Statistical Causality from a Decision-Theoretic Perspective

A. Philip Dawid

Centre for Mathematical Sciences, University of Cambridge, Cambridge CB3 0WB, United Kingdom; email: apd@statslab.cam.ac.uk

Annu. Rev. Stat. Appl. 2015. 2:273–303

First published online as a Review in Advance on November 12, 2014

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-010814-020105

Keywords

conditional independence, confounding, directed acyclic graph, dynamic treatment strategy, effect of treatment on the treated, moralization

Abstract

We present an overview of the decision-theoretic framework of statistical causality, which is well suited for formulating and solving problems of determining the effects of applied causes. The approach is described in detail, and it is related to and contrasted with other current formulations, such as structural equation models and potential responses. Topics and applications covered include confounding, the effect of treatment on the treated, instrumental variables, and dynamic treatment strategies.

1. INTRODUCTION

After decades of neglect, recent years have seen a flowering of the field of statistical causality, with an impressive array of both theoretical and applied developments. As yet, however, there is no one fully accepted foundational basis for this enterprise. Rather, there is a variety of formal and informal frameworks for framing and understanding causal questions, as well as hot discussion about their relationships, merits, and demerits: We might mention, among others, structural equation modeling and path analysis (Wright 1921), potential response models (Rubin 1978), functional models (Pearl 2009), and various forms of graphical representation. This plethora of putative foundations leaves statistical causality in much the same state of confusion as probability theory before Kolmogorov.

This review aims to add to the confusion by describing one particular approach, based on decision-theoretic principles, that its author considers superior to the others (in this of course he is fully aligned with others' attitudes toward their own works). In it, I present the main features of the approach, relate it to and compare it with some other approaches, and show how it works in some simple applications. I consider this approach more straightforward philosophically and mathematically (it requires only a very small extension to standard statistical methods), and easier to comprehend and manipulate, than other approaches that introduce new ingredients and structures, such as potential responses or deterministic functional relationships. Although I do not expect wholesale conversion to this point of view, I hope that readers already knowledgeable in causal inference will, at the very least, find it helpful to look at familiar topics through a fresh pair of spectacles.

For further details and developments of the material in this article, the reader is referred to Berzuini et al. 2012b,c; Dawid 2000, 2002, 2003, 2007a, 2010a,b, 2011, 2012; Dawid & Constantinou 2014; Dawid & Didelez 2008, 2010, 2012; Didelez et al. 2006; Geneletti & Dawid 2011; and Guo & Dawid 2010. The lecture notes by Dawid (2007b) contain a fuller exposition of the decision-theoretic approach.

1.1. Causality and Agency

The concept of causality has been a focus of interest for philosophers for millennia, but, as befits any worthwhile philosophical conundrum, this attention has not resulted in a settled approach to understanding it. In a previous article (Dawid 2010b), I reviewed a variety of philosophical conceptions and theories, focusing particularly on that which is most germane to my own approach: the agency theory of causality (Hausman 1998; Price 1991; Woodward 2003, 2013). This theory interprets causality as being all about how an external manipulation would affect a system: for example, how the quality of a chemical product would respond to adjustments of the lever that controls the pressure in the production process. Much of statistical science—in particular, the whole subfield of experimental design—aims to address exactly these kinds of questions about the effects of interventions on a system. Such questions are indeed a major object of all scientific inquiry.

An important advantage of the agency approach is its clear separation of cause and effect variables, which results in the elimination of definitional ambiguities associated with the possibility of reverse causation or common causes. But having clean definitions is not enough for practical purposes: We must be able to relate those definitions to properties of the empirical world. Whereas drawing such relationships is relatively unproblematic in cases involving genuine experimentation, doing so becomes a major headache when we can observe a system only in its natural habitat and are unable to apply to it the interventions that are essential to understanding its causal properties. The importance of making a clear distinction between intervening and

merely observing has been stressed by numerous authors, including Meek & Glymour (1994), Pearl (2009), Rubin (1978), and Spirtes et al. (2000). Most of the recent emphasis of the statistical causality enterprise focuses on observational situations: Such work attempts to identify conditions under which one can extract causal conclusions from these situations and to develop techniques for doing so.

1.2. Effects of Causes and Causes of Effects

The emphasis in this work is on the problem of assessing the future effects of a contemplated intervention in a system: That is, it focuses on identifying the effects of causes (EoC). An entirely different problem is that of judging what might have been the cause of an observed outcome: that is to say, identifying the causes of effects (CoE). This is very much an issue in legal cases, which may seek to assign responsibility or blame.

Most of the effort to date in statistical causality has focused on CoE problems, an important exception being Pearl (2009), who explores both EoC and CoE problems in detail. However, whereas Pearl and others who have dealt with CoE have used identical mathematical machinery (in Pearl's case, based on assumed functional relationships) for both EoC and CoE, I do not consider this approach to be appropriate. I discussed the relationships and differences between EoC and CoE problems, as well as their requisite infrastructures, in some detail in another article (Dawid 2000), in which I argued that, whereas (unlike for EoC) some form of counterfactual logic appears unavoidable for assessing CoE, the difficulties in appropriately modeling a CoE problem have been underappreciated. Some discussion and analysis of CoE issues in the context of using epidemiological data to address a legal case of toxic tort can be found in articles by Dawid (2011) and Dawid et al. (2013, 2014).

1.3. Article Overview

In Section 2, I introduce a simple example that locates statistical causality firmly within the purview of classical statistical decision analysis. Section 3 then explores some variant formulations of this problem, including structural equations and potential responses. In Section 4, these approaches are explored and compared in the familiar context of statistical experimental design. Section 5 moves the discussion on to the more problematic context of causal inference from observational data and explores the meaning and representation of the important concept of no confounding: From a decision-theoretic approach, this concept is usefully described in terms of relationships between different regimes—e.g., interventional or observational—under which data can, at least in principle, be gathered. I show how this and similar requisite properties can be usefully expressed and manipulated in terms of an extension of the probabilistic notion of conditional independence to allow for both stochastic and nonstochastic variables. Section 6 develops the associated algebraic and graphical theory. In Section 7, I introduce influence diagrams as useful graphical representations of causal problems and relate these diagrams to the use of directed acyclic graph (DAG) representations as described by Pearl (2009). The remainder of the article explores, from the decision-theoretic perspective, a number of important special applications. Section 9 examines the observational identification of causal effects using sufficient covariates, propensity analysis and *do*-calculus, and the possibility of identifying the effect of treatment on the treated; Section 10 considers the use of instrumental variables; and Section 11 treats problems in which a sequence of actions can be applied over time in response to intermediate observations and outcomes. The discussion in Section 12 concludes by expressing some skepticisms about currently popular approaches that make unavoidable use of counterfactual reasoning.

2. A DECISION PROBLEM

I have a headache and am considering whether or not to take two aspirin tablets. It is generally accepted that aspirin has a beneficial effect on headaches: In some sense—and our task is to try to make the following statement more precise—taking aspirin causes headaches to get better faster. The solution to my decision problem is thus intimately bound up with the cause–effect relationship of aspirin on headaches.

This observation leads naturally on to a suspicion that at least some part (specifically, that which aims to understand the EoC; CoE problems require a different approach) of the enterprise of statistical causality might be fruitfully recast as a special application of standard statistical decision analysis. This point of view is not a currently popular one, however, and indeed there is a variety of other approaches to interpreting causality in a statistical setting. This article considers the relationships, similarities, and differences between these approaches, and in it I hope to demonstrate that the decision-theoretic approach is more natural, more straightforward, and more useful than its competitors.

To formulate the decision problem, let the binary variable X denote whether I take aspirin ($X = 1$) or not ($X = 0$), and let Y be the log time it takes for my headache to go away. I myself can choose X: It is a decision variable and does not have a probability distribution. Nevertheless, it is still meaningful to consider my conditional distribution P_x for how the eventual response Y will turn out, given that I choose $X = x$. For the moment, we assume that the distributions P_0 and P_1 are known. Where we need to be definite, we (purely for simplicity) take them to have the following normal probability density function:

$$p(y \mid X = x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp -\frac{(y - \mu_x)^2}{2\sigma^2}, \tag{1}$$

with mean μ_0 or μ_1 for $x = 0$ or 1, respectively, and variance σ^2 in either case.

We can interpret the distribution P_1 (respectively, P_0) as expressing my hypothetical uncertainty about Y, if I were to decide on action $X = 1$ (respectively, $X = 0$). These distributions can incorporate various sources and types of uncertainty, including stochastic effects of external influences arising and acting between the points of treatment application and eventual response. We need only the distributions P_1 and P_0 to address my decision problem: I simply compare the two different hypothetical distributions for Y, decide which I prefer, and make the associated decision.

One possible comparison of P_1 and P_0 might be in terms their respective means, μ_1 and μ_0. The effect of taking aspirin, rather than taking nothing, might then be quantified as the change in the expected response, $\delta := \mu_1 - \mu_0$. Alternatively, we might look at the difference of the means of $Z = e^Y$ under the two possible treatments, $e^{\sigma^2/2}(e^{\mu_1} - e^{\mu_2})$, or compare the variances of Z under the two treatments. Any such comparison of an appropriately chosen feature of the two hypothetical distributions of Y can be regarded as a summary of the causal effect of taking aspirin (relative to taking nothing).

More formally, we might apply statistical decision analysis (see, for example, Raiffa 1968) to structure and solve this decision problem. Suppose that I quantify the loss that I will suffer if my headache lasts y minutes using a real-valued loss function, $L(y)$. If I were to take the aspirin, my expected loss would be $E_{Y\sim P_1}\{L(Y)\}$; if not, it would be $E_{Y\sim P_0}\{L(Y)\}$. The principles of statistical decision analysis now direct me to choose the treatment leading to the smaller expected loss. A trivial but fundamentally important point is that, whatever loss function is used, this solution involves only the two hypothetical distributions P_1 and P_0 for Y, conditional on taking either action. The effect of treatment might be measured by the reduction in expected loss, $E_{P_0}\{L(Y)\} - E_{P_1}\{L(Y)\}$, and the correct decision is to take aspirin just when this reduction

is positive. Although there is no uniquely appropriate measure of the effect of treatment, the rest of this article focuses for simplicity on the difference of the means of the two hypothetical distributions, $\delta := E_{P_1}(Y) - E_{P_0}(Y)$.

3. ALTERNATIVE FORMULATIONS

3.1. The Decision-Theoretic Model

The essential ingredients of the decision-theoretic analysis above were the two hypothetical distributions for Y, conditional on setting $X = 0$ or $X = 1$. For simplicity only, we have specialized these distributions to be normal:

$$Y \mid X = x \sim \mathcal{N}(\mu_x, \sigma^2). \tag{2}$$

We refer to this formulation as a stochastic or decision-theoretic (DT) model. In it, the term average causal effect (ACE) simply denotes the difference of the means of the two hypothetical distributions for Y, $\mu_1 - \mu_0$.

3.2. The Simple Structural Equation Model

The assumptions of the distributional specification 2 are often expressed in the following alternative way:

$$Y = \mu_X + E, \tag{3}$$

where

$$E \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

and it is implicit that the error E is independent of X. A system of equations similar to equation 3, which may contain hundreds of relationships representing response (endogenous) variables as functions of other (endogenous and exogenous) variables as well as of external error variables such as E, together with associated explicit or implicit assumptions about the joint distribution of the error terms, constitutes a simple structural equation (SSE) model. Such models are popular in econometrics and other fields as representations of causal structures.

