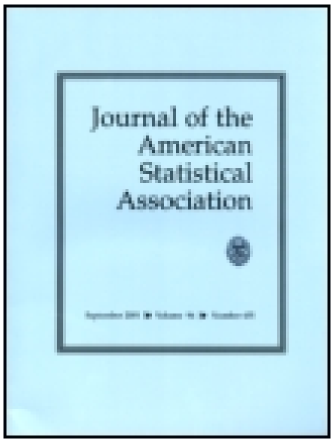

Journal of the American Statistical Association

Journal of the American Statistical Association

Publication details, including instructions for authors and subscription information: http://www.tandfonline.com/loi/uasa20

Causal Inference without Counterfactuals

A. P. Dawid [a]

[a] Department of Statistical Science , University College London , London , WC1E 6BT , U.K.

Published online: 17 Feb 2012.

To cite this article: A. P. Dawid (2000) Causal Inference without Counterfactuals, Journal of the American Statistical Association, 95:450, 407-424

To link to this article: http://dx.doi.org/10.1080/01621459.2000.10474210

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 treatment 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 $u_{i1}, u_{i2} \ldots, u_{in_i}$. Then the observed response on unit u_{ij} is $X_{ij} := Y_i(u_{ij})$. The collection $(X_{ij}\colon i = t, c; j = 1, \ldots, n_i)$ 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, τ = 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.

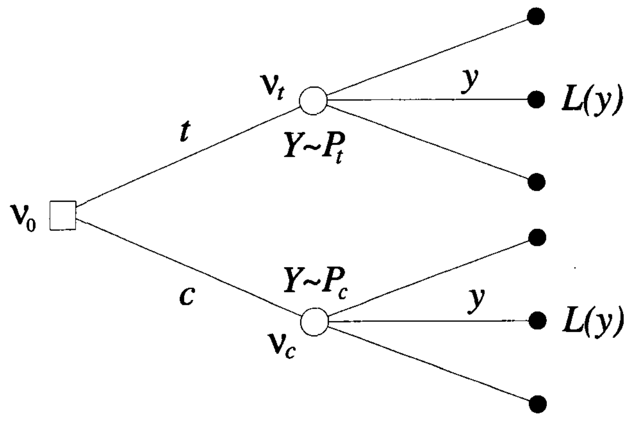

Figure 1. Decision Tree.

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, $\text{var}(\alpha_i) = \sigma_\alpha^2$, 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

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

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

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

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 $\rho_{ct\cdot K} \geq 0$)

$$\gamma_K \leq \rho \leq 1 \qquad (20)$$

(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 involved, 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