Health Services & Outcomes Research Methodology 2:189–220, 2001

Causal Inference in the Health Sciences: A Conceptual Introduction

JUDEA PEARL
Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA 90024
E-mail: judea@cs.ucla.edu

Received February 15, 2001; revised December 18, 2001; accepted January 3, 2002

Abstract. This paper provides a conceptual introduction to causal inference, aimed to assist health services researchers benefit from recent advances in this area. 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 underlie 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, corrections for noncompliance, and a symbiosis between counterfactual and graphical methods of analysis.

Keywords: structural equation models, confounding, noncompliance, graphical methods, counterfactuals

1. Introduction

The research questions that motivate most studies in the health sciences are causal in nature. For example, what is the efficacy of a given drug in a given population? What fraction of deaths from given disease could have been avoided by a given treatment or 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, exposures, preconditions, and outcomes. Good statisticians have always known that the elucidation of causal relationships from observational studies must be shaped by knowledge (or assumptions) about how the data were generated; such assumptions are crucial to causal inference. This paper introduces useful language and tools for clarifying such assumptions and 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 health 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 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 health research 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 Sorbon, 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 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. Nancy 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 health research 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 (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 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 $P(disease \mid symptom)$ 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.

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's (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. A co-reviewer on a paper I have read recently offered the following objection to the method exemplified by the author:

> "A thoughtful and knowledgeable epidemiologist could write down two or more equally plausible models that leads to different conclusions regarding confounding."

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 quantitative health science 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 practice of estimating direct effects by controlling intermediate variables can lead to biased estimates. Again, using counterfactual notation, Robins and Greenland (1989) and 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 health 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 Sewall 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]

$$y = \beta x + u \tag{1}$$

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

Equation (1) still does not properly express the causal relationship implied by this assignment process, because equations are symmetrical objects; if we re-write (1) as

$$x = (y - u)/\beta \tag{2}$$

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 Yon 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

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

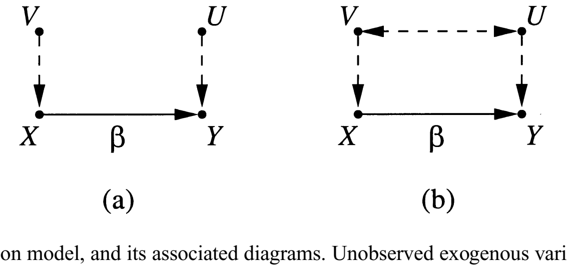

Figure 1. A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

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 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

$$Cov(X, Y) = \beta \tag{3}$$

for Figure 1(a), and

$$Cov(X, Y) = \beta + Cov(U, V) \tag{4}$$

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 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:

$$\begin{aligned} z &= f_Z(w) \\ x &= f_X(z, v) \\ y &= f_Y(x, u) \end{aligned} \tag{5}$$

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

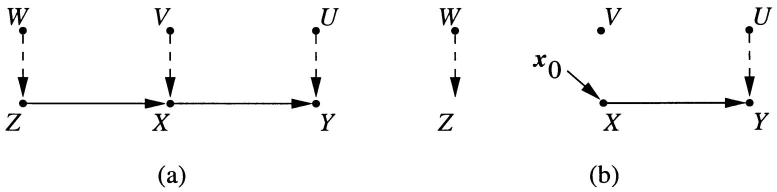

Figure 2. (a) The diagram associated with the structural model of Eq. (5). (b) The diagram associated with the modified model of Eq. (6), representing the intervention $do(X = x_0)$.

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; Koopmans, 1953).

