Forthcoming, *Sociological Methodology*.

TECHNICAL REPORT
R-355
August 2010

THE FOUNDATIONS OF CAUSAL INFERENCE

Judea Pearl*
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

August 24, 2010

Abstract

This paper reviews recent advances in the foundations of causal inference and introduces a systematic methodology for defining, estimating, and testing causal claims in experimental and observational studies. It is based on nonparametric structural equation models (SEM)—a natural generalization of those used by econometricians and social scientists in the 1950s and 1960s, which provides a coherent mathematical foundation for the analysis of causes and counterfactuals. In particular, the paper surveys the development of mathematical tools for inferring the effects of potential interventions (also called "causal effects" or "policy evaluation"), as well as direct and indirect effects (also known as "mediation"), in both linear and nonlinear systems. Finally, the paper clarifies the role of propensity score matching in

*Portions of this paper are adapted from Pearl (2000a, 2009a,b, 2010a); I am indebted to Peter Bentler, Felix Elwert, David MacKinnon, Stephen Morgan, Patrick Shrout, Christopher Winship, and many bloggers on the UCLA Causality Blog (http://www.mii.ucla.edu/causality/) for reading and commenting on various segments of this manuscript, and to two anonymous referees for their thorough editorial input. This research was supported in parts by NIH grant #1R01 LM009961-01, NSF grant #IIS-0914211, and ONR grant #N000-14-09-1-0665.

causal analysis, defines the relationships between the structural and potential-outcome frameworks, and develops symbiotic tools that use the strong features of both.

1 Introduction

The questions that motivate most studies in the social and behavioral sciences are causal, not statistical. For example, what is the efficacy of a given social program in a given community? Can data prove an employer guilty of hiring discrimination? What fraction of past crimes could have been prevented by a given policy? Why did one group of students succeed where others failed? What can a typical public school student gain by switching to a private school? These are *causal* questions because they require some knowledge of the data-generating process; they cannot be computed from the data alone, regardless of sample size.

Remarkably, although much of the conceptual and algorithmic tools needed for tackling such problems are now well established, and although these tools invoke structural equations—a modeling tool developed by social scientists—they are hardly known among rank and file researchers. The barrier has been cultural; formulating causal problems mathematically requires certain extensions to the standard mathematical language of statistics, and these extensions are not generally emphasized in the mainstream literature and education. As a result, the common perception among quantitative social scientists is that causality is somehow "controversial" or "ill understood" or requiring esoteric assumptions, or demanding extreme caution and immense erudition in the history of scientific thought. Not so.

The paper introduces basic principles and simple mathematical tools that are sufficient for solving most (if not all) problems involving causal and counterfactual relationships. The principles are based on the nonparametric structural equation model (SEM)—a natural generalization of the models used by econometricians and social scientists in the 1950s and 1960s, yet cast in new mathematical underpinnings, liberated from the parametric blindfolds that have conflated regression with causation and thus obscured the causal content of traditional SEMs. This semantical framework, enriched with a few ideas from logic and graph theory, gives rise to a general, formal, yet friendly calculus of causes and counterfactuals that resolves many long-standing problems in sociological methodology.

To this end, Section 2 (based on Pearl 2009a, pp. 38–40)begins by illuminating two conceptual barriers that impede the transition from statistical to causal analysis: (1) coping with untested assumptions and (2) acquiring new mathematical notation; it is then followed by a brief historical account of how these barriers have impeded progress in social science methodology. Crossing these barriers, Section 3.1 (based on Pearl 2009a, ch. 1) then introduces the fundamentals of the structural theory of causation, with emphasis on the formal representation of causal assumptions, and formal definitions of causal effects, counterfactuals, and joint probabilities of counterfactuals. Section 3.2 (based on Pearl 2009a, ch. 3) uses these modeling fundamentals to represent interventions and develops mathematical tools for estimating causal effects (Section 3.3) and counterfactual quantities (Section 3.4). Sections 3.3.2 and 3.5 introduce new results, concerning the choice of measurements (3.3.2) and the limits of analytical tools in coping with heterogeneity (3.5).

The tools described in Section 3 permit investigators to communicate causal assumptions formally using diagrams, then to 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.

4. identify the testable implications (if any) of the model's assumptions, and devise ways of testing the assumptions behind each causal claim.

5. decide, prior to taking any data, what measurements ought to be taken, whether one set of measurements is as good as another, and which measurements tend to bias our estimates of the target quantities.

Section 4 outlines a general methodology to guide problems of causal inference. It is structured along five major steps: (1) define, (2) assume, (3) identify, (4) test, and (5) estimate. Each step benefits from the tools developed in Section 3. This five-step methodology is an expansion of the

one presented in Pearl (2010a) and clarifies the role of local testing (4.3.1), propensity score matching (4.3.2), and approximation methods (4.3.3).

Section 5 relates these tools to those used in the potential-outcome framework, and offers a formal mapping between the two frameworks and a symbiosis (based on Pearl, 2009a, pp. 231–34) that exploits the best features of both and demystifies enigmatic terms such as "potential outcomes," "ignorability," "treatment assignment," and more. Finally, the benefit of this symbiosis is demonstrated in Section 6, in which the structure-based logic of counterfactuals is harnessed to estimate causal quantities that cannot be defined within the paradigm of controlled randomized experiments. These include direct and indirect effects, or "mediation," a topic with long tradition in social science research, which only recently has been given a satisfactory formulation in nonlinear systems (Pearl, 2001, 2010b).

2 From Association to Causation

2.1 The Basic Distinction and Its Implications

The aim of standard statistical analysis, typified by regression, estimation, and hypothesis testing techniques, is to assess parameters of a distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables, estimate probabilities of past and future events, as well as update those probabilities 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 not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under *changing conditions*—for example, changes induced by treatments, new policies, or other external interventions.

This distinction implies that causal and associational concepts do not mix. 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. This information must be provided by causal assumptions that

identify relationships that remain invariant when external conditions change.

A useful demarcation line that makes the distinction between associational and causal concepts crisp 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 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, propensity score, risk ratio, odds ratio, marginalization, Granger causality, conditionalization, "controlling for," and so on. Examples of causal concepts are randomization, influence, effect, confounding, "holding constant," disturbance, error terms, structural coefficients, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, and attribution. The former can, while the latter cannot, be defined in term of distribution functions.

This demarcation line is extremely useful in tracing the assumptions that are needed for substantiating various types of scientific claims. Every claim invoking causal concepts must rely on some premises that invoke such concepts; it cannot be inferred from or even defined in terms statistical associations alone.

This principle, though it goes back to the late ninteenth century, has far reaching consequences that are not generally recognized in the standard literature. Wright (1923), for example, specifically declared that "prior knowledge of the causal relations is assumed as prerequisite" before one can draw causal conclusions from path diagrams. The same understanding overrides the classical works of Blalock (1964) and Duncan (1975). And yet, even today, it is not uncommon to find "sociologists [who] routinely employ regression analysis and a variety of related statistical models to draw causal inferences from survey data" (Sobel, 1996, p. 353). More subtly, it is not uncommon to find seasoned sociologists wondering why an instrumental variable is a causal concept while a propensity score would not be.[1] Such confusions may tempt one to define the former in terms of the latter, or to ignore the untestable causal assumptions that are necessary for the former.

This association/causation demarcation line further implies that causal relations cannot be expressed in the language of probability and, hence, that

[1]The answer of course is that the defining conditions for an instrumental variable invoke unobserved variables (see Pearl, 2009a, p. 247–48) while the propensity score is defined in terms of the conditional probability of observed variables (see equation 31). I am grateful to one reviewer for demonstrating this prevailing confusion.

any mathematical approach to causal analysis must acquire new notation for expressing causal relations—probability calculus is insufficient. 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.

2.2 Untested Assumptions and New Notation

The preceding two requirements—to commence causal analysis with untested,[2] theoretically or judgmentally based assumptions, and to extend the syntax of probability calculus—constitute the two main obstacles to the acceptance of causal analysis among professionals with traditional training in statistics.

Associational assumptions, even untested, are testable in principle, given sufficiently large samples and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference stands out 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 prior causal assumptions, say that treatment does not change gender, remains substantial regardless of sample size.

This makes it doubly important that the notation we use for expressing causal assumptions be cognitively meaningful and unambiguous so that we can clearly judge the plausibility or inevitability of the assumptions articulated. Analysts can no longer ignore the mental representation in which scientists store experiential knowledge, since it is this representation, and the language used to access it that determine the reliability of the judgments upon which the analysis so crucially depends.

How do we recognize causal expressions in the social science literature? Those versed in the potential-outcome notation (Neyman, 1923; Rubin, 1974; Holland, 1988; Sobel, 1996) can recognize such expressions through the subscripts that are attached to counterfactual events and variables—for exam-

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

ple, $Y_x(u)$ or Z_{xy}. (Some authors use parenthetical expressions such as $Y(0)$, $Y(1)$, $Y(x,u)$, 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 $P(Y_x = y)$ (see Section 5 for a formal definition). Alternatively, Pearl (1995) 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.[3] Still a third notation that distinguishes causal expressions is provided by graphical models, where the arrows convey causal directionality, or structural equations, in which the equality signs (=) represent right-to-left assignment operators (:=) (Pearl, 2009a, p. 138).[4]

2.3 SEM and Causality: A Brief History[5]

Quantitative sociological researchers have chosen structural equation models and their associated causal diagrams as the primary language for causal analysis. Influenced by the pioneering work of Sewall Wright (1923) and early econometricians (Haavelmo, 1943; Simon, 1953; Marschak, 1950; Koopmans, 1953), Blalock (1964) and Duncan (1975) considered SEM a mathematical tool for drawing causal conclusions from a combination of observational data and theoretical assumptions. They were explicit about the importance of the latter, and about the unambiguous causal reading of the model parameters, once the assumptions are substantiated.

In time, however, the proper causal reading of structural equation models and the theoretical basis on which it rests became suspect of ad hockery, even

[3]Clearly, $P(Y = y|do(X = x))$ is equivalent to $P(Y_x = y)$. This 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.

[4]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.

[5]A more comprehensive account of the history of SEM and its causal interpretations is given in Pearl (1998). Pearl (2009a, pp. 368–74) further devotes a whole section of his book *Causality* to advise SEM students on the causal reading of SEM and how do defend it against the skeptics.

to seasoned workers in the field. This occurred partially due to the revolution in computer power, which made sociological workers "lose control of their ability to see the relationship between theory and evidence" (Sørensen, 1998, p. 241). But it was also due to a steady erosion of the basic understanding of what SEMs stand for.

In his critical paper Freedman (1987, p. 114) challenged the causal interpretation of SEM as "self-contradictory," and none of the 11 discussants of his paper were able to articulate the correct, noncontradictory interpretation of the example presented by Freedman. Instead, SEM researchers appeared willing to live with the contradiction. In his highly cited commentary on SEM, Chin (1998) writes: "researchers interested in suggesting causality in their SEM models should consult the critical writing of Cliff (1983), Freedman (1987), and Baumrind (1993)."

This, together with the steady influx of statisticians into the field, has left SEM researchers in a quandary about the meaning of the SEM parameters, and has caused some to avoid causal vocabulary altogether and to regard SEM as an encoding of parametric family of density functions, void of causal interpretation. Muthén (1987), for example, wrote "It would be very healthy if more researchers abandoned thinking of and using terms such as cause and effect" (Muthén, 1987). Many SEM textbooks have subsequently considered the word "causal modeling" to be an outdated misnomer (e.g., Kelloway, 1998, p. 8), giving clear preference to causality-free nomenclature such as "covariance structure," "regression analysis," or "simultaneous equations."

The confusion between regression and structural equations has further eroded confidence in the latter adequacy to serve as a language for causation. Sobel (1996), for example, states that the interpretation of the parameters of the model as effects "do not generally hold, even if the model is correctly specified and a causal theory is given," and "the way sociologists use structural equation models to draw causal inferences is problematic in both experimental and nonexperimental work." Comparing structural equation models to the potential-outcome framework, Sobel (2008) further states that "In general (even in randomized studies), the structural and causal parameters are not equal, implying that the structural parameters should not be interpreted as effect." In Section 3 of this paper we show the opposite: structural and causal parameters are one and the same thing, and they should *always* be interpreted as effects.

Another advocate of the potential-outcome framework is Holland (1995, p. 54), who explains the source of the confusion: "I am speaking, of course,

about the equation: $\{y = a+bx+\epsilon\}$. What does it mean? The only meaning I have ever determined for such an equation is that it is a shorthand way of describing the conditional distribution of $\{y\}$ given $\{x\}$." We will see that the structural interpretation of the equation above has in fact nothing to do with the conditional distribution of $\{y\}$ given $\{x\}$; rather, it conveys counterfactual information that is orthogonal to the statistical properties of $\{x\}$ and $\{y\}$ (see footnote 18).

We will further see (Section 4) that the SEM language in its nonparametric form offers a mathematically equivalent and conceptually superior alternative to the potential-outcome framework that Holland and Sobel advocate for causal inference. It provides in fact the formal mathematical basis for the potential-outcome framework and a friendly mathematical machinery for a general cause-effect analysis.

3 Structural Models, Diagrams, Causal Effects, and Counterfactuals

This section provides a gentle introduction to the structural framework and uses it to present the main advances in causal inference that have emerged in the past two decades. We start with recursive linear models,[6] in the style of Wright (1923), Blalock (1964), and Duncan (1975) and, after explicating carefully the meaning of each symbol and the causal assumptions embedded in each equation, we advance to nonlinear and nonparametric models with latent variables, and we show how these models facilitate a general analysis of causal effects and counterfactuals.

