consider such concepts as "pure randomness," "objective probability," or "all nontreatment causes of the outcome" as absolutes, but rather as meaningful only relative to a specified information base.

The concept of information base is closely related to, but not identical with, that of "context" in Section 14 of my article. For example, for inference about effects of causes in the restricted context of the completely homogeneous population, as introduced in my Section 5, the basic model provides for a stable probabilistic dependence of outcome on treatment alone. As long as future treatment decisions will be restricted to new units drawn from the same ho-

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

mogeneous population, with no observation of any further properties of those units, there is no necessity to introduce any more detail into the modeling. As soon as these conditions fail, a more detailed model, perhaps involving covariates, will be required to relate past and future situations. However, it will make no significant difference if one does build and use models incorporating an unnecessarily refined level of detail, because one will just end up averaging over the unused deeper levels (as in Case 2 in my Sec. 8).

2. POSITIVISM

Casella and Schwartz claim that Popperian positivism is out of fashion among philosophers, whereas counterfactual analysis is "in." I trust, however, that my arguments will be considered on their own merits, rather than on whether they are fashionable.

Casella and Schwartz are also wrong to confuse Jeffreys's law with the likelihood principle. Example 3 of Dawid (1984) illustrates Jeffreys's law in a purely predictivist, noncausal setting. I hope that this helps clarify the issue.

Pearl says: "If our conclusions have no practical consequences, then the sensitivity to invalid assumptions is totally harmless, and Dawid's warning is empty. If, on the other hand, our conclusions do have practical consequences, then the sensitivity to assumptions automatically makes those assumptions testable, and Dawid's warning turns contradictory." I am in total agreement with this, as with his statement that "many counterfactual modeling assumptions do have testable implications." As long as attention is confined to such testable aspects, no problem arises. My point is that the models I criticize also have *untestable* implications, and that (unless one takes great care) it is all too easy to use them to make "inferences" that are sensitive to purely arbitrary and untestable choices that may be made for ingredients in these models. (The issue is not exactly that of identifiability in the usual technical sense).

Casella and Schwartz have misunderstood my use of the term "fatalism." In a counterfactual context, this goes way beyond the "simple realist" view that the world is just out there, to embrace the idea that there are many parallel worlds just out there, each waiting to be conjured into existence by some independent agent's choice of action. This conception is neither simple nor realist.

3. INSTRUMENTALISM

As long as counterfactual models are used only instrumentally, in accordance with Jeffreys's law, for empirically meaningful purposes, my objections to them are reasonably muted. Although I do say in the article that "I remain to be persuaded" of their usefulness for inferential purposes, I cannot definitively rule this out.

Wasserman, and Robins and Greenland, essentially confine counterfactual models to instrumental use, arguing that it is merely necessary to ensure that one does not attempt inference about nonidentifiable (or, more correctly, empirically nondeterminable) aspects of the model. They regard it as a selling point for counterfactual modeling that it allows one to make this distinction between its terms. Well, maybe so. But it is a distinction that has eluded some extremely able statisticians, such as Neyman, and Wilk and Kempthorne, and has many extremely subtle aspects (as discussed, for example, at the end of my Sec. 9.1). I would prefer to build on firmer ground than this, using models that do not allow empirically meaningless statements and inferences, whenever this is possible (which I currently believe is always).

Cox points out that a counterfactual model incorporating TUA implies the observable property that the distributions in different treatment groups differ only by translation, and regards this as a good reason to care about that property. But if the translation property holds for one scale of measurement, it will not do so for another (e.g., after taking logarithms), and it beggars belief that I have been lucky enough to find the unique scale on which this property holds. Or if one works with counterfactuals, why should I imagine that I have found the unique scale on which TUA holds? In either case, models such as these need to be regarded as crude jobbing assumptions, rather than believable assertions about reality.