3.2.1. 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 Eq. (5) with $x = x_0$, and obtain a new model, M_{x_0},

$$\begin{aligned} z &= f_Z(w) \\ x &= x_0 \\ y &= f_Y(x, u) \end{aligned} \tag{6}$$

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 Eq. (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

$$E(Y|do(x_0')) - E(Y|do(x_0)) \tag{7}$$

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

$$E(Y|do(x_0'))/E(Y|do(x_0)). \tag{8}$$

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 (1995a, 2000) (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 1. (*The Causal Markov Condition*)

Any distribution generated by a Markovian model M can be factorized as:

$$P(v_1,\ v_2, \ldots, v_n) = \prod_i P(v_i|pa_i) \tag{9}$$

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

$$P(z,\ y,\ x) = P(z)P(x|z)P(y|x) \tag{10}$$

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

Corollary 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

$$P(v_1, v_2, \ldots, v_k|do(x_0)) = \prod_{i|V_i \notin X} P(v_i|pa_i)|_{x=x_0} \tag{11}$$

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

Corollary 1 instructs us to remove from the product of Eq. (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 1, it must generate a distribution that is factorized according to the modified graph, yielding the truncated product of Corollary 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 Eq. (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.

3.2.2. Deriving Causal Effects

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

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x) \tag{12}$$

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

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0) \tag{13}$$

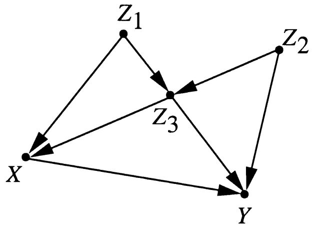

Figure 3. Markovian model illustrating the derivation of the causal effect of *X* on *Y*, Eq. (14). Error terms are not shown explicitly.

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

$$P(y|do(x_0)) = \sum_{z_1,z_2,z_3} P(z_1)P(z_2)P(z_3|z_1,z_2)P(y|z_2,z_3,x_0) \tag{14}$$

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

$$P(z_1,z_3,y|do(x),do(z_2)) = P(z_1)P(z_3|z_1,z_2)P(y|z_2,z_3,x) \tag{15}$$

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

$$P(y|do(x),\ do(z_2)) \sum_{z_1,z_3} P(z_1)P(z_3|z_1,z_2)P(y|z_2,z_3,x) \tag{16}$$

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).

3.2.3. *Coping with Unmeasured Confounders*

Things are more complicated when we face unmeasured confounders. For example, it is not immediately clear whether the formula in Eq. (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

$$P(y|do(x_0)) = \sum_{z_2,z_3} P(z_2)P(z_3|z_2)P(y|z_2, z_3, x_0) \tag{17}$$

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.

3.2.4. *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 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 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 each of the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, and $\{W_2, Z_3\}$ 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 to 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

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

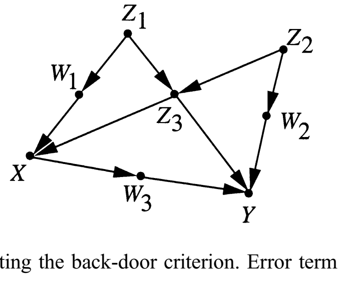

Figure 4. Markovian model illustrating the back-door criterion. Error terms are not shown explicitly.

and

$$P(Y = y|do(X = x)) = \sum_{s} P(Y = y|X = x, S = s)P(S = s) \tag{18}$$

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 Eq. (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).

3.2.5. 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

$$P(y|do(x)) = \sum_{w_3} P(w_3|do(x))P(y|do(w_3)) \tag{19}$$

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 3–4).

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 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 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 Eq. (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 Eq. (6), formed by the intervention $do(X = x_0)$ (Fig. 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 $P(Y_{x_1} = y_1 | Y = y_0, X = x_0)$ 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 Fig. 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_{x1} must take on 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 and Greenland's 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

3.4.1. Formulating the Assumptions

Consider the model of Figure 5(a) and Eq. (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 Eq. (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.

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 Eq. (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 are governed by the modified model of Eq. (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.

$$P(Y = y|do(x_0)) = \sum_v P(Y = y|X = x_0, V = v)P(V = v) \tag{20}$$

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 (Balka 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 Eq. (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]

$$y = f_Y(x, u) = \beta x + u; \tag{21}$$

multiplying this equation by z and taking expectations, gives

$$\beta = Cov(Z, Y)/(Cov(Z, X) \tag{22}$$

which reduces β to correlations among observed measurements. Eq. (22) is known as the *instrumental variable* estimand (Bowden and Turkington, 1984).

Similarly, Angrist et al (1996) have shown that certain nonlinear restrictions of the function f_X and f_Y may render the causal effect identifiable.