3.1 Introduction to Structural Equation Models

How can we express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such relationship mathematically was made in the 1920s 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

[6] By "recursive" we mean systems free of feedback loops. We allow however correlated errors, or latent variables that create such correlations. Most of our results, with the exception of Sections 3.2.3 and 3.3 are valid for nonrecursive systems, allowing reciprocal causation.

stands for a certain symptom of the disease, Wright would write a linear equation:

$$y = \beta x + u_Y, \tag{1}$$

where x stands for the level (or severity) of the disease, y stands for the level (or severity) of the symptom, and u_Y stands for all factors, other than the disease in question, that could possibly affect Y when X is held constant.[7] In interpreting this equation we should think of a physical process whereby nature *examines* the values of x and u and, accordingly, *assigns* to variable Y the value $y = \beta x + u_Y$. Similarly, to "explain" the occurrence of disease X, we could write $x = u_X$, where U_X stands for all factors affecting X.

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

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

it might be misinterpreted to mean that the symptom influences the disease. To express the directionality of the underlying process, Wright augmented the equation with a diagram, later called "path diagram," in which arrows are drawn from (perceived) causes to their (perceived) effects, and more importantly, the absence of an arrow makes the empirical claim that Nature assigns values to one variable irrespective of another. In Figure 1, for example, the absence of arrow from Y to X represents the claim that symptom Y is not among the factors U_X that affect disease X. Thus, in our example, the complete model of a symptom and a disease would be written as in Figure 1: The diagram encodes the possible existence of (direct) causal influence of X on Y, and the absence of causal influence of Y on X, while the equations encode the quantitative relationships among the variables involved, to be determined from the data. The parameter β in the equation is called a "path coefficient," and it quantifies the (direct) causal effect of X on Y. Once we commit to a particular numerical value of β, the equation claims that a unit increase for X would result in β units increase of Y regardless of the values taken by other variables in the model, and regardless of whether the increase in X originates from external or internal influences.

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

The variables U_X and U_Y 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 U_X, for example, represents factors that contribute to the disease X, which may or may not be correlated with U_Y (the factors that influence the symptom Y). Thus, background factors in structural equations differ fundamentally from residual terms in regression equations. The latters, usually denoted by letters ϵ_X, ϵ_Y, are artifacts of analysis which, by definition, are uncorrelated with the regressors. The formers are part of physical reality (e.g., genetic factors, socioeconomic conditions), which are responsible for variations observed in the data; they are treated as any other variable, though we often cannot measure their values precisely and must resign ourselves to merely acknowledging their existence and assessing qualitatively how they relate to other variables in the system.

If correlation is presumed possible, it is customary to connect the two variables, U_Y and U_X, by a dashed double arrow, as shown in Figure 1(b). By allowing correlations among omitted factors, we allow in effect for the presence of latent variables affecting both X and Y, as shown explicitly in Figure 1(c), which is the standard representation in the SEM literature (e.g., Bollen, 1989). In contrast to traditional latent variable models, however, our attention will not be focused on the connections among such latent variables but, rather, on the causal effects that those variables induce among the observed variables. In particular, we will not be interested in the causal effect of one latent variable on another and, therefore, there will be no reason to distinguish between correlated errors and causally related latent variables; it is only the distinction between correlated and uncorrelated errors (e.g., between Figure 1(a) and (b)) that need to be made. Moreover, when the error terms are uncorrelated, it is often more convenient to eliminate them altogether from the diagram (as in Figure 3, Section 3.2.3), with the understanding that every variable, X, is subject to the influence of an independent disturbance U_X.

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, the arrow in $X \rightarrow Y$ designates X as a parent of Y and Y as a child of X. A "path" is any consecutive sequence of edges, solid

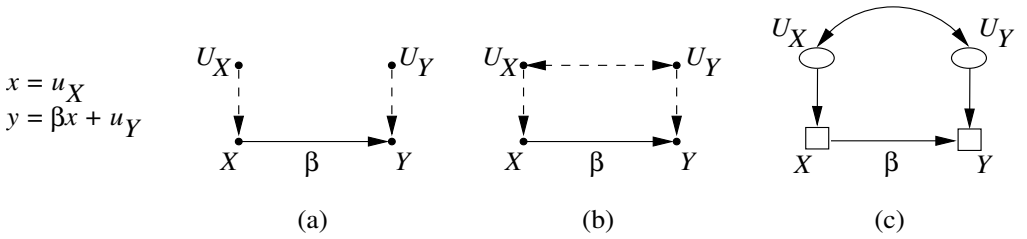

Figure 1: A simple structural equation model, and its associated diagrams, showing (a) independent unobserved exogenous variables (connected by dashed arrows), (b) dependent exogenous variables, and (c) an equivalent, more traditional notation, in which latent variables are enclosed in ovals.

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, U_X, U_Y, 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, we can immediately write the relations

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

for Figure 1(a), and

$$Cov(X, Y) = \beta + Cov(U_Y, U_X) \tag{4}$$

for Figure 1(b)–(c). (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 $Cov(U_Y, U_X) = 0$),such relationships may allow us to solve for the path coefficients in terms of observed covariance terms only, and this amounts to inferring the magnitude of (direct) causal effects from observed, nonexperimental associations, assuming of course that we are 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 represents a claim of zero

influence, while a missing double arrow represents a claim of zero covariance. In Figure 1(a), for example, the assumptions that permit us to identify the direct effect β are encoded by the missing double arrow between U_X and U_Y, indicating $Cov(U_Y, U_X)$=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 $Cov(U_Y, U_X) = 0$, 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_Y and U_X, 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* (Pearl, 1988).

Definition 1 (*d*-separation) *A set S of nodes is said to block a path p if either (1) p contains at least one arrow-emitting node that is in S, or (2) p contains at least one collision node that is outside S and has no descendant in S. If S blocks* all *paths from X to Y, it is said to "d-separate X and Y," and then, X and Y are independent given S, written $X \perp\!\!\!\perp Y|S$.*

To illustrate, the path $U_Z \rightarrow Z \rightarrow X \rightarrow Y$ is blocked by $S = \{Z\}$ and by $S = \{X\}$, since each emits an arrow along that path. Consequently we can infer that the conditional independencies $U_Z \perp\!\!\!\perp Y|Z$ and $U_Z \perp\!\!\!\perp Y|X$ will be satisfied in any probability function that this model can generate, regardless of how we parametrize the arrows. Likewise, the path $U_Z \rightarrow Z \rightarrow X \leftarrow U_X$ is blocked by the null set $\{\emptyset\}$, but it is not blocked by $S = \{Y\}$ since Y is a descendant of the collision node X. Consequently, the marginal independence $U_Z \perp\!\!\!\perp U_X$ will hold in the distribution, but $U_Z \perp\!\!\!\perp U_X|Y$ may or may not hold. This special handling of collision nodes (or *colliders*, e.g., $Z \rightarrow X \leftarrow U_X$) reflects a general phenomenon known as *Berkson's paradox* (Berkson, 1946), whereby observations on a common consequence

of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail.

The conditional independencies entailed by d-separation constitute the main opening through which the assumptions embodied in structural equation models can confront the scrutiny of nonexperimental data. In other words, almost all statistical tests capable of invalidating the model are entailed by those implications.[8] In addition, d-separation further defines con-

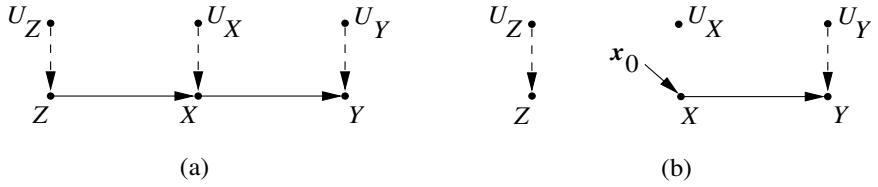

Figure 2: The diagrams associated with (a) the structural model of equation (5) and (b) the modified model of equation (6), representing the intervention $do(X = x_0)$.

ditions for model equivalence (Verma and Pearl, 1990; Ali et al., 2009) that are mathematically proven and should therefore supercede the heuristic (and occasionally false) rules prevailing in social science research (Lee and Hershberger, 1990).

3.2 From Linear to Nonparametric Models and Graphs

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, with only a few attempts (e.g., Muthén, 1984; Winship and Mare, 1983; Bollen, 1989, ch. 9) to extend its capabilities to models involving discrete variables, nonlinear dependencies, and heterogeneous effect modifications.[9] A central requirement for any such extension

[8] Additional implications called "dormant independence" (Shpitser and Pearl, 2008) may be deduced from some semi-Markovian models, i.e., graphs with correlated errors (Verma and Pearl, 1990).

[9] These attempts were limited to ML estimation of regression coefficients in specific nonlinear functions but failed to relate those coefficients to causal effects among the observed variables (see Section 6.5).

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, was proposed in Haavelmo (1943); Strotz and Wold (1960); Spirtes et al. (1993); Pearl (1993a, 2000a); and Lindley (2002), and it 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 nonparametric interpretation of the diagram in Figure 2(a) corresponds to a set of three functions, each corresponding to one of the observed variables:

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_Y), \end{aligned} \tag{5}$$

where in this particular example U_Z, U_X and U_Y 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 the values on the right variables (inputs). The absence of a variable from the right-hand side of an equation encodes the assumption that nature ignores that variable in the process of determining the value of the dependent variable. For example, the absence of variable Z from the arguments of f_Y conveys the empirical claim that variations in Z will leave Y unchanged, as long as variables U_Y 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).

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 with a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $do(x_0)$ that holds X constant (at $X = x_0$) in model

M of Figure 2(a), we replace the equation for x in equation (5) with $x = x_0$, and obtain a new model, M_{x_0},

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

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 postintervention distribution of variables Y and Z (also called "controlled" or "experimental" distribution), to be distinguished from the preintervention distribution, $P(x, y, z)$, associated with the original model of equation (5). For example, if X represents a treatment variable, Y a response variable, and Z some covariate that affects the amount 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 situation in which treatment $X = x_0$ is administered uniformly to the population.

In general, we can formally define the postintervention distribution by the equation

$$P_M(y|do(x)) \triangleq P_{M_x}(y) \tag{7}$$

In words: In the framework of model M, the postintervention distribution of outcome Y is defined as the probability that model M_x assigns to each outcome level $Y = y$.

From this distribution, we are 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

$$E(Y|do(x_0')) - E(Y|do(x_0)), \tag{8}$$

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

$$E(Y|do(x_0'))/E(Y|do(x_0)). \tag{9}$$

The variance $Var(Y|do(x_0))$, or any other distributional parameter, may also enter the 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 (2000a, 1995) (see footnote 3). The central question in the analysis of causal effects is the question of *identification*: Can the controlled (postintervention) distribution, $P(Y = y|do(x))$, be estimated from data governed by the preintervention distribution, $P(z, x, y)$?

The problem of *identification* has received considerable attention in econometrics (Hurwicz, 1950; Marschak, 1950; Koopmans, 1953) and social science (Duncan, 1975; Bollen, 1989), usually in linear parametric settings, where it reduces to asking whether some model parameter, β, has a unique solution in terms of the parameters of P (the distribution of the observed variables). In the nonparametric formulation, identification is more involved, since the notion of "has a unique solution" does not directly apply to causal quantities such as $Q(M) = P(y|do(x))$, which have no distinct parametric signature and are defined procedurally by simulating an intervention in a causal model M, as in eauation (6). The following definition provides the needed extension:

Definition 2 (identifiability (Pearl, 2000a, p. 77)) *A quantity $Q(M)$ is identifiable, given a set of assumptions A, if for any two models M_1 and M_2 that satisfy A, we have*

$$P(M_1) = P(M_2) \Rightarrow Q(M_1) = Q(M_2) \tag{10}$$

In words, the details of M_1 and M_2 do not matter; what matters is that the assumptions in A (e.g., those encoded in the diagram) would constrain the variability of those details in such a way that equality of P's would entail equality of Q's. When this happens, Q depends on P only and should therefore be expressible in terms of the parameters of P. The next subsections exemplify and operationalize this notion.

3.2.2 Estimating the Effect of Interventions

To understand how hypothetical quantities such as $P(y|do(x))$ or $E(Y|do(x_0))$ can be estimated from actual data and a partially specified model, let us begin with a simple demonstration on the model of Figure 2(a). We will see that, despite our ignorance of f_X, f_Y, f_Z and $P(u)$, $E(Y|do(x_0))$ is nevertheless identifiable and is given by the conditional expectation $E(Y|X = x_0)$. We do this by deriving and comparing the expressions for these two quantities, as defined by equations (5) and (6), respectively. The mutilated model in equation (6) dictates

$$E(Y|do(x_0)) = E(f_Y(x_0, u_Y)), \tag{11}$$

whereas the preintervention model of equation (5) gives

$$\begin{aligned} E(Y|X = x_0)) &= E(f_Y(x, u_Y)|X = x_0) \\ &= E(f_Y(x_0, u_Y)|X = x_0) \\ &= E(f_Y(x_0, u_Y)) \end{aligned} \tag{12}$$

which is identical to (11). Therefore,

$$E(Y|do(x_0)) = E(Y|X = x_0)) \tag{13}$$

Using a similar though somewhat more involved derivation, we can show that $P(y|do(x))$ is identifiable and given by the conditional probability $P(y|x)$.

We see that the derivation of (13) was enabled by two assumptions: (1) that Y is a function of X and U_Y only, and (2) that U_Y is independent of $\{U_Z, U_X\}$, hence of X. The latter assumption parallels the celebrated "orthogonality" condition in linear models, $Cov(X, U_Y) = 0$, which has been used routinely, often thoughtlessly, to justify the estimation of structural coefficients by regression techniques.