The assumptions of the SSE model (equation 3) clearly imply the distributional properties of the DT model (equation 1). Does this mean that the SSE and DT models are equivalent? No: Making this assumption would mean ignoring the additional algebraic structure of equation 3, whereby Y is represented as a deterministic mathematical function of the two variables X and E. Unlike the distributional formulation of equation 1, in equation 3 all of the uncertainty is compressed into the single variable E, via equation 4. If we take equation 3 and its ingredients seriously, we can get more out of it.

It is common, and indeed seems very natural, to interpret equation 3 as follows: The values of X and E are assigned separately (by the decision maker and by nature, respectively), and Y is then determined by the equation. In this case, given that E takes value e, then Y will take the value $y_x := \mu_x + e$ if I set X to x. That is, I will then observe the variable $Y_x := \mu_x + E$. We can regard Y_x as the potential response to the hypothetical setting $X = x$. It will become the actual response, Y, when indeed $X = x$: Thus, $Y = Y_X = \mu_X + E$. Note that, in this interpretation, when in fact I set $X = 1$, the counterfactual (because it is predicated on a hypothesis that runs counter to known facts) response $Y_0 = \mu_0 + E$ to $X = 0$ is still a well-defined function of the ingredients of the model given by equation 3.

Given the above interpretations, switching my decision from $X = 1$ to $X = 0$ would result in Y switching from $Y_1 = \mu_1 + E$ to $Y_0 = \mu_0 + E$, with the identical E. The causal effect of

this switch might then be measured by $Y_1 - Y_0$, which in this case is the constant $\mu_1 - \mu_0$. This comparison is purely algebraic and is unrelated to the stochastic properties of the model. It may be termed an individual causal effect (ICE). Note that all of these manipulations rely fundamentally on the implicit assumption that the value of E remains fixed, regardless of which decision I make. Such an assumption simply has no counterpart in the DT model (equation 2).

3.3. The Extended Structural Equation Model

An extension of the SSE model (equation 3) is given by the following expression:

$$Y = \mu_X + E_X, \tag{5}$$

where we now have a pair of error variables, $\boldsymbol{E} = (E_0, E_1)$, that has a bivariate distribution and is assumed to be independent of X. In particular, when $X = 0$ we have $Y = Y_0 := \mu_0 + E_0$, with E_0 having its initially assigned distribution. Similarly, we have $Y = Y_1 := \mu_1 + E_1$ when $X = 1$. And $Y = Y_X$.

Suppose we model the pair of errors $\boldsymbol{E}$ as being bivariate normal as follows:

$$\boldsymbol{E} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{6}$$

where $\boldsymbol{\Sigma}(2 \times 2)$ has diagonal entries σ^2 and off-diagonal entries $\rho\sigma^2$. Then (no matter what the correlation ρ may be), we obtain the DT model (equation 1) for Y given X. But again, if we take the algebraic structure seriously, and if we further suppose that the value of $\boldsymbol{E}$ is unaffected by the choice made for X, we can define potential responses $Y_x := \mu_x + E_x$, as well as the ICE, $Y_1 - Y_0$, which in this case is a random quantity, $\mu_1 - \mu_0 + E_1 - E_0$.

It is important to note that the relationship between the extended structural equation (ESE) model and the induced DT model is many to one: The dependence structure (here embodied in the correlation ρ) does not enter into the induced DT model.

3.4. The Potential Response Model

As seen above in Sections 3.2 and 3.3, starting from a structural equation model (simple or extended) allows us to define the pair of potential responses $\boldsymbol{Y} = (Y_0, Y_1)$ and derive its bivariate distribution in terms of the ingredients of that model. For the ESE model given by equation 5 and distributional assumption 6, the implied distribution of $\boldsymbol{Y}$ is

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{7}$$

where $\boldsymbol{\mu} := (\mu_0, \mu_1)$. Both the value and the distribution of $\boldsymbol{Y}$ are regarded as independent of the applied treatment X.

We might alternatively start at this point, simply taking the pair $\boldsymbol{Y}$ as a primitive ingredient of our model with a bivariate distribution (for example, that of equation 7) and again assuming that we can change the value of X without changing either the value or distribution of $\boldsymbol{Y}$. This is the general potential response (PR) model. The underlying philosophical conception is that both potential responses Y_0 and Y_1 are real and coexist—even though it is logically impossible to observe both of them together.

If we start from a PR model, we can recover a DT model: In particular, if we start with the bivariate distribution 7, we recover the marginal distributions 2. But, again, this relationship is many to one.

3.5. The Functional Model

Mathematically, the models introduced in Sections 3.2, 3.3, and 3.4 all have the following common functional form:

$$Y = f(X, U), \tag{8}$$

where X is a decision variable that represents the cause of interest; Y is the effect of interest; U is an additional extraneous random variable, the value and distribution of which are taken as being independent of X; and f is a deterministic function of its arguments.

In the ESE model given by equation 5, we can take $U = \boldsymbol{E}$ and $f(x, (e_0, e_1)) = \mu_x + e_x$; the SSE model of equation 3 is the degenerate case of this having $U = E$ and $f(x, e) = \mu_x + e$.

In the case of a PR model, we can formally take U to be the pair (Y_0, Y_1) and the function f to be given by

$$f(x, (y_0, y_1)) = y_x. \tag{9}$$

In all of the above applications, the variable U typically represents a somewhat imaginary quantity that does not correspond to any variable observable in the empirical world. Indeed, if U were required to correspond to a real-world variable, the application of functional models would be limited to the very special situation of complete determinism, in which the pair of real variables (X, U) fully determines Y.

A general functional model of the form given in equation 8 is mathematically equivalent to a PR model, on defining $Y_0 = f(0, U)$, $Y_1 = f(1, U)$. We thus see that (mathematically if not necessarily in terms of their interpretation) PR models, ESE models, and general functional models need not be distinguished. Further, any functional model determines a DT model: Under the model given by equation 8, the relevant distribution of Y given $X = x$ is simply the marginal distribution of $f(x, U)$. Conversely, given any DT model $Y \mid X = x \sim P_x$, we can construct a functional model corresponding to it. One simple way is as a PR model (equation 9), in which the marginal distribution of Y_x is P_x. However, in contrast to the essentially unique cross-correspondence between the other models considered above, the functional representation of a DT model is far from unique—as can again be seen, for example, from the arbitrariness of the dependence parameter ρ in the PR representation of the stochastic model; this parameter can never be identified from data.

4. CAUSAL INFERENCE FROM EXPERIMENTAL STUDIES

Having set out a variety of formulations of my basic decision problem, I now address the following question: How might I gather and use data to help identify the required ingredients? From the DT perspective, I need to assess my hypothetical distributions for Y, P_0 and P_1, under either treatment choice. These assessments should be informed by (conditioned on) whatever relevant information I may have, such as, for example, the responses observed for similar headaches (my own or those of other people) that received one or the other of the two treatments.

We initially restrict attention to the simplest case. Suppose I can observe two groups of people. Each group consists of individuals I can regard as similar to (technically, exchangeable with) me in all features relevant to their development of headaches and reaction to treatment. All members of the first group are assigned the active treatment (aspirin), $X = 1$, whereas members of the second

group receive the control treatment (no aspirin), $X = 0$.[1] Finally, I observe the responses of all individuals: Let Y_{xi} denote the response of the ith individual receiving treatment x.

4.1. The Decision-Theoretic Approach

Under the above assumptions, I can model the responses of the treated individuals as being randomly drawn from P_1, and, similarly, I can model the responses of the untreated individuals as being drawn from P_0. I can then use completely standard statistical methods to estimate and compare (in any way I choose) the two distributions P_1 and P_0. In particular, I have access to all of the ingredients required for my stochastic decision problem.

For example, under the distributional model 2, we would take

$$Y_{xi} \sim \mathcal{N}(\mu_x, \sigma^2) \tag{10}$$

all independently and can then base inference about the difference $\delta = \mu_1 - \mu_0$ on Student's t-distribution. This is the bread and butter of the most elementary courses in statistics. Here, however, we have emphasized—as may be done more rarely—the assumptions needed to justify the relevance of this inference to the decision problem I face.

4.2. Other Approaches

Suppose we now take the ESE model (equation 5) seriously [this model includes the simpler SSE model (equation 3) as the special case $E_0 = E_1$ or, equivalently, $\rho = 1$]. Because that model implies all of the distributional properties of equation 2, I can still do all that I did in Section 4.1, thereby using the data to help solve my decision problem. Now, however, I might want to do more. For example, I might want to say something about the distribution of my individual causal effect, ICE $= \mu_1 - \mu_0 + E_1 - E_0$. Can I use the data to help me do this?

The mean of the ICE is $\delta = \mu_1 - \mu_0$, which I can estimate. But its variance is $2(1 - \rho)\sigma^2$. Although I can estimate σ^2 from my data, the dependence on the correlation ρ is problematic: I could estimate ρ only if I had observations on the bivariate pairs $\boldsymbol{Y} = (Y_0, Y_1)$, thus observing both potential outcomes for the same individual. Because each individual receives just one of the two treatments, the full pair $\boldsymbol{Y}$ is, logically, never fully observable, so there is simply no way I can estimate ρ. In particular, I have no way of distinguishing observationally between the general ESE model (equation 5) and the SSE model (equation 3). However, these models have different implications for var(ICE), which is $2(1 - \rho)\sigma^2$ for the ESE model and 0 for the SSE model. In a similar fashion, if I were interested in the mean of the ratio ICE, namely $\mathrm{E}(Y_1/Y_0)$, or in the estimate of my ICE after having taken aspirin and observed response $Y = y$, which is $\mathrm{E}(Y_1 - Y_0 \mid X = 1, Y = y) = (1 - \rho)y + (\rho\mu_1 - \mu_0)$, these seemingly innocuous queries could not be addressed by any data I could ever collect, as they all depend on the unknowable value of ρ.

As the ESE model is a special case of a PR model or a general functional model, all of the above caveats also apply to those models. We can thus divide putative causal inferences from such a model into "sheep" and "goats." Sheep are inferences that depend on only the marginal distributions, P_0 and P_1, of the individual potential responses, Y_0 and Y_1, and are thus identifiable

[1]The usual operational method is first to form a single group of individuals who are like me and then to randomize the assignment of treatment to its individuals. Then both the resulting treatment and control groups will still be exchangeable with me on their pretreatment characteristics. The second stage by itself ensures internal validity: The treated and untreated groups should be comparable with each other. Without the first stage, however, we do not have external validity, which permits generalization beyond the data to external cases of interest—in this case, myself.

from data. In contrast, goats do not have this simple dependency and so are not identifiable. For example, any putative causal inference that makes essential use of var(ICE) is a goat.

This distinction is all too easy to neglect, resulting in attempts to make goat-like inferences. Given a fully specified ESE, PR, or functional model, making such inferences will be mathematically possible, and one may not notice that the answer would be different for a mathematically distinct model that is observationally entirely equivalent to this one because it has the same marginal distributions. Such an apparent causal inference is, to say the least, misleading.

