Sociedad Española de Estadística
e Investigación Operativa

Test

Volume 12, Number 2. December 2003

Statistics and Causal Inference: A Review

Judea Pearl
Cognitive Systems Laboratory, Computer Science Department,
University of California, Los Angeles, U.S.A.

Sociedad de Estadística e Investigación Operativa
Test (2003) *Vol.* **12**, *No.* 2, pp. 281–345

Sociedad de Estadística e Investigación Operativa
Test (2003) *Vol.* **12**, *No.* 2, pp. 281–345

Statistics and Causal Inference: A Review

Judea Pearl[*]
Cognitive Systems Laboratory, Computer Science Department,
University of California, Los Angeles, U.S.A.

Abstract

This paper aims at assisting empirical researchers benefit from recent advances in causal inference. The paper stresses the paradigmatic shifts that must be undertaken in moving from traditional statistical analysis to causal analysis of multivariate data. Special emphasis is placed on the assumptions that underly all causal inferences, the languages used in formulating those assumptions, and the conditional nature of causal claims inferred from nonexperimental studies. These emphases are illustrated through a brief survey of recent results, including the control of confounding, the assessment of causal effects, the interpretation of counterfactuals, and a symbiosis between counterfactual and graphical methods of analysis.

Key Words: Structural equation models, confounding, noncompliance, graphical methods, counterfactuals.

AMS subject classification: 68T30

1 Introduction

Almost two decades have passed since Paul Holland published his seminal paper, Holland (1986), by the same title. Our understanding of causal inference has since increased several folds, due primarily to advances in three areas:

1. Nonparametric structural equations.

2. Graphical models.

[*] Correspondence to: Judea Pearl, Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA 90024. Email: judea@cs.ucla.edu

Preliminary version of this paper was published in *Health Services and Outcomes Research Methodology*, 2:189–220, 2001.

Received: March 2003; Accepted: March 2003

3. Symbiosis between counterfactual and graphical methods.

This paper aims at summarizing and exemplifying these advances.

These advances are central to the empirical sciences because the research questions that motivate most studies in the health, social and behavioral sciences are not statistical but causal in nature. For example, what is the efficacy of a given drug in a given population? Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual, in a specific incident? Not surprisingly, the central target of such studies is the elucidation of cause-effect relationships among variables of interests, for example, treatments, policies, preconditions and outcomes. While good statisticians have always known that the elucidation of causal relationships from observational studies must be shaped by assumptions about how the data were generated, the relative roles of assumptions and data has been a subject of numerous controversies. This paper settles these controversies by introducing useful language for formulating such assumptions and tools for analyzing empirical data in light of these assumptions.

In order to express causal assumptions mathematically, certain extensions are required in the standard mathematical language of Statistics, and these extensions are not generally emphasized in the mainstream literature and education. As a result, large segments of the statistical research community find it hard to appreciate and benefit from the many theoretical results that causal analysis has produced in the past two decades. These include advances in graphical models (Pearl, 1988; Lauritzen, 1996; Cowell et al., 1999), counterfactual or "potential outcome" analysis (Rosenbaum and Rubin, 1983; Robins, 1986; Manski, 1995; Angrist et al., 1996; Greenland et al., 1999b), structural equation models Heckman and Smith (1998), and a more recent formulation, which unifies these approaches under a single interpretation (Pearl, 1995a, 2000).

This paper aims at making these advances more accessible to the general research community[1]. To this end, Section 2 begins by illuminating two conceptual barriers that impede the transition from statistical to causal analysis: (i) coping with untested assumptions and (ii) acquiring

[1] Excellent introductory expositions can also be found in Kaufman and Kaufman (2001) and Robins (2001).

new mathematical notation. Crossing these barriers, Section 3.1 then introduces the fundamentals of causal modeling from a perspective that is relatively new to the statistical literature. It is based on *structural equation models* (SEM), which have been used extensively in economics and the social sciences (Goldberger, 1972; Duncan, 1975; Joreskog and Sorbom, 1978), even though the causal content of these models has been obscured significantly since their inception (Muthen, 1987; Chou and Bentler, 1995) (see Freedman, 1987, for critique and Pearl, 2000, Chapter 5 for historical perspective). Section 3.2 uses these modeling fundamentals to develop simple mathematical tools for estimating causal effects and for the control of confounding. These tools permit investigators to communicate causal assumptions formally using diagrams, then inspect the diagram and

1. Decide whether the assumptions made are sufficient for obtaining consistent estimates of the target quantity;
2. Derive (if the answer to item 1 is affirmative) a closed-form expression for the target quantity in terms of distributions of observed quantities; and
3. Suggest (if the answer to item 1 is negative) a set of observations and experiments that, if performed, would render a consistent estimate feasible.

Section 4 relates these tools to procedures that are used in the potential outcome approach. Finally, Section 4.3, offers a symbiosis that exploits the best features of the two approaches—structural models and potential outcome.

2 From associational to causal analysis: Distinctions and barriers

2.1 The basic distinction: coping with change

The aim of standard statistical analysis, typified by regression and other estimation techniques, is to infer parameters of a distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer aspects of the data generation process. With the help of such aspects, one can deduce not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions.* This capability includes predicting the effects of interventions (e.g., treatments or policy decisions) and spontaneous changes (e.g., epidemics or natural disasters), identifying causes of reported events, and assessing responsibility and attribution (e.g., whether event x was necessary (or sufficient) for the occurrence of event y).

This distinction implies that causal and associational concepts do not mix. Associations characterize static conditions, while causal analysis deals with changing conditions. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change—say from observational to experimental setup—because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified[2].

Drawing analogy to visual perception, the information contained in a probability function is analogous to a geometrical description of a three-dimensional object; it is sufficient for predicting how that object will be viewed from any angle outside the object, but it is insufficient for predicting how the object will be deformed if manipulated and squeezed by external forces. The additional information needed for making such predictions (e.g., the object's resilience or elasticity) is analogous to the information that causal assumptions provide in various forms—graphs, structural equations or plain English. The role of this information is to identify those aspects of the world that remain invariant when external conditions change, say due to treatments or policy decisions.

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: one cannot substantiate

[2] Even the theory of stochastic processes, which provides probabilistic characterization of certain dynamic phenomena, assumes a fixed density function over time-indexed variables. There is nothing in such a function to tell us how it would be altered if external conditions were to change; for example, restricting a variable to a certain value, or forcing one variable to track another.

causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies. Cartwright (1989) expressed this principle as “no causes in, no causes out”, meaning we cannot convert statistical knowledge into causal knowledge.

2.2 Formulating the basic distinction

A useful demarcation line that makes the distinction between associational and causal concepts unambiguous and easy to apply, can be formulated as follows. An associational concept is any relationship that can be defined in terms of a joint distribution (be it personal or frequency-based) of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone. Examples of associational concepts are: correlation, regression, dependence, conditional independence, likelihood, collapsibility, risk ratio, odd ratio, marginalization, conditionalization, “controlling for,” and so on[3]. Examples of causal concepts are: randomization, influence, effect, confounding, “holding constant,” disturbance, spurious correlation, instrumental variables, intervention, explanation, attribution, and so on. The purpose of this demarcation line is not to exclude these causal concepts from the province of statistical analysis[4] but, rather, to make it easy for investigators to trace the assumptions that are needed for substantiating various types of scientific claims. Every claim invoking causal concepts must be traced to some premises that invoke such concepts; it cannot be derived or inferred from statistical associations alone.

2.3 Ramifications of the basic distinction

This principle has far reaching consequences that are not generally recognized in the standard statistics literature. Many researchers, for example, are convinced that confounding is solidly founded in standard, frequentist statistics, and that it can be given an associational definition saying

[3] The term ‘risk ratio’ and ‘risk factors’ have been used ambivalently in the literature; some authors insist on a risk factor having causal influence on the outcome, and some embrace factors that are merely associated with the outcome.

[4] Pearl (2000) termed this distinction “causal vs. statistical” to reflect the overwhelming emphasis on associational concepts in the statistical literature. The term “causal vs. associational” is used here as an invitation for statisticians to correct past neglects.

(roughly): "U is a potential confounder for examining the effect of treatment X on outcome Y when both U and X and U and Y are not independent." That this definition and all its many variants must fail, is obvious from basic considerations:

1. Confounding deals with the discrepancy between an association measured in an observational study and an association that would prevail under ideal experimental conditions.
2. Associations prevailing under experimental conditions are causal quantities because they cannot be inferred from the joint distribution alone. Therefore, confounding is a causal concept; its definition cannot be based on statistical associations alone, since these *can* be derived from the joint distribution.

Indeed, one can construct simple examples showing that the associational criterion is neither necessary nor sufficient, that is, some confounders may not be associated with X nor with Y and some non-confounders may be associated with both X and Y (Pearl, 2000, pp. 185–186); see also Section 3.1[5]. This further implies that confounding bias cannot be detected or corrected by statistical methods alone, not even by the most sophisticated techniques that purport to "control for confounders", such as stepwise selection Kleinbaum et al. (1998) or collapsibility-based methods Grayson (1987). One must make some assumptions regarding causal relationships in the problem, in particular about how the potential "confounders" affect other covariates in the problem, before an adjustment can safely correct for confounding bias. It follows that the rich epidemiological literature on the control of confounding must be predicated upon some tacit causal assumptions and, since causal vocabulary has generally been avoided in much of that literature (e.g., Bishop, 1971; Whittemore, 1978; Grayson, 1987; Hauck et al., 1991; Becher, 1992),[6] major efforts would be required to assess the relevance of this impressive literature to the modern conception

[5] Similar arguments apply to the concepts of "randomization" and "instrumental variables" which are commonly thought to have associational definitions. Our demarcation line implies that they don't, and this implication guides us toward explicating the causal assumptions upon which these concepts are founded (see Section 3.4). Randomization, for example, is based on the assumption that the outcome of a fair coin is not "causally influenced" by any variable that can be measured on a macroscopic level.

[6] Notable exception is the analysis of Greenland and Robins (1986).

of confounding as *effect bias* Greenland et al. (1999b)[7].

Another ramification of the sharp distinction between associational and causal concepts is that any mathematical approach to causal analysis must acquire new notation for expressing causal assumptions and causal claims. The vocabulary of probability calculus, with its powerful operators of conditionalization and marginalization, is insufficient for expressing causal information. To illustrate, the syntax of probability calculus does not permit us to express the simple fact that "symptoms do not cause diseases", let alone draw mathematical conclusions from such facts. All we can say is that two events are dependent—meaning that if we find one, we can expect to encounter the other, but we cannot distinguish statistical dependence, quantified by the conditional probability $P(disease|symptom)$ from causal dependence, for which we have no expression in standard probability calculus[8]. Scientists seeking to express causal relationships must therefore supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the relation "symptoms cause disease" is distinct from the symbolic representation of "symptoms are associated with disease." Only after achieving such a distinction can we label the former sentence "false," and the latter "true", so as to properly incorporate causal information in the design and interpretation of statistical studies.

The preceding two requirements: (1) to commence causal analysis with untested[9], theoretically or judgmentally based assumptions, and (2) to extend the syntax of probability calculus, constitute, in my experience, the two main obstacles to the acceptance of causal analysis among statisticians and among professionals with traditional training in statistics. We shall now explore in more detail the nature of these two barriers, and why they have been so tough to cross.

[7] Although the confounding literature has permitted one causal assumption to contaminate its vocabulary—that the adjusted confounder must not be "affected by the treatment", Cox (1958)—this condition alone is insufficient for determining which variables need be adjusted for (Pearl, 2000, pp. 182–9).

[8] Attempts to define causal dependence by adding temporal information and conditioning on the entire past (e.g., Suppes, 1970) violate the statistical requirement of limiting the analysis to "observed variables", and encounter other insurmountable difficulties (see, Eells, 1991; Pearl, 2001, pp. 249–257).

[9] By "untested" I mean untested using frequency data in nonexperimental studies.

2.4 The barrier of untested assumptions

There are three fundamental differences between associational and causal assumptions. First, associational assumptions, even untested, are testable in principle, given sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference is especially accentuated in Bayesian analysis. Though the priors that Bayesians commonly assign to statistical parameters are untested quantities, the sensitivity to these priors tends to diminish with increasing sample size. In contrast, sensitivity to priors of causal parameters, say those measuring the effect of smoking on lung cancer, remains non-zero regardless of sample size.

Second, associational assumptions can be expressed in the familiar language of probability calculus, and thus assume an aura of scholarship and scientific respectability. Causal assumptions, as we have seen before, are deprived of that honor, and thus become immediate suspect of informal, anecdotal or metaphysical thinking. Again, this difference becomes illuminated among Bayesians, who are accustomed to accepting untested, judgmental assumptions, and should therefore invite causal assumptions with open arms—they don't. A Bayesian is prepared to accept an expert's judgment, however esoteric and untestable, so long as the judgment is presented as a probability expression. Bayesians turn apprehensive when that same judgment is cast in plain causal English, as in "treatment does not change gender." A typical example can be seen in Lindley and Novick (1981) treatment of confounding, in the context of Simpson's paradox (see Pearl, 2000, pp. 174–182 for details).

The third resistance to causal (vis-à-vis associational) assumptions stems from their intimidating clarity. Assumptions about abstract properties of density functions or about conditional independencies among variables are, cognitively speaking, rather opaque, hence they tend to be forgiven, rather than debated. In contrast, assumptions about how variables cause one another are shockingly transparent, and tend therefore to invite counter-arguments and counter-hypotheses. Ironically, it is the latter feature that often deters researchers from articulating assumptions in causal vocabulary. Indeed, since the bulk of scientific knowledge is organized in causal schema, scientists are incredibly creative in constructing competing alternatives to any causal hypothesis, however plausible. Statistical hypotheses in contrast, having been several levels removed from our store of knowledge, are relatively protected from such challenges, and offer therefore a safer ride toward the conclusion.