Together with Cox (and perhaps Hume), Casella and Schwartz regard the deterministic formulation of TUA as "just a convenient simplification," the real action being at the population-average level. It is certainly true that if one is careful, one can do sensible analyses at this level using a model at the individual level that one does not even pretend to believe. If this approach could never lead one into trouble, then I would have no real objection. However, in the light of the arguments in my Sections 9.1 and 9.2, I feel that it passes over treacherous quicksands.

Cox and Pearl argue that a rigorous positivist approach would have excluded many of the most important and successful scientific theories and advances. Casella and Schwartz observe that features such as elegance and simplicity are important in a theory (although they miss the irony in their extract from my 1976 Barndorff-Nielsen discussion). Certainly it has been fruitful to incorporate terms for unobserved entities, such as quarks, into scientific theories. Sometimes these terms are purely instrumental, allowing for a simpler and more elegant reformulation of a theory, without any observable consequences. Sometimes they widen the scope of application of the theory. And sometimes they allow one to make new predictions, which can in principle be checked. However, I do not feel that the counterfactual approach to causal inference has, as yet, provided any of these advantages.

These discussants implicitly suppose that counterfactual language is strictly richer than decision-analytic language. Pearl says "Giving up this richness is the price we would pay for Dawid's insurance." I would instead regard this additional "richness" as a dangerous embarrassment of riches, as the only new possibilities for "inference" that it opens up have no empirical content and should be avoided. However, in some respects decision-analytical language is richer than counterfactual language. For example, as soon as one posits the existence of "the value that Y would take were the unit to receive treatment t," one has constructed a rigid and unbreakable link between observational and interventional

situations. When this is not appropriate, a counterfactual description becomes impossible. A decision-analytic approach can model much more flexible relationships between these situations without difficulty.

4. PARTIAL COMPLIANCE

Pearl and Wasserman dispute my claims in Section 10.2 that counterfactuals are unnecessary and unhelpful for analyzing the problem of imperfect treatment compliance. Wasserman's question "whether most people could have derived the bounds without counterfactual reasoning" is (if not itself counterfactual) of no logical importance. The point is that there are mathematical techniques, at least as simple and straightforward as those used by Balke and Pearl, that produce the required inequalities as outputs without requiring counterfactuals as inputs. Pearl's assertion that "when we examine the conditional probabilities that achieve those bounds, we find that they represent subjects with deterministic behaviour" would be of doubtful significance if true, but is in any case false, as I have shown by a simple counterexample that Pearl has not disputed. Further details of both these points will be submitted for publication in due course.

5. COMPETING RISKS

Rubin and Robins and Greenland both raise the same problem of causal inference in a competing risks framework. Both contributions argue that some kind of counterfactual modeling (perhaps not deterministic) is required to yield "meaningful causal contrasts," even in a fully randomized setting when we care only about effects of causes. I am not sure that I have fully absorbed the Robins–Greenland analysis, which proceeds through several stages of refinement, starting with, but rapidly leaving behind, that suggested by Rubin. My own attempt at understanding their final suggestion, for the "fully stochastic" case, leads to $\Phi_{tc} = E\{E(Z|Y = 1, K, t) - E(Z|Y = 1, K, c)\}$, where K denotes the covariate ("level of detail") conditional on which the assumption of "pure randomness" for the Y's is being invoked. This can be estimated from experiments in which K is measured, but its value will depend on the specification of K, which I find troubling. (As I have attempted to explain in Sec. 1 here, I cannot accept the idea of an ultimate, or uniquely appropriate, level of detail, which might impart unambiguous "objectivity" to the parameters (θ_t, θ_c) figuring in their account; and even if I could, I would have no idea how to make meaningful statements, or express meaningful opinions, about them.) In particular, in the fully randomized setting I am not sure what considerations should prevent me from simply taking K to be trivial, thus obtaining $\Phi_{tc} = E(Z|Y = 1, t) - E(Z|Y = 1, c)$—at any rate, this is easily estimated.