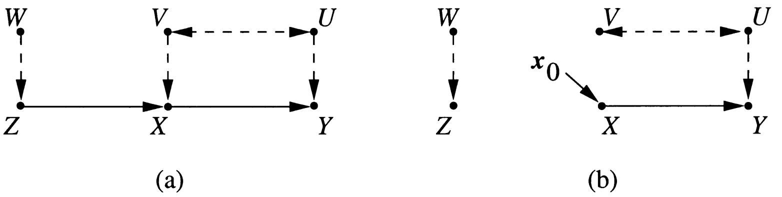

Figure 5. (a) Causal diagram representing a clinical trial with imperfect compliance. (b) A diagram representing interventional treatment control.

3.4.2. 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 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 Eq. (7) subject to the constraint provided by the observed distribution

$$\begin{aligned} P(x, y|z) &= \sum_{v,u} P(x, y, v, u|z) \\ &= \sum_{v,u} P(y|x, u, v)P(x|z, v)P(u, v) \end{aligned} \tag{23}$$

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

$$P(y|do(x')) - P(y|do(x'')) = \sum_{u} [P(y|x', u) - P(y|x'', u)]P(u) \tag{24}$$

Our task is then to bound the expression in Eq. (24) given the observed probabilities $P(y, x|z)$ as expressed in Eq. (23). This task amounts to a constrained optimization exercise of finding the highest and lowest values of Eq. (24) subject to the equality constraint in Eq. (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.

3.4.3. 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 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

$$\max_z \sum_y \left[\max_z P(x, y|z)\right] \leq 1 \tag{25}$$

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, Eq. (25) reduces to the four inequalities

$$\begin{aligned} P(Y=0, X=0|Z=0) + P(Y=1, X=0|Z=1) \leq 1 \\ P(Y=0, X=1|Z=0) + P(Y=1, X=1|Z=1) \leq 1 \\ P(Y=1, X=0|Z=0) + P(Y=0, X=0|Z=1) \leq 1 \\ P(Y=1, X=1|Z=0) + P(Y=0, X=1|Z=1) \leq 1 \end{aligned} \tag{26}$$

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 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 (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 Eq. (26) can be tightened (Balke and Pearl, 1997) to give

$$\begin{aligned} P(y, X = 1|Z = 1) \geq P(y, X = 1|Z = 0) \\ P(y, X = 0|Z = 0) \geq P(y, X = 0|Z = 1) \end{aligned} \tag{27}$$

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 $X \to Y \to X$ 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

$$x = g(z, y, v) \tag{28}$$

can be replaced with

$$x = g'(z, v') \tag{29}$$

such that v' is still independent of z. The nonparametric nature of the structural equations in Eq. (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

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 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

$$X = x \Rightarrow Y_x = Y, \tag{30}$$

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 (30).

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 (30) are automatically satisfied in the structural interpretation and, moreover, that investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \quad \text{for all } y \text{ and } z \tag{31}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x \text{ and } z \tag{32}$$

Eq. (31) 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 (32) generalizes (30) 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 (Fig. 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

$$Y_x \perp\!\!\!\perp X | Z \tag{33}$$

(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

$$\begin{aligned} P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\ &= \sum_z P^*(Y_x = y|x, z)P(z) \quad \text{(using (33))} \\ &= \sum_z P^*(Y = y|x, z)P(z) \quad \text{(using (30))} \\ &= \sum_z P(y|x, z)P(z). \end{aligned} \tag{34}$$

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 Eq. (18).

We see that the assumption of conditional ignorability (33) 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.