Naturally, if we were to apply this derivation to the linear models of Figure 1(a) or (b), we would get the expected dependence between Y and the intervention $do(x_0)$:

$$\begin{aligned} E(Y|do(x_0)) &= E(f_Y(x_0, u_Y)) \\ &= E(\beta x_0 + u_Y) \\ &= \beta x_0 \end{aligned} \tag{14}$$

This equality endows β with its causal meaning as "effect coefficient." It is extremely important to keep in mind that in structural (as opposed to regressional) models, β is not "interpreted" as an effect coefficient but is "proven" to be one by the derivation above. β will retain this causal interpretation regardless of how X is actually selected (through the function f_X in Figure 2(a)) and regardless of whether U_X and U_Y are correlated (as in Figure 1(b)) or uncorrelated (as in Figure 1(a)). Correlations may only impede our ability to estimate β from nonexperimental data, but it will not change its definition as given in (14). Accordingly, and contrary to endless confusions in the literature (see footnote 18), structural equations say absolutely nothing about the conditional expectation $E(Y|X = x)$. Such connection may exist under special circumstances—for example, if $cov(X, U_Y) = 0$, as in equation

(13)—but, it is otherwise irrelevant to the definition or interpretation of β as effect coefficient, or to the empirical claims of equation (1).

Section 3.2.3 will circumvent these derivations altogether by reducing the identification problem to a graphical procedure. Indeed, since graphs encode all the information that nonparametric structural equations represent, they should permit us to solve the identification problem without resorting to algebraic analysis.

3.2.3 Causal Effects from Data and Graphs

Causal analysis in graphical models begins with the realization that all causal effects are identifiable 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 too can be determined from the graph structure (Section 3.3). The key to these results rests with the following basic theorem.

Theorem 1 (the causal Markov condition) *Any distribution generated by a Markovian model M can be factorized as:*

$$P(v_1, v_2, \ldots, v_n) = \prod_i P(v_i|pa_i) \tag{15}$$

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

$$P(z, y, x) = P(z)P(x|z)P(y|x), \tag{16}$$

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*

$$P(v_1, v_2, \ldots, v_k|do(x_0)) = \prod_{i|V_i \notin X} P(v_i|pa_i)\,|_{x=x_0} \tag{17}$$

where $P(v_i|pa_i)$ are the preintervention conditional probabilities.[10]

Corollary 1 instructs us to remove from the product of equation (15) those factors that quantify how the intervened variables (members of set X) are influenced by their preintervention parents. This removal follows from the fact that the postintervention 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 preintervention distribution of equation (16). As expected, the distribution of Z is not affected by the intervention, since[11]

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

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

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

This example demonstrates how the (causal) assumptions embedded in the model M permit us to predict the postintervention distribution from the preintervention 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 (specifically, the identity of X's parents) that permits the derivation to go through.

The truncated factorization formula enables us to derive causal quantities directly, without dealing with equations or equation modification as in equations (11)–(13). Consider, for example, the model shown in Figure 3, in

[10] A simple proof of the causal Markov theorem is given in Pearl (2000a, p. 30). This theorem was first presented in Pearl and Verma (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 it is also implicit in the G-computation formula of Robins (1987); see also Lauritzen (2001).

[11] Throughout this paper, summation signs should be understood to represent integrals whenever the summed variables are continuous.

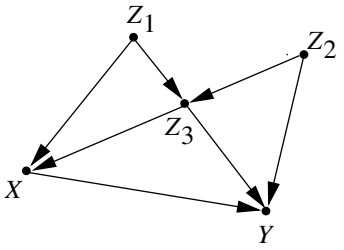

Figure 3: A Markovian model illustrating the derivation of the causal effect of X on Y, as shown in equation (20). Error terms are not shown explicitly.

which the error variables are kept implicit. Instead of writing down the corresponding five nonparametric equations, we can write the joint distribution directly as

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x), \quad (18)$$

where each marginal or conditional probability on the right-hand side is directly estimable from the data. Now suppose we intervene and set variable X to x_0. The postintervention distribution can readily be written (using the truncated factorization formula (17)) as

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0), \quad (19)$$

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

$$P(y|do(x_0)) = \sum_{z_1,z_2,z_3} P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0) \quad (20)$$

Note that this formula corresponds precisely with 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 (Pearl and Robins, 1995; Arjas and Parner, 2004). For example, if X and Z_2 are both treatment variables, and Z_1 and Z_3 are measured covariates, then the postintervention distribution would be

$$P(z_1, z_3, y|do(x), do(z_2)) = P(z_1)P(z_3|z_1, z_2)P(y|z_2, z_3, x), \tag{21}$$

and the causal effect of the treatment sequence $do(X = x), do(Z_2 = z_2)$[12] would be

$$P(y|do(x), do(z_2)) = \sum_{z_1,z_3} P(z_1)P(z_3|z_1, z_2)P(y|z_2, z_3, x) \tag{22}$$

This expression coincides with the G-computation formula in Robins (1987), 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).

3.3 Coping with Latent Confounders

Things are more complicated when we face latent confounders—that is, unmeasured factors that affect two or more observed variables. For example, it is not immediately clear whether the formula in equation (20) can be estimated if any of Z_1, Z_2, and Z_3 is not measured. A few but challenging algebraic steps would reveal that we can perform the summation over Z_2 to obtain

$$P(y|do(x_0)) = \sum_{z_1,z_3} P(z_1)P(z_3|z_1)P(y|z_1, z_3, x_0), \tag{23}$$

which means that we need only adjust for Z_1 and Z_3 without ever measuring Z_2. In general, it can be shown (Pearl, 2000a, p. 73) that whenever the graph is Markovian the postinterventional distribution $P(Y = y|do(X = x))$ is given by the expression

$$P(Y = y|do(X = x)) = \sum_{t} P(y|t, x)P(t), \tag{24}$$

where T is the set of direct causes of X (also called "parents") in the graph.[13] This allows us to write (23) directly from the graph, thus skipping the algebra

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

[13] The operation described in equation (24) is known as "adjusting for T" or "controlling for T." In linear analysis, the problem amounts to finding an appropriate set of regressors.

that led to (23). It further implies that, no matter how complicated the model, the parents of X are the only variables that need to be measured to estimate the causal effects of X.

It is not immediately clear however whether other sets of variables beside X's parents suffice for estimating the effect of X, whether some algebraic manipulation can further reduce equation (23), 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 from a graphical criterion to be discussed next.

3.3.1 Covariate Selection—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

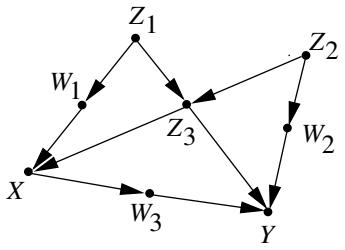

Figure 4: A Markovian model illustrating the back-door criterion. Error terms are not shown explicitly.

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 versus 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 "admissible set" for adjustment. The problem of defining an admissible set, let alone finding one, has baffled epidemiologists and social scientists for decades (see Greenland et al. (1999) and Pearl (1998) for a review).

The following criterion, named "back-door" in Pearl (1993a), settles this problem by providing a graphical method of selecting admissible sets of factors for adjustment.

Definition 3 (admissible sets—the back-door criterion) *A set S is admissible (or "sufficient") 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, "blocking" is interpreted as in Definition 1. 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, are in S, and the collision node Z_3 is not outside S.

Based on this criterion we see, for example, that the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$ are each 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 causatinamely, it correctly represents the target quantity: the causal effect of X on Y. The reason for excluding descendants of X (e.g., W_3 or any of its descendants) is given in (Pearl, 2009a, p. 338–41).

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

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

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

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

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

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

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

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

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

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

and

$$P(Y = y|do(X = x)) = \sum_s P(Y = y|X = x, S = s)P(S = s) \tag{25}$$

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

An equivalent expression for the causal effect (25) can be obtained by multiplying and dividing by the conditional probability $P(X = x|S = s)$, giving

$$P(Y = y|do(X = x)) = \sum_s \frac{P(Y = y, X = x, S = s)}{P(X = x|S = s)} \tag{26}$$

from which the name "Inverse Probability Weighting" has evolved (Pearl, 2000a, pp. 73, 95).

Interestingly, it can be shown that any irreducible 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.

In linear analysis, finding a sufficient set S is tantamount to finding a set S of regressors such that the total effect of X on Y is given by the potential regression coeffcient of Y on X, given S. A similar criterion applies to the identification of path coeffcients (Pearl, 2009a, p. 150).

The back-door criterion allows us to write equation (25) directly, by selecting a sufficient set S directly 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 covariates—namely, a set S that minimizes measurement cost or sampling variability (Tian et al., 1998).

All in all, one can safely state that, armed with the back-door criterion, causality has removed "confounding" from its store of enigmatic and controversial concepts.

3.3.2 Confounding Equivalence—A Graphical Test

Another problem that has been given graphical solution recently is that of determining whether adjustment for two sets of covariates would result in the same confounding bias (Pearl and Paz, 2010). The reasons for posing this question are several. First, an investigator may wish to assess, prior to taking any measurement, whether two candidate sets of covariates, differing substantially in dimensionality, measurement error, cost, or sample variability, are equally valuable in their bias-reduction potential. Second, assuming that the structure of the underlying DAG is only partially known, we may wish to test, using adjustment, which of two hypothesized structures is compatible with the data. Structures that predict equal response to adjustment for two sets of variables must be rejected if, after adjustment, such equality is not found in the data.

Definition 4 (*c*-equivalence) *Define two sets of covariates, T and Z, as c-*equivalent*, (c connotes "confounding") if the following equality holds:*

$$\sum_t P(y|x,t)P(t) = \sum_z P(y|x,z)P(z) \qquad \forall x, y \tag{27}$$

Definition 5 (Markov boundary) *For any set of variables S in a DAG G, and any variable $X \notin S$, the* Markov boundary S_m *of S (relative to X) is the minimal subset of S that d-separates X from all other members of S.*

In Figure 4, for example, the Markov boundary of $S = \{W_1, Z_1, Z_2, Z_3\}$ is $S_m = \{W_1, Z_3\}$, while the Markov boundary of $X = \{W_3, Z_3, Y\}$ is $S_m = S$.

Theorem 2 *(Pearl and Paz, 2010)*
Let Z and T be two sets of variables in G, containing no descendant of X. A necessary and sufficient condition for Z and T to be c-equivalent is that at least one of the following conditions holds:

1. *$Z_m = T_m$, (i.e., the Markov boundary of Z coincides with that of T).*
2. *Z and T are admissible (i.e., satisfy the back-door condition).*

For example, the sets $T = \{W_1, Z_3\}$ and $Z = \{Z_3, W_2\}$ in Figure 4 are c-equivalent, because each blocks all back-door paths from X to Y. Similarly, the nonadmissible sets $T = \{Z_2\}$ and $Z = \{W_2, Z_2\}$ are c-equivalent, since their Markov boundaries are the same ($T_m = Z_m = \{Z_2\}$). In contrast, the sets $\{W_1\}$ and $\{Z_1\}$, although they block the same set of paths in the graph, are not c-equivalent; they fail both conditions of Theorem 2.

Tests for c-equivalence (27) are fairly easy to perform, and they can also be assisted by propensity score methods. The information that such tests provide can be as powerful as conditional independence tests. The statistical ramification of such tests is explicated in Pearl and Paz (2010).

3.3.3 General Control of Confounding

Adjusting for covariates is only one of many methods that permits us to estimate causal effects in nonexperimental studies. Pearl (1995) 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 multistage 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 unblocked 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

$$P(y|do(x)) = \sum_{w_3} P(w_3|do(x))P(y|do(w_3)). \tag{28}$$

In this example, the variable W_3 acts as a "mediating instrumental variable" (Pearl, 1993b; Chalak and White, 2006; Morgan and Winship, 2007).

The analysis used in the derivation and validation of such results invokes mathematical rules 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 nonexperimental data. When such a transformation is feasible, we can be sure that the causal quantity is identifiable.

Applications of this calculus to problems involving multiple interventions (e.g., time-varying treatments), conditional policies, and surrogate experiments were developed in Pearl and Robins (1995), Kuroki and Miyakawa (1999), and Pearl (2000a, chs. 3–4).

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

Theorem 3 (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.*[14]

For example, if W_3 is the only observed covariate in the model of Figure 4, $P(y|do(x))$ can be estimated since every path from X to W_3 (the only child of X) traces either the arrow $X \rightarrow W_3$, or the arrow $W_3 \rightarrow Y$, both emanating from a measured variable (W_3).

Shpitser and Pearl (2006) have further extended this theorem by (1) presenting a *necessary* and sufficient condition for identification, and (2) extending the condition from causal effects to any counterfactual expression. The corresponding unbiased estimands for these causal quantities are readable directly from the diagram.

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

Graph-based methods for effect identification under measurement errors are discussed in (Pearl, 2010c; Hernán and Cole, 2009; Cai and Kuroki, 2008).

3.4 Counterfactual Analysis in Structural Models

Not all questions of causal character can be encoded in $P(y|do(x))$ type expressions, thus implying that not all causal questions can be answered from experimental studies. For example, questions of attribution or susceptibility (e.g., what fraction of test failure cases are *due to* a specific educational program?) cannot be answered from experimental studies, and naturally this kind of question cannot be expressed in $P(y|do(x))$ notation.[15] 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 make a minimal modification in the current model, which may have assigned X a different value, say $X = x'$, so as to ensure the specified condition $X = x$. Such a minimal modification amounts to replacing the equation for X by a constant x, as we have done in equation (6). This replacement permits the constant x to differ from the actual value of X (namely $f_X(z, u_X)$) without rendering the system of equations inconsistent, thus yielding a formal interpretation of counterfactuals in multistage models, where the dependent variable in one equation may be an independent variable in another.

Definition 6 (unit-level counterfactuals—the "surgical" definition (Pearl, 2000a, p. 98)**)** *Let M be a structural model and M_x a modified version of M, with the equa-*

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

[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), using surgeries (equation 29), thus permitting endogenous variables to serve as counterfactual antecedents. The "surgery definition" was used in Pearl (2000a, p. 417) and criticized by Cartwright (2007) and Heckman (2005); see Pearl (2009a, pp. 362–63, 374–79) for rebuttals.

tion(s) of X replaced by $X = x$. Denote the solution for Y in the equations of M_x by the symbol $Y_{M_x}(u)$. The counterfactual $Y_x(u)$ (Read: "The value of Y in unit u, had X been x") is given by

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{29}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x.