4.3. Treatment-Unit Additivity

The ESE model (equation 5) would represent the data as

$$Y_{xi} = \mu_x + E_{xi}. \tag{11}$$

For the special case of an SSE model (equation 3), this representation becomes

$$Y_{xi} = \mu_x + E_i, \tag{12}$$

a sum of one term, μ_x, that depends only on the treatment x applied, and another term, E_i, that depends only on the unit i to which it is applied. This property is termed treatment-unit additivity (TUA) and is entirely equivalent to the property that the ICE, $Y_1 - Y_0$, has the identical value (namely, $\mu_1 - \mu_0$) across all individuals.

One reason we might like the TUA assumption is as follows: So far, I have had to assume that the individuals on whom I have data are like me—in particular, that we all have the same distribution for our error term E. But this assumption is often unrealistic. For example, a clinical trial typically has stringent recruitment criteria that I would not satisfy. I can, however, relax the model given by equation 12 in such a way as to not require that my own E be drawn from the same distribution as the (E_i) in the data (for example, my E could have a higher mean or variance). But because my own ICE is still $\mu_1 - \mu_0$, which is estimable from the data, causal inference about it is unaffected by this relaxation of the SSE model (although it does still rely on TUA).

An alternative DT analysis, which thus does not require TUA, is as follows [see section 8.1 of Dawid (2000) for a more detailed account]: Because of different selection criteria, my personal hypothetical distribution P_x for my response Y if I take treatment $X = x$ is allowed to differ from P_x^*, the distribution of Y for the individuals in my data set who receive treatment $X = x$. But suppose (for example) I can model the mean μ_x of my distribution P_x as related to the mean μ_x^* of P_x^* by $\mu_x = \mu_x^* + \gamma$ for some γ that does not depend on x. Then my own ACE, $\mu_1 - \mu_0 = \mu_1^* - \mu_0^*$, can be estimated from the data.

4.3.1. Neyman and Fisher. As there is no observational way of distinguishing between the SSE model, for which TUA holds, and the more general ESE model, for which it does not, the arguments in Section 4.2 would classify any attempt at causal inference that is dependent on the assumption of TUA as a goat. In this light, it is interesting to revisit the ill-tempered debate between Neyman and Fisher presented in the discussion of the paper by Neyman (1935).

The essence of Neyman's model[2] involves an experiment in which various treatments $t = 1, \ldots, T$ can be applied to various experimental units $u = 1, \ldots, U$ (for example, plots in a field) that might themselves be described in more detail [for example, in a randomized blocks layout $u = (i, j)$ $(i = 1, \ldots . I; j = 1, \ldots, J)$ for the jth plot in the ith row]. The treatments are applied

[2]Note that I use different notation and ignore certain elaborations that are irrelevant for current purposes.

to the units according to a randomization scheme that takes into account their structure. In what is considered to be the first use of a potential response formulation in statistics, Neyman (1935) introduces y_{tu} to represent the response that would be observed on unit u if it were to receive treatment t. The values (y_{tu}) for fixed u, as t varies, are regarded as having simultaneous existence, even though at most one can be observed. Neyman regards the collection of the (y_{tu}), for all t and u, as the unknown parameter and bases statistical inference on the distribution of the observed responses brought about by the random assignment of treatments to units.

Neyman introduces the following null hypothesis:

H_0^* : The value of $y_{t.}$ does not depend on t,

where $y_{t.}$ denotes the average of the y_{tu} over the U units in the experiment. That is, $y_{t.}$ is the average response that would be obtained if treatment t were applied to all of the experimental units, and Neyman's null hypothesis is that this average response would be the same for every treatment—allowing, however, for the possibility of unit-level differences that just happen to average out.[3] Neyman's analysis [corrected and extended by Wilk & Kempthorne (1955)] shows that for certain designs, such as the Latin square, the standard F-test is a valid test of his H_0^* only under the assumption of TUA. Note that under the TUA assumption, H_0^* becomes equivalent to

H_0^{**} : For each unit u, the value of y_{tu} does not depend on t.

Fisher criticized H_0^*, and the associated test, as being based on an entirely inappropriate formulation of the phrase "no differences between the treatments."[4] From the point of view of the DT approach, it is troubling that, according to Neyman, the standard F-test is or is not valid according to whether or not we assume TUA—a distinction without any empirically observable consequences. Neyman's analysis must therefore be classified as a goat.

5. OBSERVATIONAL STUDIES AND CONFOUNDING

Thus far, we have treated X as a decision variable under the control of a human agent rather than nature—not only for the actual decision problem I myself face, but also for the experimental individuals used to supply inputs for my problem. Often, however, genuine experimentation is impossible, and we have to rely on data already collected in circumstances over which we had no control over who received which treatment. Such observational studies raise serious problems of interpretation and relevance, and great care is needed in drawing conclusions from them (Madigan et al. 2014, Rosenbaum 2010).

Suppose I have data on a group of individuals whom I can regard as exchangeable with me, but whose treatments have already been assigned without my knowing how. For each individual, I have information (say, for one headache episode) on the treatment applied, X, and on its duration, Y (and, typically, on some other relevant variables as well). Because I did not have the option to choose which treatment to apply, X is no longer a decision variable: It has become a random variable.

A natural question is as follows: Can I still use (an estimate of) the distribution of Y for those individuals who received treatment $X = 1$ as a proxy for my own hypothetical distribution P_1 (and similarly for $X = 0$)? To do so in this case, I must be able to regard the treated patients as being

[3] But they need not do so if averaged over some other collection of units.

[4] Fisher's arguments are characteristically intuitive rather than formal—although they are no less compelling for that—and he is often taken as having favored H_0^{**}, which is phrased in terms of potential responses, as the appropriate null hypothesis. My own reading of his remarks, however, does not find any clear commitment to a PR interpretation.

similar to (exchangeable with) me, with respect to relevant features existing prior to treatment choice. However, even when this exchangeability can be assumed to hold at the level of the whole group, it need not hold for the subgroup of those who received treatment, as the treatment decision itself may have been correlated with these features. For example, aspirin may have been given only for really bad headaches. This state of affairs is referred to as confounding and obstructs straightforward causal interpretation of observational data. We shall have no confounding only when I can, simultaneously, consider myself as exchangeable both with the patients who received aspirin and with those who received none. When this is so, those two groups of patients must be exchangeable with each other, which in turn requires that treatment application was oblivious to (independent of) any features of the individuals that could be relevant to their reactions to treatment. This is most easily ensured by randomization. In cases where randomization was not carried out, we can sometimes (albeit rarely) attempt to argue that the data can nevertheless be treated as if they had been randomized.

For a functional model $Y = f(X, U)$, the defining property of no confounding is typically taken as requiring independence (in the observational data) between X and U: In the notation of Dawid (1979) (see Section 6 below),

$$X \perp\!\!\!\perp U. \tag{13}$$

This property is trivially equivalent to $U \perp\!\!\!\perp X$, which says that the observational distribution of U given $X = x$ does not depend on x—thus mimicking a property already assumed for the case that X is my decision variable. For an ESE model, the above requirement translates as $X \perp\!\!\!\perp \boldsymbol{E}$, and for a PR model, it translates as $X \perp\!\!\!\perp \boldsymbol{Y}$. (However, as U, $\boldsymbol{E}$, and $\boldsymbol{Y}$ typically do not correspond to any empirically observable variables, the mental exercise required to assess whether the above independence properties hold can be perplexing.) The next section discusses just why equation 13 might be considered as expressing no confounding and extends the analysis to the DT interpretation of this concept.

5.1. Regimes

A helpful way to think about confounding (or the absence of it) is in terms of different data-generating regimes and their relationships. In the above example, we can distinguish three such regimes: One is the observational regime under which the available data have been observed, and the others are the two interventional regimes that correspond to the circumstance in which an external intervention is made to impose one of the two treatments. It is helpful to introduce a nonstochastic variable F_X, which has possible values 0, 1, and $\emptyset$ (read "idle"). Here, $F_X = 0$ labels the interventional regime that sets $X = 0$, $F_X = 1$ labels the interventional regime that sets $X = 1$, and $F_X = \emptyset$ labels the observational regime. There will be a joint distribution of all relevant variables for each of these regimes. Thus, F_X has the status of a statistical parameter that indexes which distribution we are referring to. Note that (assuming intervention is perfectly successful) $X = x$ with probability 1 in regime $F_X = x$ ($x = 0, 1$), whereas X will be a genuinely stochastic variable in regime $F_X = \emptyset$.

We have previously interpreted a functional model as given by equation 8, as well as specializations of it such as ESE or PR models, as incorporating an implicit assumption of stability: The relevant variable U should have the same value, and hence the same distribution, no matter which treatment is applied. In fact this assumption is not quite sufficient: We also need to assume that U has the same distribution, regardless of which regime is operating. This is necessary if we are to justify transfer (under suitable conditions) of information from the observational regime to the interventional regimes.

Suppose property 13 holds. We desire to compute and contrast the distributions of the response Y under the two interventional regimes. Under the intervention with active treatment $Y = f(1, U)$, where U has its marginal distribution under $F_X = 1$. In the observational regime, we can estimate the conditional distribution of Y given $X = 1$, which is that of $f(1, U)$ given $X = 1$ in regime $F_X = \emptyset$. On account of property 13 (which we suppose to apply to the observational regime), this conditional distribution is the same as the marginal distribution of $f(1, U)$ in regime $F_X = \emptyset$. Under our extended stability assumption, however, U has the same distribution in all regimes, so this distribution is indeed the same as the desired distribution of $Y = f(1, U)$ in regime $F_X = 1$. We have thus shown that, taking together property 13 and the assumptions of stability, we can deduce no confounding, interpreted as

$$(Y \mid F_X = x) \approx (Y \mid X = x; F_X = \emptyset) \quad (x = 0, 1), \tag{14}$$

where the symbol $\approx$ denotes "has the same distribution as."

We now note that the left-hand side of relation 14 refers to what we have termed the hypothetical distribution, P_x of Y, under an intervention to set $X = x$. The right-hand side refers to a conditional distribution that is, in principle, estimable from observational data. All of the special ingredients of the functional model have evaporated, and we are left with an expression that is fully meaningful within the DT framework. Moreover, within that framework we can simply and directly take property 14 (however it may be justified) as the appropriate expression of no confounding.

5.2. Conditional Independence

An alternative way of expressing property 14 is as follows: First note that, because $X = x$ with probability 1 in regime $F_X = x$, property 14 is equivalent to

$$(Y \mid X = x; F_X = x) \approx (Y \mid X = x; F_X = \emptyset) \quad (x = 0, 1). \tag{15}$$

This relation expresses the conditional distribution of Y, given $X = x$, as a modular component, that can be transferred without change between observational and interventional settings. This modular interpretation of causality offers a useful pragmatic take on a slippery philosophical concept.

The distributional identity 15 can also be considered as an expression of the conditional independence property (Dawid 1979, 1980):

$$Y \perp\!\!\!\perp F_X \mid X, \tag{16}$$

which says that the distribution of Y, given information on the value of X and on the regime F_X under which that value arose, is in fact the same for all of the regimes. In this way, we have converted a causal property into a probabilistic one (albeit involving the nonrandom regime variable F_X). Because the theory of conditional independence is well established (see Section 6 below), this is a fruitful reinterpretation that will be particularly helpful for describing and manipulating causal properties. Thus, we work with property 16 and its DT interpretation as our formal expression of no confounding throughout the remainder of this review.

