Causal Inference in Statistics: A Gentle Introduction*

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
judea@cs.ucla.edu

Abstract

This paper provides a conceptual introduction to causal inference, aimed to assist 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 underly all causal inferences, the languages used in formulating those assumptions, and the conditional nature of causal claims inferred from nonexperimental studies. These emphases are illustrated through a brief survey of recent results, including the control of confounding, 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 statistics-based sciences are causal in nature. For example, what is the efficacy of a given drug in a given population? what fraction of crimes could have been avoided by a given educational 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. Yet until very recently, the dominant methodology has been based almost exclusively on statistical analysis which, traditionally, has excluded causal vocabulary both from its mathematical language and from its mainstream educational program. As a result, large segments of the research community find it extremely hard to appreciate and benefit from the many theoretical results that causal analysis has produced the past two decades. These include advances in graphical models [Pearl, 1988; Lauritzen, 1996], counterfactual or "potential outcome" analysis [Rubin and Rosenbaum, 1983; Robins, 1986; Manski, 1995; 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 by illuminating certain conceptual problems that I have found

---

to be major barriers in the transition from statistical to causal analysis.[1] To this end, I will introduce the fundamentals of causal modeling from a perspective that is relatively new to the statistical literature. It is based on *structural equation models* (SEM), which have been used extensively in economics and the social sciences [Goldberger, 1972; Duncan, 1975], even though the causal content of these models has been obscured significantly since their inception (see [Pearl, 2000, Chapter 5] for historical perspective). I will use these modeling fundamentals to develop simple mathematical tools for the estimation of causal effects and the control of confounding, and relate these tools to procedures that are used in the potential outcome approach. Finally, I will offer a symbiosis that exploits the best features of the two approaches—structural models and potential outcome.

2 From Statistical 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 population. 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 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 statistical concepts do not mix. Statistics deals 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 not cure the latter. More generally, there is nothing in a distribution function that would 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

---

[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.

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 statistical and causal concepts unambiguous and easy to apply, can be formulated as follows. A statistical concept is any concept that can be defined in terms of a distribution (be it personal or frequency-based) of observed variables, and a causal concept is any concept concerning changes in variables that cannot be defined from the distribution alone. Examples of statistical concepts are: mean, variance, correlation, regression, dependence, conditional independence, association, 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 claims alone.

2.3   Ramifications of the basic distinction

This principle has consequences that are not generally recognized in the standard 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 a discrepancy between some proposed measure of association 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,

---

[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] The term "causal vs. association" distinction would perhaps be more inviting to statisticians, though the concepts of "mean" and "variance" do not involve association

> 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 3.2].[5] This further implies that confounding bias cannot be 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" tie to 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, [6] major efforts would be required to assess the relevance of this rich literature to the modern conception of confounding as *effect bias* [Greenland et al., 1999].[7]

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

The preceding two requirements: (1) to commence causal analysis with untested,[9] 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

---

[5]Similar arguments apply to the concepts of "randomization" and "instrumental variables" which are commonly thought to have statistical 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 3.3).

[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. 182–9].

[8]Attempts to define causal dependence by adding temporal information and conditioning on the entire past (e.g., [Suppes, 1970]) violate the statistical requirement of limiting the analysis to "observed variables", and encounter other insurmountable difficulties (see Eells [1991], Pearl [2000], pp. 249–257).

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

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 barriers of causal assumptions and causal notation

All statistical studies are based on some untested assumptions. For examples, we often assume that variables are multivariate normal, that the density function has certain smoothness properties, or that a certain parameter falls in a given range. The question thus arises why causal assumptions, say, that symptoms do not cause disease or that treatment does not change gender, invite mistrust and resistance among statisticians.

Ironically, the answer is primarily notational. Statistical 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. Statisticians, especially Bayesians, are prepared to accept an expert's judgment, however untestable, so long as the judgment is presented as a probability expression. Statisticians turn apprehensive when that same judgment is cast in plain causal English (see Pearl [2000, pp. 177-180] for examples.)

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

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,

---

[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 to readers confronting various definitions of concepts such as confounding, randomization or instrumental variables; any definition that is cast in standard probability expressions, void of graphs, counterfactual subscripts or $do(*)$ operators, can safely be discarded as inadequate.

1981]. However, as problem complexity grew, 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 history, causal analysis has remained enigmatic to most rank and file researchers, and its potentials still lay grossly underutilized in statistical research. The reason for this, I am firmly convinced, can be traced to the cumbersome counterfactual notation in which causal analysis has been presented to the research community. The difficulties that have hindered wider acceptance of this presentation include: its reliance on esoteric, hypothetical and seemingly "counter factual" relationships among events [Dawid, 2000], its estrangement from common understanding of cause-effect relationships, the absence of compact mathematical models for representing counterfactual relationships, and the elaborate judgmental effort that is required both for formulating assumptions and for justifying derivational steps in this notation. The next section provides a conceptualization that alleviates these difficulties; it offers both a mathematical approach to causal 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 Fig. 1: The diagram encodes the possible

---

[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.

existence of (direct) causal influence of X on Y, and the absence of causal influence of Y on X, while the equations encode the quantitative relationships among the variables involved, to be determined from the data. The parameter β in the equation is called a "path coefficient" and it quantifies the (direct) causal effect of X on Y; given the numerical value of β, the equation claims that a unit increase in X would result in β units increase of Y. The variables V and U are called "exogenous," (also "disturbances" or "errors") they represent background factors, often unobserved, that influence but are not influenced by the other variables (called "endogenous") in the model. Variable V, for example, represents factors that contribute to the disease X, which may or may not be correlated with U (the factors that influence the symptom Y). If correlation is presumed possible, it is customary to connect the two variables, U and V, by a dashed double arrow, as shown in Fig. 1(b).