We see that the unit-level counterfactual $Y_x(u)$, which in the Neyman-Rubin approach is treated as a primitive, undefined quantity, is actually a derived quantity in the structural framework. We use the same subscripted notation for both, because they represent the same physical entity: the response Y of experimental unit u under the hypothetical condition $X = x$. The fact that we equate the experimental unit u with a vector of background conditions, $U = u$, in M, reflects the understanding that the name of a unit or its identity do not matter; it is only the vector $U = u$ of attributes characterizing a unit that determines its behavior or response. As we go from one unit to another, the laws of nature, as they are reflected in the functions f_X, f_Y, etc., remain invariant; only the attributes $U = u$ vary from individual to individual.

To illustrate, consider the solution of Y in the modified model M_{x_0} of equation (6), which Definition 6 endows with the symbol $Y_{x_0}(u_X, u_Y, u_Z)$. This entity has a clear counterfactual interpretation, for it stands for the way an individual with characteristics (u_X, u_Y, u_Z) would respond, had the treatment been x_0, rather than the treatment $x = f_X(z, u_X)$ actually received by that individual. In our example, since Y does not depend on u_X and u_Z, we can write

$$Y_{x_0}(u) = Y_{x_0}(u_Y, u_X, u_Z) = f_Y(x_0, u_Y). \tag{30}$$

In a similar fashion, we can derive

$$Y_{z_0}(u) = f_Y(f_X(z_0, u_X), u_Y),$$

$$X_{z_0,y_0}(u) = f_X(z_0, u_X),$$

and so on. These examples reveal the counterfactual reading of each individual structural equation in the model of equation (5). The equation $x = f_X(z, u_X)$, for example, advertises the empirical claim that, regardless

of the values taken by other variables in the system, had Z been z_0, X would take on no other value but $x = f_X(z_0, u_X)$.

Clearly, the distribution $P(u_Y, u_X, u_Z)$ 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 linear equations (as in Figure 1),

$$x = u_X \qquad y = \beta x + u_X,$$

the conditioning events $Y = y_0$ and $X = x_0$ yield $U_X = x_0$ and $U_Y = y_0 - \beta x_0$, and we can conclude that, with probability one, Y_{x_1} must take on the value $Y_{x_1} = \beta x_1 + U_Y = \beta(x_1 - x_0) + y_0$. 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 $\{U_X, U_Y\}$ and, for that reason, attributional queries are generally not identifiable in nonparametric models (see Pearl (2000a, ch. 9)).

In general, if x and x' are incompatible, then Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$."[17] Such concerns have been a source of objections to treating counterfactuals as jointly distributed random variables (Dawid, 2000). The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels neutralizes these objections (Pearl, 2000b), since the contradictory joint statement is mapped into an ordinary event, one where the background variables satisfy both statements simultaneously, each in its own distinct submodel; such events have well-defined probabilities.

The surgical definition of counterfactuals given by (29), provides the conceptual and formal basis for the Neyman-Rubin potential-outcome framework, an approach to causation that takes a controlled randomized trial (CRT) as its ruling paradigm, assuming that nothing is known to the experimenter about the science behind the data. This "black-box" approach, which

[17] For example, "The probability is 80% that Joe belongs to the class of patients who will be cured if they take the drug and die otherwise."

has thus far been denied the benefits of graphical or structural analyses, was developed by statisticians who found it difficult to cross the two mental barriers discussed in Section 2.2. Section 5 establishes the precise relationship between the structural and potential-outcome paradigms, and outlines how the latter can benefit from the richer representational power of the former.

3.5 Remarks on Heterogeneity

The distinction between general, or population-level causes (e.g., "Drinking hemlock causes death") and singular or unit-level causes (e.g., "Socrates' drinking hemlock caused his death"), which many philosophers have regarded as irreconcilable (Eells, 1991), introduces no tension at all in the structural theory. The two types of sentences differ merely in the level of situation-specific information that is brought to bear on a problem—that is, in the specificity of the evidence e that enters the quantity $P(Y_x = y|e)$. When e includes *all* factors u, we have a deterministic, unit-level causation on our hand; when e contains only a few known attributes (e.g., age, income, occupation) while others are assigned probabilities, a population-level analysis ensues.

The inherently nonlinear nature of nonparametric structural equations permits us to go beyond constant-coefficient models and encode the way causal effects may vary across individuals having differing characteristics, a pervasive condition known as "effect heterogeneity (Xie, 2007; Elwert and Winship, 2010). This does not mean of course that we are able to quantify the degree of heterogeneity due to totally unknown (hence unobserved) individual variations. No analysis can recover individual-level effects from a one-time population-level study, be it observational or experimental. In a population where some individuals respond positively and some negatively, it is quite possible to find an average causal effect of zero (Pearl, 2009a, p. 36) without knowing which subpopulation a given individual belongs to, or whether such subpopulations exist.

What structural modeling enables us to do is, first, account for individual variations whenever they are due to observed characteristics (say income, occupation, age, etc.), second, estimate average causal effects despite variation in unobserved characteristics, whenever they are known not to influence certain variables in the analysis (as in Section 3.4), and, finally, assess, by simulation, the extent to which regression type estimators would yield biased results when the parametric form used misspecifies the nonlinearities

involved (VanderWeele and Robins, 2007; Elwert and Winship, 2010).

4 Methodological Principles of Causal Inference

The structural theory described in the previous sections dictates a principled methodology that eliminates much of the confusion concerning the interpretations of study results as well as the ethical dilemmas that this confusion tends to spawn. The methodology dictates that every investigation involving causal relationships (and this entails the vast majority of empirical studies in the health, social, and behavioral sciences) should be structured along the following five-step process:

1. **Define:** Express the target quantity Q as a function $Q(M)$ that can be computed from any model M.

2. **Assume:** Formulate causal assumptions using ordinary scientific language and represent their structural part in graphical form.

3. **Identify:** Determine if the target quantity is identifiable (i.e., expressible in terms of estimable parameters).

4. **Test:** Identify the testable implications of M (if any) and test those that are necessary for the identifiability of Q.

5. **Estimate:** Estimate the target quantity if it is identifiable, or approximate it, if it is not.

4.1 Defining the Target Quantity

The definitional phase is the most neglected step in current practice of quantitative analysis. The structural modeling approach insists on defining the target quantity, be it "causal effect," "mediated effect," "effect on the treated," or "probability of causation" before specifying any aspect of the model, without making functional or distributional assumptions and prior to choosing a method of estimation.

The investigator should view this definition as an *algorithm* that receives a model M as an input and delivers the desired quantity $Q(M)$ as the output.

Surely, such algorithm should not be tailored to any aspect of the input M nor to the interpretation of the variables in V; it should be general, and ready to accommodate any conceivable model M whatsoever. Moreover, the investigator should imagine that the input M is a completely specified model, with all the functions $f_X, f_Y, \ldots$ and all the U variables (or their associated probabilities) given precisely. This is the hardest step for statistically trained investigators to make; knowing in advance that such details will never be estimable from the data, the definition of $Q(M)$ appears like a futile exercise in fantasyland—but it is not.

For example, the formal definition of the interventional distribution $P(y|do(x))$, as given in equation (7), is universally applicable to all models, parametric as well as nonparametric, through the formation of a submodel M_x. This definition remains the same regardless of whether X stands for treatment, gender, or the gravitational constant; manipulation restrictions do not enter the definitional phase of the study (Pearl, 2009a, pp. 361, 375). By defining causal effect procedurally, thus divorcing it from its traditional parametric representation, the structural theory avoids the many pitfalls and confusions that have plagued the interpretation of structural and regressional parameters for the past half century.[18]

4.2 Explicating Causal Assumptions

This is the second most neglected step in causal analysis. In the past, the difficulty has been the lack of a language suitable for articulating causal assumptions which, aside from impeding investigators from explicating assumptions, also inhibited them from giving causal interpretations to their findings.

Structural equation models, in their counterfactual reading, have removed this lingering difficulty by providing the needed language for causal analysis. Figures 3 and 4 illustrate the graphical component of this language,

[18]Note that β in equation (1), the incremental causal effect of X on Y, is defined procedurally by

$$\beta \triangleq E(Y|do(x_0+1)) - E(Y|do(x_0)) = \frac{\partial}{\partial x}E(Y|do(x)) = \frac{\partial}{\partial x}E(Y_x).$$

Naturally, all attempts to give β statistical interpretation have ended in frustrations (Holland, 1988; Whittaker, 1990; Wermuth, 1992; Wermuth and Cox, 1993), some persisting well into the twenty-first century (Sobel, 2008).

where assumptions are conveyed through the missing arrows in the diagram. If numerical or functional knowledge is available, for example, linearity or monotonicity of the functions $f_X, f_Y, \ldots$, those are stated separately, and applied in the identification and estimation phases of the study. Today we understand that the longevity and natural appeal of structural equations stem from the fact that they permit investigators to communicate causal assumptions formally and in the very same vocabulary in which scientific knowledge is stored.

Unfortunately, however, this understanding is not shared by all causal analysts; some analysts vehemently oppose the re-emergence of structure-based causation and insist, instead, on articulating causal assumptions exclusively in the unnatural (though formally equivalent) language of "potential outcomes," "ignorability," "missing data," "treatment assignment," and other metaphors borrowed from clinical trials. This modern assault on structural models is perhaps more dangerous than the regressional invasion that suppressed the causal readings of these models in the late 1970s (Richard, 1980). While sanctioning causal inference in one narrow style of analysis, the modern assault denies validity to any other style, including structural equations, thus discouraging investigators from subjecting models to the scrutiny of scientific knowledge.

This exclusivist attitude is manifested in passages such as: "The crucial idea is to set up the causal inference problem as one of missing data" or "If a problem of causal inference cannot be formulated in this manner (as the comparison of potential outcomes under different treatment assignments), it is not a problem of inference for causal effects, and the use of 'causal' should be avoided," or, even more bluntly, "the underlying assumptions needed to justify any causal conclusions should be carefully and explicitly argued, not in terms of technical properties like "uncorrelated error terms," but in terms of real world properties, such as how the units received the different treatments" (Wilkinson et al., 1999).

The methodology expounded in this paper testifies against such restrictions. It demonstrates the viability and scientific soundness of the traditional structural equation paradigm, which stands diametrically opposed to the "missing data" paradigm. It renders the vocabulary of "treatment assignment" stifling and irrelevant (e.g., there is no "treatment assignment" in sex discrimination cases). Most importantly, it strongly prefers the use of "uncorrelated error terms," (or "omitted factors") over its "strong ignorability" alternative as the proper way of articulating causal assumptions. Even

the most devout advocates of the "strong ignorability" language use "omitted factors" when the need arises to defend assumptions (e.g., Sobel, 2008).

4.3 Identification, Tests, Estimation, and Approximation

Having unburdened itself from parametric representations, the identification process in the structural framework proceeds either in the space of assumptions (i.e., the diagram) or in the space of mathematical expressions, after translating the graphical assumptions into a counterfactual language, as demonstrated in Section 5.3. Graphical criteria such as those of Definition 3 and Theorem 3 permit the identification of causal effects to be decided entirely within the graphical domain, where it can benefit from the guidance of scientific understanding. Identification of counterfactual queries, on the other hand, often require a symbiosis of both algebraic and graphical techniques. The nonparametric nature of the identification task (Definition 1) makes it clear that contrary to traditional folklore in linear analysis, it is not the model that need be identified but the query Q—the target of investigation. It also provides a simple way of proving nonidentifiability: the construction of two parameterization of M, agreeing in P and disagreeing in Q, is sufficient to rule out identifiability.

4.3.1 Testing the Relevant Assumptions

When Q is identifiable, the structural framework also delivers an algebraic expression for the estimand $EST(Q)$ of the target quantity Q, examples of which are given in equations (24) and (25), and estimation techniques are then unleashed as discussed in Section 4.3.2. A prerequisite part of this estimation phase is a test for the testable implications, if any, of those assumptions in M that render Q identifiable—there is no point in estimating $EST(Q)$ if the data proves those assumptions false and $EST(Q)$ turns out to be a misrepresentation of Q. The testable implications of any given model are vividly advertised by its associated graph G. Each d-separation condition in G corresponds to a conditional independence test that can be tested in the data to support the validity of M. These can easily be enumerated by attending to each missing edge in the graph. For example, in Figure 3, the missing edges are $Z_1 - Z_2, Z_1 - Y$, and $Z_2 - X$. Accordingly, the testable

implications of M are

$$\begin{aligned} Z_1 &\perp\!\!\!\perp Z_2 \\ Z_1 &\perp\!\!\!\perp Y|\{X_1, Z_2, Z_3\} \\ Z_2 &\perp\!\!\!\perp X|\{Z_1, Z_3\}. \end{aligned}$$

In linear systems, these conditional independence constraints translate into zero coefficients in the proper regression equations. For example, the three implications above translate into $a = 0, b_1 = 0$, and $c_1 = 0$ in the following regressions:

$$\begin{aligned} Z_1 &= aZ_2 + \epsilon \\ Z_1 &= b_1Y + b_2X + b_3Z_2 + b_4Z_3 + \epsilon' \\ Z_2 &= c_1X + c_3Z_1 + c_4Z_3 + \epsilon''. \end{aligned}$$

Such tests are easily conducted by routine regression techniques, and they provide valuable diagnostic information for model modification, in case any of them fail (see Pearl, 2009a, pp. 143–45). Software for automatic detection of all such tests, as well as other implications of graphical models, are reported in Kyono (2010).

If the model is Markovian (i.e., acyclic with no unobserved confounders), then the d-separation conditions are the ONLY testable implications of the model. If the model contains unobserved confounders, then additional constraints can be tested, beyond the d-separation conditions (see footnote 8).