6. CONDITIONAL INDEPENDENCE AND GRAPHS

This section recapitulates various aspects of the mathematical theory of conditional independence that will be useful for manipulating causal concepts. For further detail, the reader is referred to

papers by Dawid (1979, 1980, 2002) and Dawid & Constantinou (2014), as well as the thesis by Constantinou (2013).

For random variables $X, Y, \ldots$, having joint distribution P, we say X is independent of Y given Z (written $X \perp\!\!\!\perp Y \mid Z$) to mean that the distribution of X given $(Y, Z) = (y, z)$ depends only on the value z of Z. More formally,

> **Definition 1 (Conditional independence):** We say X is conditionally independent of Y given Z, and write $X \perp\!\!\!\perp Y \mid Z$, if, for any measurable set A in the range of X, there exists a function $w(Z)$ of Z alone such that $P(X \in A \mid Y, Z) = w(Z)$ [P-almost surely].

When we need to specify explicitly the underlying joint distribution P, we write $X \perp\!\!\!\perp Y \mid Z[P]$. Independence, $X \perp\!\!\!\perp Y$, is the special case of conditional independence for which the conditioning variable Z is trivial.

6.1. Axioms of Conditional Independence

Some general properties of probabilistic conditional independence (CI) are as follows (Dawid 1979), where we write $W \preceq Y$ to mean that W is a function of Y:

$$
\begin{array}{lll}
\text{P1} & \text{“Symmetry”} & : \; X \perp\!\!\!\perp Y \mid Z \qquad\qquad\qquad \Rightarrow \; Y \perp\!\!\!\perp X \mid Z \\
\text{P2} & & : \; X \perp\!\!\!\perp Y \mid X \\
\text{P3} & \text{“Decomposition”} & : \; X \perp\!\!\!\perp Y \mid Z, \quad W \preceq Y \quad \Rightarrow \; X \perp\!\!\!\perp W \mid Z \\
\text{P4} & \text{“Weak union”} & : \; X \perp\!\!\!\perp Y \mid Z, \quad W \preceq Y \quad \Rightarrow \; X \perp\!\!\!\perp Y \mid (W, Z) \\
 & & \left.\begin{array}{l} X \perp\!\!\!\perp Y \mid Z \\ \text{and} \\ X \perp\!\!\!\perp W \mid (Y, Z) \end{array}\right\} \\
\text{P5} & \text{“Contraction”} & : \qquad\qquad\qquad\qquad\qquad \Rightarrow \; X \perp\!\!\!\perp (Y, W) \mid Z.
\end{array}
$$

[The descriptive terms “Symmetry,” “Decomposition,” “Weak union,” and “Contraction” are those given in chapter 3 of the book by Pearl (1988).] Further properties of CI can be derived by regarding P1–P5 as axioms for a logical system, rather than calling on more specific properties of probability distributions.

6.2. Extension to Nonstochastic Variables

For Definition 1 to make sense, we must be able to talk about distributions for X; thus, X must be a random variable. Subject to appropriate interpretation of the “almost sure” qualification, however, Y and Z need not be. In particular, this is the case for property 16, our interpretation of no confounding, which involves the nonstochastic regime indicator variable F_X.

We must exercise a little care when applying the notation and theory of Section 6.1 to nonstochastic variables to ensure that these variables always appear, explicitly or implicitly, as conditioning variables. Nevertheless, suitably interpreted, properties P1–P5 do still hold (Dawid 1980, Constantinou 2013). In fact, any deduction made using these properties will be valid, so long as, in both premises and conclusions, no nonstochastic variables appear in the leftmost term in a conditional independence statement (although we are allowed to violate this condition in the intermediate steps of an argument). Thus, we can apply P1–P5 freely, even in the presence of nonstochastic variables, provided that we do not attempt to derive any obviously meaningless assertion.

6.3. Graphical Representation

There is a remarkable and technically valuable analogy between (*a*) CI properties holding between random variables and (*b*) certain separation properties of a directed acyclic graph (DAG)

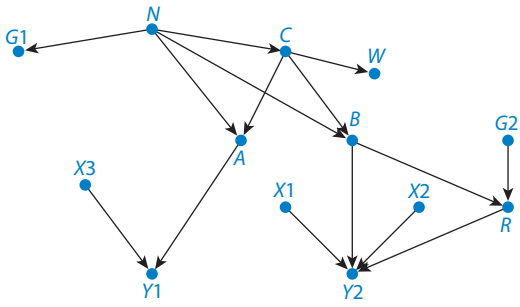

Figure 1

Directed acyclic graph (DAG) $\mathcal{D}$ for criminal trial evidence (Dawid & Evett 1997). The nodes represent relevant hypothesis and evidence variables, and the graph encodes the assumed probabilistic dependence and independence relationships between them.

(Lauritzen et al. 1990). This analogy enables us to use graphical methods to streamline probabilistic manipulations.

The graphical analog of probabilistic CI is the following somewhat complex separation property: Let A, B, and C be sets of nodes of the DAG $\mathcal{D}$. We first form the subgraph $\mathcal{D}'$ of $\mathcal{D}$ that contains only the nodes in A, B, and C, together with all of their ancestors in $\mathcal{D}$ and all of their connecting arrows. The resulting graph, $\mathcal{D}'$, is the relevant ancestral DAG. Next, we insert an undirected edge between any two nodes in $\mathcal{D}'$ that have a common child but that are not already joined by an arrow (said to be unmarried). We then convert all remaining edges to be undirected by dropping the arrowheads, producing the moralized ancestral graph, $\mathcal{G}'$. Finally, in the undirected graph $\mathcal{G}'$, we check whether every connected path from a node in A to one in B intersects C. If so, we say that C d-separates A from B[5], and we write $A \perp_{\mathcal{D}} B \mid C$. It can be shown that at a purely formal level, and with $\preceq$ now interpreted as meaning "is a subset of," $\subseteq$, this separation property satisfies axioms P1–P5 in Section 6.1.

Now suppose that each node v of $\mathcal{D}$ has an associated random variable X_v. Denote $(X_v : v \in A)$ by X_A. We say that a joint distribution P for all of these variables satisfies the local directed Markov property with respect to $\mathcal{D}$ if, for every node v,

$$X_v \perp\!\!\!\perp X_{\mathrm{nd}(v)} \mid X_{\mathrm{pa}(v)}, \tag{17}$$

where (using a self-explanatory analogy with a genetic pedigree) pa(v) denotes the set of parents of node v, and nd(v) denotes its nondescendents in $\mathcal{D}$. In this case it can be shown that, whenever we find $A \perp_{\mathcal{D}} B \mid C$ (by inspection of the DAG), we can deduce the probabilistic conditional independence property $X_A \perp\!\!\!\perp X_B \mid X_C [P]$. We term this the moralization criterion.

As an example, the DAG $\mathcal{D}$ depicted in **Figure 1** describes the relationships between the evidence and other variables figuring in a criminal trial (Dawid & Evett 1997). This graph is constructed so that each node corresponds to a variable in the problem, and the assumed dependence structure of the variables satisfies the following property (termed the local directed Markov property): Each variable is supposed to be probabilistically conditionally independent of its nondescendents in the graph, conditional on its graph parents. For example, the distribution of $Y1$ (measured properties of a tuft of fibers found at the crime scene), given all other variables,

[5]The name refers to an alternative but equivalent way of expressing this separation property, as described by Pearl (1986) and by Verma & Pearl (1990).

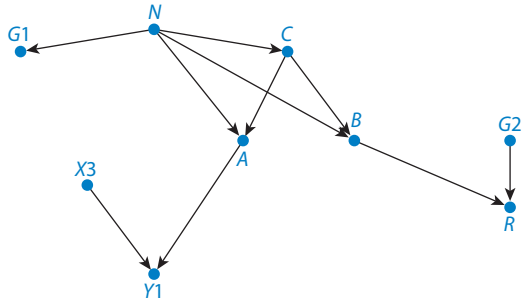

Figure 2

Ancestral subgraph $\mathcal{D}'$ of directed acyclic graph (DAG) $\mathcal{D}$ (Dawid & Evett 1997) for querying the conditional independence property $(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N)$. Only nodes that are ancestral in $\mathcal{D}$ to these variables have been retained.

is supposed to be fully determined by the values of $X3$ (properties of the suspect's sweater) and those of A (an indicator of whether or not the fibers came from the suspect's sweater). Similarly, the distribution of B (denoting the query: Who left blood on the sweater?), given all variables other than $Y2$ (the type of blood) and R (whether or not the blood pattern was a spray), in fact depends only on the values of N (the number of offenders) and of C (whether or not the suspect was an offender). Such assessments can often be made at a qualitative level, before attempting numerical specification of probabilities. In turn, that specification is simplified because we need to describe only the conditional distribution for each variable given its graph parents.

Suppose we now wish to query whether $(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N)$. The relevant ancestral graph $\mathcal{D}'$ is shown in **Figure 2**, and its moralized version, $\mathcal{G}'$, is shown in **Figure 3**. We note the impossibility of tracing a path in $\mathcal{G}'$ from either B or R to either $G1$ or $Y1$ without passing through either A or N. Thus $(B, R) \perp_{\mathcal{D}} (G1, Y1) \mid (A, N)$. From this, we deduce the probabilistic CI property $(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N)$.

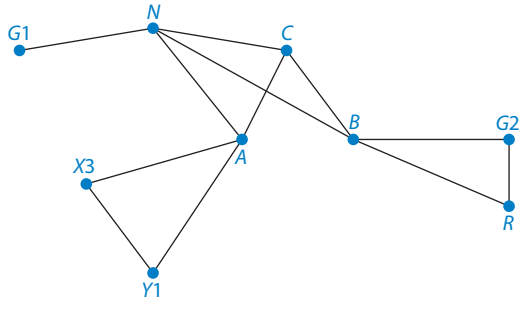

Figure 3

Moralized ancestral subgraph $\mathcal{G}'$ of directed acyclic graph (DAG) $\mathcal{D}$ (Dawid & Evett 1997) for querying the conditional independence property $(B, R) \perp\!\!\!\perp (G1, Y1) \mid (A, N)$. Unmarried parents of a common child in ancestral subgraph $\mathcal{D}'$ have been joined, then arrowheads have been removed. Because every path in $\mathcal{G}'$ linking (B, R) to $(G1, Y1)$ intersects (A, N), we can deduce that the queried conditional independence property does indeed hold.

Caution. Although every DAG thus describes some collection of CI properties and can be used to manipulate them, by no means can every such collection be represented by a DAG. In full generality, we may need to use algebraic manipulations, successively applying the CI axioms P1–P5 to derive the implicit consequences of an assumed collection of conditional independencies.

Markov equivalence. Distinct DAGs can have identical separation properties and so represent identical collections of conditional independencies. These graphs are then said to be Markov equivalent.

The skeleton of a DAG $\mathcal{D}$ is the undirected graph obtained by ignoring the directions of the arrows on the edges of $\mathcal{D}$. An immorality in $\mathcal{D}$ is a configuration of the form $a \to c \leftarrow b$, where a and b are both parents of a common child c but neither $a \to b$ nor $b \to a$.