It is important to emphasize, therefore, that causal analysis does not deal with defending modeling assumptions, in much the same way that differential calculus does not deal with defending the physical validity of a differential equation that a physicist chooses to use. In fact no analysis void of experimental data can possibly defend causal assumptions. Instead, causal analysis deals with the conclusions that logically follow from the combination of data and a given set of assumptions, just in case one is prepared to accept the latter. Thus, all causal inferences are necessarily *conditional*, and the most one can demand from such analysis is:

1. That the premises be amenable to mathematical analysis.
2. That the premises be articulated in a meaningful and unambiguous language for one to judge their plausibility or inevitability.

The structural equation language introduced in Section 3 will be shown to have these two features.

2.5 The barrier of new notation

The need to adopt a new notation, foreign to the province of probability theory, has been traumatic to most persons trained in statistics; partly because the adaptation of a new language is difficult in general, and partly because statisticians—this author included—have been accustomed to assuming that all phenomena, processes, thoughts, and modes of inference can be captured in the powerful language of probability theory.

How does one recognize causal expressions in the statistical literature? Those versed in the potential-outcome notation (Neyman, 1923; Rubin, 1974; Holland, 1988), can recognize such expressions through the subscripts that are attached to counterfactual events and counterfactual variables, e.g. $Y_x(u)$ or Z_{xy}. (Some authors use parenthetical expressions, e.g. $Y(x,u)$ or $Z(x,y)$.) The expression $Y_x(u)$, for example, stands for the value that outcome Y would take in individual u, had treatment X been at level x. If u is chosen at random, Y_x is a random variable, and one can talk about the probability that Y_x would attain a value y in the population, written

$P(Y_x = y)$. Alternatively, Pearl (1995a) and Kaufman and Kaufman (2001) used expressions of the form $P(Y = y|set(X = x))$ or $P(Y = y|do(X = x))$ to denote the probability (or frequency) that event $(Y = y)$ would occur if treatment condition $X = x$ were enforced uniformly over the population[10]. Still a third notation that distinguishes causal expressions is provided by graphical models, where the arrows convey causal directionality[11].

However, in the bulk of the statistical literature, causal claims rarely appear in the mathematics. They surface only in the verbal interpretation that investigators occasionally attach to certain associations, and in the verbal description with which investigators justify assumptions. For example, the assumption that a covariate is not affected by a treatment, a necessary assumption for the control of confounding Cox (1958), is expressed in plain English, not in a mathematical expression.

The absence of notational distinction between causal and statistical relationships at first seemed harmless, because investigators were able to keep such distinctions implicitly in their heads, and managed to confine the mathematics to conventional, conditional probability expressions (Breslow and Day, 1980; Miettinen and Cook, 1981). However, as problem complexity grew, the notational inadequacy of probability calculus began to surface, and intense controversies ensued in the 1980-90's between writers using conventional statistical notation and the few who endeavored to enrich probability calculus with causal vocabulary. Robins (1986, 1987), for example, showed that conventional methods of estimating survival distributions under time-dependent treatments, (e.g., time-dependent Cox regression) may be biased. Greenland and Robins (1986) showed (using counterfactual analysis) that conventional definitions that equated confounding to noncollapsibility would generally lead to biased effect estimates. Holland and Rubin (1988) came to similar conclusions. Using diagrams for guidance, Weinberg (1993) noted that epidemiologists who follow established practices and informal criteria often adjust for the wrong set of covariates. Likewise, Robins and Greenland (1992) proved that the then prevailing

[10] Clearly, $P(Y = y|do(X = x))$ is equivalent to $P(Y_x = y)$, which is what we normally assess in a controlled experiment, with X randomized, in which the distribution of Y is estimated for each level x of X.

[11] These notational clues should be useful for detecting inadequate definitions of causal concepts; any definition of confounding, randomization or instrumental variables that is cast in standard probability expressions, void of graphs, counterfactual subscripts or $do(*)$ operators, can safely be discarded as inadequate.

practice of estimating direct effects by controlling intermediate variables can lead to biased estimates. Again, using counterfactual notation, Robins and Greenland (1989); Greenland (1999) showed that conventional criteria for deciding legal responsibility (for exposure-induced damages), which were based on risk ratio instead of probability of causation, can be severely biased relative to judicial standards. Thus, the notational inadequacy of standard statistics, which was first tolerated and glossed over, took a heavy toll before explicit causal notation brought it to light.

Remarkably, despite this record of success, the mathematics of causal analysis has remained enigmatic to most rank and file researchers, and its potentials still lay grossly underutilized in the statistics based sciences. The reason for this, I am firmly convinced, can be traced to the unfriendly and ad-hoc notation in which causal analysis has been presented to the research community. The next section provides a conceptualization that overcomes these mental barriers; it offers both a friendly mathematical machinery for cause-effect analysis and a formal foundation for counterfactual analysis.

3 The language of diagrams and structural equations

3.1 Linear structural equation models

How can one express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such relationship mathematically was made in the 1920's by the geneticist Wright (1921). Wright used a combination of equations and graphs to communicate causal relationships. For example, if X stands for a disease variable and Y stands for a certain symptom of the disease, Wright would write a linear equation[12] :

$$y = \beta x + u, \tag{3.1}$$

where x stand for the level (or severity) of the disease, y stands for the level (or severity) of the symptom, and u stands for all factors, other than the disease in question, that could possibly affect Y. In interpreting this equation one should think of a physical process whereby Nature examines

[12] Linear relations are used here for illustration purposes only; they do not represent typical disease-symptom relations but illustrate the historical development of path analysis. Additionally, we will use standardized variables, that is, zero mean and unit variance.

the values of x and u and, accordingly, *assigns* to variable Y the value $y = \beta x + u$.

Equation (3.1) still does not properly express the causal relationship implied by this assignment process, because equations are symmetrical objects; if we re-write (3.1) as

$$x = (y - u)/\beta, \tag{3.2}$$

it might be misinterpreted to mean that the symptom influences the disease, against the understanding that no such influence exists. To prevent such misinterpretations, Wright augmented the equation with a diagram, later called "path diagram", in which arrows are drawn from (perceived) causes to their (perceived) effects, and the absence of an arrow encodes the absence of direct causal influence between the corresponding variables. Thus, in our example, the complete model of a symptom and a disease would be written as in Figure 1: The diagram encodes the possible existence of (direct) causal influence of X on Y, and the absence of causal influence of Y on X, while the equations encode the quantitative relationships among the variables involved, to be determined from the data. The parameter β in the equation is called a "path coefficient" and it quantifies the (direct) causal effect of X on Y; given the numerical value of β, the equation claims that a unit increase in X would result in β units increase of Y. The variables V and U are called "exogenous"; they represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence but are not influenced by the other variables (called "endogenous") in the model. Unobserved exogenous variables are sometimes called "disturbances" or "errors", they represent factors omitted from the model but judged to be relevant for explaining the behavior of variables in the model. Variable V, for example, represents factors that contribute to the disease X, which may or may not be correlated with U (the factors that influence the symptom Y). If correlation is presumed possible, it is customary to connect the two variables, U and V, by a dashed double arrow, as shown in Figure 1(b).

In reading path diagrams, it is common to use kinship relations such as parent, child, ancestor, and descendent, the interpretation of which is usually self evident. For example, an arrow $X \rightarrow Y$ designates X as a parent of Y and Y as a child of X. By convention, only observed variables qualify as "parents", thus, in Figure 1(a), only X qualifies as a parent of Y, since U is unobserved (as indicated by the dashed arrow). Likewise, the

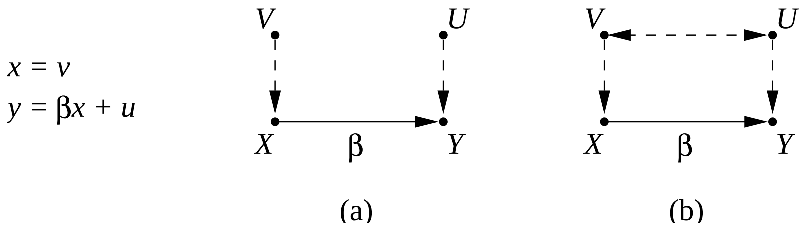

Figure 1: *A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.*

ancestors (respectively, descendants) of a given node, Y, are those variables that can be traced from Y going against (respectively, along) the solid arrows in the diagram. A "path" is any consecutive sequence of edges, solid or dashed. For example, there are two paths between X and Y in Figure 1(b), one consisting of the direct arrow $X \rightarrow Y$ while the other tracing the nodes X, V, U and Y.

Wright's major contribution to causal analysis, aside from introducing the language of path diagrams, has been the development of graphical rules for writing down the covariance of any pair of observed variables in terms of path coefficients and of covariances among the error terms. In our simple example, one can immediately write the relations

$$Cov(X, Y) = \beta \tag{3.3}$$

for Figure 1(a), and

$$Cov(X, Y) = \beta + Cov(U, V) \tag{3.4}$$

for Figure 1(b) (these can be derived of course from the equations, but, for large models, algebraic methods tend to obscure the origin of the derived quantities). Under certain conditions, (e.g. if $Cov(U, V) = 0$), such relationships may allow one to solve for the path coefficients in term of observed covariance terms only, and this amounts to inferring the magnitude of (direct) causal effects from observed, nonexperimental associations, assuming of course that one is prepared to defend the causal assumptions encoded in the diagram.

It is important to note that, in path diagrams, causal assumptions are encoded not in the links but, rather, in the missing links. An arrow merely

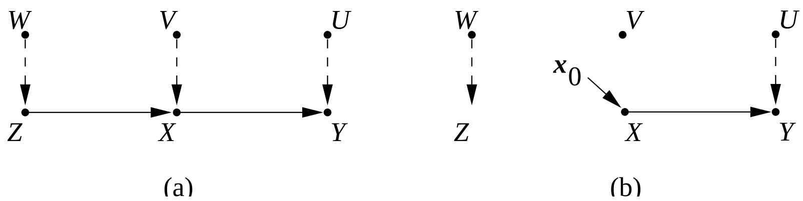

Figure 2: *(a) The diagram associated with the structural model of Equation (3.5). (b) The diagram associated with the modified model of Equation (3.6), representing the intervention* $do(X = x_0)$.

indicates the possibility of causal connection, the strength of which remains to be determined (from data); a missing arrow makes a definite commitment to a zero-strength connection. In Figure 1(a), for example, the assumptions that permits us to identify the direct effect β is encoded by the missing double arrow between V and U, indicating $Cov(U,V)$=0, together with the missing arrow from Y to X. Had any of these two links been added to the diagram, we would not have been able to identify the direct effect β. Such additions would amount to relaxing the assumption $Cov(U,V) = 0$, or the assumption that Y does not effect X, respectively. Note also that both assumptions are causal, not associational, since none can be determined from the joint density of the observed variables, X and Y; the association between the unobserved terms, U and V, can only be uncovered in an experimental setting; or (in more intricate models, as in Figure 5) from other causal assumptions.

Although each causal assumption in isolation cannot be tested, the sum total of all causal assumptions in a model often has testable implications. The chain model of Figure 2(a), for example, encodes seven causal assumptions, each corresponding to a missing arrow or a missing double-arrow between a pair of variables. None of those assumptions is testable in isolation, yet the totality of all those assumptions implies that Z is unassociated with Y in every stratum of X. Such testable implications can be read off the diagrams using a graphical criterion known as *d-separation* (see Pearl, 2000, pp. 16–19), and these constitute the only opening through which the assumptions embodied in structural equation models can confront the scrutiny of nonexperimental data. In other words, every conceivable statistical test capable of invalidating the model is entailed by those implications.

3.2 From linear to nonparametric models

Structural equation modeling (SEM) has been the main vehicle for effect analysis in Economics and the Behavioral and Social Sciences (Goldberger, 1972; Duncan, 1975; Bollen, 1989). However, the bulk of SEM methodology was developed for linear analysis and, until recently, no comparable methodology has been devised to extend its capabilities to models involving dichotomous variables or nonlinear dependencies. A central requirement for any such extension is to detach the notion of "effect" from its algebraic representation as a coefficient in an equation, and redefine "effect" as a general capacity to transmit *changes* among variables. Such an extension, based on simulating hypothetical interventions in the model, is presented in Pearl (1995a, 2000) and has led to new ways of defining and estimating causal effects in nonlinear and nonparametric models (that is, models in which the functional form of the equations is unknown).

The central idea is to exploit the invariant characteristics of structural equations without committing to a specific functional form. For example, the non-parametric interpretation of the diagram of Figure 2(a) corresponds to a set of three functions, each corresponding to one of observed variables:

$$\begin{aligned} z &= f_Z(w) \\ x &= f_X(z, v) \\ y &= f_Y(x, u), \end{aligned} \tag{3.5}$$

where W, V and U are assumed to be jointly independent but, otherwise, arbitrarily distributed. Each of these functions represents a causal process (or mechanism) that determines the value of the left variable (output) from those on the right variables (inputs). The absence of a variable on the right of an equations encodes the assumption that it has no direct effect on the left variable. For example, the absence of variable Z from the arguments of f_Y indicates that variations in Z will leave Y unchanged, as long as variables U, and X remain constant. A system of such functions are said to be *structural* if they are assumed to be autonomous, that is, each function is invariant to possible changes in the form of the other functions (Simon, 1953; Holland, 1953).

Representing interventions

This feature of invariance permits us to use structural equations as a basis for modeling causal effects and counterfactuals. This is done through a mathematical operator called $do(x)$ which simulates physical interventions by deleting certain functions from the model, replacing them by a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $do(x_0)$ that holds X constant (at $X = x_0$) in model M of Figure 2(a), we replace the equation for x in Equation (3.5) with $x = x_0$, and obtain a new model, M_{x_0},

$$\begin{aligned} z &= f_Z(w) \\ x &= x_0 \\ y &= f_Y(x, u), \end{aligned} \tag{3.6}$$

the graphical description of which is shown in Figure 2(b).