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 (32) or (30), 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 Eq. (33), the key to the derivation of Eq. (34), 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 Holland (1988), Angrist et al. (1996), and 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 next section 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

$$Y_{pa_Y} = Y_{pa_Y,s}. \tag{35}$$

2. *Independence restrictions:* If $Z_1, \ldots, Z_k$ is 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

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1\,pa_1}, \ldots, Z_{k\,pa_k}\}. \tag{36}$$

The exclusion restrictions expresses the fact that each parent states include *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:

$$PA_z = \{\emptyset\}, PA_X = \{Z\}, PA_Y = \{X\}. \tag{37}$$

Consequently, the exclusion restrictions translate into:

$$\begin{aligned} X_z &= X_{yz} \\ Z_y &= Z_{xy} = Z_x = Z \\ Y_x &= Y_{xz} \end{aligned} \tag{38}$$

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

$$Z \perp\!\!\!\perp \{Y_x, X_z\}. \tag{39}$$

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 Ya 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 Eq. (34), 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 Eqs. (38)–(39) 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 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 health scientists a powerful methodology for empirical research.

Acknowledgments

Sander Greenland has provided valuable suggestions on the first draft of this paper. Thorough comments by three anonymous referees have led to many improvements in the presentation, and have reinforced the views expressed in Section 2. This research was supported in parts by grants from NSF, ONR (MURI) and AFOSR.

Notes

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

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.
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).
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. 182–189).
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 [2000], pp. 249–257).
9. By "untested" I mean untested using frequency data in nonexperimental studies.
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.
12. Linear relations are used 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.
13. A simple proof of the Causal Markov Theorem is given Pearl (2000, p. 30). This theorem was first stated in Verma and Pearl (1991), but it is implicit in the works of Kiiveri et al. (1984) and others. Corollary 1 was named "Manipulation Theorem" in Spirtes et al. (1993), and is also implicit in Robins' (1987) *G*-computation formula. See Lauritzen (1999).
14. For clarity, we drop the (superfluous) subscript 0 from x_0 and z_{2_0}.
15. The terms "arrow-emitting node" and "collision node" are to be interpreted literally as illustrated by the examples given.
16. Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of *Y*.
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 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.
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).

20. Looser bounds were derived earlier by Robins (1989) and Manski (1990).
21. The inequality is sharp in the sense that every distribution $P(x, y, z)$ satisfying Eq. (25) can be generated by the model defined in Figure 5(a).
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).

References

J. D. Angrist, G. W. Imbens and D. B. Rubin, "Identification of causal effects using instrumental variables (with comments)," *Journal of the American Statistical Association*, 91(434), pp. 444–472, 1996.

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

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

H. Becher, "The concept of residual confounding in regression models and some applications," *Statistics in Medicine*, 11, pp. 1747–1758, 1992.

Y. M. M. Bishop, "Effects of collapsing multidimensional contingency tables," *Biometrics*, 27, pp. 545–562, 1971.

K. A. Bollen. *Structural Equations with Latent Variables*, John Wiley, New York, 1989.

B. Bonet, "Instrumentality tests revisited," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, CA, pp. 48–55, 2001.

R. J. Bowden and D. A. Turkington. *Instrumental Variables*, Cambridge University Press, Cambridge, England, 1984.

N. E. Breslow and N. E. Day. *Statistical Methods in Cancer Research; Vol. 1, The Analysis of Case-Control Studies*, IARC, Lyon, 1980.

N. Cartwright. *Nature's Capacities and Their Measurement*, Clarendon Press, Oxford, 1989.

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

R. G. Cowell, A. P. Dawid, S. L. Lauritzen and D. J. Spielgelhalter. *Probabilistic Networks and Expert Systems*, Springer Verlag, New York, NY, 1999.

D. R. Cox. *The Planning of Experiments*, John Wiley and Sons, NY, 1958.

A. P. Dawid, "Conditional independence in statistical theory," *Journal of the Royal Statistical Society, Series B*, 41(1), pp. 1–31, 1979.

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

E. Eells. *Probabilistic Causality*, Cambridge University Press, Cambridge, MA, 1991.

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

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

A. S. Goldberger, "Structural equation models in the social sciences," *Econometrica: Journal of the Econometric Society*, 40, pp. 979–1001, 1972.

D. A. Grayson, "Confounding confounding," *American Journal of Epidemiology*, 126, pp. 546–553, 1987.

S. Greenland and J. M. Robins, "Identifiability, exchangeability, and epidemiological confounding," *International Journal of Epidemiology*, 15(3), pp. 413–419, 1986.