Theorem 1 (Frydenberg 1990, Verma & Pearl 1991): Two DAGs $\mathcal{D}_0$ and $\mathcal{D}_1$ on the same vertex set V are Markov equivalent if and only if they have the same skeleton and the same immoralities.

Example 1: There are just three possible DAGs on two nodes:

(a) $A \to B$
(b) $A \leftarrow B$
(c) $A \quad B$.

Because DAGs (a) and (b) have the same skeleton and neither has any immoralities, these DAGs are Markov equivalent: Indeed, they embody no nontrivial CI properties whatsoever. However, DAG (c), which has a different skeleton, embodies the marginal independence property $A \perp\!\!\!\perp B$.

Example 2: Consider the following DAGs on three nodes:

(a) $A \to B \to C$
(b) $A \leftarrow B \leftarrow C$
(c) $A \leftarrow B \to C$
(d) $A \to B \leftarrow C$.

These all have the same skeleton. However, whereas DAGs (a), (b), and (c) have no immoralities, (d) has one immorality. Consequently, although (a), (b), and (c) are Markov equivalent to each other, (d) is not Markov equivalent to them. Indeed, (a), (b), and (c) all express the conditional independence property $A \perp\!\!\!\perp C \mid B$, whereas ($d$) expresses the marginal independence property $A \perp\!\!\!\perp C$.

7. CAUSAL INTERPRETATIONS OF DIRECTED ACYCLIC GRAPHS

It is common, and appears very natural, to want to interpret an arrow $a \to b$ in a DAG as representing some kind of direct causal dependence of b on a. This interpretation is potentially dangerous, however, because there is nothing in the DAG semantics, as presented above, to justify it. We prefer to introduce causality into a DAG in a different way: by explicitly representing regime indicators[6] and applying the moralization criterion to the resulting influence diagram (ID), a DAG containing both stochastic and nonstochastic variables. As a simple example, the no confounding property 16 is represented by the ID of **Figure 4**.

[6]This develops on an idea introduced by Spirtes et al. (2000); the reader is also referred to the book chapter by Lauritzen (2000) and the book by Pearl (2009).

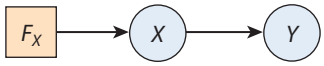

Figure 4

No confounding. The node F_X represents a nonstochastic variable whose value indicates the observational or interventional regime determining X. The directed acyclic graph (DAG) encodes the causal no confounding property $Y \perp\!\!\!\perp F_X \mid X$: The probabilistic dependence of Y on X is the same, regardless of whether or how X is set by intervention or allowed to arise naturally.

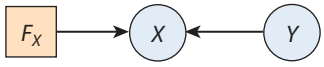

Figure 5

X does not cause Y. The node F_X represents a nonstochastic variable whose value indicates the observational or interventional regime determining X. The directed acyclic graph (DAG) encodes the causal property $Y \perp\!\!\!\perp F_X$: The probabilistic behavior of Y is the same, regardless of whether or how X is set by intervention or allowed to arise naturally.

Consider now the effect of reversing the arrow from X to Y, as shown in **Figure 5**. Without the intervention node F_X, the two graphs would have been Markov equivalent [as was the case for graphs (*a*) and (*b*) in Example 1]. Now, however, we can easily see that these graphs no longer represent equivalent assumptions, as although they have the same skeleton, they have different immoralities. **Figure 5** expresses the marginal independence property $Y \perp\!\!\!\perp F_X$, thereby making it explicit that the marginal distribution of Y is the same, regardless of whether or how X is subjected to intervention. That is, X has no effect on Y in any regime.

8. PEARLIAN DIRECTED ACYCLIC GRAPHS

Consider the DAG depicted in **Figure 6**. Interpreted purely stochastically, this graph is nothing but a representation of the CI properties $A \perp\!\!\!\perp B$, $D \perp\!\!\!\perp (A, B, C)$, and $E \perp\!\!\!\perp (A, B) \mid (C, D)$, together with all other properties, such as $E \perp\!\!\!\perp B \mid (A, C)$, that are deducible from these using P1–P5 (or, equivalently, readable off the DAG using the moralization criterion).

In the approach of Pearl (2009), a DAG such as that in **Figure 6** is taken to represent causal properties. A helpful way of understanding Pearl's interpretation is to consider the DAG to be a shorthand for the influence diagram depicted in **Figure 7**, in which a nonstochastic intervention node has been associated with every stochastic node. Using the moralization criterion, we can read

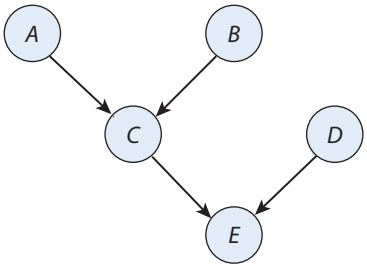

Figure 6

A directed acyclic graph (DAG) representing properties of probabilistic conditional independence between the stochastic variables A, B, C, D, and E. A Pearlian interpretation of the same DAG would further represent causal relationships among these variables.

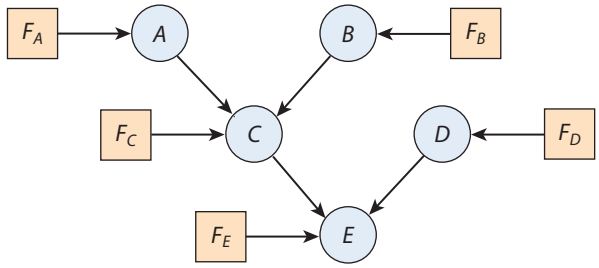

Figure 7

For each of the variables in **Figure 6**, an intervention node has been added. By applying the moralization criterion to the resulting augmented directed acyclic graph (DAG), the causal properties that are only implicit in the Pearlian interpretation of **Figure 6** are made fully explicit.

from this augmented DAG that, for example, $C \perp\!\!\!\perp (D, F_A, F_B, F_D, F_E) \,|\, (A, B, F_C)$. For $F_C = \emptyset$ (the only nontrivial case), this expression says that the natural conditional distribution of C, given A and B, is not further affected by additional conditioning on the value of D, nor is it affected by whether any or all of A, B, D, or E arose naturally or by intervention. Similar properties hold for the other domain variables. In particular, we can see that the conditional distribution for a node, given its domain parents, when it is allowed to arise naturally, remains unchanged when its parents are set by intervention (and is thus a modular component, invariant across different regimes). The augmented DAG thus automatically encodes (via moralization semantics) the assumptions made externally by Pearl, without requiring any new ingredients or concepts. Further, this type of DAG makes it easy to read off the implications of such assumptions directly. It also makes it clear that, when endowed with Pearl's causal interpretation, DAGs that are prima facie Markov equivalent (such as $X \rightarrow Y$ and $X \leftarrow Y$) are not causally equivalent, because their augmented forms will not be Markov equivalent. For all of these reasons, we prefer the explicit use of augmented DAGs over Pearl's shorthand form, which in any case courts confusion with the purely stochastic interpretation of a DAG.

Caution. A Pearlian DAG model, or its augmented DAG equivalent, is justified only to the extent that it models the actual the behavior of the world in the setting to which it is intended to apply. In particular, we must ask whether or not the various interventional situations are indeed related to the noninterventional one in the specific way represented by the DAG. As such considerations necessarily involve cross-regime comparisons, no assessment of their appropriateness can be made on the basis of purely observational data.

9. IDENTIFYING CAUSAL EFFECTS

Suppose we are interested in the causal effect of a treatment variable T on a response variable Y. In the DT framework, this concern requires us to identify, and to contrast, the two interventional distributions: P_1, for Y in regime $F_T = 1$, and P_0, for Y in regime $F_T = 0$. For simplicity, we again confine attention to the ACE

$$\text{ACE} := \text{E}(Y \mid F_T = 1) - \text{E}(Y \mid F_T = 0). \tag{18}$$

With only observational data, gathered in regime $F_T = \emptyset$, we will not be in a position to directly assess these interventional distributions of Y. We will thus need to make assumptions to justify and guide computation of the ACE from such data. Because any such assumptions must relate distributions across distinct regimes, they will not be empirically testable if we have only

observational data. It will, however, be important to present some sort of convincing argument for the suitability of any assumptions imposed.

At the simplest level, we might assume no confounding: $Y \perp\!\!\!\perp F_T \mid T$. In this case, we could simply estimate the observational conditional distribution of $Y \mid T = t, F_T = \emptyset$ and take that as the desired interventional distribution of $Y \mid F_T = t$ ($t = 0, 1$). Thus, under this assumption we have

$$\text{ACE} = \text{E}(Y \mid T = 1, F_T = \emptyset) - \text{E}(Y \mid T = 1, F_T = \emptyset), \tag{19}$$

which is straightforwardly estimable from observational data.

However, in many realistic contexts, the no confounding property is simply unbelievable: We will have confounding: $Y \not\perp\!\!\!\perp F_T \mid T$, and equation 19 might fail. Note that this definition of confounding does not require the existence of what are often called confounding variables, or confounders. But to make progress in identifying ACE, we typically have to introduce further variables with appropriate properties.

9.1. Sufficient Covariates

A variable U is a pretreatment variable if it exists and is (in principle) observable prior to the point at which the treatment decision is made. In this case, its value, and therefore its distribution, must be the same under both interventional regimes, $F_T = 0$ and $F_T = 1$. It is frequently (though not invariably) the case that this common distribution of U also holds in the relevant observational regime $F_T = \emptyset$, giving us

$$U \perp\!\!\!\perp F_T. \tag{20}$$

Such a variable is termed a covariate.

9.2. Unconfounders

When we cannot assume no confounding, we might be able to tell an alternative and more convincing story in terms of a (typically multivariate) covariate U, claiming that we will have no residual confounding after conditioning on U. Formally,

$$Y \perp\!\!\!\perp F_T \mid (U, T). \tag{21}$$

For example, if our data arise from an observational study of patients treated by a certain doctor who might be allocating treatment according to his own observations U of the general health of each patient, it could be reasonable to suppose that, conditionally on U, we would have no residual confounding. If we can observe U, we can then use the observational distribution of Y given $(U, T = t)$ as the distribution of Y given U in the interventional regime $F_T = t$.

A variable satisfying both properties 20 and 21 is often called a confounder, although unconfounder might be a more appropriate term. We shall call such a variable a sufficient covariate. Properties 20 and 21 are represented by the ID of **Figure 8**.

Functional model. Suppose our starting point was a functional model $Y = f(T, U)$ [which includes (E)SE and PR models]. Because the same function is supposed to apply irrespective of the regime operating, property 21 holds trivially. We have further assumed that U has the same value, and hence the same distribution, in all regimes, so property 20 also holds. Thus, formally at least, U is an unconfounder. However, the variable U is typically unobservable in such a formulation (this being a logical necessity in the PR approach, in which U represents the pair $\boldsymbol{Y}$ of potential responses), limiting the operational usefulness of this observation.

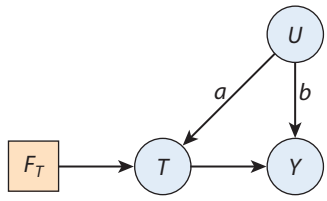

Figure 8

A directed acyclic graph (DAG) representing U as a sufficient covariate to control the probabilistic causal dependence of Y on T. When either arrow a or arrow b is absent, there is no confounding of the effect of T on Y.