The joint distribution associated with the modified model, denoted $P(z, y|do(x_0))$ describes the post-intervention distribution of variables Y and Z (also called "controlled" or "experimental" distribution), to be distinguished from the pre-intervention distribution, $P(x, y, z)$, associated with the original model of Equation (3.5). For example, if X represents a treatment variable, Y a response variable, and Z some covariate that affects the amount of of treatment received, then the distribution $P(z, y|do(x_0))$ gives the proportion of individuals that would attain response level $Y = y$ and covariate level $Z = z$ under the hypothetical treatment $X = x_0$ that is administered uniformly to the population.

From this distribution, one is able to assess treatment efficacy by comparing aspects of this distribution at different levels of x_0. A common measure of treatment efficacy is the average difference

$$E(Y|do(x'_0)) - E(Y|do(x_0)), \tag{3.7}$$

where x'_0 and x_0 are two levels (or types) of treatment selected for comparison. Another measure is the ratio

$$E(Y|do(x'_0))/E(Y|do(x_0)). \tag{3.8}$$

The variance $Var(Y|do(x_0))$, or any other distributional parameter, can also serve as a basis for comparison; all these measures can be obtained from the controlled distribution function $P(Y = y|do(x)) = \sum_z P(z, y|do(x))$ which was called "causal effect" in Pearl (2000, 1995a) (see footnote 10). The central question in the analysis of causal effects is the question of *identification*: Can the controlled (post-intervention) distribution, $P(Y = y|do(x))$, be estimated from data governed by the pre-intervention distribution, $P(z, x, y)$? This is the problem of *identification* which has received considerable attention by causal analysts.

A fundamental theorem in causal analysis states that such identification would be feasible whenever the model is *Markovian*, that is, the graph is acyclic (i.e., containing no directed cycles) and all the error terms are jointly independent. Non-Markovian models, such as those involving correlated errors (resulting from unmeasured confounders), permit identification only under certain conditions, and these conditions can be determined from the graph structure using the following basic theorem.

Theorem 3.1 (The Causal Markov Condition). *Any distribution generated by a Markovian model M can be factorized as*

$$P(v_1, v_2, \ldots, v_n) = \prod_i P(v_i|pa_i), \tag{3.9}$$

where $V_1, V_2, \ldots, V_n$ are the endogenous variables in M, and pa_i are (values of) the endogenous parents of V_i in the causal diagram associated with M.

For example, the distribution associated with the model in Figure 2(a) can be factorized as

$$P(z, y, x) = P(z)P(x|z)P(y|x), \tag{3.10}$$

since X is the (endogenous) parent of Y, Z is the parent of X, and Z has no parents.

Corollary 3.1 (Truncated factorization). *For any Markovian model, the distribution generated by an intervention $do(X = x_0)$ on a set X of endogenous variables is given by the truncated factorization*

$$P(v_1, v_2, \ldots, v_k|do(x_0)) = \prod_{i|V_i \notin X} P(v_i|pa_i)\,|_{x=x_0}\ , \tag{3.11}$$

where $P(v_i|pa_i)$ are the pre-intervention conditional probabilities[13]*.*

Corollary 3.1 instructs us to remove from the product of Equation (3.9) all factors associated with the intervened variables (members of set X). This follows from the fact that the post-intervention model is Markovian as well, hence, following Theorem 3.1, it must generate a distribution that is factorized according to the modified graph, yielding the truncated product of Corollary 3.1. In our example of Figure 2(b), the distribution $P(z, y|do(x_0))$ associated with the modified model is given by

$$P(z, y|do(x_0)) = P(z)P(y|x_0),$$

where $P(z)$ and $P(y|x_0)$ are identical to those associated with the pre-intervention distribution of Equation (3.10). As expected, the distribution of Z is not affected by

the intervention, since

$$P(z|do(x_0)) = \sum_y P(z, y|do(x_0)) = P(z) \sum_y P(y|do(x_0)) = P(z),$$

while that of Y is sensitive to x_0, and is given by

$$P(y|do(x_0)) = P(y|x_0).$$

This example demonstrates how the (causal) assumptions embedded in the model M permit us to predict the post-intervention distribution from the pre-intervention distribution, which further permits us to estimate the causal effect of X on Y from nonexperimental data, since $P(y|x_0)$ is estimable from such data. Note that we have made no assumption whatsoever on the form of the equations or the distribution of the error terms; it is the structure of the graph alone that permits the derivation to go through.

Deriving causal effects

The truncated factorization formula enables us to derive causal quantities directly, without dealing with equations or equation modification as

[13] A simple proof of the Causal Markov Theorem is given in Pearl (2000, p. 30). This theorem was first stated in Pearl and Verma (1991), but it is implicit in the works of Kiiveri et al. (1984) and others. Corollary 3.1 was named "Manipulation Theorem" in Spirtes et al. (1993), and is also implicit in Robins (1987) G-computation formula. See Lauritzen (1999).

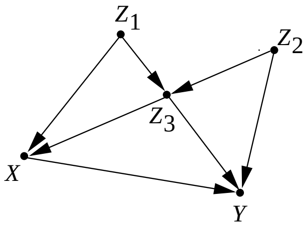

Figure 3: Markovian model illustrating the derivation of the causal effect of X on Y, Equation (3.14). Error terms are not shown explicitly.

in Equation (3.6). Consider, for example, the model shown in Figure 3,in which the error variables are kept implicit. Instead of writing down the corresponding five nonparametric equations, we can write the join distribution directly as

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x), \quad (3.12)$$

where each marginal or conditional probability on the right hand side is directly estimatable from the data. Now suppose we intervene and set variable X to x_0. The post-intervention distribution can readily be written (using the truncated factorization formula) as

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0), \quad (3.13)$$

and the causal effect of X on Y can be obtained immediately by marginalizing over the Z variables, giving

$$P(y|do(x_0)) = \sum_{z_1,z_2,z_3} P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0). \quad (3.14)$$

Note that this formula corresponds precisely to what is commonly called "adjusting for Z_1, Z_2 and Z_3" and, moreover, we can write down this formula by inspection, without thinking on whether Z_1, Z_2 and Z_3 are confounders, whether they lie on the causal pathways, and so on. Though such questions can be answered explicitly from the topology of the graph, they are dealt with automatically when we write down the truncated factorization formula and marginalize.

Note also that the truncated factorization formula is not restricted to interventions on a single variable; it is applicable to simultaneous or sequential interventions such as those invoked in the analysis of time varying treatment with time varying confounders Robins (1986). For example, if X and Z_2 are both treatment variables, and Z_1 and Z_3 are measured covariates, then the post-intervention distribution would be

$$P(z_1, z_3, y|do(x), do(z_2)) = P(z_1)P(z_3|z_1, z_2)P(y|z_2, z_3, x), \qquad (3.15)$$

and the causal effect of the treatment sequence $do(X = x), do(Z_2 = z_2)$[14] would be

$$P(y|do(x), do(z_2)) = \sum_{z_1, z_3} P(z_1)P(z_3|z_1, z_2)P(y|z_2, z_3, x). \qquad (3.16)$$

This expression coincides with Robins (1987) G-computation formula, which was derived from a more complicated set of (counterfactual) assumptions. As noted by Robins, the formula dictates an adjustment for covariates (e.g., Z_3) that might be affected by previous treatments (e.g., Z_2).

Coping with unmeasured confounders

Things are more complicated when we face unmeasured confounders. For example, it is not immediately clear whether the formula in Equation (3.14) can be estimated if any of Z_1, Z_2 and Z_3 is not measured. A few algebraic steps would reveal that one can perform the summation over Z_1 (since Z_1 and Z_2 are independent) to obtain

$$P(y|do(x_0)) = \sum_{z_2, z_3} P(z_2)P(z_3|z_2)P(y|z_2, z_3, x_0), \qquad (3.17)$$

which means that we need only adjust for Z_2 and Z_3 without ever observing Z_1. But it is not immediately clear that no algebraic manipulation can further reduce the resulting expression, or that measurement of Z_3 (unlike Z_1, or Z_2) is necessary in any estimation of $P(y|do(x_0))$. Such considerations become transparent in the graphical representation, to be discussed next.

Selecting covariates for adjustment (the back-door criterion)

Consider an observational study where we wish to find the effect of X on Y, for example, treatment on response, and assume that the factors deemed

[14] For clarity, we drop the (superfluous) subscript 0 from x_0 and z_{2_0}.

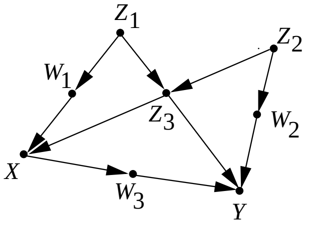

Figure 4: *Markovian model illustrating the back-door criterion. Error terms are not shown explicitly.*

relevant to the problem are structured as in Figure 4; some are affecting the response, some are affecting the treatment and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or life style, others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment, namely, that if we compare treated vs. untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a “sufficient set” or a set “appropriate for adjustment”.

The following criterion, named “back-door” in Pearl (1993), provides a graphical method of selecting such a set of factors for adjustment. It states that a set S is appropriate for adjustment if two conditions hold:

1. No element of S is a descendant of X.
2. The elements of S “block” all “back-door” paths from X to Y, namely all paths that end with an arrow pointing to X.

In this criterion, a set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S[15]. For example, the set $S = \{Z_3\}$ blocks the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$, because the arrow-emitting node Z_3 is in S. However, the set $S = \{Z_3\}$ does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$, because

[15] The terms “arrow-emitting node” and “collision node” are to be interpreted literally as illustrated by the examples given.

none of the arrow-emitting nodes, Z_1 and Z_2, is in S, and the collision node Z_3 is not outside S.

Based on this criterion we see, for example, that the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, and $\{W_2, Z_3\}$, each is sufficient for adjustment, because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because, as explained above, it does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The intuition behind the back-door criterion is as follows. The back-door paths in the diagram carry spurious associations from X to Y, while the paths directed along the arrows from X to Y carry causative associations. Blocking the former paths (by conditioning on S) ensures that the measured association between X and Y is purely causative, namely, it correctly represents the target quantity: the causal effect of X on Y.

Formally, the implication of finding a sufficient set S is that, stratifying on S is guaranteed to remove all confounding bias relative the causal effect of X on Y. In other words, the risk difference in each stratum of S gives the correct causal effect in that stratum. In the binary case, for example, the risk difference in stratum s of S is given by

$$P(Y = 1|X = 1, S = s) - P(Y = 1|X = 0, S = s),$$

while the causal effect (of X on Y) at that stratum is given by

$$P(Y = 1|do(X = 1), S = s) - P(Y = 1|do(X = 0), S = s).$$

These two expressions are guaranteed to be equal whenever S is a sufficient set, such as $\{Z_1, Z_3\}$ or $\{Z_2, Z_3\}$ in Figure 4. Likewise, the average stratified risk difference, taken over all strata,

$$\sum_s [P(Y = 1|X = 1, S = s) - P(Y = 1|X = 0, S = s)]P(S = s),$$

gives the correct causal effect of X on Y in the entire population

$$P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)).$$

In general, for multivalued variables X and Y, finding a sufficient set S permits us to write

$$P(Y = y|do(X = x), S = s) = P(Y = y|X = x, S = s),$$

and

$$P(Y = y|do(X = x)) = \sum_s P(Y = y|X = x, S = s)P(S = s). \qquad (3.18)$$

Since all factors on the right hand side of the equation are estimable (e.g., by regression) from the pre-interventional data, the causal effect can likewise be estimated from such data without bias.

Interestingly, it can be shown that any sufficient set, S, taken as a unit, satisfies the associational criterion that epidemiologists have been using to define "confounders". In other words, S must be associated with X and, simultaneously, associated with Y, given X. This need not hold for any specific members of S. For example, the variable Z_3 in Figure 4, though it is a member of every sufficient set and hence a confounder, can be unassociated with both Y and X (Pearl, 2000, p. 195).

The back-door criterion allows us to write Equation (3.18) directly, by selecting a sufficient set S from the diagram, without manipulating the truncated factorization formula. The selection criterion can be applied systematically to diagrams of any size and shape, thus freeing analysts from judging whether "X is conditionally ignorable given S", a formidable mental task required in the potential-response framework Rosenbaum and Rubin (1983). The criterion also enables the analyst to search for an optimal set of covariate—namely, a set S that minimizes measurement cost or sampling variability (Tian et al., 1998).

General control of confounding

Adjusting for covariates is only one of many methods that permits us to estimate causal effects in nonexperimental studies. Pearl (1995a) has presented examples in which there exists no set of variables that is sufficient for adjustment and where the causal effect can nevertheless be estimated consistently. The estimation, in such cases, employs multi-stage adjustments. For example, if W_3 is the only observed covariate in the model of Figure 4, then there exists no sufficient set for adjustment (because no set of observed covariates can block the paths from X to Y through Z_3), yet $P(y|do(x))$ can be estimated in two steps; first we estimate $P(w_3|do(x)) = P(w_3|x)$ (by virtue of the fact that there exists no back-door path from X to W_3), second we estimate $P(y|do(w_3))$ (since X constitutes a sufficient set for the effect of W_3 on Y) and, finally, we combine the two effects together and obtain

$$P(y|do(x)) = \sum_{w_3} P(w_3|do(x))P(y|do(w_3)). \tag{3.19}$$

The analysis used in the derivation and validation of such results invokes mathematical means of transforming causal quantities, represented by expressions such as $P(Y = y|do(x))$, into *do*-free expressions derivable from $P(z, x, y)$, since only *do*-free expressions are estimable from non-experimental data. When such a transformation is feasible, we are ensured that the causal quantity is identifiable.

General graphical methods for the identification and control of confounders, were presented in Galles and Pearl (1995), while extensions to problems involving multiple interventions (e.g., time varying treatments) were developed in Pearl and Robins (1995), Kuroki and Miyakawa (1999), and (Pearl, 2000, Chapters 3 and 4).