Both the Rubin and Robins–Greenland approaches lead to inferences sensitive to arbitrary and untestable features of their counterfactual models, as well as to arbitrary magical ingredients (such as "objective probabilities") in fairy stories cleverly disguised as mathematics. They thus fall squarely into my "goat" category. (For the record, and using the Robins–Greenland notation, my own attitude is that, as there is no difficulty in determining an empirically meaningful probability structure for the observable (Y, Z) given treatment—even though this is defined over an unusual space, where Z automatically takes the value 'undefined' whenever $Y = 0$—why should one create a problem where none exists? The real problem is how to define a sensible utility measure on this outcome space.)

6. VAGUENESS AND CLARITY

Robins and Greenland admit that counterfactuals are subject to inherent vagueness. Now there are at least three distinct ways in which vagueness might enter into causal analyses. The first is because the way in which the theoretical terms in the model are to be formally combined and manipulated has not been clearly defined or understood. Counterfactual models do not usually suffer from this problem. However, I feel that this purely mathematical clarity all too often has been misinterpreted as all that is required, when instead it is the clarity of the relationship between the theory and the world that is of far more importance.

A common cause of vagueness at this interpretive level is simple sloppiness: not making clear the intended relationship between terms in the model and features of the external world, or questions being put to the world. Vagueness of this sort is easily overcome. Shafer's example, of the difficulty of identifying the causal effect of taking aspirin when the outcome also depends on the taking of water, is relevant. Questions of the type "was it because . . .?" are just too vague to be meaningful, and need to be pinned down further by rephrasing the causal query in whatever way appears most relevant to situation at hand (e.g., one might decide to compare the actual outcome with that pursuant to taking the water without the aspirin).

The third, and most insidious, form of vagueness is when it is simply not logically possible to give a clear external interpretation for some of the terms in the models. I find this vagueness pervading most talk of "potential responses," and particularly apparent in the analyses of the competing risks problem in Section 5 here. One does not have to be a card-carrying logical positivist to ask: "What exactly is it that you think you are talking about?" I am left quite breathless by Robins and Greenland's conclusion (in the face of cogent contrary evidence that they themselves have presented, both here and elsewhere) that counterfactuals constitute "a powerful tool for eliminating vagueness."

7. CAUSES OF EFFECTS

I am surprised at how little of the discussion relates to my suggestions for inference about "causes of effects," which I had expected to be the most controversial. Perhaps this is because, as Shafer points out with some disgust, my analysis here already concedes a good deal to the counterfactual school.

I am puzzled by Pearl's account of what he understands by counterfactuals. Like me, he seems to care that our theoretical inferences be closely linked to the empirical world. This leads him to interpret his Q_{II} in terms of a prediction

of the response to aspirin in the *next* headache episode, given information about what happened in the current one. I have no objection to thinking about this, and do not consider that there is anything counterfactual about it. I wonder if the other discussants do? My own understanding of a counterfactual outcome is that it refers to the *same* unit (in this case, headache episode) that has already been observed, under (purely hypothetical) different treatment conditions. Pearl attempts to link these two distinct ideas by means of an assumption that "a person's characteristics do not change over time" (this is related to what I have termed "uniformity"). Are we to infer that all my headaches will respond in exactly the same way to the same treatment? This assumption is indeed "testable," and not just "in principle." I would normally expect it to be falsified. How would Pearl proceed then?

Cox does not see the need for a sharp distinction between the treatments of effects of causes and of causes of effects. He draws attention to a number of practical differences, but I do not see that these are relevant at the more philosophical level of my own treatment. Put simply, at this level the fundamental difference is that inference about causes of effects is necessarily sensitive to arbitrary assumptions about the joint distribution of the metaphysical collection of potential responses, whereas these can be avoided for inference about effects of causes.

Shafer is the only discussant to seriously address my tentative suggestions regarding inference for causes of effects, and he is merciless in searching out their weak points and driving a dagger into them. The more I think about these issues, the more I am tempted to accept Shafer's strictures and disown my own suggestions! But perhaps something can be rescued.