9.3. Nonconfounding

Specializations of the above structure are obtained when we can assume that either of the arrow a or arrow b in **Figure 8** is absent. It can readily be checked that in either case we will have $Y \perp\!\!\!\perp F_T \mid T$ (no confounding). We might call such a sufficient covariate U a nonconfounder, and we can safely forget that it ever existed: We can simply apply Equation 19.

The ID with arrow a absent represents the additional property (i.e., over and above properties 20 and 21) $T \perp\!\!\!\perp U \mid F_T : T$ is independent of U in every regime. As this condition holds trivially for the interventional regimes, in which T is constant, it merely requires that T be independent of U in the observational regime—that is, that the variables U that putatively might have affected the doctor's decision did not in fact do so. This property would be perfectly believable if the doctor had tossed a coin to determine his decision, which is why randomized studies can directly address causal queries. But in the case of an observational study, we need to make some alternative convincing case for this property. Then (and only then), we can treat the study as if it had been randomized. This argument is similar to that of Section 5.1 for functional models, but because it involves a real variable U rather than a fictitious one, and stochastic rather than deterministic relationships, it supplies a more operational justification for assuming no confounding.

The ID with b absent represents the additional property $Y \perp\!\!\!\perp U \mid T$, which says that the conditional distribution of Y given (T, U) (which, by condition 21, has already been supposed the same in all regimes) does not in fact depend on U—that is, U is not predictive of outcome. In that case, even if U is associated with treatment assignment, this association will not generate confounding.

9.4. Deconfounding

More generally, suppose U is a sufficient covariate that is observed in the observational regime. We then define

$$\mathrm{SCE}_U := \mathrm{E}(Y \mid U, F_T = 1) - \mathrm{E}(Y \mid U, F_T = 0), \tag{22}$$

the specific causal effect of treatment, given U. This is a random variable, a function of U, whose value $\mathrm{SCE}_U(u)$ when $U = u$ is the average treatment effect in the subgroup of individuals having $U = u$.

Now $T = t$ with probability 1 under F_t. Then using property 21 we find that $\mathrm{E}(Y \mid U, F_T = t) = \mathrm{E}(Y \mid U, T = t, F_T = t) = \mathrm{E}(Y \mid U, T = t, F_T = \emptyset)$.[7] We deduce

$$\mathrm{SCE}_U = \mathrm{E}(Y \mid U, T = 1, F_T = \emptyset) - \mathrm{E}(Y \mid U, T = 0, F_T = \emptyset), \tag{23}$$

[7]More accurately, these identifications require an additional positivity condition (Guo & Dawid 2010), which will typically be satisfied.

showing that SCE_U is estimable from observational data. This is a reflection of the fact that we have no confounding conditional on U.

Also, by the "extension of the conversation" rule of probability, we have

$$\begin{aligned}\mathrm{E}(Y \mid F_T = t) &= \mathrm{E}\{E(Y \mid U, F_T = t) \mid F_T = t\} \\ &= \mathrm{E}\{E(Y \mid U, F_T = t) \mid F_T = \emptyset\}\end{aligned}$$

by property 20, and it follows that

$$\mathrm{ACE} = \mathrm{E}(\mathrm{SCE}_U \mid F_T = \emptyset). \tag{24}$$

That is, for any sufficient covariate U, the overall ACE is the observational expectation of the associated specific causal effect. As, by equation 23, SCE_U is itself an observationally estimable quantity, equation 24 allows us to estimate ACE whenever we can observe a sufficient covariate.

Note that in the PR framework, in which we take $U = \boldsymbol{Y}$, SCE becomes $Y_1 - Y_0$, termed the individual causal effect, ICE. Then equation 24 shows that ACE = E(ICE). Because ICE is necessarily unobservable, however, this formal identity has no operational content.

9.5. Effect of Treatment on the Treated

Suppose that I am thinking of taking aspirin and regard myself as exchangeable with those individuals in the data who did in fact receive aspirin—though not necessarily with those who did not. I can then use the treated group to assess my hypothetical expected response, $\mathrm{E}(Y \mid F_T = 1)$ for Y, were I to take the aspirin. It seems, however, as though I am not in a position to assess the contrasting hypothetical expectation, $\mathrm{E}(Y \mid F_T = 0)$, and therefore I cannot assess my personal effect of treatment. But in the presence of a sufficient covariate U—even if not observed—I may be able to do so.

We define the effect of treatment on the treated as

$$\mathrm{ETT} := \mathrm{E}(\mathrm{SCE}_U \mid T = 1, F_T = \emptyset). \tag{25}$$

That is, ETT is the average, in the observational regime, of the specific causal effect (defined relative to U), over those individuals who did in fact receive the asprin, $T = 1$—and are thus like me.

It might appear that, given a choice over which sufficient covariate U to use in equation 25, the choice of U might affect the value of ETT. Fortunately it turns out that this is not so, on account of the following result (Geneletti & Dawid 2011):

> **Theorem 2:** Suppose $\Pr(T = 1 \mid F_T = \emptyset) > 0$. Then, for any sufficient covariate U, ETT defined by equation 25 satisfies the following:
>
> $$\mathrm{ETT} = \frac{\mathrm{E}(Y \mid F_T = \emptyset) - \mathrm{E}(Y \mid F_T = 0)}{\Pr(T = 1 \mid F_T = \emptyset)}. \tag{26}$$

We have previously noted that, within the PR framework, we can formally regard the pair $\boldsymbol{Y}$ of potential responses as a sufficient covariate. In that case, the SCE becomes the ICE, $Y_1 - Y_0$, and equation 25 delivers $\mathrm{ETT} = \mathrm{E}(Y_1 - Y_0 \mid T = 1, F_T = \emptyset)$, which is the usual PR definition of ETT. However, the above argument shows that the PR framework is not essential for defining this quantity.

Formula 26 shows that we can identify ETT whenever we can observe the response Y in the observational regime ($F_T = \emptyset$), as well as in a sample of people from whom the treatment was withheld ($F_T = 0$). In addition, although the definition of ETT supposes the existence of some sufficient covariate, it is not necessary to have observations on it.

9.6. Reduction of a Sufficient Covariate

Suppose U is a sufficient covariate. A function V of U is a sufficient reduction of U if V is itself a sufficient covariate. As property 20 for V follows immediately from the same property for U, we need only investigate whether property 27 holds for V: that is,

$$Y \perp\!\!\!\perp F_T \mid (V, T). \tag{27}$$

We can impose various additional conditions to ensure this, such as the following:

Condition 1 (Treatment-sufficient reduction):

$$T \perp\!\!\!\perp U \mid (V, F_T = \emptyset). \tag{28}$$

(In the observational regime, the choice of treatment depends on U only through the value of V.)

Note that this condition does not involve the outcome variable Y, except for the essential requirement that the starting variable U itself be a sufficient covariate for the effect of T on Y. Also, because T is constant in any interventional regime, property 28 is equivalent to

$$T \perp\!\!\!\perp U \mid (V, F_T). \tag{29}$$

Further, as V is a function of U, we trivially have

$$V \perp\!\!\!\perp F_T \mid U \tag{30}$$

and

$$Y \perp\!\!\!\perp V \mid (U, T, F_T). \tag{31}$$

Theorem 3 below now follows on applying the moralization criterion to the ID of **Figure 9**, which faithfully represents the CI properties 20, 21, 29, 30, and 31 used to deduce property 27.

Theorem 3: Suppose U is a sufficient covariate, and let be V be a function of U such that Condition 1 holds. Then V is a sufficient covariate.

Propensity score. An alternative description of treatment-sufficient reduction is as follows: Using P1, the defining property 28 can be expressed as

$$U \perp\!\!\!\perp T \mid (V, F_T = \emptyset). \tag{32}$$

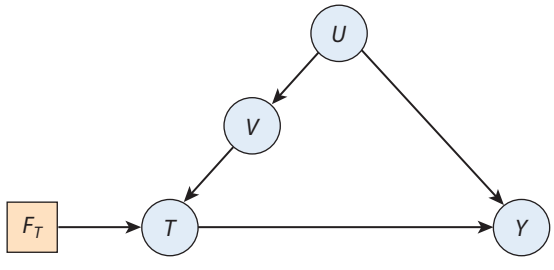

Figure 9

A directed acyclic graph (DAG) representing V, a function of the sufficient covariate U, as a treatment-sufficient reduction of U, making V itself a sufficient covariate.

This asserts that, in the observational regime, the conditional distribution of U given V is the same, whether further conditioned on $T = 0$ or on $T = 1$; that is, V is a balancing score for U (Rosenbaum & Rubin 1983). Property 32 can also be fruitfully interpreted as follows: Consider the family $\mathcal{Q} = \{Q_0, Q_1\}$ comprising the pair of observational conditional distributions for U given, respectively, $T = 0$ and $T = 1$. Property 32 asserts that V is a sufficient statistic (in the usual Fisherian sense) for this family. In particular, a minimal treatment-sufficient reduction is obtained as a minimal sufficient statistic for $\mathcal{Q}$: That is, any $(1, 1)$-function of the likelihood ratio statistic $\Lambda := q_1(X)/q_0(X)$. We might term such a minimal treatment-sufficient covariate a propensity variable, as one form for it is the treatment-assignment probability

$$\Pi := \Pr(T = 1 \mid U, F_T = \emptyset) = \pi\Lambda/(1 - \pi + \pi\Lambda) \tag{33}$$

[where $\pi := \Pr(T = 1 \mid F_T = \emptyset)$], which is known as the propensity score (Rosenbaum & Rubin 1983). Either Λ or Π supplies a one-dimensional sufficient reduction of the original, perhaps highly multivariate, sufficient covariate U.[8]

9.7. *do*-Calculus

We here make use of the notation of Pearl (2009), in which, for example, $p(y \mid x, \check{z})$ refers to $\Pr(Y = y \mid X = x, F_Z = z)$, it being implicit that $z \neq \emptyset$ and all unmentioned intervention variables are idle.

Let X, Y, Z, and W be arbitrary sets of variables in a problem also involving intervention variables. The following rules follow immediately from the definition of conditional independence.[9]

Rule 1 (Insertion/deletion of observations): If $Y \perp\!\!\!\perp Z \mid (X, F_X \neq \emptyset, W)$, then

$$p(y \mid \check{x}, z, w) = p(y \mid \check{x}, w). \tag{34}$$

Rule 2 (Action/observation exchange): If $Y \perp\!\!\!\perp F_Z \mid (X, F_X \neq \emptyset, Z, W)$, then

$$p(y \mid \check{x}, \check{z}, w) = p(y \mid \check{x}, z, w). \tag{35}$$

Rule 3 (Insertion/deletion of actions): If $Y \perp\!\!\!\perp F_Z \mid (X, F_X \neq \emptyset, W)$, then

$$p(y \mid \check{x}, \check{z}, w) = p(y \mid \check{x}, w). \tag{36}$$

Successive application of these rules, coupled with the property $F_X = x \Rightarrow X = x$ and the laws of probability, can sometimes allow one to express a causal expression in purely observational terms. This was the essence of the argument in Section 9.2, which (assuming for simplicity that all variables are discrete) can be expressed in general terms as follows:

Theorem 4 (Back-door formula): Suppose that

$$Z \perp\!\!\!\perp F_X \tag{37}$$

$$Y \perp\!\!\!\perp F_X \mid (X, Z). \tag{38}$$

[8]However, this property may not be as useful as it may first appear (Guo & Dawid 2010).