A recent analysis, Tian and Pearl (2002), further shows that the key to identifiability lies not in blocking paths between X and Y but, rather, in blocking paths between X and its immediate successors on the pathways to Y. All existing criteria for identification are special cases of the one defined in the following theorem:

Theorem 3.2 (Tian and Pearl (2002)). *A sufficient condition for identifying the causal effect $P(y|do(x))$ is that every path between X and any of its children traces at least one arrow emanating from a measured variable.*[16]

3.3 Counterfactual analysis in structural models

Not all questions of causal character can be encoded in $P(y|do(x))$ type expressions, in much the same way that not all causal questions can be answered from experimental studies. For example, questions of attribution (e.g., what fraction of death cases are *due to* specific exposure?) or of susceptibility (what fraction of some healthy unexposed population would have gotten the disease had they been exposed?) cannot be answered from experimental studies, and naturally, this kind of questions cannot be expressed

[16] Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of Y.

in $P(y|do(x))$ notation[17]. To answer such questions, a probabilistic analysis of counterfactuals is required, one dedicated to the relation "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. Remarkably, unknown to most economists and philosophers, structural equation models provide the formal interpretation and symbolic machinery for analyzing such counterfactual relationships[18].

The key idea is to interpret the phrase "had X been x" as an instruction to modify the original model and replace the equation for X by a constant x, as we have done in Equation (3.6). This replacement permits the constant x to differ from the actual value of X (namely $f_X(z, v)$) without rendering the system of equations inconsistent, thus yielding a formal interpretation of counterfactuals in multi-stage models, where the dependent variable in one equation may be an independent variable in another.

To illustrate, consider again the modified model M_{x_0} of Equation (3.6), formed by the intervention $do(X = x_0)$ (Figure 2(b)). Call the solution of Y in model M_{x_0} the *potential response* of Y to x_0, and denote it by the symbol $Y_{x_0}(u, v, w)$. This entity can be given a counterfactual interpretation, for it stands for the way an individual with characteristics (u, v, w) would respond, had the treatment been x_0, rather than the treatment $x = f_X(z, v)$ actually received by that individual. In our example, since Y does not depend on v and w, we can write:

$$Y_{x_0}(u, v, w) = Y_{x_0}(u) = f_Y(x_0, u).$$

Clearly, the distribution $P(u, v, w)$ induces a well defined probability on the counterfactual event $Y_{x_0} = y$, as well as on joint counterfactual events, such as '$Y_{x_0} = y$ AND $Y_{x_1} = y'$,' which are, in principle, unobservable if $x_0 \neq x_1$. Thus, to answer attributional questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1|Y = y_0, X = x_0)$

[17] The reason for this fundamental limitation is that no death case can be tested twice, with and without treatment. For example, if we measure equal proportions of deaths in the treatment and control groups, we cannot tell how many death cases are actually attributable to the treatment itself; it is quite possible that many of those who died under treatment would be alive if untreated and, simultaneously, many of those who survived with treatment would have died if not treated.

[18] Connections between structural equations and a restricted class of counterfactuals were first recognized by Simon and Rescher (1966). These were later generalized by Balke and Pearl (1995) to permit counterfactual conditioning on dependent variables.

which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model. For example, assuming a linear equation for Y (as in Figure 1),

$$y = \beta x + u,$$

the conditions $Y = y_0$ and $X = x_0$ yield $V = x_0$ and $U = y_0 - \beta x_0$, and we can conclude that, with probability one, Y_{x_1} must take the value: $Y_{x_1} = \beta x_1 + U = \beta(x_1 - x_0) + y_0$. In other words, if X were x_1 instead of x_0, Y would increase by β times the difference $(x_1 - x_0)$. In nonlinear systems, the result would also depend on the distribution of U and, for that reason, attributional queries are generally not identifiable in nonparametric models (Pearl, 2000, Chapter 9).

This interpretation of counterfactuals, cast as solutions to modified systems of equations, provides the conceptual and formal link between structural equation modeling and the Neyman-Rubin potential-outcome framework, as well as Robins' extensions, which will be discussed in Section 4. It ensures us that the end results of the two approaches will be the same; the choice is strictly a matter of convenience or insight.

3.4 An example: Non-compliance in clinical trials

Formulating the assumptions

Consider the model of Figure 5(a) and Equation (3.5), and assume that it represents the experimental setup in a typical clinical trial with partial compliance. Let Z, X, Y be observed variables, where Z represents a randomized treatment assignment, X is the treatment actually received, and Y is the observed response. The U term represents all factors (unobserved) that influence the way a subject responds to treatments; hence, an arrow is drawn from U to Y. Similarly, V denotes all factors that influence the subject's compliance with the assignment, and W represents the random device used in deciding assignment. The dependence between V and U allows for certain factors (e.g., socio economic status or predisposition to disease and complications) to influence both compliance and response. In Equation (3.5), f_X represents the process by which subjects select treatment level and f_Y represents the process that determines the outcome Y. Clearly, perfect compliance would amount to setting $f_X(z, v) = z$ while any dependence on v represents imperfect compliance.

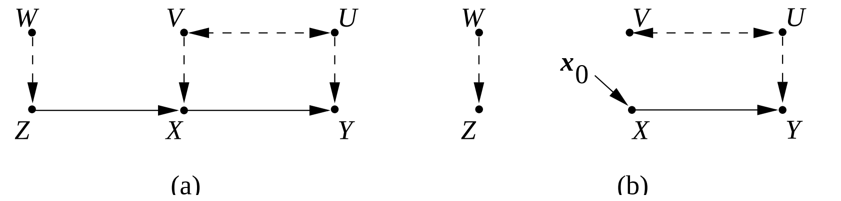

Figure 5: *(a) Causal diagram representing a clinical trial with imperfect compliance. (b) A diagram representing interventional treatment control.*

The graphical model of Figure 5(a) reflects two assumptions.

1. The assignment Z does not influence Y directly but rather through the actual treatment taken, X. This type of assumption is called "exclusion" restriction, for it excludes a variable (Z) from being a determining argument of the function f_Y.

2. The variable Z is independent of U and V; this is ensured through the randomization of Z, which rules out a common cause for both Z and U (as well as for Z and V).

By drawing the diagram of Figure 5(a) an investigator encodes an unambiguous specification of these two assumptions, and permits the technical part of the analysis to commence, under the interpretation provided by Equation (3.5).

The target of causal analysis in this setting is to estimate the causal effect of the treatment (X) on the the outcome (Y). This effect is defined as the response of the population in hypothetical experiment in which we administer treatment at level $X = x_0$ uniformly to the entire population and let x_0 take different values on hypothetical copies of the population. Such hypothetical experiments is governed by the modified model of Equation (3.6) and the corresponding distribution $P(y|do(x_0))$. An inspection of the diagram in Figure 5(a) reveals immediately that this distribution is not identifiable by adjusting for confounders. The graphical criterion for such adjustment requires the existence of observed covariates on the "back-door" path $X \leftarrow V \leftrightarrow U \rightarrow Y$, so as to block (by stratification) the spurious associations created by that path. Had V (or U) been observable, the treatment effect would have been obtained by stratification on the levels of V

$$P(Y=y|do(x_0)) = \sum_v P(Y=y|X=x_0, V=v)P(V=v), \qquad (3.20)$$

thus yielding an estimable expression that requires no measurement of U and no assumptions relative the dependence between U and V. However, since V (and U) are assumed to be unobserved, and since no other blocking covariates exist, the investigator can conclude that confounding bias cannot be removed by adjustment. Moreover, it can be shown that, in the absence of additional assumptions, the treatment effect in such graphs cannot be identified by any method whatsoever Balke and Pearl (1997); one must therefore resort to approximate methods of assessment.

It is interesting to note that it is our insistence on allowing arbitrary functions in Equation (3.5) that curtails our ability to infer the treatment effect from nonexperimental data (when V and U are unobserved). In linear systems, for example, the causal effect of X on Y is identifiable, as can be seen by writing[19] :

$$y = f_Y(x,u) = \beta x + u; \qquad (3.21)$$

multiplying this equation by z and taking expectations, gives

$$\beta = Cov(Z,Y)/(Cov(Z,X)), \qquad (3.22)$$

which reduces β to correlations among observed measurements. Equation (3.22) is known as the *instrumental variable* estimand (Bowden and Turkington, 1984).

Similarly, Imbens and Angrist (1994) have shown that certain nonlinear restrictions of the functions f_X and f_Y may render the causal effect identifiable.

Bounding causal effects

When conditions for identification are not met, the best one can do is derive *bounds* for the quantities of interest—namely, a range of possible values that

[19] Note the β represents the incremental causal effect of X on Y, defined by

$$\beta \triangleq E(Y|do(x_0+1)) - E(Y|do(x_0)).$$

Naturally, all attempts to give β statistical interpretation have ended in frustration (Whittaker, 1990; Wermuth, 1992; Wermuth and Cox, 1993).

represents our ignorance about the data-generating process and that cannot be improved with increasing sample size. In our example, this amounts to bounding the average difference of Equation (3.7) subject to the constraint provided by the observed distribution

$$\begin{aligned} P(x,y|z) &= \sum_{v,u} P(x,y,v,u|z) \\ &= \sum_{v,u} P(y|x,u,v)P(x|z,v)P(u,v), \end{aligned} \tag{3.23}$$

where the product decomposition is licensed by the conditional independencies shown in Figure 5(a). Likewise, since the causal effect is governed by the modified model of Figure 5(b), it can be written

$$P(y|do(x')) - P(y|do(x'')) = \sum_{u} [P(y|x',u) - P(y|x'',u)]P(u). \tag{3.24}$$

Our task is then to bound the expression in Equation (3.24) given the observed probabilities $P(y,x|z)$ as expressed in Equation (3.23). This task amounts to a constrained optimization exercise of finding the highest and lowest values of Equation (3.24) subject to the equality constraint in Equation (3.23), where the maximization ranges over all possible functions $P(u,v)$, $P(y|x,u,v)$ and $P(x|z,u)$ that satisfy those constraints.

Using linear-programming techniques, Balke and Pearl (1997) have derived closed-form solutions for these bounds[20] and showed that despite the imperfection of the experiments, the derived bounds can yield significant and sometimes accurate information on the treatment efficacy. Chickering and Pearl (1997) further used Bayesian techniques (with Gibbs sampling) to investigate the sharpness of these bounds as a function of sample size.

Testable implications

The two assumptions embodied in the model of Figure 5(a), that Z is randomized and has no direct effect on Y, are untestable in general (Bonet, 2001). However, if the treatment variable may take only a finite number of values, the combination of these two assumptions yields testable implications, and these can be used to alert investigators to possible violations

[20] Looser bounds were derived earlier by Robins (1989) and Manski (1990)

of these assumptions. The testable implications take the form of inequalities which restrict aspects of the observed conditional distribution $P(x, y|z)$ from exceeding certain bounds Pearl (1995b).

One specially convenient form that these restrictions assume is given by the inequality

$$\max_x \sum_y [\max_z P(x, y|z)] \leq 1. \tag{3.25}$$

Pearl (1995b) called this restriction an *instrumental inequality*, because it constitutes a necessary condition for any variable Z to qualify as an instrument relative to the pair (X, Y). This inequality is sharp for binary valued X, but becomes loose when the cardinality of X increases[21].

If all observed variables are binary, Equation (3.25) reduces to the four inequalities

$$\begin{aligned}
P(Y=0, X=0|Z=0) &+ P(Y=1, X=0|Z=1) \leq 1 \\
P(Y=0, X=1|Z=0) &+ P(Y=1, X=1|Z=1) \leq 1 \\
P(Y=1, X=0|Z=0) &+ P(Y=0, X=0|Z=1) \leq 1 \\
P(Y=1, X=1|Z=0) &+ P(Y=0, X=1|Z=1) \leq 1.
\end{aligned} \tag{3.26}$$

We see that the instrumental inequality is violated when the controlling instrument Z manages to produce significant changes in the response variable Y while the direct cause, X, remains constant.

The instrumental inequality can be used in the detection of undesirable side-effects. Violations of this inequality can be attributed to one of two possibilities: either there is a direct causal effect of the assignment (Z) on the response (Y), unmediated by the treatment (X), or there is a common causal factor influencing both variables. If the assignment is carefully randomized, then the latter possibility is ruled out and any violation of the instrumental inequality (even under conditions of imperfect compliance) can safely be attributed to some direct influence of the assignment process on subjects' response (e.g., psychological aversion to being treated). Alternatively, if one can rule out any direct effects of Z on Y, say through

[21] The inequality is sharp in the sense that every distribution $P(x, y, z)$ satisfying Equation (3.25) can be generated by the model defined in Figure 5(a).

effective use of a placebo, then any observed violation of the instrumental inequality can safely be attributed to spurious dependence between Z and V, namely, to selection bias.

The instrumental inequality (3.25) can be tightened appreciably if we are willing to make additional assumptions about subjects' behavior—for example, that increasing recommended dosage Z would induce no individual to decrease the actual dosage X or, mathematically, that for all v we have

$$f_X(z_1, v) \geq f_X(z_2, v),$$

whenever $z_1 \geq z_2$. In the binary case, such an assumption amounts to having no contrarians in the population, namely, no individual who would consistently act contrary to his or her assignment. Under this assumption, which Imbens and Angrist (1994)call monotonicity, the inequalities in Equation (3.26) can be tightened (Balke and Pearl, 1997) to give

$$\begin{array}{rcl} P(y, X = 1|Z = 1) & \geq & P(y, X = 1|Z = 0) \\ P(y, X = 0|Z = 0) & \geq & P(y, X = 0|Z = 1) \end{array} \tag{3.27}$$

for all $y \in \{0, 1\}$. Violation of these inequalities now means either selection bias or a direct effect of Z on Y or the presence of contrarian subjects.

It is also interesting to note that the analysis of noncompliance presented in this section is valid under more general conditions than those shown in the graph of Figure 5(a). If an arrow from Y to X is added to the graph, a cyclic graph containing the feedback loop $X \rightarrow Y \rightarrow X$ is obtained. Such a loop may represent, for example, patients deciding on dosage X by continuously monitoring their response Y. Nonetheless, the structural equation model will not change, because, under the assumption that the process is at equilibrium, y is a unique function of x and u, and an equation of the form

$$x = g(z, y, v) \tag{3.28}$$

can be replaced with

$$x = g'(z, v'), \tag{3.29}$$

such that v' is still independent of z. The nonparametric nature of the structural equations in Equation (3.5) permits us to make such transformations without affecting the results of the analysis. Consequently, testable implications and nonparametric bounds obtained from the analysis of the acyclic model are still valid for the cyclic case.