S. Greenland, J. Pearl and J. M Robins, "Causal diagrams for epidemiologic research," *Epidemiology*, 10(1), pp. 37–48, 1999a.

S. Greenland, J. M. Robins and J. Pearl, "Confounding and collapsibility in causal inference," *Statistical Science*, 14(1), pp. 29–46, 1999b.

S. Greenland, "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, pp. 1166–1169, 1999.

W. W. Hauck, J. M. Heuhaus, J. D. Kalbfleisch and S. Anderson, "A consequence of omitted covariates when estimating odds ratios," *Journal Clinical Epidemiology*, 44(1), pp. 77–81, 1991.

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

P. W. Holland and D. B. Rubin, "Causal inference in retrospective studies," *Evaluation Review*, 13, pp. 203–231, 1988.

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

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

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

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

H. Kiiveri, T. P. Speed and J. B. Carlin, "Recursive causal models," *Journal of Australian Math Society*, 36, pp. 30–52, 1984.

D. G. Kleinbaum, L. L. Kupper, K. E. Muller and A. Nizam. *Applied Regression Analysis and Other Multivariable Methods*, Duxbury Press, Pacific Grove, third edition, 1998.

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

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

S. L. Lauritzen. *Graphical Models*, Clarendon Press, Oxford, 1996.

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

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

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

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

O. S. Miettinen and E. F. Cook, "Confounding essence and detection," *American Journal of Epidemiology*, 114, pp. 593–603, 1981.

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

J. Neyman, "On the application of probability theory to agricultural experiments," Essay on principles. Section 9. *Statistical Science*, 5(4), pp. 465–480, 1923.

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

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

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

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

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

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

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

J. M. Robins, "The analysis of randomized and non-randomized aids treatments trials using a new approach to casual inference in longitudinal studies," in *Health Service Research Methodology: A Focus on AIDS* (L. Sechrest, H. Freeman, and A. Mulley, eds.), U.S. Public Health Service, Washington D.C., pp. 113–159, 1989a.

J. M. Robins and S. Greenland, “The probability of causation under a stochastic model for individual risk,” *Biometrics*, 45, pp. 1125–1138, 1989b.

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

J. M. Robins, “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, pp. 1393–1512, 1986.

J. M. Robins, “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), pp. 139S–161S, 1987.

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

P. Rosenbaum and D. Rubin, “The central role of propensity score in observational studies for causal effects,” *Biometrica*, 70, pp. 41–55, 1983.

D. B. Rubin, “Estimating causal effects of treatments in randomized and nonrandomized studies,” *Journal of Educational Psychology*, 66, pp. 688–701, 1974.

H. A. Simon and N. Rescher, “Cause and counterfactual,” *Philosophy and Science*, 33, pp. 323–340, 1966.

H. A. Simon, “Causal ordering and identifiability,” in *Studies in Econometric Method* (Wm. C. Hood and T. C. Koopmans, eds.), Wiley and Sons Inc., pp. 49–74, 1953.

M. E. Sobel, “Causal inference in statistical models of the process of socioeconomic achievement,” *Sociological Methods & Research*, 27(2), pp. 318–348, 1998.

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

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

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

J. Tian and J. Pearl, “On the identification of causal effects,” in *Proceedings of the American Association of Artificial Intelligence*, AAAI Press/The MIT Press, Menlo Park, CA, 2002.

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

C. R. Weinberg, “Toward a clearer definition of confounding,” *American Journal of Epidemiology*, 137, pp. 1–8, 1993.

N. Wermuth and D. Cox, “Linear dependencies represented by chain graphs,” *Statistical Science*, 8(3), pp. 204–218, 1993.

N. Wermuth, “On block-recursive regression equations (with discussion),” *Brazilian Journal of Probability and Statistics*, 6, pp. 1–56, 1992.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*, John Wiley, Chichester, England, 1990.

A. S. Whittemore, “Collapsibility of multidimensional contingency tables,” *Journal of the Royal Statistical Society, B*, 40(3), pp. 328–340, 1978.

S. Wright, “Correlation and causation,” *Journal of Agricultural Research*, 20, pp. 557–585, 1921.