Investigators should be reminded, however, that only a fraction, called "kernel," of the assumptions embodied in M are needed for identifying Q (Pearl, 2004), the rest may be violated in the data with no effect on Q. In Figure 2, for example, the assumption $\{U_Z \perp\!\!\!\perp U_X\}$ is not necessary for identifying $Q = P(y|do(x))$; the kernel $\{U_Y \perp\!\!\!\perp U_Z, U_Y \perp\!\!\!\perp U_X\}$ (together with the missing arrows) is sufficient. Therefore, the testable implication of this kernel, $Z \perp\!\!\!\perp Y|X$, is all we need to test when our target quantity is Q; the assumption $\{U_Z \perp\!\!\!\perp U_X\}$ need not concern us.

More importantly, investigators must keep in mind that only a tiny fraction of any kernel lends itself to statistical tests; the bulk of it must remain untestable, at the mercy of scientific judgment. In Figure 2, for example, the assumption set $\{U_X \perp\!\!\!\perp U_Z, U_Y \perp\!\!\!\perp U_X\}$ constitutes a sufficient kernel for $Q = P(y|do(x))$ (see equation 28) yet it has no testable implications whatsoever. The prevailing practice of submitting an entire structural equation

model to a "goodness of fit" test (Bollen, 1989) in support of causal claims is at odds with the logic of structural modeling (see Pearl, 2000a, pp. 144–45. Statistical tests can be used for rejecting certain kernels in the rare cases where such kernels have testable implications, but passing these tests does not prove the validity of any causal claim; one can always find alternative causal models that make a contradictory claim and, yet, possess identical statistical implications.[19] The lion's share of supporting causal claims falls on the shoulders of untested causal assumptions.[20]

Some researchers consider this burden to be a weakness of structural models and would naturally prefer a methodology in which claims are less sensitive to judgmental assumptions; unfortunately, no such methodology exists. The relationship between assumptions and claims is a universal one—namely, for every set A of assumptions (knowledge) there is a unique set of conclusions C that one can deduce from A, given the data, regardless of the method used. The completeness results of Shpitser and Pearl (2006) imply that structural modeling operates at the boundary of this universal relationship; no method can do better.

4.3.2 Estimation and Propensity Score Matching

The mathematical derivation of causal effect estimands, like equations (25) and (28) is merely a first step toward computing quantitative estimates of those effects from finite samples, using the rich traditions of statistical estimation and machine learning, Bayesian as well as non-Bayesian. Although the estimands derived in (25) and (28) are nonparametric, this does not mean that we should refrain from using parametric forms in the estimation phase of the study. Parameterization is in fact necessary when the dimensionality of a problem is high. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in (28) can be converted to the product $E(Y|do(x)) = r_{W_3X} r_{YW_3 \cdot X} x$,

[19] This follows logically from the demarcation line of Section 2.1. The fact that some social scientists were surprised by the discovery of contradictory equivalent models (see (Pearl, 2009a, p. 148) suggests that these scientists did not take very seriously the ramifications of the causal-statistical distinction, or that they misunderstood the conditional nature of all causal claims drawn from observational studies (see Pearl, 2009a, pp. 369–73.

[20] The methodology of "causal discovery" (Spirtes et al. 2000; Pearl 2000a, ch. 2) is likewise based on the causal assumption of "faithfulness" or "stability"—a problem-independent assumption that constrains the relationship between the structure of a model and the data it may generate. We will not assume stability in this paper.

where $r_{YZ \cdot X}$ is the (standardized) coefficient of Z in the regression of Y on Z and X. More sophisticated estimation techniques are the "marginal structural models" of Robins (1999), and the "propensity score" method of Rosenbaum and Rubin (1983), which were found to be particularly useful when dimensionality is high and data are sparse (see Pearl (2009a, pp. 348–52)).

The method of propensity score (Rosenbaum and Rubin, 1983), or propensity score matching (PSM), is the most developed and popular strategy for causal analysis in observational studies (Morgan and Winship, 2007; D'Agostino, Jr., 1998); it deserves therefore a separate discussion. PSM is based on a simple, yet ingenious, idea of purely statistical character. Assuming a binary action (or treatment) X, and an arbitrary set S of measured covariates, the propensity score $L(s)$ is the probability that action $X = 1$ will be chosen by a participant with characteristics $S = s$, or

$$L(s) = P(X = 1|S = s). \tag{31}$$

Rosenbaum and Rubin showed us that, viewing $L(s)$ as a function of S (hence, as a random variable) X and S are independent given $L(s)$—or $X \perp\!\!\!\perp S|L(s)$. In words, all units that map into the same value of $L(s)$ are comparable, or "balanced," in the sense that, within each stratum of L, treated and untreated units have the same distribution of characteristics S.[21]

Let us assume, for simplicity, that $L(s)$ can be estimated separately from the data and approximated by discrete strata $L = \{l_1, l_2, \ldots, l_k\}$. The conditional independence $X \perp\!\!\!\perp S|L(s)$, together with the functional mapping $S \rightarrow L$, renders S and L c-equivalent in the sense defined in Section 3.3.2, equation (27)—namely, for any Y,

$$\sum_s P(y|s, x)P(s) = \sum_l P(y|l, x)P(l). \tag{32}$$

This follows immediately by writing

$$\begin{aligned} \textstyle\sum_l P(y|l, x)P(l) &= \textstyle\sum_s \sum_l P(y|l, s, x)P(l)P(s|l, x) \\ &= \textstyle\sum_s \sum_l P(y|s, x)P(l)P(s|l) \\ &= \textstyle\sum_s P(y|s, x)P(s). \end{aligned}$$

[21] This independence emanates from the special nature of the function $L(s)$ and is not represented in the graph; i.e., if we depict L as a child of S, L would not in general d-separate S from X.

The c-equivalence of S and L implies that, if for any reason we wish to estimate the "adjustment estimand" $\sum_s P(y|s,x)P(s)$, with S and Y two arbitrary sets of variables, then, instead of summing over a high-dimensional set S, we might as well sum over a one-dimensional vector $L(s)$. The asymptotic estimate, in the limit of a very large sample, would be the same in either method.

This c-equivalence further implies that if we choose to approximate the interventional distribution $P(y|do(x))$ by the adjustment estimand $E_sP(y|s,x)$, then, asymptotically, the same approximation can be achieved using the estimand $E_lP(y|l,x)$, where the adjustment is performed over the strata of L. The latter has the advantage that, for finite samples, each of the strata is less likely to be empty and each is likely to contain both treated and untreated units for comparison.

The method of propensity score can thus be seen as an efficient estimator of the adjustment estimand, formed by an arbitrary set of covariates S; it makes no statement regarding the appropriateness of S, nor does it promise to correct for any confounding bias, or to refrain from creating new bias where none exists.

In the special case where S is *admissible*, that is,

$$P(y|do(x)) = E_sP(y|s,x), \tag{33}$$

L would be admissible as well, and we would then have an unbiased estimand of the causal effect,[22]

$$P(y|do(x)) = E_lP(y|l,x),$$

accompanied by an efficient method of estimating the right-hand side. Conversely, if S in inadmissible, L would be inadmissible as well, and all we can guarantee is that the bias produced by the former would be faithfully and efficiently reproduced by the latter.

The simplicity of PSM methods and the strong endorsement they received from prominent statisticians (Rubin, 2007), social scientists (Morgan and Winship, 2007; Berk and de Leeuw, 1999), health scientists (Austin, 2008), and economists (Heckman, 1992) has increased the popularity of the

[22]Rosenbaum and Rubin (1983) proved the c-equivalence of S and L only for admissible S, which is unfortunate; it gave users the impression that propensity score matching somehow contributes to bias reduction vis-à-vis ordinary adjustment.

method to the point where some federal agencies now expect program evaluators to use this approach as a substitute for experimental designs (Peikes et al., 2008). This move reflects a general tendency among investigators to play down the cautionary note concerning the required admissibility of S, and to interpret the mathematical proof of Rosenbaum and Rubin as a guarantee that, in each strata of L, matching treated and untreated subjects somehow eliminates confounding from the data and contributes therefore to overall bias reduction. This tendency was further reinforced by empirical studies (Heckman et al., 1998; Dehejia and Wahba, 1999) in which agreement was found between propensity score analysis and randomized trials, and in which the agreement was attributed to the ability of the former to "balance" treatment and control groups on important characteristics. Rubin (2007) has encouraged such interpretations by stating: "This application uses propensity score methods to create subgroups of treated units and control units...as if they had been randomized. The collection of these subgroups then 'approximate' a randomized block experiment with respect to the observed covariates."

Subsequent empirical studies, however, have taken a more critical view of propensity scores, noting with disappointment that a substantial bias is sometimes measured when careful comparisons are made to results of clinical studies (Smith and Todd, 2005; Luellen et al., 2005; Peikes et al., 2008).

The reason for these disappointments lie in a popular belief that adding more covariates can cause no harm (Rosenbaum, 2002, p. 76), which seems to absolve one from thinking about the causal relationships among those covariates, the treatment, the outcome and, most importantly, the confounders left unmeasured (Rubin, 2009).

This belief stands contrary to the conclusions of the structural theory of causation. The admissibility of S can be established only by appealing to causal knowledge, and such knowledge, as we know from d-separation and the back-door criterion, makes bias reduction a nonmonotonic operation—that is, eliminating bias (or imbalance) due to one confounder may awaken and unleash bias due to dormant, unmeasured confounders. Examples abound where adding a variable to the analysis not only is not needed but would introduce irreparable bias (Pearl, 2009a; Shrier, 2009; Sjölander, 2009). In Figure 3, for example, if the arrows emanating from Z_3 are weak, then no adjustment is necessary; adjusting for Z_3 or matching with the propensity score $L(z_3) = P(X = 1|Z = z_3)$ would introduce bias by opening the back-door path

$$X \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow Y.$$

Another general belief that stands contrary to the structural theory is that the bias-reducing potential of propensity score methods can be assessed experimentally by running case studies and comparing effect estimates obtained by propensity scores to those obtained by controlled randomized experiments (Shadish and Cook, 2009). Such comparisons would be informative for problems governed by the same graph structures and the same choice of S. In general, however, such comparison tells us very little about the performance of PSM methods in problems that differ from the one in the randomized trial. Measuring significant bias reduction in one problem instance (say, an educational program in Oklahoma) does not preclude a bias increase in another (say, crime control in Arkansas), even under identical statistical distributions $P(x, s, y)$.

It should be emphasized, though, that contrary to conventional wisdom (e.g., Rubin, 2007, 2009), propensity score methods are merely efficient estimators of the right-hand side of (25); they entail the same asymptotic bias and cannot be expected to reduce bias in the event that the set S does not satisfy the back-door criterion (Pearl, 2000a, 2009c,d). Consequently, the prevailing practice of conditioning on as many pretreatment measurements as possible is dangerously misguided; some covariates (e.g., Z_3 in Figure 3) may actually increase bias if included in the analysis (see footnote 28). Using simulation and parametric analysis, Heckman and Navarro-Lozano (2004) and Bhattacharya and Vogt (2007) indeed confirmed the bias-raising potential of certain covariates in propensity score methods. In particular, such covariates include: (1) colliders, (2) variables on the pathways from X to Y, or descendants thereof (Pearl, 2009a, pp. 339–40), and (3) instrumental variables and variables that affect X more strongly than they affect Y (Bhattacharya and Vogt, 2007; Pearl, 2010d).[23] The graphical tools presented in this section unveil the character of these covariates and show precisely what covariates should and should not be included in the conditioning set for propensity score matching.

[23]Contrary to prevailing practice (documented in Bhattacharya and Vogt (2007)), adding an instrumental variable as a predictor in the propensity score tends to amplify bias (if such exists) despite the improvement in prediction of the so called "treatment assisgnment." This is one of several bad practices that graph-based analysis may help rectify.

4.3.3 Bounds and Approximations

When conditions for identification are not met, the best we can do is derive *bounds* for the quantities of interest—namely, a range of possible values of Q that represents our ignorance about the details of the data-generating process M and that cannot be improved with increasing sample size. A classical example of a nonidentifiable model that has been approximated by bounds, is the problem of estimating causal effect in experimental studies marred by noncompliance, the structure of which is given in Figure 5.

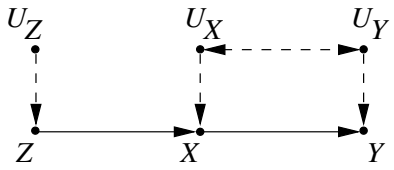

Figure 5: Causal diagram representing the assignment (Z), treatment (X), and outcome (Y) in a clinical trial with imperfect compliance.

Our task in this example is to find the highest and lowest values of Q

$$Q \triangleq P(Y = y|do(x)) = \sum_{u_X} P(Y = y|X = x, U_X = u_X)P(U_X = u_X) \quad (34)$$

subject to the equality constraints imposed by the observed probabilities $P(x, y, |z)$, where the maximization ranges over all possible functions $P(u_Y, u_X)$, $P(y|x, u_X)$, and $P(x|z, u_Y)$ that satisfy those constraints.

Realizing that units in this example fall into 16 equivalence classes, each representing a binary function $X = f(z)$ paired with a binary function $y = g(x)$, Balke and Pearl (1997) were able to derive closed-form solutions for these bounds.[24] They showed that, in certain cases, the derived bounds can yield significant 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. Kaufman and colleagues (2009) used this technique to bound direct and indirect effects (see Section 6).

[24]These equivalence classes were later called "principal stratification" by Frangakis and Rubin (2002). Looser bounds were derived earlier by Robins (1989) and Manski (1990).