4 The language of potential outcomes and counterfactuals

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y_x(u)$, read: "the value that Y would obtain in unit u, had X been x" (Neyman, 1923; Rubin, 1974). In Section 3.3 we saw that this counterfactual entity has the natural interpretation as representing the solution for Y in a modified system of equation, where *unit* is interpreted a vector u of background factors that characterize an experimental unit. Each structural equation model thus provides a compact representation for a huge number of counterfactual claims. The potential outcome framework lacks such compact representation. In the potential outcome framework, $Y_x(u)$ is taken as primitive, that is, an undefined quantity in terms of which other quantities are defined. Thus, the structural interpretation of $Y_x(u)$ can be regarded as the formal basis for the potential outcome approach. In particular, this interpretation forms a connection between the opaque English phrase "the value that Y would obtain in unit u, had X been x" and a mathematical model that simulates hypothetical changes in X. The formation of the submodel M_x explicates mathematically how the hypothetical condition "had X been x" could be realized, by pointing to and replacing the equation that is violated in making $X = x$ a reality. The logical consequence of such hypothetical conditions can then be derived mathematically.

4.1 Formulating assumptions

The distinct characteristic of the potential outcome approach is that, although investigators must think and communicate in terms of undefined, hypothetical quantities such as $Y_x(u)$, the analysis itself is conducted almost entirely within the axiomatic framework of the probability theory. This is accomplished, by postulating a "super" probability function on both hypothetical and real events. If U is treated as a random variable then the value of the counterfactual $Y_x(u)$ becomes a random variable as well, denoted as Y_x. The potential-outcome analysis proceeds by treating the observed distribution $P(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function P^* defined over both observed and counterfactual variables. Queries about causal effects (written $P(y|do(x))$ in the structural analysis) are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y_x = y)$. The new hypothetical entities Y_x are treated as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence. Moreover, these hypothetical entities are not entirely whimsy, but are assumed to be connected to observed variables via consistency constraints Robins (1986) such as

$$X = x \implies Y_x = Y, \tag{4.1}$$

which states that, for every u, if the actual value of X turns out to be x, then the value that Y would take on if X were x is equal to the actual value of Y. For example, a person who chose treatment x and recovered, would also have recovered if given treatment x by design.

The main conceptual difference between the two approaches is that, whereas the structural approach views the intervention $do(x)$ as an operation that changes the distribution but keeps the variables the same, the potential-outcome approach views the variable Y under $do(x)$ to be a different variable, Y_x, loosely connected to Y through relations such as (4.1).

Pearl (2000, Chapter 7) shows, using the structural interpretation of $Y_x(u)$, that it is indeed legitimate to treat counterfactuals as jointly distributed random variables in all respects, that consistency constraints like (4.1) are automatically satisfied in the structural interpretation and, moreover, that investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \quad \text{for all } y \text{ and } z \tag{4.2}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x \text{ and } z. \tag{4.3}$$

Equation (4.2) ensures that the interventions $do(Y = y)$ results in the condition $Y = y$, regardless of concurrent interventions, say $do(Z = z)$, that are applied to variables other than Y. Equation (4.3) generalizes (4.1) to cases where Z is held fixed, at z.

To communicate substantive causal knowledge, the potential-outcome analyst must express causal assumptions as constraints on P^*, usually in the form of conditional independence assertions involving counterfactual variables. For instance, in our example of a randomized clinical trial with imperfect compliance (Figure 5(a)), to communicate the understanding that the treatment assignment (Z) is randomized (hence independent of both the way subjects react to treatments and how subjects comply with the assignment), the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{X_z, Y_x\}$[22]. To further formulate the understanding that Z does not affect Y directly, except through X, the analyst would write a, so called, "exclusion restriction" $Y_{xz} = Y_x$.

4.2 Performing inferences

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest; in other cases, only bounds on the solution can be obtained. For example, if one can plausibly assume that a set Z of covariates satisfies the conditional independence

$$Y_x \perp\!\!\!\perp X | Z \tag{4.4}$$

(an assumption that was termed "conditional ignorability" by Rosenbaum and Rubin (1983) then the causal effect $P^*(Y_x = y)$ can readily be evaluated to yield

$$\begin{aligned} P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\ &\overset{(\text{using } (4.4))}{=} \sum_z P^*(Y_x = y|x, z)P(z) \\ &\overset{(\text{using } (4.1))}{=} \sum_z P^*(Y = y|x, z)P(z) \\ &= \sum_z P(y|x, z)P(z). \end{aligned} \tag{4.5}$$

The last expression contains no counterfactual quantities (thus permitting us to drop the asterisk from P^*) and coincides precisely with the standard covariate-adjustment formula of Equation (3.18).

We see that the assumption of conditional ignorability (4.4) qualifies Z as a sufficient covariate for adjustment, and is equivalent therefore to the graphical criterion (called "back door" in Section 3.2) that qualifies such covariates by tracing paths in the causal diagram.

[22] The notation $Y \perp\!\!\!\perp X|Z$ stands for the conditional independence relationship $P(Y = y, X = x|Z = z) = P(Y = y|Z = z)P(X = x|Z = z)$, Dawid (1979).

The derivation above may explain why the potential outcome approach appeals to mathematical statisticians; instead of constructing new vocabulary (e.g., arrows), new operators ($do(x)$) and new logic for causal analysis, almost all mathematical operations in this framework are conducted within the safe confines of probability calculus. Save for an occasional application of rule (4.3) or (4.1), the analyst may forget that Y_x stands for a counterfactual quantity—it is treated as any other random variable, and the entire derivation follows the course of routine probability exercises.

However, this mathematical convenience often comes at the expense of conceptual clarity, especially at a stage where causal assumptions need be formulated. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability Equation (4.4), the key to the derivation of Equation (4.5), holds in any familiar situation, say in the experimental setup of Figure 5(a). This assumption reads: "the value that Y would obtain had X been x, is independent of X, given Z". Paraphrased in experimental metaphors, and applied to variable V, this assumption reads: The way an individual with attributes V would react to treatment $X = x$ is independent of the treatment actually received by that individual. Such assumptions of conditional independence among counterfactual variables are not straightforward to comprehend or ascertain, for they are cast in a language far removed from ordinary understanding of cause and effect. When counterfactual variables are not viewed as byproducts of a deeper, process-based model, it is also hard to ascertain whether *all* relevant counterfactual independence judgments have been articulated, whether the judgments articulated are redundant, or whether those judgments are self-consistent. The need to express, defend, and manage formidable counterfactual relationships of this type explain the slow acceptance of causal analysis among epidemiologists and statisticians, and why economists and social scientists continue to use structural equation models instead of the potential-outcome alternatives advocated in Angrist et al. (1996); Holland (1988); Sobel (1998).

On the other hand, the algebraic machinery offered by the potential-outcome notation, once a problem is properly formalized, can be extremely powerful in refining assumptions (Angrist et al., 1996), deriving consistent estimands (Robins, 1986), bounding probabilities of necessary and sufficient causation (Tian and Pearl, 2000), and combining data from experimental and nonexperimental studies (Pearl, 2000). The Section (4.3) presents a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams, translating these assumptions into potential outcome notation, performing the mathematics in the algebraic language of counterfactuals and, finally, interpreting the result in plain causal language.

4.3 Combining graphs and algebra

The formulation of causal assumptions using graphs was discussed in Section 3. In this subsection we will systematize the translation of these assumptions from graphs to counterfactual notation.

Structural equation models embody causal information in both the equations and the probability function $P(u)$ assigned to the error variables; the former is encoded as missing arrows in the diagrams the latter as missing (double arrows) dashed arcs. Each parent-child family (PA_i, X_i) in a causal diagram G corresponds to an equation in the model M. Hence, missing arrows encode exclusion assumptions, that is, claims that adding excluded variables to an equation will not change the outcome of the hypothetical experiment described by that equation. Missing dashed arcs encode independencies among error terms in two or more equations. For example, the absence of dashed arcs between a node Y and a set of nodes $\{Z_1, \ldots, Z_k\}$ implies that the corresponding background variables, U_Y and $\{U_{Z_1}, \ldots, U_{Z_k}\}$, are independent in $P(u)$.

These assumptions can be translated into the potential-outcome notation using two simple rules (Pearl, 1995a, p. 704); the first interprets the missing arrows in the graph, the second, the missing dashed arcs.

1. *Exclusion restrictions:* For every variable Y having parents PA_Y and for every set of endogenous variables S disjoint of PA_Y, we have

$$Y_{pa_Y} = Y_{pa_Y, s}. \tag{4.6}$$

2. *Independence restrictions:* Let $Z_1, \ldots, Z_k$ be any set of nodes not connected to Y via dashed arcs, and let $PA_1, \ldots, PA_k$ be their respective sets of parents. We have

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1\ pa_1}, \ldots, Z_{k\ pa_k}\}. \tag{4.7}$$

The exclusion restrictions expresses the fact that each parent set includes *all* direct causes of the child variable, hence, fixing the parents of

Y, determines the value of Y uniquely, and intervention on any other set S of (endogenous) variables can no longer affect Y. The independence restriction translates the independence between U_Y and $\{U_{Z_1}, \ldots, U_{Z_k}\}$ into independence between the corresponding potential-outcome variables. This follows from the observation that, once we set their parents, the variables in $\{Y, Z_1, \ldots, Z_k\}$ stand in functional relationships to the U terms in their corresponding equations.

As an example, the model shown in Figure 5(a) displays the following parent sets

$$PA_{_Z} = \{\emptyset\},\ PA_{_X} = \{Z\},\ PA_{_Y} = \{X\}. \tag{4.8}$$

Consequently, the exclusion restrictions translate into

$$\begin{array}{rcl} X_z & = & X_{yz} \\ Z_y & = & Z_{xy} = Z_x = Z \\ Y_x & = & Y_{xz}. \end{array} \tag{4.9}$$

The absence of any dashed arc between Z and $\{Y, X\}$ translates into the independence restriction

$$Z \perp\!\!\!\perp \{Y_x, X_z\}. \tag{4.10}$$

This is precisely the condition of randomization; Z is independent of all its non-descendants, namely independent of U and V which are the exogenous parents of Y and X, respectively. (Recall that the exogenous parents of any variable, say Y, may be replaced by the counterfactual variable Y_{pa_Y}, because holding PA_Y constant renders Y a deterministic function of its exogenous parent U_Y.)

The role of graphs is not ended with the formulation of causal assumptions. Throughout an algebraic derivation, like the one shown in Equation (4.5), the analyst may need to employ additional assumptions that are entailed by the original exclusion and independence assumptions, yet are not shown explicitly in their respective algebraic expressions. For example, it is hardly straightforward to show that the assumptions of Equations (4.9) and (4.10) imply the conditional independence $(Y_x \perp\!\!\!\perp Z | \{X_z, X\})$ but do not imply the conditional independence $(Y_x \perp\!\!\!\perp Z | X)$. These are not easily derived by algebraic means alone. Such implications can, however, easily be tested in the graph of Figure 5(a) using the graphical criterion for conditional independence, called d-separation (see Greenland et al., 1999a; Pearl, 2000, pp. 16–17, 213–215). Thus, when the need arises to employ independencies in the course of a derivation, the graph may assist the procedure by vividly displaying the independencies that logically follow from our assumptions.

5 Conclusions

Statistics is strong in devising ways of describing data and inferring distributional parameters from sample. Causal inference require two addition ingredients: a science-friendly language for articulating causal knowledge, and a mathematical machinery for processing that knowledge, combining it with data and drawing new causal conclusions about a phenomena. This paper introduces nonparametric structural equations models as a formal and meaningful language for formulating causal assumptions, and for explicating many concepts used in scientific discourse. These include: randomization, intervention, direct and indirect effects, confounding, counterfactuals, and attribution. The algebraic component of the structural language coincides with the potential-outcome framework, and its graphical component embraces Wright's method of path diagrams. When unified and synthesized, the two components offer statistical investigators a powerful methodology for empirical research.

Acknowledgments

Portions of this paper are based on my book Causality (Pearl, 2000). This research was supported in parts by grants from NSF, ONR (MURI) and AFOSR.

DISCUSSION

Stephen E. Fienberg and Amelia M. Haviland
Department of Statistics
Carnegie Mellon University
Pittsburgh, U.S.A.

Introduction: Can we measure the causal effect of discrimination?

In his excellent review of recent statistical approaches to causal inference, Pearl proposes the following causal question motivating research in the social sciences: "Whether data can prove an employer guilty of hiring discrimination?" Here, we explore whether one can in fact apply Pearl's causal inference tools in the context of discrimination. Our main concern relates to one expressed by Freedman (2003), namely the extent to which we can engage in causal inference when the "treatment" variables of interest are concomitants such as gender, race and age, which are not manipulable. We examine the causal meaning of discrimination and use both the causal framework and the explicit formulation of counterfactuals that are of interest to the study of discrimination.

What is discrimination?

Discrimination is usually taken to mean the differential treatment of individuals based on a perceived characteristic or group membership. In the United States and elsewhere, there is often a legal definition of discrimination that is of relevance to our discussion: An action is said to be discriminatory, e.g., with respect to race, if the treatment of an individual would be different had that person been of a different race. Thus, in the context of employment, we might wish to say that an employer's actions are discriminatory if he/she treats employees or applicants for positions differently "because of" their race. But race itself is not the "cause" of labor market discrimination, nor is it the cause of differences in access to education, family wealth, or health outcomes. The same could be said of gender and of age (if we exclude the possibility of sex change operations). Unlike age, but similarly to gender, race is a characteristic of a person that is both a social construct and may have some distinctive physical attributes, e.g., related to skin color. At this level race is not a manipulable variable. In Pearl's notation, we cannot set $do(X = x)$ where X is race. This is the Freedman (2003) argument referred to above.