[9]We assume throughout any positivity conditions required to ensure that the relevant conditional probabilities are well defined.

Then

$$p(y \mid \check{x}) = \sum_z p(y \mid Z = z, X = x)p(Z = z). \tag{39}$$

The most usual application of this *do*-calculus is for a model represented by a Pearlian DAG. However, it is easiest to work with the augmented DAG.[10] We first note that conditioning on $F_X \neq \emptyset$ has the effect of removing all arrows incoming to the set X other than those from F_X. The resulting reduced DAG can then be interrogated, using the usual moralization criterion, to deduce conditional independence properties that can be used as input to Rules 1–3 (in fact Rule 1 is now redundant). In this context, it can be shown constructively (Huang & Valorta 2006; Shpitser & Pearl 2006a,b) that whenever there exists a reduction of a causal expression to purely observational terms, it can be found by applying the *do*-calculus.

10. INSTRUMENTAL VARIABLES

In the presence of an unobserved sufficient covariate U, it is typically not possible to estimate the ACE of a treatment variable X on a response variable Y from observational data. Some progress can be made if we can assume the existence of an observable instrumental variable Z, which can be thought of as an imperfect proxy for an intervention. The assumptions required in such a case are typically expressed in informal terms such as the following (Martens et al. 2006):

(*a*) Z has a causal effect on X
(*b*) Z affects the outcome Y only through X ("no direct effect of Z on Y")
(*c*) Z does not share common causes with the outcome Y ("no confounding of the effect of Z on Y").

These assumptions might be formalized as observational conditional independence properties, such as the following:

$$X \not\perp\!\!\!\perp Z \tag{40}$$

$$U \perp\!\!\!\perp Z \tag{41}$$

$$Y \perp\!\!\!\perp Z \mid (X, U). \tag{42}$$

Note the analogy between properties 41 and 20, as well as that between properties 42 and 21, where (with T relabeled as X) Z takes the place of F_X. However, unlike the case of an imposed intervention, Z does not determine the value of X, but merely has some association with it, as described by property 40. These assumptions are represented by the DAG of **Figure 10** (where for property 40 we need to assume that this is a faithful representation).

For all that this might be a fruitful analogy, the requirements represented by properties 40–42, as well as **Figure 10**, leave something to be desired: Because they relate solely to the observational regime, they cannot, of themselves, have any causal consequences—at best such consequences are left implicit, leaving room for confusion. It is far better to make the requisite causal assumptions explicit. We do this by elaborating **Figure 10** to explicitly include the nonstochastic regime indicator F_X for X, as in **Figure 11**. For $F_X = \emptyset$, doing so recovers the assumptions encoded in **Figure 10**. In addition, the inclusion of the regime indicator relates the observational structure to

[10] Pearl's analysis, similar to its precursor in an article by Spirtes et al. (1999), works with equivalent, somewhat more complex, formulations in terms of unaugmented DAGs.

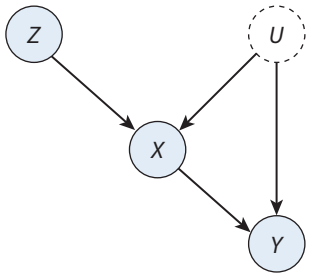

Figure 10

In the presence of an unobserved sufficient covariate U, Z acts as an instrumental variable for the causal effect of X on Y. The directed acyclic graph (DAG) encodes conditional independence properties assumed to hold in the observational regime.

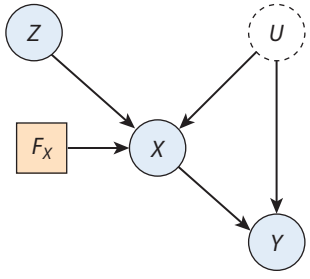

Figure 11

The directed acyclic graph (DAG) depicted in **Figure 10** has been augmented with the regime indicator node F_X. It now relates the observational and interventional regimes, therby making the causal assumptions explicit.

what would happen under an intervention to set X. In particular, it clarifies that U is assumed to be a sufficient covariate for the effect of X on Y, and it further encodes the following properties[11]:

$$U \perp\!\!\!\perp Z \,|\, F_X \tag{43}$$

$$Y \perp\!\!\!\perp Z \,|\, (X, U, F_X). \tag{44}$$

Properties 43 and 44 extend properties 41 and 42 to apply under intervention, as well as observationally.

10.1. Linear Model

Suppose now that all of the observables are univariate and we can describe the dependence of Y on (X, U) (which we have assumed the same in all regimes) by a linear model:

$$\mathrm{E}(Y \,|\, X, U, F_X) = W + \beta X \tag{45}$$

for some function W of U.

[11] **Figure 11** also encodes the additional but inessential property $Z \perp\!\!\!\perp F_X$.

We deduce

$$\mathrm{E}(Y \mid F_X = x) = w_0 + \beta x,$$

where $w_0 := \mathrm{E}(W \mid F_X = x)$ is a constant independent of x because $U \perp\!\!\!\perp F_X$. Thus β can be interpreted causally, as describing how the mean of Y changes in response to the manipulation of X. Our aim is to identify β.

By property 44, equation 45 is also $\mathrm{E}(Y \mid X, Z, U, F_X = \emptyset)$. Then

$$\mathrm{E}(Y \mid Z, F_X = \emptyset) = \mathrm{E}(W \mid Z, F_X = \emptyset) + \beta \mathrm{E}(X \mid Z, F_X = \emptyset).$$

By property 43, however, the first term on the right-hand side is constant. Thus,

$$\mathrm{E}(Y \mid Z, F_X = \emptyset) = \text{constant} + \beta \mathrm{E}(X \mid Z, F_X = \emptyset). \tag{46}$$

Equation 46 relates two functions of Z, each of which can be identified from observational data. Consequently (as long as neither side is constant), we can identify the causal parameter β from such data. Indeed it readily follows from equation 46 that (in the observational regime) $\beta = \mathrm{Cov}(Y, Z)/\mathrm{Cov}(X, Z)$, which can be estimated by the ratio of the coefficients of Z in the sample linear regressions of Y on Z and of X on Z.

10.2. Binary Variables

When all of the observable variables Z, X, and Y are binary, we cannot fully identify the causal probability $P(Y = 1 \mid F_X = x)$ from observational data without making further assumptions. However, we can develop inequalities it must satisfy. This approach was instigated by Manski (1990). His inequalities were refined by Balke & Pearl (1997), under the strong additional condition of deterministic dependence[12] of X on (Z, U) and of Y on (X, U). This condition was shown to be unnecessary by Dawid (2003), who developed a fully stochastic DT approach. In either approach, the analysis involves subtle convex duality arguments.

11. DYNAMIC TREATMENT STRATEGIES

In the ID of **Figure 12**, L_1 and L_2 represent attributes of a patient, T_1 and T_2 represent treatments that can be applied, and Y represents a response of interest. These variables are supposed to be generated in the order shown, each in response to all its predecessors. The nonstochastic regime indicator node σ can take value $\emptyset$, indicating the observational regime; otherwise, a value $\sigma = s$ describes a hypothetical treatment strategy specifying how treatment T_1 should be chosen in response to the observation of L_1, as well as how T_2 should be chosen in response to observation of (L_1, T_1, L_2). Typically such a strategy will prescribe deterministic choices, but there is no difficulty in allowing further randomization. Our task is to infer the consequence, $\mathrm{E}(Y \mid \sigma = s)$, of such a hypothetical strategy from properties of the observational regime $\sigma = \emptyset$.

Figure 12 encodes the following conditional independencies:

$$L_1 \perp\!\!\!\perp \sigma \tag{47}$$

$$L_2 \perp\!\!\!\perp \sigma \mid (L_1, T_1) \tag{48}$$

$$Y \perp\!\!\!\perp \sigma \mid (L_1, T_1, L_2, T_2). \tag{49}$$

[12] An alternative interpretation of this condition is in terms of potential outcomes.

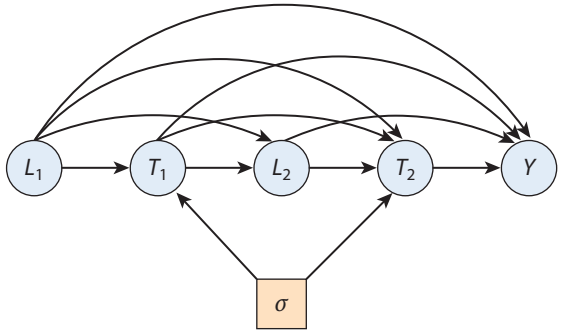

Figure 12

Stochastic variables L_1, T_1, L_2, T_2, and Y are observed in sequence. The node σ represents a nonstochastic variable whose values indicate a treatment strategy describing either how T_1 and T_2 are to be assigned in light of previous observations or a purely observational regime. The directed acyclic graph (DAG) encodes the causal property of sequential ignorability.

The condition given by property 49, for example, says that the conditional distribution of Y given the previous variables (L_1, T_1, L_2, T_2), in the observational regime $\sigma = \emptyset$, would also apply under the operation of an imposed strategy $\sigma = s$. This is a no residual confounding type of assumption that might or might not be appropriate. When conditions 47–49 apply, we say we have sequential ignorability.

We will always have

$$p(l_1, t_1, l_2, t_2, y \mid \sigma = s) = p(l_1 \mid \sigma = s) \tag{50}$$

$$\times\, p(t_1 \mid l_1, \sigma = s) \tag{51}$$

$$\times\, p(l_2 \mid l_1, t_1, \sigma = s) \tag{52}$$

$$\times\, p(t_2 \mid l_1, t_1, l_2, \sigma = s) \tag{53}$$

$$\times\, p(y \mid l_1, t_1, l_2, t_2, \sigma = s). \tag{54}$$

Now, the terms 51 and 53 are specified by the strategy s. Also, under sequential ignorability, we can replace $\sigma = s$ by $\sigma = \emptyset$ in the terms 50, 52, and 54, so that those terms are estimable from observational data. We thus have all the ingredients needed to identify the joint distribution of all variables under the strategy $\sigma = s$, and we can identify the desired consequence, $\mathrm{E}(Y \mid \sigma = s)$ by marginalization. This computation, which can be effectively restructured as a recursion (Dawid & Didelez 2010), reduces to the g-computation formula posed by Robins (1986). That article (see also Chakraborty & Murphy 2014) set the problem up in a PR framework, assuming the simultaneous existence of potential responses (L_{1s}, L_{2s}, Y_s) for each possible strategy $\sigma = s$, subject to certain consistency requirements. Sequential ignorability was then expressed as a conditional independence property involving these potential responses. Our DT approach is more straightforward to interpret, justify, and implement, and it allows for randomized strategies.

It will often be unrealistic to impose the no residual confounding assumption of sequential ignorability. Such an assumption might become more reasonable when additional variables are added to the system, even though these variables are not used by the considered strategy $\sigma = s$. In such a case, one can add further conditions, generalizing those of Section 9.3, which when acceptable would imply that we will indeed have sequential ignorability (for further details, see Dawid & Didelez 2010, Dawid & Constantinou 2014).