Figure 1: A simple structural equation model, and its associated diagrams. Exogenous variables (usually unobserved) are connected by dashed arrows.

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 Fig. 1(a), and

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

for Fig. 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 Fig. 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 statistical, 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 the analysis of instrumental variables) 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 Fig. 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 falsifying the model is entailed by those implications.

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 $do(X = x_0)$.

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]. 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 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 Fig. 2(a) corresponds to a set of three functions, each corresponding to one of observed variables:

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

8

$$y \;\; = \;\; f_Y(x, u)$$

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 (input). The absence of a variable on the right of an equations encodes the assumption that it has no direct effect on the left variable. For example, the absence of variable Z from the arguments of f_Y indicates that variations in Z will leave Y unchanged, as long as variables U, and X remain constant. A system of such functions are said to be *structural* if they are assumed to be autonomous, that is, each function is invariant to possible changes in the form of the other functions [Simon, 1953; Koopmans, 1953].

Representing interventions

This feature of invariance permits us to use structural equations as a basis for modeling causal effects and counterfactuals. This is done through a mathematical operator called $do(x)$ which simulates physical interventions by deleting certain functions from the model, replacing them by a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $do(x_0)$ that holds X constant (at $X = x_0$) in the model of Fig. 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 Fig. 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. 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]. Thus, a central problem in the analysis of causal effects is whether we can estimate the post-intervention distribution from data governed by the pre-intervention distribution. This is the problem of *identification* which has received considerable attention by causal analysts.

A fundamental theorem in causal analysis states that, in general, such identification would be feasible whenever the model is *Markovian*, that is, the graph is acyclic 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:

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

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 Fig. 2(a) can be factorized as

$$P(z, y, x) = P(z)P(x|z)P(y|x) \qquad (10)$$

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

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

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 Fig. 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

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

---

[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].

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

Deriving Causal Effects

The truncated factorization formula enables us to derive causal quantities directly, without dealing with equations or equation modification as in Eq. (6). Consider, for example, the model shown in Fig. 3, in which the error variables are kept implicit.

Figure 3: Markovian model illustrating the derivation of the causal effect of X on Y, Eq. (14), and the back-door criterion. Error terms are not shown explicitly.

Instead of writing down the corresponding five nonparametric equations, we can write the join distribution directly as

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x) \qquad (12)$$

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

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0) \qquad (13)$$

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

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

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

Note also that the truncated factorization formula is not restricted to interventions on a single variable; it is applicable to simultaneous or sequential interventions

such as those invoked in the analysis of time varying treatment with time varying confounders [Robins, 1986]. For example, if X and Z_2 are both treatment variables, and Z_1 and Z_3 are measured covariates, then the post-intervention distribution would be

$$P(z_1, z_3, y | do(x), do(z_2)) = P(z_1)P(z_3|z_1, z_2)P(y|z_2, z_3, x) \qquad (15)$$

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

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

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

Coping with unmeasured confounders

Things are a bit 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

$$P(y|do(x_0)) = \sum_{z2, z3} P(z_2)P(z_3|z_2)P(y|z_2, z_3, x_0) \qquad (17)$$

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

Selecting covariates for adjustment (the back-door criterion)

Consider an observational study where we wish to find the effect of X on Y, for example, treatment on response, and assume that the factors deemed relevant to the problem are structured as in Fig. 3; 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 subjects under the same value of those measurements and average, we get the correct causal effect of treatment on response in the population.

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.

---

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

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. For example, the set $S = \{Z_3\}$ blocks the path $X \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$, because the arrow-emitting node Z_3 is in S. However, the set $S = \{Z_3\}$ does not block the path $X \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow Y$, 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 that the sets $\{Z_1, Z_2, Z_3\}, \{Z_1, Z_3\}$, and $\{Z_2, Z_3\}$, each qualifies 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 $X \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow Y$.

Interestingly, it can be shown that any minimal 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 Fig. 3, 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. (17) directly, by inspection, without manipulating the truncated factorization formula. The criterion can be tested systematically on diagrams of any size and shape, thus freeing analysts from judging whether "X is conditionally ignorable given Z," 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 Z that minimizes measurement cost or sampling variability [Tian et al., 1998].

Intervention calculus

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 nonstandard adjustments. The analysis used in these cases 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 say that the causal quantity is identifiable. The calculus developed for performing such transformations [Pearl, 1995] permits the investigator to inspect the causal diagram and

1. Decide whether the assumptions embodied in the model are sufficient to obtain 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.