Because the specification of "context" is essential to the implementation of my approach, if that approach is to have any applicability and usefulness it must be supplemented with rules and reasons for selecting one context rather than another. Those reasons will normally combine both scientific understanding of the world and more "political" concerns.

With regard to the scientific aspects, I would usually be happy to answer "yes" to the question: "Had I brought my umbrella with me today, would it still have rained?," because, even though meteorology is a very inexact science, I believe that (notwithstanding chaos theory and the butterfly effect) enough is known about its uniformities to justify this assertion. When Shafer asks: "Had I instead bet on tails, would the toss come out the same way?," the "obvious" answer is again "yes." But my call might have been delayed a little longer, the coin tosser's grasp on the coin might have changed, or other influential factors might have altered. So the obvious answer is not scientifically determined, but rather involves one's own, essentially arbitrary, conception of what parallel universes are relevant.

The "political" dimension to the choice of context is particularly clear in the case of the United States Government versus the tobacco industry. Shafer introduces the counterfactual question: "How much less would the government have spent on caring for the ill had the companies not taken these actions?" I agree with him that this question compounds scientific uncertainty with nonscientific arbitrariness. But (as Shafer seems to grant) there may be nonscientific but nevertheless compelling reasons to narrow down that arbitrariness in particular ways. For example, the argument has been put that had the tobacco companies discouraged smoking when they first had evidence of its dangers, more people would have given it up, leading them to live longer and thereby end up as a *greater* overall expense on the government. There have been attempts to have this line of argument ruled out by the courts. In such a case one would be obliged, for legal rather than scientific reasons, to attempt a comparison of the health-care expenditure in the real world with that in a parallel world in which (say) people all lived for the same length of time, but did not suffer from smoking-related illness. If one can describe this intended alternative world clearly enough, specifying exactly which variables in the two worlds should be regarded as identical and which as conditionally independent, one can proceed to make inference about causes of effects. But it must be very clear just how far such inference is from being based on scientific understanding or interest alone.

Past Conditionals. Shafer contrasts the aforementioned counterfactual question with what he terms a "past conditional," which is a historical exercise in looking forward. The distinction is important, but I believe that both kinds of question have their place. In thinking about the tobacco companies' culpability, one should try to put oneself into their shoes at the time, and ask whether, in light of what they then knew or could reasonably ascertain, the decisions they made were legal, ethical, and prudent. However, conditional on their culpability, their liability for damages should depend on the (then unknown) way in which things turn out between that time and now, in both the real and the relevant counterfactual universe. One cannot always expect to be insulated from the consequences of one's actions, even when one could not reasonably predict what those would be.

8. FINAL COMMENTS

After this lengthy article, discussion, and rejoinder, it may be helpful for the reader to attempt to plot each contributor in a multivariate space of attitudes to counterfactuals, having the following dimensions:

Fact–Fiction. Are counterfactuals to be regarded as genuine features of the external world, or are they purely theoretical terms?

Real–Instrumental. Can any inferences based on counterfactuals be allowed, or should they be restricted to those that could in principle be formulated without mention of counterfactuals?

Clear–Vague. Do counterfactual terms in a model have a clear relationship with meaningful aspects of the problem addressed? Can counterfactual constructions and arguments help to clarify understanding?

Helpful–Dangerous. Can use of counterfactuals streamline thinking and assist analyses, or do they promote misleading lines of argument and false conclusions?

Clearly, I am an outlier from most of the discussants on most of these dimensions—and I must confess that my position in the space has hardly budged at all (at any rate, in the directions it was meant to) in response to the discussion. I fear that my rejoinder may have an equally contrary effect on the discussants. But the interchange will have been worthwhile if it encourages even a few readers, coming to these issues fresh and without preconceptions, to pay serious attention to the arguments underpinning the various views exchanged, before settling comfortably and immovably into their own preferred positions on the use of counterfactuals for causal inference.

ADDITIONAL REFERENCE

Dawid, A. P. (1982), "The Well-Calibrated Bayesian" (with discussion), *Journal of the American Statistical Association*, 77, 604–613.