12. DISCUSSION

The DT language for causality has sometimes been criticized for not being as rich as that of alternative approaches, such as PR models, that can make statements in their own mathematical terms that simply have no DT counterpart. I regard this as a strength of DT, not a weakness: Formal mathematical expressions (for example, the variance of the ICE discussed in Section 3.4) that do not relate directly to features of the real world are at best unnecessary and at worst dangerously misleading. Within the DT framework, we are not plagued with "the fundamental problem of causal inference" (Holland 1986, p. 947), which is only a self-created problem of the PR approach. The DT approach also fosters healthy skepticism of other methods, such as principal stratification (Frangakis & Rubin 2002), that depend crucially on the philosophically perplexing assumption of the real simultaneous existence of potential response pairs (Dawid & Didelez 2012), together with necessarily untestable assumptions about their properties. Within the ambit of problems that are well posed, the DT framework has all the expressive power necessary, uncluttered by unnecessary and distracting formal mathematical ingredients.

SUMMARY POINTS

1. The traditional machinery of statistical inference does not distinguish between causal relationships and mere associations.
2. Although several formal frameworks have been proposed for describing and analyzing causal processes involving uncertainty, these mostly interpret causal relationships as essentially deterministic.
3. In contrast, the DT approach interprets statistical causality as expressing invariance of probabilistic properties across different regimes: for example, interventional and observational scenarios.
4. This approach involves only a small extension of traditional probability theory: the introduction of a nonstochastic regime indicator variable.
5. Causality properties can then be expressed using the existing formal language of conditional independence. In many cases, these properties can also be described and manipulated using a graphical representation, with associated formal semantics.
6. The DT formalism supplies a straightforward and intuitive approach to formulating and analyzing questions about the effects of applied causes. In particular, it focuses on the assumptions that have to be made and justified in order to license causal inferences.

FUTURE ISSUES

1. The fresh perspective offered by the DT approach should be applied to elucidate further subtle issues, such as defining and untangling direct and indirect effects.
2. Study of the causes of observed effects will require extensions to the theory.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I should like to thank Stephen Fienberg for encouraging me to write this review article and the Department of Statistics at the University of Pennsylvania for its hospitality during its preparation.

LITERATURE CITED

Balke AA, Pearl J. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Am. Stat. Assoc.* 92:1172–76

Berzuini C, Dawid AP, Bernardinelli L, eds. 2012a. *Causality: Statistical Perspectives and Applications*. Chichester, UK: John Wiley & Sons

Berzuini C, Dawid AP, Bernardinelli L. 2012b. An overview of statistical causality. See Berzuini et al. 2012a, pp. xvii–xxv

Berzuini C, Dawid AP, Didelez V. 2012c. Assessing dynamic treatment strategies. See Berzuini et al. 2012a, chapter 8, pp. 85–100

Chakraborty B, Murphy SA. 2014. Dynamic treatment regimes. *Annu. Rev. Stat. Appl.* 1:447–64

Constantinou P. 2013. *Conditional independence and applications in statistical causality*. PhD Thesis, University of Cambridge, Cambridge, UK

Dawid AP. 1979. Conditional independence in statistical theory. *J. R. Stat. Soc. B* 41:1–15; discussion pp. 16–31

Dawid AP. 1980. Conditional independence for statistical operations. *Ann. Stat.* 8:598–617

Dawid AP. 2000. Causal inference without counterfactuals. *J. Am. Stat. Assoc.* 95:407–24; discussion pp. 424–48

Dawid AP. 2002. Influence diagrams for causal modelling and inference. *Int. Stat. Rev.* 70:161–89. Corrigenda. 2002. *Int. Stat. Rev.* 70:437

Dawid AP. 2003. Causal inference using influence diagrams: the problem of partial compliance (with discussion). In *Highly Structured Stochastic Systems*, ed. PJ Green, NL Hjort, S Richardson, pp. 45–81. Oxford, UK: Oxford Univ. Press

Dawid AP. 2007a. Counterfactuals, hypotheticals and potential responses: A philosophical examination of statistical causality. In *Causality and Probability in the Sciences*, Texts in Philosophy, vol. 5, ed. F Russo, J Williamson, pp. 503–32. London: College Publications

Dawid AP. 2007b. *Fundamentals of statistical causality*. Research Report 279, Department of Statistical Science, University College London. 94 pp. **http://www.ucl.ac.uk/statistics/research/pdfs/rr279.pdf**

Dawid AP. 2010a. Beware of the DAG! *J. Mach. Learn. Res. Workshop Conf. Proc.* 6:59–86. **http://tinyurl.com/33va7tm**

Dawid AP. 2010b. Seeing and doing: the Pearlian synthesis. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, ed. R Dechter, H Geffner, JY Halpern, pp. 305–29. London: College Publications

Dawid AP. 2011. The role of scientific and statistical evidence in assessing causality. In *Perspectives on Causation*, ed. R Goldberg, pp. 133–47. Oxford, UK: Hart Publishing

Dawid AP. 2012. The decision-theoretic approach to causal inference. See Berzuini et al. 2012a, pp. 25–42

Dawid AP, Constantinou P. 2014. A formal treatment of sequential ignorability. *Stat. Biosci.* 6:166–88

Dawid AP, Didelez V. 2008. Identifying optimal sequential decisions. In *Proc. Twenty-Fourth Conf. Uncertain. Artif. Intell.*, ed. D McAllester, P Myllymaki, pp. 113–20. Corvallis, OR: AUAI Press. **http://uai2008.cs.helsinki.fi/UAI_camera_ready/dawid.pdf**

Dawid AP, Didelez V. 2010. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Stat. Surv.* 4:184–231

Dawid AP, Didelez V. 2012. "Imagine a can opener"—the magic of principal stratum analysis. *Int. J. Biostat.* 8(1):Article 19

Dawid AP, Evett IW. 1997. Using a graphical method to assist the evaluation of complicated patterns of evidence. *J. Forensic Sci.* 42:226–31

Dawid AP, Faigman DL, Fienberg SE. 2013. Fitting science into legal contexts: Assessing effects of causes or causes of effects? *Sociol. Methods Res.* 43:359–90; discussion pp. 391–421

Dawid AP, Musio M, Fienberg SE. 2014. From statistical evidence to evidence of causality. arXiv:1311.7513 [math.ST]
Didelez V, Dawid AP, Geneletti SG. 2006. Direct and indirect effects of sequential treatments. In *Proc. Twenty-Second Conf. Uncertain. Artif. Intell.*, ed. R Dechter, T Richardson, pp. 138–46. Arlington, VA: AUAI Press
Frangakis CE, Rubin DB. 2002. Principal stratification in causal inference. *Biometrics* 58:21–29
Frydenberg M. 1990. The chain graph Markov property. *Scand. J. Stat.* 17:333–53
Geneletti S, Dawid AP. 2011. Defining and identifying the effect of treatment on the treated. In *Causality in the Sciences*, ed. PM Illari, F Russo, J Williamson, pp. 728–49. Oxford, UK: Oxford Univ. Press
Glymour C, Cooper GF, eds. 1999. *Computation, Causation and Discovery*. Menlo Park, CA: AAAI Press
Guo H, Dawid AP. 2010. Sufficient covariates and linear propensity analysis. *J. Mach. Learn. Res. Workshop Conf. Proc.* 9:281–88. **http://jmlr.csail.mit.edu/proceedings/papers/v9/guo10a/guo10a.pdf**
Hausman DM. 1998. *Causal Asymmetries*. Cambridge, UK: Cambridge Univ. Press
Holland PW. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81:945–60; discussion pp. 960–70
Huang Y, Valtorta M. 2006. Pearl's calculus of intervention is complete. In *Proc. Twenty-Second Conf. Uncertain. Artif. Intell.*, ed. R Dechter, T Richardson, pp. 217–24. Arlington, VA: AUAI Press
Lauritzen SL. 2000. Causal inference from graphical models. In *Complex Stochastic Systems*, ed. OE Barndorff-Nielsen, DR Cox, C Klüppelberg, pp. 63–107. London: CRC Press
Lauritzen SL, Dawid AP, Larsen BN, Leimer HG. 1990. Independence properties of directed Markov fields. *Networks* 20:491–505
Madigan D, Stang PE, Berlin JA, Schuemie M, Overhage M, et al. 2014. A systematic statistical approach to evaluating evidence from observational studies. *Annu. Rev. Stat. Appl.* 1:11–39
Manski CF. 1990. Nonparametric bounds on treatment effects. *Am. Econ. Rev. Pap. Proc.* 80:319–23
Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. 2006. Instrumental variables: applications and limitations. *Epidemiology* 17:260–67
Meek C, Glymour C. 1994. Conditioning and intervening. *Br. J. Philos. Sci.* 45:1001–21
Neyman J. 1935. Statistical problems in agricultural experimentation. *J. R. Stat. Soc. Suppl.* 2:107–54; discussion pp. 154–80
Pearl J. 1986. A constraint–propagation approach to probabilistic reasoning. In *Proc. First Conf. Uncertain. Artif. Intell.*, ed. LN Kanal, JF Lemmer, pp. 357–70. Amsterdam: North-Holland
Pearl J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann
Pearl J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.
Price H. 1991. Agency and probabilistic causality. *Br. J. Philos. Sci.* 42:157–76
Raiffa H. 1968. *Decision Analysis*. Reading, MA: Addison-Wesley
Robins JM. 1986. A new approach to causal inference in mortality studies with sustained exposure periods—application to control of the healthy worker survivor effect. *Math. Model.* 7:1393–512
Rosenbaum PR. 2010. *Design of Observational Studies*. New York: Springer
Rosenbaum PR, Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6:34–68
Shpitser I, Pearl J. 2006a. Identification of conditional interventional distributions. In *Proc. Twenty-Second Conf. Uncertain. Artif. Intell.*, ed. R Dechter, T Richardson, pp. 437–44. Arlington, VA: AUAI Press
Shpitser I, Pearl J. 2006b. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proc. Twenty-First Conf. Uncertain. Artif. Intell.*, ed. F Bacchus, T Jaakkola, pp. 1219–26. Arlington, VA: AUAI Press
Spirtes P, Glymour C, Scheines R. 2000. *Causation, Prediction and Search*. New York: Springer-Verlag. 2nd ed.
Spirtes P, Glymour C, Scheines R, Meek C, Fienberg S, Slate E. 1999. Prediction and experimental design with graphical causal models. See Glymour & Cooper 1999, pp. 65–93
Verma T, Pearl J. 1990. Causal networks: semantics and expressiveness. In *Proc. Fourth Conf. Uncertain. Artif. Intell.*, ed. RD Shachter, TS Levitt, LN Kanal, JF Lemmer, pp. 69–76. Amsterdam: North-Holland
Verma T, Pearl J. 1991. Equivalence and synthesis of causal models. In *Proc. Sixth Conf. Uncertain. Artif. Intell.*, ed. PP Bonissone, M Henrion, LN Kanal, JF Lemmer, pp. 255–68. Amsterdam: North-Holland

Wilk MB, Kempthorne O. 1955. Fixed, mixed and random models. *J. Am. Stat. Assoc.* 50:1144–67
Woodward J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford, UK: Oxford Univ. Press
Woodward J. 2013. Causation and manipulability. *The Stanford Encyclopedia of Philosophy*, ed. EN Zalta. **http://plato.stanford.edu/archives/win2013/entries/causation-mani/**
Wright SS. 1921. Correlation and causation. *J. Agric. Res.* 20:557–85