The issue of manipulability is less clear if one considers race to be purely a social construct because then at least theoretically it could be manipulated; however, we can manipulate and even randomize the "perception" of race. For example, Bertrand and Mullainathan (2003) used real job resumes but randomly assigned distinctively African American or white names to the resumes. (An unknown here is whether names that are distinctively black or white carry other connotations beyond those associated with only race, such as class.) Similar ideas have been used in psychology experiments, e.g., where researchers randomly assign pictures of people of different races as treatment effects to elicit subjects' responses in different settings.

There are also situations where we can manipulate information about concomitant variables without randomization. For example, Goldin and Rouse (2000) consider how information about the gender of applicants to symphony orchestras was removed through the creation of blind auditions. In this case gender was not randomized but the information about it became unavailable. Thus it was possible to obtain data on the effect of having the information versus not having it on the proportions of women and men who were hired or promoted. But then we must focus our attention on knowledge of the hiring process, so that we can eliminate alternative explanations for the effect of information on gender. Goldin and Rouse (2000) attempted to do this through various forms of generalized linear models, although there is some issue about the extent to which they succeeded.

To summarize, even if we cannot manipulate concomitant variables, there are situations where we can randomize the perception of the concomitant variables or we can assess the effect of a shift from having the information available to not having it available. In these cases, we are manipulating perception of race or gender. In our first example, the manipulation is $do(X = x)$ where X involves randomly assigning a racially identifiable name to a resume. In the second, the manipulation is the access to information about gender, the "do command" creates an environment in which discrimination based on gender is not possible because the information is unavailable. This case is closer to the counterfactual we are most interested in exploring in the context of labor markets, i.e., the expected outcome (e.g., with regard to hiring or wages) if there were no labor market discrimination.

Counterfactuals at work

An example of the counterfactuals of interest in labor market discrimination is the wages of individuals and groups without discrimination. However, when the question is one of attribution, i.e., the existence of labor market discrimination or the mechanism under which it operates, we clearly cannot observe these counterfactuals. Thus we ask: What is the causal relationship among race, labor market discrimination, and wages? As we mentioned above, race certainly does not cause labor market discrimination. And labor market discrimination does not cause race, although this idea does correspond with the notion that race and gender as social constructs only have meaning through their consequences. Perhaps it can be thought of as an interaction effect-discrimination acts on race (or gender). Due to the lack of direct data on discrimination, instead of using discrimination in statistical models, we use race (or gender) with the idea that if there were no labor market discrimination then we would expect people with similar productivity to earn equal wages. Thus, in the absence of discrimination, and with complete information on productivity, race should not be associated with wages. The problem with implementing this logic is that we do not have complete information on productivity and thus we cannot discern whether an observed association between race and wages is due to unobserved differences in productivity or is evidence of labor market discrimination. Conversely, suppose an association between race and wages is not observed once productivity variables are controlled for. This again could be due to unobserved differences in productivity (in the opposite direction), or it could be evidence of a lack of labor market discrimination.

Using the potential outcome notation for discrimination

Pearl suggests that we use the notion of potential outcomes in the absence of a potential experiment in order to attribute a cause to an effect. In the context of discrimination, we observe that race (or sex) is associated with wages, and we would like to determine how much of this difference by race can be attributed to labor market discrimination. This has been the focus of an extended research effort in the economics literature on discrimination, mainly involving the functional forms used to assign these attributions.

Suppose, as in Haviland (2003), our goal is to assess how much of the observed wage difference between demographic groups, highly educated members of different racial and ethnic groups, is attributable to discrimination based on demographic group membership after adjusting for differences in qualifications. Let the unadjusted wage gap be defined by the difference in conditional expected values,

$$\Gamma(G_j) = E(y_1|G_j = 1) - E(y_0|G_j = 0), \tag{1}$$

where y_1 and y_0 are the natural logarithm of wages as though one is treated as a member of the demographic group of interest or the base comparison group respectively, $G_j = 1$ indicates that respondents are a member of the demographic group j (black, Hispanic, or Asian men), $G_j = 0$ indicates that respondents are a members of the base comparison group (non-Hispanic white men), and $\Gamma(G_j)$ is the wage gap for group G_j. Besides demographic group membership, however, there are other characteristics that affect wages and whose distributions differ between the groups. In attempting to isolate the effect of market discrimination, the goal is to control for pre-market characteristics in the estimates of the wage gap and decompose the total wage gap into an amount associated with differences in these characteristics and the amount remaining. Ideally, we would have wages for each person, with his own characteristics besides group membership, as though he were a white male and as though he were a member of his own demographic group. (Similarly, we would want to have wages for each white male as he would be paid were he a member of each other demographic group.) *It is these missing counterfactuals that we need to estimate in order to obtain an estimate of the average wage gap not associated with differences in the distributions of the covariates.*

Researchers have focused on obtaining a sufficient set of confounders to estimate these missing counterfactuals consistently. The potential outcome literature refers to this as the strong ignorability criterion: given the set of covariates, group membership is independent of y_0 (Heckman et al., 1998). Note that there are several versions of this condition depending on functional form and methodology (this form is specific to estimating just the effect of 'treatment on the treated' and so refers only to y_0). We can address this type of assumption well by using Pearl's suggested path diagrams. If we assume that we have obtained such a set, then we can decompose the unadjusted gap into a portion that is associated with differences in the distributions of the observed covariates and a portion that either is associated with differences in the returns to these covariates or that is not associated with the covariates.

A consequence of the assumption of strong ignorability of group assignment given the covariates stated previously is that $E(y_0|G_j = 1, X = x) = E(y_0|G_j = 0, X = x)$. This assumption, in combination with the rest, allows us to consistently estimate the decomposition of the unadjusted wage gap into explainable and unexplainable portions. To demonstrate, let the overall average wages in group G_j be expressed as

$$E(y_1|G_j = 1) = \sum_X p_{jx} E(y_1|G_j = 1, X = x), \tag{2}$$

where $E(y_1|G_j = 1, X = x)$ is the expected earnings in group G_j with characteristic $X = x$, and p_{jx} is the proportion of members of group G_j with characteristic $X = x$. We can consistently estimate these values using standard data sources and estimation techniques. Similarly, for white men we can write

$$E(y_0|G_j = 0) = \sum_X p_{Wx} E(y_0|G_j = 0, X = x). \tag{3}$$

Substituting equations (2) and (3) into equation (1) yields

$$\begin{aligned}
&E(y_1|G - j = 1) - E(y_0|G - j = 0)\\
&= \sum_X p_{jx} E(y_1|G_j = 1, X = x) - \sum_X p_{Wx} E(y_0|G_j = 0, X = x)\\
&= \sum_X p_{jx} \left[E(y_1|G_j = 1, X = x) - E(y_0|G_j = 0, X = x)\right]\\
&\quad - \sum_X \left[p_{Wx} - p_{jx}\right] E(y_0|G_j = 0, X = x).
\end{aligned} \tag{4}$$

Our strong ignorability assumption allows us to use equation (4), which may be observed, to estimate

$$\begin{aligned}
&E(y_1|G-j=1) - E(y_0|G-j=0) \\
&\quad = \sum_X p_{jx}\left[E(y_1|G_j=1, X=x) - E(y_0|G_j=1, X=x)\right] \\
&\qquad - \sum_X \left[p_{Wx} - p_{jx}\right] E(y_0|G_j=0, X=x), \qquad (5)
\end{aligned}$$

where the second summation gives the portion of the unadjusted gap associated with differences in the distribution of confounders, and the first is the portion associated with unexplained differences in the returns to these confounders.

These models are typically estimated parametrically, with separate linear regressions for each racial group (or each sex). In this framework the expected values are regression coefficients, and what is referred to in the example above as the probability of having a particular characteristic is replaced by the mean of the characteristic in each group. Three problems with these parametric models have been noted in the literature. First, there is often a substantial lack of support in the data to make comparisons over large portions of the union of the domains of the data for the groups of interest. This problem may be exacerbated by the second problem, which is incorrect functional form. The final, related problem is that the parametric version of the decomposition (referred to as the Blinder-Oaxaca model and traditionally used in studies of gender, racial, and ethnic wage gaps, e.g., see the review by Altonji and Blank (1999)) makes different predictions depending on whether the coefficients from the regression on the control group, the group of interest, or a pooled group are used to decompose the unadjusted gap. These points and their potentially substantial effects on estimates are demonstrated in Barsky et al. (2002).

One nonparametric alternative, matching, provides an intuitively clear method for estimating the missing counterfactuals while avoiding the pitfalls of parametric models in this context. To estimate the missing counterfactual for a 32-year-old Hispanic man with a master's degree in business administration, i.e., his expected wage if he were paid as a white male, Haviland (2003) uses the mean of the wages of white men of the same age with the same highest degree and field. Assuming these counterfactuals can be estimated for each member of the demographic group, the mean gap conditional on age and education can be estimated by averaging over the gaps for each individual in the group of interest. This estimate is often referred to as the effect of 'treatment on the treated' where in this case 'treatment' is demographic group membership under the condition of labor market discrimination. Similarly, Pearl (2001) uses the term 'natural direct effect' to describe the change we would expect in men's employment if they were treated as women by employers which is how an "effect of treatment on the untreated" would be interpreted in this context.

The nonparametric method described above has the additional advantage of making problems of support transparent. In fact, the matching method of estimating the missing counterfactuals is only consistent if it is used over the intersection of the supports of both distributions (Heckman et al., 1998). It also provides an estimate of the full distribution of wage gaps instead of a single point estimate for the mean.

As Pearl discusses, the fact that there are difficulties with the parametric models here has consequences for attributing causal effects. He suggests that assigning attribution is only possible with parametric models where the form of a specific effect of a variable or set of variables is considered known and thus can be applied outside the range of that data. It is clear from the econometric literature that, even assuming the set of confounders selected to be in the model is sufficient to meet the strong ignorability criterion, we cannot justify the parametric models. The nonparametric alternatives require us to make estimates only within the range of the data, and thus we must make all assumptions for estimating parameters not directly encountered in the data explicitly and not in the form of a parametric model. Indeed, Pearl suggests that with nonparametric models it is not possible to attribute causes, at least not in the areas where there is no overlap of the data. This is similar to the problems with propensity score models when the probability of being in the treated group is zero or one in some range of the variables that are being controlled for (Rosenbaum and Rubin, 1983).

In summary, we highlight three sets of problems. First, we have to use race or gender as a proxy for discrimination because we typically cannot observe the mechanisms through which labor force discrimination occurs. To make direct inference about discrimination, we need other information about the data generating mechanism, e.g., how does labor market discrimination operate to effect wages. Without this information we need other untestable assumptions to carry over from one circumstance to another. Second, even when our goal is simply to partition observed differences in the outcome to known and unknown reasons, there are a host of problems with what variables to include in the partition. Third, functional forms (parametric versus nonparametric and the form of the parametric models) are problematic and affect whether or not we can make causal attributions.

Pearl recommends the use of path diagrams to encode causal assumptions and determine the conditions under which causal inference is possible. Given the preceding discussion, it is unclear how to represent the labor market discrimination and non-labor market discrimination factors affecting wages in a causal path diagram. Pearl notes this difficulty in circumstances where an experiment is not possible; this is the case with discrimination when the observable variable, race, is not manipulable. Under these circumstances the researcher's goal is to determine a sufficient set of covariates whose distributions differ by race but are not due to labor market discrimination in order to partition any difference in outcome by race. Pearl's 'back-door' criterion, while difficult to apply in detail without a path diagram, is intuitively unlikely to be met by observational studies of discrimination because the factors that affect wages and differ by race but are not affected by market discrimination are difficult or impossible to measure. On the other hand, these tools may all be used when race is replaced in the model by the perception of race, which can be manipulated.

Back doors and the resume experiment

Bertrand and Mullainathan (2003) solved the problems with the 'back door' links in a path diagram for discrimination by randomizing a proxy variable for race. They used real resumes with all personal identification removed and matched these resumes to help-wanted advertisements in Chicago and Boston. Then they randomized "African American sounding names" and "white sounding names" to the pairs of resumes matched to each advertisement. This randomization breaks the links between the education and work experience on the resume (what an employer observes when making a decision to call back a potential employee for an interview or not) and race. This is what the typical observational studies cannot ensure: no matter how many characteristics we attempt to control for, we may not have the correct functional form and/or there may be other unobserved confounders.

There are both positive and negative consequences of their focus on a particular early stage of the employment process. It is unclear how to relate the observed experimental outcome (call backs) to more typical labor market outcomes such as employment or wages. On the other hand, this focus makes it possible to randomize the perception of race across hypothetical people. In addition, the specificity makes this experiment similar to a small-scale case study in that it can get closer to the actual mechanism by which labor market discrimination occurs, the racial distinctiveness of a name affects the probability of call backs. Thus it breaks the ties between the usual host of confounders between race and the indicators of productivity that potential employers observe, and it comes closer to identifying the mechanism through which labor market discrimination operates. (This assumes that having a racially distinctive name does not carry any other information or connotations besides race itself.) Both the randomization and the identification of a mechanism are important for making a causal attribution and thus making it possible to make predictions outside the current situation. For instance, it may be possible to estimate the effect of particular policy changes such as removing names from resumes before call back decisions are made or having employment agencies use codes or some other technique.

Concluding remarks

Pearl's paper provides us with an excellent review of a number of issues from the recent literature on causal modeling and the ways that they link to the literature on counterfactuals and potential outcomes. We have attempted in our discussion to show the direct relevance of these ideas to an important policy problem to which Pearl actually refers in his opening paragraph. We have explained not only why the issue of discrimination is of interest in the context of causal models but also why it is far more complex than many authors have hitherto suggested. The tools of causal inference are, we believe, essential to such a discussion.

David Heckerman
Microsoft Research
Redmond, U.S.A.
Ross Shachter
Stanford University
Stanford, U.S.A.