5 The Potential-Outcome Framework

This section compares the structural theory presented in Sections 1–3 to the potential-outcome framework, usually associated with the names of Neyman (1923) and Rubin (1974), which takes the randomized experiment as its ruling paradigm and has appealed therefore to researchers who do not find that paradigm overly constraining. This framework is not a contender for a comprehensive theory of causation for it is subsumed by the structural theory and excludes ordinary cause-effect relationships from its assumption vocabulary. We here explicate the logical foundation of the Neyman-Rubin framework, its formal subsumption by the structural causal model, and how it can benefit from the insights provided by the broader perspective of the structural theory.

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y_x(u)$, which stands for "the value that outcome Y would obtain in experimental unit u, had treatment X been x." Here, *unit* may stand for an individual patient, an experimental subject, or an agricultural plot. In Section 3.4 (equation (29)) we saw that this counterfactual entity has a natural interpretation in structural model; it is the solution for Y in a modified system of equations, where *unit* is interpreted as a vector u of background factors that characterize an experimental unit. Each structural equation model thus carries a collection of assumptions about the behavior of hypothetical units, and these assumptions permit us to derive the counterfactual quantities of interest. In the potential-outcome framework, however, no equations are available for guidance and $Y_x(u)$ is taken as primitive, that is, an undefined quantity in terms of which other quantities are defined; not a quantity that can be derived *from* the model. In this sense the structural interpretation of $Y_x(u)$ given in (29) provides the formal basis for the potential-outcome approach; the formation of the submodel M_x explicates mathematically how the hypothetical condition "had X been x" is realized, and what the logical consequences are of such a condition.

5.1 The "Black-Box" Missing-Data Paradigm

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.

Naturally, these hypothetical entities are not entirely whimsy. They are assumed to be connected to observed variables via consistency constraints (Robins, 1986) such as

$$X = x \implies Y_x = Y, \tag{35}$$

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 (Pearl, 2010e).[25] For example, a person who chose treatment x and recovered, would also have recovered if given treatment x by design. When X is binary, it is sometimes more convenient to write (35) as

$$Y = xY_1 + (1 - x)Y_0$$

Whether additional constraints should tie the observables to the unobservables is not a question that can be answered in the potential-outcome framework, for it lacks an underlying model to define such constraints.

The main conceptual difference between the two approaches is that, whereas the structural approach views the intervention $do(x)$ as an operation that changes a 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 (35) but remaining unobserved whenever $X \neq x$. The problem of inferring probabilistic properties of Y_x then becomes one of "missing-data" for which estimation techniques have been developed in the statistical literature.

[25] Note that we are using the same subscript notation Y_x for counterfactuals in both the "missing data" and the "structural" paradigms to emphasize their formal equivalence and the fact that the "surgery" definition of equation (29) is the mathematical basis for both.

Pearl (2000a, ch. 7) uses the structural interpretation of $Y_x(u)$ to show that it is indeed legitimate to treat counterfactuals as jointly distributed random variables in all respects, that consistency constraints like (35) are automatically satisfied in the structural interpretation, and moreover, that in recursive models investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \quad \text{for all } y, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{36}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{37}$$

Equation (36) ensures that the interventions $do(Y = y)$ result in the condition $Y = y$, regardless of concurrent interventions, say $do(Z = z)$, that may be applied to variables other than Y. Equation (37) generalizes (35) to cases where Z is held fixed, at z. (See Halpern (1998) for proof of completeness.)

5.2 Problem Formulation and the Demystification of "Ignorability"

The main drawback of this black-box approach surfaces in problem formulation— namely, the phase where a researcher begins to articulate the "science" or "causal assumptions" behind the problem of interest. Such knowledge, as we have seen in Section 1, must be articulated at the onset of every problem in causal analysis— causal conclusions are only as valid as the causal assumptions upon which they rest.

To communicate scientific knowledge, the potential-outcome analyst must express assumptions as constraints on P^*, usually in the form of conditional independence assertions involving counterfactual variables. For instance, in the example shown in Figure 5, the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{Y_{z_1}, Y_{z_2}, \ldots, Y_{z_k}\}$ to communicate the understanding that Z is randomized (hence independent of U_X and U_Y).[26] To further formulate the understanding that Z does not affect Y directly, except through X, the analyst would write a so called "exclusion restriction": $Y_{xz} = Y_x$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest. For example, if we can

[26]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).

plausibly assume that in Figure 4 a set Z of covariates satisfies the conditional independence

$$Y_x \perp\!\!\!\perp X|Z \tag{38}$$

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

$$\begin{aligned} P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\ &= \sum_z P^*(Y_x = y|x,z)P(z) \quad \text{(using (38))} \\ &= \sum_z P^*(Y = y|x,z)P(z) \quad \text{(using (35))} \\ &= \sum_z P(y|x,z)P(z). \end{aligned} \tag{39}$$

The last expression contains no counterfactual quantities (thus permitting us to drop the asterisk from P^*) and coincides precisely with the standard covariate-adjustment formula of equation (25).

We see that the assumption of conditional ignorability (38) qualifies Z as an admissible covariate for adjustment; it mirrors therefore the "back-door" criterion of Definition 3, which bases the admissibility of Z on an explicit causal structure encoded in the 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 (37) or (35), 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.

This orthodoxy exacts a high cost: Instead of bringing the theory to the problem, the problem must be reformulated to fit the theory; all background knowledge pertaining to a given problem must first be translated into the language of counterfactuals (e.g., ignorability conditions) before analysis can commence. This translation may in fact be the hardest part of the problem. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability (38), the key to the derivation of (39),

holds in any familiar situation, say in the experimental setup of Figure 2(a). This assumption reads: "the value that Y would obtain had X been x, is independent of X, given Z." Even the most experienced potential-outcome expert would be unable to discern whether any subset Z of covariates in Figure 4 would satisfy this conditional independence condition.[27] Likewise, to derive equation (28) in the language of potential-outcome (see Pearl, 2000a, p. 223), we would need to convey the structure of the chain $X \rightarrow W_3 \rightarrow Y$ using the cryptic expression $W_{3_x} \perp\!\!\!\perp \{Y_{w_3}, X\}$, read: "the value that W_3 would obtain had X been x is independent of the value that Y would obtain had W_3 been w_3 jointly with the value of X." Such assumptions are cast in a language so far removed from ordinary understanding of scientific theories that, for all practical purposes, they cannot be comprehended or ascertained by ordinary mortals. As a result, researchers in the graphless potential-outcome camp rarely use "conditional ignorability" (38) to guide the choice of covariates; they view this condition as a hoped-for miracle of nature rather than a target to be achieved by reasoned design.[28]

Replacing "ignorability" with a conceptually meaningful condition (i.e., back-door) in a graphical model permits researchers to understand what conditions covariates must fulfill before they eliminate bias, what to watch for and what to think about when covariates are selected, and what experiments we can do to test, at least partially, if we have the knowledge needed for covariate selection.

Aside from offering no guidance in covariate selection, formulating a problem in the potential-outcome language encounters three additional hurdles when counterfactual variables are not viewed as byproducts of a deeper, process-based model: it is hard to ascertain (1)whether *all* relevant judgments have been articulated, (2) whether the judgments articulated are *redundant*, and (3) whether those judgments are *self-consistent.* The need to

[27]Inquisitive readers are invited to guess whether $X_z \perp\!\!\!\perp Z|Y$ holds in Figure 2(a), then reflect on why causality is so slow in penetrating statistical education.

[28]The opaqueness of counterfactual independencies explains why many researchers within the potential-outcome camp are unaware of the fact that adding a covariate to the analysis (e.g., Z_3 in Figure 4, Z in Figure 5) may actually *increase* confounding bias in propensity score matching. According to Rosenbaum (2002, p. 76) for example, "there is little or no reason to avoid adjustment for a true covariate, a variable describing subjects before treatment." Rubin (2009) goes as far as stating that refraining from conditioning on an available measurement is "nonscientific ad hockery" for it goes against the tenets of Bayesian philosophy (see Pearl (2009c,d) and Heckman and Navarro-Lozano (2004) for a discussion of this fallacy).

express, defend, and manage formidable counterfactual relationships of this type explain the slow acceptance of causal analysis among health scientists and statisticians, and why most economists and social scientists continue to use structural equation models (Wooldridge, 2002; Stock and Watson, 2003; Heckman, 2008) instead of the potential-outcome alternatives advocated in Angrist et al. (1996); Holland (1988); Sobel (1998); and Sobel (2008).

On the other hand, the algebraic machinery offered by the counterfactual notation, $Y_x(u)$, once a problem is properly formalized, can be extremely powerful in refining assumptions (Angrist et al., 1996; Heckman and Vytlacil, 2005), 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, 2000a, p. 302). The next subsection (5.3) presents a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams; translating these assumptions into counterfactual notation; performing the mathematics in the algebraic language of counterfactuals, using (35), (36), and (37); and, finally, interpreting the result in graphical terms or plain causal language. The mediation problem of Section 6 illustrates how such symbiosis clarifies the definition and identification of direct and indirect effects, a task deemed insurmountable, "deceptive" and "ill-defined" by advocates of the structureless potential-outcome approach (Rubin, 2004, 2005).

5.3 Combining Graphs and Potential Outcomes

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 exogenous 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 manipulating variables that are excluded from 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, 2000a, p. 232); 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

$$Y_{pa_Y} = Y_{pa_Y, s}. \tag{40}$$

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

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1\ pa_1}, \ldots, Z_{k\ pa_k}\}. \tag{41}$$

The exclusion restrictions expresses the fact that each parent set includes *all* direct causes of the child variable; hence, fixing the parents of Y determines the value of Y uniquely, and intervention on any other set S of (endogenous) variables can no longer affect Y. The independence restriction translates the independence between U_Y and $\{U_{Z_1}, \ldots, U_{Z_k}\}$ into independence between the corresponding potential-outcome variables. This follows from the observation that, once we set their parents, the variables in $\{Y, Z_1, \ldots, Z_k\}$ stand in functional relationships to the U terms in their corresponding equations.

As an example, consider the model shown in Figure 5, which serves as the canonical representation for the analysis of instrumental variables (Angrist et al., 1996; Balke and Pearl, 1997). This model displays the following parent sets:

$$PA_Z = \{\emptyset\},\ PA_X = \{Z\},\ PA_Y = \{X\}. \tag{42}$$

Consequently, the exclusion restrictions translate into

$$\begin{aligned} X_z &= X_{yz} \\ Z_y &= Z_{xy} = Z_x = Z \\ Y_x &= Y_{xz}, \end{aligned} \tag{43}$$

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

$$Z \perp\!\!\!\perp \{Y_x, X_z\}. \tag{44}$$

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

The role of graphs is not ended with the formulation of causal assumptions. Throughout an algebraic derivation, such as the one shown in equation (39), the analyst may need to employ additional assumptions that are entailed by the original exclusion and independence assumptions yet are not shown explicitly in their respective algebraic expressions. For example, it is hardly straightforward to show that the assumptions of equations (43)–(44) 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 using the graphical reading for conditional independence (Definition 1). (See Pearl, 2000a, pp. 16–17, 213–15.) Thus, when the need arises to employ independencies in the course of a derivation, the graph may assist the procedure by vividly displaying the independencies that logically follow from our assumptions.

6 Mediation: Direct and Indirect Effects

6.1 Direct Versus Total Effects

The causal effect we have analyzed so far, $P(y|do(x))$, measures the *total* effect of a variable (or a set of variables) X on a response variable Y. In many cases, this quantity does not adequately represent the target of investigation and attention is focused instead on the *direct* effect of X on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or

race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

From a policymaking viewpoint, an investigator may be interested in decomposing effects to quantify the extent to which racial salary disparity is due to educational disparity, or, more generally, the extent to which sensitivity to a given variable can be reduced by eliminating sensitivity to an intermediate factor, standing between that variable and the outcome. Often, the decomposition of effects into their direct and indirect components carries theoretical scientific importance, for it tells us "how nature works" and, therefore, enables us to predict behavior under a rich variety of conditions and interventions.

Structural equation models provide a natural language for analyzing path-specific effects and, indeed, considerable literature on direct, indirect, and total effects has been authored by SEM researchers (Alwin and Hauser, 1975; Graff and Schmidt, 1981; Sobel, 1987; Bollen, 1989)), for both recursive and nonrecursive models. This analysis usually involves sums of powers of coefficient matrices, where each matrix represents the path coefficients associated with the structural equations.

Yet despite its ubiquity, the analysis of mediation has long been a thorny issue in the social and behavioral sciences (Judd and Kenny, 1981; Baron and Kenny, 1986; Muller et al., 2005; Shrout and Bolger, 2002; MacKinnon et al., 2007a) primarily because structural equation modeling in those sciences were deeply entrenched in linear analysis, where the distinction between causal parameters and their regressional interpretations can easily be conflated. The difficulties were further amplified in nonlinear models, where sums and products are no longer applicable. As demands grew to tackle problems involving categorical variables and nonlinear interactions, researchers could no longer define direct and indirect effects in terms of structural or regressional coefficients, and all attempts to extend the linear paradigms of effect decomposition to nonlinear systems produced distorted results (MacKinnon et al., 2007b). These difficulties have accentuated the need to redefine and derive causal effects from first principles, uncommitted to distributional assumptions or a particular parametric form of the equations. The structural methodology presented in this paper adheres to this philosophy and it has produced indeed a principled solution to the mediation problem, based on the counterfactual reading of structural equations (29). The subsections,

that follow summarize the method and its solution.