Graphical methods of performing some of these tasks, extending emphasizing the identification and control of confounders, are presented in Galles and Pearl [1995]; 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, 2001] 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, 2001*]
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.[15]

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. 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.[16]

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 modification permits the constant x to differ from the actual value of X (namely $f_X(z, v)$) without creating logical contradiction, and thus enables cascaded inference in models where the antecedent of one counterfactual is a consequence of 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 $x_0 \neq x_1$.

This interpretation of counterfactuals, 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.

---

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

[16] 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.

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." In section 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 not derived from a formal model but, rather, it 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 governs 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 Assumption

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

$$X = x \implies Y_x = Y, \tag{18}$$

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 (18).

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

$$Y_{yz} = y \quad \text{for all } y \text{ and } z \tag{19}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x \text{ and } z \tag{20}$$

Eq. (19) 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 (20) generalizes (18) 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 the example of Fig. 2(a), to communicate the understanding that Z is randomized (hence independent of both V and U), the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{X_z, Y_x\}$. 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

$$Y_x \perp\!\!\!\perp X | Z \tag{21}$$

(an assumption that was termed "conditional ignorability" by [Rosenbaum and Rubin, 1983] then the causal effect $P^*(Y_x = y)$ can readily be evaluated, using (18) and (21), to yield

$$
\begin{aligned}
P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\
&= \sum_z P^*(Y_x = y|x, z)P(z) \\
&= \sum_z P^*(Y = y|x, z)P(z) \\
&= \sum_z P(y|x, z)P(z).
\end{aligned}
\tag{22}
$$

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. (14).

We see that the assumption of conditional ignorability (21) qualifies Z as a sufficient covariate for adjustment, and is equivalent therefore to the graphical criterion (called "back door" in Section 3.1) 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 (20), 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. (21), the key to the derivation of Eq. (22), holds in any familiar situation, say in the experimental setup of Fig. 2(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 thought of having to express, defend, and manage formidable counterfactual relationships of this type may explain why the potential-outcome enterprise is currently viewed with such awe and despair among rank-and-file epidemiologists and statisticians—and why economists and social scientists continue to use structural equations 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, Chapter 9]. Pearl [2000, pp. 213–5, 231–4] presents a way that should help researchers combine 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.

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 this knowledge, combining it with data and drawing new causal conclusions about a phenomena. This paper introduces 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, attribution, and explanations [Pearl, 2000]. 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 syn-

thesized, the two components offer scientists a powerful methodology for empirical research.

References

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

[Balke and Pearl, 1995] A. Balke and J. Pearl. Counterfactuals and policy analysis in structural models. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 11–18. Morgan Kaufmann, San Francisco, 1995.

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

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

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

[Dawid, 2000] A.P. Dawid. Causal inference without counterfactuals (with comments and rejoinder). *Journal of the American Statistical Association*, 95(450):407–448, June 2000.

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

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

[Galles and Pearl, 1995] D. Galles and J. Pearl. Testing identifiability of causal effects. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 185–195. Morgan Kaufmann, San Francisco, 1995.

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

[Grayson, 1987] D.A. Grayson. Confounding confounding. *American Journal of Epidemiology*, 126:546–553, 1987.

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

[Greenland et al., 1999] S. Greenland, J.M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46, February 1999.

[Heckman and Smith, 1998] J.J. Heckman and J. Smith. Evaluating the welfare state. In S. Strom, editor, *Econometric and Economic Theory in the 20th Century*, pages 1–60. Cambridge University Press, Cambridge, England, 1998.

[Heckman, 2001] J.J. Heckman. Econometrics and empirical economics. *Journal of Econometrics*, 100(1):c–5, 2001.

[Holland, 1988] P.W. Holland. Causal inference, path analysis, and recursive structural equations models. In C. Clogg, editor, *Sociological Methodology*, pages 449–484. American Sociological Association, Washington, D.C., 1988.

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

[Kiiveri et al., 1984] H. Kiiveri, T.P. Speed, and J.B. Carlin. Recursive causal models. *Journal of Australian Math Society*, 36:30–52, 1984.

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

[Koopmans, 1953] T.C. Koopmans. Identification problems in econometric model construction. In W.C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 27–48. Wiley, New York, 1953.

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

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

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

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

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

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

[Pearl and Robins, 1995] J. Pearl and J.M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

[Pearl and Verma, 1991] J. Pearl and T. Verma. A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pages 441–452. Morgan Kaufmann, San Mateo, CA, 1991.

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

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

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

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

[Robins, 1986] 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*, 7:1393–1512, 1986.

[Robins, 1987] 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):139S–161S, 1987.

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

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

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

[Simon and Rescher, 1966] H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33:323–340, 1966.

[Simon, 1953] H.A. Simon. Causal ordering and identifiability. In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. Wiley and Sons, Inc., 1953.

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

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

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

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

[Tian and Pearl, 2001] J. Tian and J. Pearl. On the identification of causal effects. Technical Report R-290, University of California, Los Angeles, CA, 2001.

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

[Wright, 1921] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.