We have been working with causal graphical models for over a decade and believe that many statisticians and engineers would benefit from using this approach, recognizing the crucial distinction between passive inference and active intervention. Consequently, we are delighted to see Professor Pearl summarize his causal framework in this publication.

His framework is quite elegant and builds on seminal works in statistics and decision making. His "*do*" graph surgery for causal models, for example, was first proposed by Strotz and Wold (1960), and "*do*" itself is a classic decision, albeit with some strong but subtle assumptions. Nonetheless, as Pearl describes in this contribution, statisticians have been slow to adopt his work. Pearl gives two possible reasons why this is so. Here, we offer and address another reason, which we find more compelling.

We first saw Pearl's causal framework in 1993. Although the framework seemed quite promising to us, we were puzzled by the meanings of and assumptions underlying a critical component of his framework—namely, the causal model.

Pearl gave (and still gives) a clear definition of cause-effect in terms of the *do* operator (e.g., Pearl, 2000, page 204) and a clear definition of the *do* operator in terms of a given causal model (e.g., Pearl, 2000, page 70). However, his definition of causal model was (and still is) unclear. For example, on page 44 of his book, Pearl writes:

> **Definition (Causal Structure).** *A* causal structure *of a set of variables V is a directed acyclic graph (DAG) in which each node corresponds to a distinct element of V, and each link represents a direct functional relationship among the corresponding variables.*
> **Definition (Causal Model).** *A* causal model *is a pair $M =< D, \Theta_D >$ consisting of a causal structure D and a set of parameters Θ_D compatible with D. The parameters Θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each $X_i \in V$ and a probability measure $P(u_i)$ to each u_i, where PA_i are the parents of X_i in D and where each U_i is a random disturbance distributed according to $P(u_i)$, independently of all other u.*

Although it appears to be mathematically precise, we find this definition to be confusing. What is the source or semantics of the "direct functional relationship among the corresponding variables" and the "random disturbances"? What assumptions are we making when building such a model?

To answer these questions, we developed a foundation for Pearl's framework based on decision theory—in particular, the work of Savage (1954). That is, we defined Pearl's causal model as well as the *do* operator and cause and effect in terms of Savage's primitives (Heckerman and Shachter, 1995). As an additional benefit, we showed how Pearl's causal model can be equivalently represented as an influence diagram—a graphical representation of decision making under uncertainty used now for almost three decades (Miller et al., 1976).

Once we found this path from Savage to Pearl, we became quite comfortable with Pearl's framework and the assumptions necessary to apply it. Unfortunately, Pearl has downplayed the strong connections between his work and decision theory as well as the suitability of the influence diagram as a representation of causal interactions. On the contrary, we believe that people who are familiar with decision theory will find comfort, as we have, in these connections. Of course, some statisticians are not familiar with decision theory and will need to understand new concepts whether studying Pearl's ideas directly or through an initial study of decision theory. In such cases, we believe these researchers might come to an appreciation of Pearl's work most easily by first gaining an understanding of the implicit decision theory.

So how do we understand Pearl's work in terms of decision theory? Consider Savage's framework, which begins with a description of his primitives *act*, *consequence*, and *possible state of the world.* Savage (1954, pages 13–14) describes and illustrates these concepts as follows:

> To say that a decision is to be made is to say that one or more acts is to be chosen, or decided on. In deciding on an act, account must be taken of the possible states of the world, and also of the consequences implicit in each act for each possible state of the world. A *consequence* is anything that may happen to the person.
>
> Consider an example. Your wife has just broken five good eggs into a bowl when you come in and volunteer to finish making the omelet. A sixth egg, which for some reason must either be used for the omelet or wasted altogether, lies unbroken beside the bowl. You must decide what to do with this unbroken egg. Perhaps it is not too great an oversimplification to say that you must decide among three acts only, namely, to break it into

state of the world	act		
	break into bowl	break into saucer	throw away
good egg	six-egg omelet	six-egg omelet and a saucer to wash	five-egg omelet and one good egg destroyed
bad egg	no omelet and five good eggs destroyed	five-egg omelet and a saucer to wash	five-egg omelet

Table 1: *An example illustrating acts, possible states of the world, and consequences. (Taken from Savage, 1954.)*

> the bowl containing the other five, to break it into a saucer for inspection, or to throw it away without inspection. Depending on the state of the egg, each of these three acts will have some consequence of concern to you, say that indicated by Table 1.

There are two points to emphasize. First, like Pearl but with different language, Savage distinguishes between that which we can choose—namely, acts—and that which we can see—namely, consequences. Second, once we choose an act, the consequence that occurs is logically determined by the state of the world. That is, the consequence is a deterministic function of the act and the state of the world[23]. Of course, the consequences can be (and usually are) uncertain, and this uncertainty is captured by uncertainty in the state of the world.

With this understanding, the relationship between Savage and Pearl is not difficult to see. The instances of Pearl's U correspond to Savage's states of the world; the instances of Pearl's V correspond to Savage's consequences; the instances of Pearl's *do* operators correspond to Savage's acts; and Pearl's functional relationships correspond to Savage's deterministic mapping from acts and states of the world to consequences. Of course, there are many details left unsaid; and we encourage the reader to explore them in Heckerman and Shachter (1995).

One important detail worth mentioning is the special nature of the *do* operator. Given some variable X, the set of acts corresponding to $do(x)$ can not be a set that arbitrarily affects X. In particular, $do(x)$ must be a set of acts that affects X in an "atomic" way—a way that affects only X directly.

[23] Savage (1954) *defines* an act to be "a function attaching a consequence to each state of the world." Equivalently, we think of acts and a function mapping acts and states of the world to consequences as separate entities. This is a common practice in decision theory.

For example, if X corresponds to a person's wealth, the act of giving the person tax-free money might qualify as a $do(x)$, whereas the act of giving the person stolen (and marked) money would not, due to its additional (presumably undesirable) consequences. In Heckerman and Shachter (1995) we give a precise meaning to the notion of an atomic decision in terms of Savage's primitives, and hence define the *do* operator in terms of these primitives.

In summary, we believe Pearl's framework is not unlike a beautiful island filled with delights and riches. As Pearl presents it, however, there is no boat to the island. At best, one has to swim through perilous waters to get there. We offer our work as a safe and comfortable transport to his paradise.

Joseph B. Kadane
Department of Statistics
Carnegie Mellon University,
Pittusburg, U.S.A.

This is an excellent paper that reviews and updates the notable progress that has been made by the author and others in recent decades in understanding and unrevealing the mysteries of causation. I especially commend the author for his clarity of exposition; I would recommend this paper to someone wanting a friendly introduction to what has been up to now a daunting, difficult and fragmented literature.

I am struck by how compatible this work is with the Bayesian perspective. As presented by Professor Pearl, the most useful representation of a causative model is graphical; the absence of a link implies an assumption. It may not be obvious to a reader which model to assume. Indeed a reader may wish to entertain a variety of models, each representing a possible state of the world. Attaching subjective probabilities to each of these models, the whole Bayesian machinery can be used to derive not only consequences of each model separately, but also posterior probabilities of the models themselves.

It may be objected that there will be examples in which the data are not informative about which model obtains, and this can be foretold with certainty before the data are available or are examined. This is neither upsetting nor a tragedy, as it also happens in models not involving a causal issue. What it comes down to is that certain functions of the parameter vector are not identified. But even in such a circumstance, a legitimate posterior results and can be calculated. It is a problem for those who wish to use improper prior distributions for parameters taking an infinite number of possible values, but wish to insist on a proper posterior distribution. However for those with proper prior distributions, lack of identification is not a difficult or serious issue. That even an infinite number of observations of a particular type will not settle certain questions (i.e., consistency) need not be disturbing, both because one rarely has such a data set, and because often we have questions that particular data sets do not and cannot resolve. (For more about my views on identification in a Bayesian framework, see Kadane, 1975).

How do we deal socially with such situations? For example, what has happened to R.A.Fisher's contention that there was no proof that smoking causes cancer? There still is no such proof, if by proof we mean the conduct of an (obviously unethical) experiment in which some persons would be randomized to smoking and be forced to smoke, while others would be randomized not to smoke, and would be forced not to. Nonetheless, both policy makers and the general public have come to the conclusion that smoking is really bad for people, and cancer is one of the reasons. What has occurred in the intervening 50 years? Fisher (1957) claimed that perhaps incipient cancer caused people to smoke or, perhaps more plausibly, that there was some unobserved variable that led people to be more likely to smoke, and be more likely to get cancer. Cornfield et al. (1959) challenge many of the claims made on behalf of various hypotheses that would exonerate smoking as a cause of lung cancer. They also give an analysis to show that an unmeasured covariate causing both smoking and cancer would have to have a relative risk greater than the 8 observed for smoking to be an exonerating explanation of the data, thus anticipating some of the work on bounding correlations that Pearl and his colleagues have extended.

Perhaps there is such a variable, but it is not unreasonable to ask what it is, and to take the view that, after 50 years, the burden of producing evidence is on those who wish to continue to take Fisher's side of the argument. Thus socially our priors have shifted, even without crisp proof.

We owe Professor Pearl our heartfelt thanks for making these fundamental issues much easier to understand. $Do(read\ Pearl)$.

Serafín Moral
Dpto. de Ciencias de la Computación
e Inteligencia Artificial
Universidad de Granada, Spain.

This is an excellent exposition of recent advances in causality by Prof. Judea Pearl. In the eighties he made very important contributions to probabilistic graphical models and wrote his influential book (Pearl, 1988) which has been a source of inspiration and knowledge for so many researchers around the world. Then, in the nineties he has concentrated in one of the most elusive and important concepts in scientific research: causality. In Bayesian networks causality was always present, but it was only seen through the induced independence relationships among the involved variables. Now, Pearl presents an approach to causal inference, which makes use of previous advances in graphical models. He shows the difficulties of measuring causal relationships using the classical statistics language and supports its extension in order to clearly specify causal assumptions and their implications. With it, former difficult and somewhat obscure reasonings can be formalized in a simple and neat way. The importance and implications of this work will be seen without doubt in the near future. Personally, I would like to comment two aspects in which causality should play a more important role.

The first one is related with the field of learning Bayesian networks. In the paper, it is claimed that causality can not be discovered from statistical knowledge, but this is precisely the ultimate objective of learning the structure of a Bayesian network, though this is done by seeing the independencies or by optimizing a measure of adjustment of the network to the data (score). My experience tells me that when there is a causal mechanism that generates the data, then with usual learning algorithms and if the sample is large enough, it is possible to discover the original structure or an equivalent one. However, when we try to learn from real data in which there are correlations but not causal relationships relating all the variables, we obtain networks in which arcs can not be given a causal interpretation. These are, in general, complex graphs that can be seen as the basis for approximating a joint probability distribution involving all the problem variables by means of products of smaller distributions. However, even if we can not assume 'a priori' causal assumptions, procedures are based in some implications of causality. For example, Bayesian scoring metrics assume that we have ‘a priori’ distributions about the parameters which are independent for different variables: the parameters for X are independent of the parameters of Y given X. Can this be maintained if the errors associated to X and to Y given X are not independent? This causal interpretation is more evident in procedures using observational and experimental data at the same time (Cooper and Yoo, 1999), in which the truncated factorization is used in the score derivation.

The discovering of causal relationships from observational data has been discussed in learning literature. But I think that we need more research effort for determining what can we expect when there is not initial causal relationships in the variables. It really would be useful to have tools to determine when a learned arc can be given a causal interpretation. Do we need some type of ‘a priori’ hypothesis?

The second comment is of different nature and it is related with the role of imprecise probability in causal inference. Very often Bayesians claim that it is always possible to assign initial ‘a priori’ probabilities, so that a precise probability can be computed for every event of interest. Afterwards, if these values are updated by a large sample of independent observations, then the sensitivity to the initial ‘a priori’ values is small. Here, we have an example in which this is not true. We have situations in which, even if we have precise probabilities for all the observed variables, then if we want to measure the strength of causal relationships, we can only determine bounds for these values. To be able of determining precise values, we should assign ‘a priori’ probabilities for non observable error variables, which would really looks something difficult for any expert. In the hypothetical case in which they were asserted, we should be extremely careful about the conclusions as they can have strong bias due to ‘a priori’ values even if we have a very large sample.

I believe that one important lesson of this paper is that we should be very clear about initial assumptions. A theory should provide means for specifying them in a clear an unambiguous language. Then, sound inference methods should be available to get information about events of interest. I believe that it is important for a methodology not to hide underlaying hypothesis and not to force to make more assumptions than one is ready to accept by the available information. In that case, imprecision in probabilities arises in a very natural way, as initial state of knowledge can be too weak to determine precise values (Walley, 1991).

But there is another aspect in which imprecise probability can play a role in causality. It is clear that when we have two variables and a joint probability distribution there is nothing that can help us to discriminate causality from correlation. But if instead of having and only one probability distribution we have several of them, then the situation can be different. Imagine that we have two variables X and Y and that we have two possible 'a priori' distributions p_1, p_2 about X and two possible conditional distributions of Y given X, q_1, q_2, in such a way that the possible joint probability distributions are all the products $p_1.q_1, p_1.q_2, p_2.q_1, p_2.q_2$. In this case, we way that the joint imprecise information about (X, Y) decompose as a marginal about X and a conditional probability about Y given X. This does not always happen and it does not imply that we can decompose the joint information as a marginal about Y and a conditional on X given Y. The decomposition on a marginal on X and a conditional on Y given X makes sense when X is a cause of Y with uncorrelated errors. So, with imprecise probability and with only two variables we have properties that are not always verified and that are implied by the existence of a causal relationship. A more detailed exposition of this idea can be found in Moral and Cano (2002).

Rejoinder by J. Pearl

I wish to thank the discussants for taking the time to comment on my paper, and further illuminating the subtle ways in which causality enters statistical analysis.