6.2 Controlled Direct Effects

A major impediment to progress in mediation analysis has been the lack of notational facility for expressing the key notion of "holding the mediating variables fixed" in the definition of direct effect. Clearly, this notion must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, regression conditioning, matching, or adjustment. For example, consider the simple mediation models of Figure 6(a), which reads

$$\begin{aligned} x &= u_X \\ z &= f_Z(x, u_Z) \\ y &= f_Y(x, z, u_Y) \end{aligned} \tag{45}$$

and where the error terms (not shown explicitly) are assumed to be mutually

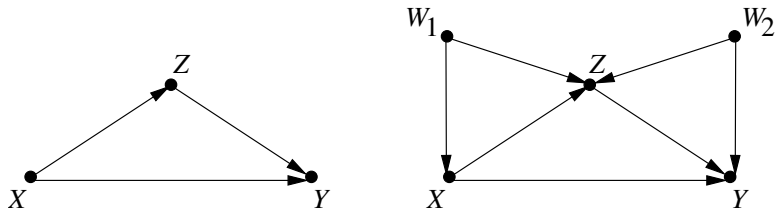

Figure 6: A generic model depicting mediation through Z (a) with no confounders and (b) with two confounders, W_1 and W_2.

independent. To measure the direct effect of X on Y it is sufficient to measure their association conditioned on the mediator Z. In Figure 6(b), however, where the error terms are dependent, it will not be sufficient to measure the association between X and Y for a given level of Z because, by conditioning on the mediator Z, which is a collision node (Definition 1), we create spurious associations between X and Y through W_2, even when there is no direct effect of X on Y (Pearl, 1998; Cole and Hernán, 2002).[29]

[29]The need to control for mediator-outcome confounders (e.g., W_2 in Figure 6(b)) was evidently overlooked in the classical paper of Baron and Kenny (1986), and has subse-

Using the $do(x)$ notation, enables us to correctly express the notion of "holding Z fixed" and to obtain a simple definition of the *controlled direct effect* of the transition from $X = x$ to $X = x'$:

$$CDE \stackrel{\Delta}{=} E(Y|do(x'), do(z)) - E(Y|do(x), do(z)).$$

Or, equivalently, we can use counterfactual notation

$$CDE \stackrel{\Delta}{=} E(Y_{x'z}) - E(Y_{xz}),$$

where Z is the set of all mediating variables. Readers can easily verify that, in linear systems, the controlled direct effect reduces to the path coefficient of the link $X \rightarrow Y$ (see footnote 18) regardless of whether confounders are present (as in Figure 6(b)) and regardless of whether the error terms are correlated or not.

This separates the task of definition from that of identification, as demanded by Section 4.1. The identification of CDE would depend, of course, on whether confounders are present and whether they can be neutralized by adjustment, but these do not alter its definition. Nor should trepidation about infeasibility of the action $do(gender = male)$ enter the definitional phase of the study. Definitions apply to symbolic models, not to human biology.[30]

Graphical identification conditions for multi-action expressions of the type $E(Y|do(x), do(z_1), do(z_2), \ldots, do(z_k))$ in the presence of unmeasured confounders were derived by Pearl and Robins (1995) (see Pearl, 2000a, ch. 4) using sequential application of the back-door conditions discussed in Section 3.2.

6.3 Natural Direct Effects

In linear systems, the direct effect is fully specified by the path coefficient attached to the link from X to Y; therefore, the direct effect is independent of the values at which we hold Z. In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to

quently been ignored by most social science researchers.

[30] In reality, it is the employer's perception of applicant's gender and his or her assessment of gender-job compatibility that renders gender a "cause" of hiring; manipulation of gender is not needed.

represent the target policy under analysis. For example, it is not uncommon to find employers who prefer males for the high-paying jobs (i.e., high z) and females for low-paying jobs (low z).

When the direct effect is sensitive to the levels at which we hold Z, it is often more meaningful to define the direct effect relative to some "natural" base-line level that may vary from individual to individual, and represents the level of Z just before the change in X. Conceptually, we can define the natural direct effect $DE_{x,x'}(Y)$[31] as the expected change in Y induced by changing X from x to x' while keeping all mediating factors constant at whatever value they *would have obtained* under $do(x)$. This hypothetical change, which Robins and Greenland (1992) conceived and called "pure" and Pearl (2001) formalized and analyzed under the rubric "natural," mirrors what lawmakers instruct us to consider in race or sex discrimination cases: "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In *Carson versus Bethlehem Steel Corp.*, 70 FEP Cases 921, 7th Cir. (1996)). Thus, whereas the controlled direct effect measures the effect of X on Y while holding Z fixed at a uniform level (z) for all units,[32] the natural direct effect allows z to vary from individual to individual to be held fixed at whatever level each individual obtains naturally, just before the change in X.

Extending the subscript notation to express nested counterfactuals, Pearl (2001) gave the following definition for the "natural direct effect":

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{46}$$

Here, Y_{x',Z_x} represents the value that Y would attain under the operation of setting X to x' and, simultaneously, setting Z to whatever value it would have obtained under the setting $X = x$. We see that $DE_{x,x'}(Y)$, the natural direct effect of the transition from x to x', involves probabilities of *nested counterfactuals* and cannot be written in terms of the $do(x)$ operator. Therefore, the natural direct effect cannot in general be identified or estimated,

[31] Pearl (2001) used the acronym NDE to denote the natural direct effect. We will delete the letter "N" from the acronyms of both the direct and indirect effect, and use DE and IE, respectively.

[32] In the hiring discrimination example, this would amount, for example, to testing gender bias by marking all application forms with the same level of schooling and other skill-defining attributes.

even with the help of ideal, controlled experiments (see footnote 15 for intuitive explanation). However, aided by the surgical definition of equation (29) and the notational power of nested counterfactuals, Pearl (2001) was nevertheless able to show that, if certain assumptions of "no confounding" are deemed valid, the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_z [E(Y|do(x', z)) - E(Y|do(x, z))]P(z|do(x)). \tag{47}$$

The intuition is simple; the natural direct effect is the weighted average of the controlled direct effect, using the causal effect $P(z|do(x))$ as a weighing function.

One condition for the validity of (47) is that $Z_x \perp\!\!\!\perp Y_{x',z}|W$ holds for some set W of measured covariates. This technical condition in itself, like the ignorability condition of (38), is close to meaningless for most investigators, as it is not phrased in terms of realized variables. The surgical interpretation of counterfactuals (29) can be invoked at this point to unveil the graphical interpretation of this condition (41). It states that W should be admissible (i.e., satisfy the back-door condition) relative to the path(s) from Z to Y. This condition, satisfied by W_2 in Figure 6(b), is readily comprehended by empirical researchers, and the task of selecting such measurements, W, can then be guided by available scientific knowledge. Additional graphical and counterfactual conditions for identification are derived in Pearl (2001), Petersen et al. (2006), and Imai et al. (2010).

In particular, it can be shown (Pearl, 2001) that expression (47) is both valid and identifiable in Markovian models (i.e., no unobserved confounders) where each term on the right can be reduced to a "*do*-free" expression using equation (24) or (25) and then estimated by regression.

For example, for the model in Figure 6(b), equation (47) reads

$$DE_{x,x'}(Y) = \sum_z \sum_{w_2} P(w_2)[E(Y|x', z, w_2)) - E(Y|x, z, w_2))] \sum_{w_1} P(z|x, w_1)P(w_1). \tag{48}$$

while for the confounding-free model of Figure 6(a) we have

$$DE_{x,x'}(Y) = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \tag{49}$$

Both (48) and (49) can easily be estimated by a two-step regression.

6.4 Natural Indirect Effects

Remarkably, the definition of the natural direct effect (46) can be turned around and provide an operational definition for the *indirect effect*—a concept shrouded in mystery and controversy, because it is impossible, using any physical intervention, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths (Pearl, 2009a, p. 355).

The *natural indirect effect*, IE, of the transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z to whatever value it would have attained had X been set to $X = x'$. Formally, this reads

$$IE_{x,x'}(Y) \triangleq E[(Y_{x,Z_{x'}}) - E(Y_x)], \tag{50}$$

which is almost identical to the direct effect (equation 46) save for exchanging x and x' in the first term (Pearl, 2001).

Indeed, it can be shown that, in general, the total effect TE of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) \triangleq E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{51}$$

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{52}$$

Since each term above is based on an independent operational definition, this equality constitutes a formal justification for the additive formula used routinely in linear systems.[33]

Note that, although it cannot be expressed in *do*-notation, the indirect effect has clear policymaking implications. For example, in the hiring discrimination context, a policymaker may be interested in predicting the gender mix in the workforce if gender bias is eliminated and all applicants are treated equally—say, the same way that males are currently treated. This quantity

[33]Some authors (e.g., VanderWeele, 2009), define the natural indirect effect as the difference $TE - DE$. This renders the additive formula a tautology of definition, rather then a theorem predicted upon the anti-symmetry $IE_{x,x'}(Y) = -IE_{x',x}(Y)$. Violation of (52) will be demonstrated in the next section.

will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent.

More generally, a policymaker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*—that is, the effect of X on Y through a selected set of paths (Avin et al., 2005).

In all these cases, the policy intervention invokes the selection of signals to be sensed rather than variables to be fixed. Pearl (2001) has therefore suggested that *signal sensing* is more fundamental to the notion of causation than *manipulation*; the latter being but a crude way of stimulating the former in an experimental setup. The mantra "No causation without manipulation" must be rejected (see Pearl, 2009a, sec. 11.4.5).

It is remarkable that counterfactual quantities like DE and IE, which could not be expressed in terms of $do(x)$ operators and therefore appear void of empirical content, can under certain conditions be estimated from empirical studies, and serve to guide policies. Awareness of this potential should embolden researchers to go through the definitional step of the study and freely articulate the target quantity $Q(M)$ in the language of science—that is, structure-based counterfactuals—despite the seemingly speculative nature of each assumption in the model (Pearl, 2000b).

6.5 The Mediation Formula: A Simple Solution to a Thorny Problem

This subsection demonstrates how the solution provided in equations (49) and (52) can be applied in assessing mediation effects in nonlinear models. We will use the simple mediation model of Figure 6(a), where all error terms (not shown explicitly) are assumed to be mutually independent, with the understanding that adjustment for appropriate sets of covariates W may be necessary to achieve this independence (as in equation 48) and that integrals should replace summations when dealing with continuous variables (Imai et al., 2010).

Combining (47) and (52), the expression for the indirect effect, IE, becomes

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)] \tag{53}$$

which provides a general formula for mediation effects, applicable to any nonlinear system, any distribution (of U), and any type of variables. Moreover, the formula is readily estimable by regression. Owing to its generality and ubiquity, I have referred to this expression as the "Mediation Formula" (Pearl, 2009b, 2010a).

The Mediation Formula represents the average increase in the outcome Y that the transition from $X = x$ to $X = x'$ is expected to produce absent any direct effect of X on Y. Though based on solid causal principles, it embodies no causal assumption other than the generic mediation structure of Figure 6(a). When the outcome Y is binary (e.g., recovery, or hiring) the ratio $(1 - IE/TE)$ represents the fraction of responding individuals who owe their response to direct paths, while $(1 - DE/TE)$ represents the fraction who owe their response to Z-mediated paths.

The Mediation Formula tells us that IE depends only on the expectation of the counterfactual Y_{xz}, not on its functional form $f_Y(x, z, u_Y)$ or its distribution $P(Y_{xz} = y)$. It calls therefore for a two-step regression which, in principle, can be performed nonparametrically. In the first step we regress Y on X and Z, and obtain the estimate

$$g(x, z) = E(Y|x, z)$$

for every (x, z) cell. In the second step we estimate the conditional expectation of $g(x, z)$ with respect to z, conditional on $X = x'$ and $X = x$, respectively, and take the difference

$$IE_{x,x'}(Y) = E_z[g(x', z) - g(x, z)].$$

Nonparametric estimation is not always practical. When Z consists of a vector of several mediators, the dimensionality of the problem might prohibit the estimation of $E(Y|x, z)$ for every (x, z) cell, and the need arises to use parametric approximation. We can then choose any convenient parametric form for $E(Y|x, z)$ (e.g., linear, logit, probit), estimate the parameters separately (e.g., by regression or maximum likelihood methods), insert the parametric approximation into (53) and estimate its two conditional expectations (over z) to get the mediated effect (VanderWeele, 2009; Pearl, 2010a).

Let us examine what the Mediation Formula yields when applied to the

linear version of Figure 6(a) (equation 45), which reads

$$\begin{aligned} x &= u_X \\ z &= b_0 + b_x x + u_Z \\ y &= c_0 + c_x x + c_z z + u_Y \end{aligned} \tag{54}$$

with u_X, u_Y, and u_Z uncorrelated, zero-mean error terms. Computing the conditional expectation in (53) gives

$$E(Y|x,z) = E(c_0 + c_x x + c_z z + u_Y) = c_0 + c_x x + c_z z$$

and yields

$$\begin{aligned} IE_{x,x'}(Y) &= \sum_z (c_x x + c_z z)[P(z|x') - P(z|x)]. \\ &= c_z[E(Z|x') - E(Z|x)] && (55) \\ &= (x' - x)(c_z b_x) && (56) \\ &= (x' - x)(b - c_x) && (57) \end{aligned}$$

where b is the total effect coefficient,

$$b = (E(Y|x') - E(Y|x))/(x' - x) = c_x + c_z b_x.$$

We thus obtained the standard expressions for indirect effects in linear systems, which can be estimated either as a difference in two regression coefficients (equation 57) or a product of two regression coefficients (equation 56), with Y regressed on both X and Z (see MacKinnon et al., 2007b). These two strategies do not generalize to nonlinear systems as shown in Pearl (2010a); direct application of (53) is necessary.

To understand the difficulty, consider adding an interaction term $c_{xz}xz$ to the model in equation (54), yielding

$$y = c_0 + c_x x + c_z z + c_{xz} xz + u_Y$$

Now assume that, through elaborate regression analysis, we obtain accurate estimates of all parameters in the model. It is still not clear what combinations of parameters measure the direct and indirect effects of X on Y, or, more specifically, how to assess the fraction of the total effect that is *explained* by mediation and the fraction that is *owed* to mediation. In linear

analysis, the former fraction is captured by the product $c_z b_x/b$ (equation 56), the latter by the difference $(b - c_x)/b$ (equation 57) and the two quantities coincide. In the presence of interaction, however, each fraction demands a separate analysis, as dictated by the Mediation Formula.