Professor Kadane makes an important observation that causal analysis would be fairly compatible with the Bayesian perspective if one is willing to attach subjective probabilities to several causal models, each representing a different configuration of mechanisms or a different vector of causal parameters. It should be remembered though that the number of causal models one normally wishes to entertain would be astronomical and the task of assigning prior probabilities to the space of causal models would thus be hindered by a difficult problem of representation. In practice, the problem is mitigated by assuming parameter independence, namely, that the parameters governing one child/parents family are independent of those governing another family. This leads to Bayesian scoring metrics mentioned by Professor Moral who points out correctly, that parameter independence, an assumption made routinely in Bayesian statistical literature, makes sense only under specific causal assumptions. For example, the parameters of the marginal of X can be assumed independent of the conditional of Y given X only if X is a cause of Y, but not if Y is a cause of X. This means that statisticians who use parameter independence in practice, are unwittingly making causal assumptions.

The functional, or mechanism-based analysis proposed in my paper explains the connection between parameter independence and causal directionality. Indeed, if each child-parent family is associated with a different physical mechanism, in the sense that changes in one mechanism can be affected independently of those in other mechanisms, it makes sense then that the parameters governing the conditional probabilities associated with those families would be independent of one another. No such independence is expected when the conditional probabilities do not characterize separate causal mechanisms.

Professor Moral has identified another area where tacit causal assumptions have been at the root of seemingly pure statistical assumptions – the composition of imprecise probabilities. Again, taking the Cartesian products of probability intervals makes sense under certain causal assumptions and not under others, and the reason is the same, Cartesian products are licensed by mechanism independence.

I welcome Fienberg and Haviland's analysis of the economical effect of discrimination within the correct framework of counterfactual analysis and structural models. One should admire econometricians for attempting to tackle such problems without a formal theory of causation.

I am delighted that Heckerman and Shachter (HS) agree with my observation that statisticians have been way too slow in embracing an analysis of causation. I do not agree however with their conclusion that statisticians would do well to approach causation via decision analysis. Quite the contrary, by insisting on decision analysis (DA) as a starting point HS exhibit the same inflexibility that has held back statistics for over a century; students of causation much break away with the confines of both associational analysis and decision analysis.

In Causality, page 110-112, I elaborate at some length on the relationships between DA and structural causal analysis. The weaknesses of the former lies precisely in Savage broad characterization of an act as a function between states of the world (the variables in U) and the set of consequences (the variables in V), without further explication of the internal nature of that function. Although correct, this characterization is not very informative; for it does not tell us how the effect of a compound actions can be derived from the effects of its components. It can be likened to the characterization of arithmetic addition as a function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$; though true, it is way too coarse, for it does not account for the axioms of addition and does not tell us, for example, how addition differs from multiplication, also a function from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{R}$.

DA assumes the existence of an oracle capable of predicting consistently the consequence of any act or combination of acts from any state of the world. Since the number of possible act-combinations is astronomical, it is inconceivable that human decision makers can make such predictions without further structure. Causal analysis explains how those predictions can be obtained from a more manageable set of primitives, structured in the form of a causal model. Given the set of functions f_i in a causal model, one can readily compute the consequence of any combination of acts, from any state of the world. These functional relations, which HS brand as "confusing" are in fact the building blocks on the basis of which Savage's oracle operates. Although it is easy to define these primitive functions in terms of Savage actions (see Causality page 205) I found this exercise to be counter productive and unnatural when, in reality, Savage actions are (judgmentally) computed from those primitive functions.

True, there is no boat to the paradise of causation if one insists on sailing the perilous waters of statistics. Even the heavy transport plane offered by HS is circular and risky. There is however a beautiful bridge to that island, friendly, safe and a walking distance away; let us not ignore the friendly way to redemption.

References

Altonji, J. and Blank, R. (1999). Gender and race in the labor market. In O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, vol. 3, pp. 3143–4259. Elsevier Science Press, New York.

ANGRIST, J. D., IMBENS, G. W., and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables (with comments). *Journal of the American Statistical Association*, 91(434):444–472.

BALKE, A. and PEARL, J. (1995). Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, eds., *Uncertainty in Artificial Intelligence*, vol. 11, pp. 11–18. Morgan Kaufmann, San Francisco.

BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176.

BARSKY, R., BOUND, J., CHARLES, K., and LUPTON, J. (2002). Accounting for the black-white wealth gap: A nonparametric approach. *Journal of the American Statistical Association*, 97:663–673.

BECHER, H. (1992). The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, 11:1747–1758.

BERTRAND, M. and MULLAINATHAN, S. (2003). Are Emily and Brendan more employable than Lakisha and Jamal? a field experiment on labor market discrimination. NBER Working paper 9873.

BISHOP, Y. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics*, 27:545–562.

BOLLEN, K. (1989). *Structural Equations with Latent Variables*. John Wiley, New York.

BONET, B. (2001). Instrumentality tests revisited. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 48–55. Morgan Kaufmann, San Francisco.

BOWDEN, R. J. and TURKINGTON, D. A. (1984). *Instrumental Variables*. Cambridge University Press, Cambridge, England.

BRESLOW, N. and DAY, N. (1980). *The Analysis of Case-Control Studies*, vol. 11 of *Statistical Methods in Cancer Research*. IARC, Lyon.

CARTWRIGHT, N. (1989). *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford.

Chickering, D. and Pearl, J. (1997). A clinician's tool for analyzing non-compliance. *Computing Science and Statistics*, 29(2):424–431.

Chou, C. and Bentler, P. (1995). Estimations and tests in structural equation modeling. In R. Hoyle, ed., *Structural Equation Modeling*, pp. 37–55. Sage, Thousand Oaks.

Cooper, G. F. and Yoo, C. (1999). Causal discovery from a mixture of experimental and observational data. In K. Laskey and H. Prade, eds., *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 116–125. Morgan Kaufmann.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203.

Cowell, R., Dawid, A., Lauritzen, S., and Spielgelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. Springer Verlag, New York.

Cox, D. (1958). *The Planning of Experiments*. John Wiley and Sons, New York.

Dawid, A. (1979). Conditional independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B*, 41(1):1–31.

Duncan, O. (1975). *Introduction to Structural Equation Models*. Academic Press, New York.

Eells, E. (1991). *Probabilistic Causality*. Cambridge University Press, Cambridge.

Fisher, R. A. (1957). Dangers of cigarette-smoking. *British Medical Journal*, 2:297–298.

Freedman, D. (1987). As others see us: A case study in path analysis (with discussion). *Journal of Educational Statistics*, 12(2):101–223.

Freedman, D. A. (2003). On specifying graphical models for causation, and the identification problem. Technical Report 601, Department of Statistics, University of California, Berkeley.

GALLES, D. and PEARL, J. (1995). Testing identifiability of causal effects. In P. Besnard and S. Hanks, eds., *Uncertainty in Artificial Intelligence*, vol. 11, pp. 185–195. Morgan Kaufmann, San Francisco.

GOLDBERGER, A. (1972). Structural equation models in the Social Sciences. *Econometrica: Journal of the Econometric Society*, 40:979–1001.

GOLDIN, C. and ROUSE, C. (2000). Orchestrating impartiality: The impact of 'blind' auditions on female musicians. *American Economic Review*, 90:715–741.

GRAYSON, D. (1987). Confounding confounding. *American Journal of Epidemiology*, 126:546–553.

GREENLAND, S. (1999). Relation of the probability of causation to the relative risk and the doubling dose: A methodologic error that has become a social problem. *American Journal of Public Health*, 89(8):1166–1169.

GREENLAND, S., PEARL, J., and ROBINS, J. (1999a). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

GREENLAND, S. and ROBINS, J. (1986). Identifiability, exchangeability, and epidemiological confounding. *International Journal of Epidemiology*, 15(3):413–419.

GREENLAND, S., ROBINS, J., and PEARL, J. (1999b). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

HAUCK, W., HEUHAUS, J., KALBFLEISCH, J., and ANDERSON, S. (1991). A consequence of omitted covariates when estimating odds ratios. *Journal Clinical Epidemiology*, 44(1):77–81.

HAVILAND, A. M. (2003). Understanding racial and gender wage gaps among the highly educated. Unpublished Ph.D. dissertation. Carnegie Mellon University, Dept. of Statistics and Heinz School of Public Policy.

HECKERMAN, D. and SHACHTER, R. (1995). Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430.

HECKMAN, J. and SMITH, J. (1998). Evaluating the welfare state. In S. Strom, ed., *Econometric and Economic Theory in the 20th Century*, pp. 1–60. Cambridge University Press, Cambridge.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an Econometric evaluation estimator. *Review of Economic Studies*, 65:261–294.

Holland, P. (1953). Identification problems in econometric model construction. In W. Hood and T. Koopmans, eds., *Studies in Econometric Method*, pp. 27–48. Wiley, New York.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Holland, P. (1988). Causal inference, path analysis, and recursive structural equations models. In C. Clogg, ed., *Sociological Methodology*, pp. 449–484. American Sociological Association, Washington.

Holland, P. and Rubin, D. (1988). Causal inference in retrospective studies. *Evaluation Review*, 13:203–231.

Imbens, G. and Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Joreskog, K. and Sorbom, D. (1978). *LISREL IV: Analysis of Linear Structural Relationships by Maximum Likelihood*. International Educational Services, Chicago.

Kadane, J. B. (1975). The role of identification in Bayesian theory. In S. E. Fienberg and A. Zellner, eds., *Studies in Bayesian Econometrics and Statistics*, pp. 175–191. North Holland Publishing, Amsterdam.

Kaufman, J. and Kaufman, S. (2001). Assessment of structured socioeconomic effects on health. *Epidemiology*, 12(2):157–167.

Kiiveri, H., Speed, T., and Carlin, J. (1984). Recursive causal models. *Journal of Australian Mathematical Society*, 36:30–52.

Kleinbaum, D., Kupper, L., K.E., M., and Nizam, A. (1998). *Applied Regression Analysis and Other Multivariable Methods*. Duxbury Press, Pacific Grove, 3rd ed.

Kuroki, M. and Miyakawa, M. (1999). Identifiability criteria for causal effects of joint interventions. *Journal of the Japan Statistical Society*, 29(2):105–117.

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lauritzen, S. L. (1999). Causal inference from graphical models. Technical Report R-99-2021, Department of Mathematical Sciences, Aalborg University, Denmark.

Lindley, D. and Novick, M. (1981). The role of exchangeability in inference. *The Annals of Statistics*, 9(1):45–58.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Harvard University Press, Cambridge, MA.

Miettinen, O. and Cook, E. (1981). Confounding essence and detection. *American Journal of Epidemiology*, 114:593–603.

Miller, A., Merkhofer, M., Howard, R., Matheson, J., and Rice, T. (1976). Development of automatic aids for decision analysis. Technical report, SRI International, Menlo Park, CA.

Moral, S. and Cano, A. (2002). Strong conditional independence for credal sets. *Annals of Mathematics and Artificial Intelligence*, 35:295–321.

Muthen, B. (1987). Response to Freedman's critique of path analysis: Improve credibility by better methodological training. *Journal of Educational Statistics*, 12(2):178–184.

Neyman, J. (1923). On the application of Probability Theory to agricultural experiments. Essay on principles. *Statistical Science*, 5(4):465–480.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo.

Pearl, J. (1993). Comment: Graphical models, causality, and intervention. *Statistical Science*, 8:266–269.

Pearl, J. (1995a). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

Pearl, J. (1995b). On the testability of causal models with latent and instrumental variables. In P. Besnard and S. Hanks, eds., *Uncertainty in Artificial Intelligence*, vol. 11, pp. 435–443. Morgan Kaufmann, San Francisco.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann, San Fransisco.

Pearl, J. and Robins, J. (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, eds., *Uncertainty in Artificial Intelligence*, vol. 11, pp. 444–453. Morgan Kaufmann, San Francisco.

Pearl, J. and Verma, T. (1991). A theory of inferred causation. In J. Allen, R. Fikes, and E. Sandewall, eds., *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, vol. 11, pp. 441–452. Morgan Kaufmann, San Mateo, CA.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – applications to control of the healthy workers survivor effect. *Mathematical Modeling*, 7:1393–1512.

Robins, J. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, 40(Suppl. 2):139S–161S.

Robins, J. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley, eds., *Health Service Research Methodology: A Focus on AIDS*, pp. 113–159. Public Health Service, Washington, DC.

Robins, J. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, 12(3):313–320.

Robins, J. and Greenland, S. (1989). The probability of causation under a stochastic model for individual risk. *Biometrics*, 45:1125–1138.

Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Rosenbaum, P. and Rubin, D. (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Savage, L. (1954). *The Foundations of Statistics*. Dover, New York.

Simon, H. (1953). Causal ordering and identifiability. In W. C. Hood and T. Koopmans, eds., *Studies in Econometric Method*, vol. 11, pp. 49–74. Wiley and Sons, Inc.

Simon, H. and Rescher, N. (1966). Cause and counterfactual. *Philosophy and Science*, 33:323–340.

Sobel, M. (1998). Causal inference in statistical models of the process of socioeconomic achievement. *Sociological Methods and Research*, 27(2):318–348.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

Strotz, R. and Wold, H. (1960). Recursive vs. nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland Publishing Co., Amsterdam.

Tian, J., Paz, A., and Pearl, J. (1998). Finding minimal separating sets. Technical Report R-254, University of California, Los Angeles.

Tian, J. and Pearl, J. (2000). Probabilities of causation: Bounds and identification. In *Proceedings of the Sixtheenth Conference on Uncertainty in Artificial Intelligence*, pp. 589–598. Morgan Kaufmann, San Francisco, CA.

Tian, J. and Pearl, J. (2002). On the identification of causal effects. In *Proceedings of the American Association of Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park.

WALLEY, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York.

WEINBERG, C. (1993). Toward a clearer definition of confounding. *American Journal of Epidemiology*, 137:1–8.

WERMUTH, N. (1992). On block-recursive regression equations. *Brazilian Journal of Probability and Statistics*, 6:1–56. With discussion.

WERMUTH, N. and COX, D. (1993). Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester.

WHITTEMORE, A. (1978). Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, Series B*, 40(3):328–340.

WRIGHT, S. (1921). Correlation and causation. *Journal of Agricultural Research*, 20(3):557–585.