To witness, substituting the nonlinear equation in (49), (52) and (53) and assuming $x = 0$ and $x' = 1$, yields the following decomposition:

$$\begin{aligned} DE &= c_x + b_0 c_{xz} \\ IE &= b_x c_z \\ TE &= c_x + b_0 c_{xz} + b_x(c_z + c_{xz}) \\ &= DE + IE + b_x c_{xz} \end{aligned}$$

We therefore conclude that the fraction of output change for which mediation would be *sufficient* is

$$IE/TE = b_x c_z/(c_x + b_0 c_{xz} + b_x(c_z + c_{xz}))$$

while the fraction for which mediation would be *necessary* is

$$1 - DE/TE = b_x(c_z + c_{xz})/(c_x + b_0 c_{xz} + b_x(c_z + c_{xz}))$$

We note that, due to interaction, a direct effect can be sustained even when the parameter c_x vanishes and, moreover, a total effect can be sustained even when both the direct and indirect effects vanish. This illustrates that estimating parameters in isolation tells us little about the effect of mediation and, more generally, mediation and moderation are intertwined and cannot be assessed separately.

If the policy evaluated aims to prevent the outcome Y by weakening the mediating pathways, the target of analysis should be the difference $TE - DE$, which measures the highest prevention effect of any such policy. If, on the other hand, the policy aims to prevent the outcome by weakening the direct pathway, the target of analysis should shift to IE, for $TE - IE$ measures the highest preventive impact of this type of policies.

The main power of the Mediation Formula shines in studies involving categorical variables, especially when we have no parametric model of the data generating process. To illustrate, consider the case where all variables are binary, still allowing for arbitrary interactions and arbitrary distributions

of all processes. The low dimensionality of the binary case permits both a nonparametric solution and an explicit demonstration of how mediation can be estimated directly from the data. Generalizations to multivalued outcomes are straightforward.

Assume that the model of Figure 6(a) is valid and that the observed data is given by Figure 7. The factors $E(Y|x,z)$ and $P(Z|x)$ can be readily

Number of Samples	X	Z	Y	$E(Y\|x,z) = \boldsymbol{g_{xz}}$	$E(Z\|x) = \boldsymbol{h_x}$
n_1	0	0	0	$\frac{n_2}{n_1+n_2} = g_{00}$	$\frac{n_3+n_4}{n_1+n_2+n_3+n_4} = h_0$
n_2	0	0	1		
n_3	0	1	0	$\frac{n_4}{n_3+n_4} = g_{01}$	
n_4	0	1	1		
n_5	1	0	0	$\frac{n_6}{n_5+n_6} = g_{10}$	$\frac{n_7+n_8}{n_5+n_6+n_7+n_8} = h_1$
n_6	1	0	1		
n_7	1	1	0	$\frac{n_8}{n_7+n_8} = g_{11}$	
n_8	1	1	1		

Figure 7: Computing the Mediation Formula for the model in Figure 6(a), with X, Y, Z binary.

estimated as shown in the two right-most columns of Figure 7 and, when substituted in (49), (52), (53), yield

$$DE = (g_{10} - g_{00})(1 - h_0) + (g_{11} - g_{01})h_0 \tag{58}$$

$$IE = (h_1 - h_0)(g_{01} - g_{00}) \tag{59}$$

$$TE = g_{11}h_1 + g_{10}(1 - h_1) - [g_{01}h_0 + g_{00}(1 - h_0)] \tag{60}$$

We see that logistic or probit regression is not necessary; simple arithmetic operations suffice to provide a general solution for any conceivable data set, regardless of the data-generating process.

In comparing these results to those produced by conventional mediation analyses we should note that conventional methods do not define direct and indirect effects in a setting where the underlying process is unknown. MacKinnon (2008, ch. 11), for example, analyzes categorical data using logistic and probit regressions and constructs effect measures using products and differences of the parameters in those regressional forms. This strategy is not

compatible with the causal interpretation of effect measures, even when the parameters are precisely known; IE and DE may be extremely complicated functions of those regression coefficients (Pearl, 2010b). Fortunately, those coefficients need not be estimated at all; effect measures can be estimated directly from the data, circumventing the parametric analysis altogether, as shown in equations (58) and (59).

In addition to providing causally sound estimates for mediation effects, the Mediation Formula also enables researchers to evaluate analytically the effectiveness of various parametric specifications relative to any assumed model (Imai et al., 2010; Pearl, 2010a). This type of analytical "sensitivity analysis" has been used extensively in statistics for parameter estimation but could not be applied to mediation analysis, owing to the absence of an objective target quantity that captures the notion of indirect effect in both linear and nonlinear systems, free of parametric assumptions. The Mediation Formula of equation (53) explicates this target quantity formally, and casts it in terms of estimable quantities.

The derivation of the Mediation Formula was facilitated by taking seriously the five steps of the structural methodology (Section 4) together with the graphical-counterfactual-structural symbiosis spawned by the surgical interpretation of counterfactuals (equation 29).

In contrast, when the mediation problem is approached from an exclusivist potential-outcome viewpoint, void of the structural guidance of equation (29), counterintuitive definitions ensue, carrying the label "principal stratification" (Rubin, 2004, 2005), which are at variance with common understanding of direct and indirect effects. For example, the direct effect is definable only in units absent of indirect effects. This means that a grandfather would be deemed to have no direct effect on his grandson's behavior in families where he has had some effect on the father. This precludes from the analysis all typical families, in which a father and a grandfather have simultaneous, complementary influences on children's upbringing. In linear systems, to take an even sharper example, the direct effect would be undefined whenever indirect paths exist from the cause to its effect. The emergence of such paradoxical conclusions underscores the wisdom, if not necessity of a symbiotic analysis, in which the counterfactual notation $Y_x(u)$ is governed by its structural definition, equation (29).[34]

[34]Such symbiosis is now standard in epidemiology research (Robins, 2001; Petersen et al., 2006; VanderWeele and Robins, 2007; Hafeman and Schwartz, 2009; VanderWeele, 2009)

7 Conclusions

Traditional statistics is strong in devising ways of describing data and inferring distributional parameters from samples. Causal inference requires two additional 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 phenomenon. This paper surveys recent advances in causal analysis from the unifying perspective of the structural theory of causation and shows how statistical methods can be supplemented with the needed ingredients. The theory invokes nonparametric structural equation models as a formal and meaningful language for defining causal quantities, formulating causal assumptions, testing identifiability, and explicating many concepts used in causal discourse. These include randomization, intervention, direct and indirect effects, confounding, counterfactuals, and attribution. The algebraic component of the structural language coincides with the potential-outcome framework, and its graphical component embraces Wright's method of path diagrams. When unified and synthesized, the two components offer statistical investigators a powerful and comprehensive methodology for empirical research.

References

Ali, R., Richardson, T. and Spirtes, P. (2009). Markov equivalence for ancestral graphs. *The Annals of Statistics* **37** 2808–2837.

Alwin, D. and Hauser, R. (1975). The decomposition of effects in path analysis. *American Sociological Review* **40** 37–47.

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

Arjas, E. and Parner, J. (2004). Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics* **31** 171–187.

Austin, P. (2008). A critical appraisal of propensity-score matching in the medical literature from 1996 to 2003. *Statistics in Medicine* **27** 2037–2049.

and is making its way slowly toward the social and behavioral sciences.

Avin, C., Shpitser, I. and Pearl, J. (2005). Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*. Morgan-Kaufmann Publishers, Edinburgh, UK.

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

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

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51** 1173–1182.

Baumrind, D. (1993). Specious causal attributions in social sciences: The reformulated stepping-stone theory of hero in use as exemplar. *Journal of Personality and Social Psychology* **45** 1289–1298.

Berk, R. and de Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association* **94** 1045–1052.

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* **2** 47–53.

Bhattacharya, J. and Vogt, W. (2007). Do instrumental variables belong in propensity scores? Tech. Rep. NBER Technical Working Paper 343, National Bureau of Economic Research, MA.

Blalock, H. (1964). *Causal Inferences in Nonexperimental Research*. University of North Carolina Press, Chapel Hill.

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

Cai, Z. and Kuroki, M. (2008). On identifying total effects in the presence of latent variables and selection bias. In *Uncertainty in Artificial Intelligence, Proceedings of the Twenty-Fourth Conference* (D. McAllester and P. Myllymäki, eds.). AUAI, Arlington, VA, 62–69.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, New York, NY.

Chalak, K. and White, H. (2006). An extended class of instrumental variables for the estimation of causal effects. Tech. Rep. Discussion Paper, UCSD, Department of Economics.

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

Chin, W. (1998). Commentary: Issues and opinion on structural equation modeling. *Management Information Systems Quarterly* **22** 7–16.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research* **18** 115–126.

Cole, S. and Hernán, M. (2002). Fallibility in estimating direct effects. *International Journal of Epidemiology* **31** 163–165.

D'Agostino, Jr., R. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* **17** 2265–2281.

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

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

Dehejia, R. and Wahba, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* **94** 1053–1063.

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

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

Elwert, F. and Winship, C. (2010). Effect heterogeneity and bias in main-effects-only regression models. In *Heuristics, Probability and Causality: A Tribute to Judea Pearl* (R. Dechter, H. Geffner and J. Halpern, eds.). College Publications, London, 327–336.

Frangakis, C. and Rubin, D. (2002). Principal stratification in causal inference. *Biometrics* **1** 21–29.

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

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

Graff, J. and Schmidt, P. (1981). A general model for decomposition of effects. In *Systems Under Indirect Observation, Part 1* (K. Jöreskog and H. Wold, eds.). North-Holland, Amsterdam, 131–148.

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

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* **11** 1–12. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis*, Cambridge University Press, 477–490, 1995.

Hafeman, D. and Schwartz, S. (2009). Opening the black box: A motivation for the assessment of mediation. *International Journal of Epidemiology* **3** 838–845.

Halpern, J. (1998). Axiomatizing causal reasoning. In *Uncertainty in Artificial Intelligence* (G. Cooper and S. Moral, eds.). Morgan Kaufmann, San Francisco, CA, 202–210. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Heckman, J. (1992). Randomization and social policy evaluation. In *Evaluations: Welfare and Training Programs* (C. Manski and I. Garfinkle, eds.). Harvard University Press, Cambridge, MA, 201–230.

Heckman, J. (2005). The scientific model of causality. *Sociological Methodology* **35** 1–97.

Heckman, J. (2008). Econometric causality. *International Statistical Review* **76** 1–27.

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

Heckman, J. and Navarro-Lozano, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics* **86** 30–57.

Heckman, J. and Vytlacil, E. (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica* **73** 669–738.

Hernán, M. and Cole, S. (2009). Invited commentary: Causal diagrams and measurement bias. *American Journal of Epidemiology* **170** 959–962.

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

Holland, P. (1995). Some reflections on Freedman's critiques. *Foundations of Science* **1** 50–57.

Hurwicz, L. (1950). Generalization of the concept of identification. In *Statistical Inference in Dynamic Economic Models* (T. Koopmans, ed.). Cowles Commission, Monograph 10, Wiley, New York, 245–257.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. *DStatistical Science* **25** 51–71.

Judd, C. and Kenny, D. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review* **5** 602–619.

Kaufman, S., Kaufman, J. and MacLenose, R. (2009). Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *Journal of Statistical Planning and Inference* **139** 3473–3487.

Kelloway, E. (1998). *Using LISREL for structural Equation Modeling*. Sage, Thousand Oaks, CA.

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

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

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

Kyono, T. (2010). Commentator: A front-end user-interface module for graphical and structural equation modeling. Tech. Rep. R-364, <http://ftp.cs.ucla.edu/pub/stat_ser/r364.pdf>, Master Thesis, Department of Computer Science, University of California, Los Angeles, CA.

Lauritzen, S. (2001). Causal inference from graphical models. In *Complex Stochastic Systems* (D. Cox and C. Kluppelberg, eds.). Chapman and Hall/CRC Press, Boca Raton, FL, 63–107.

Lee, S. and Hershberger, S. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research* **25** 313–334.

Lindley, D. (2002). Seeing and doing: The concept of causation. *International Statistical Review* **70** 191–214.

Luellen, J., Shadish, W. and Clark, M. (2005). Propensity scores: An introduction and experimental test. *Evaluation Review* **29** 530–558.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. Lawrence Erlbaum Associates, New York.

MacKinnon, D., Fairchild, A. and Fritz, M. (2007a). Mediation analysis. *Annual Review of Psychology* **58** 593–614.

MacKinnon, D., Lockwood, C., Brown, C., Wang, W. and Hoffman, J. (2007b). The intermediate endpoint effect in logistic and probit regression. *Clinical Trials* **4** 499–513.

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

MARSCHAK, J. (1950). Statistical inference in economics. In *Statistical Inference in Dynamic Economic Models* (T. Koopmans, ed.). Wiley, New York, 1–50. Cowles Commission for Research in Economics, Monograph 10.

MORGAN, S. and WINSHIP, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research (Analytical Methods for Social Research)*. Cambridge University Press, New York, NY.

MULLER, D., JUDD, C. and YZERBYT, V. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology* **89** 852–863.

MUTHÉN, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* **46** 115–132.

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

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

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

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

PEARL, J. (1993b). Mediating instrumental variables. Tech. Rep. R-210, <http://ftp.cs.ucla.edu/pub/stat_ser/R210.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.

PEARL, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods and Research* **27** 226–284.

PEARL, J. (2000a). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

PEARL, J. (2000b). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association* **95** 428–431.

PEARL, J. (2001). Direct and indirect effects. In *Uncertainty in Artificial Intelligence, Proceedings of the Seventeenth Conference*. Morgan Kaufmann, San Francisco, CA, 411–420.

PEARL, J. (2004). Robustness of causal claims. In *Proceedings of the Twentieth Conference Uncertainty in Artificial Intelligence* (M. Chickering and J. Halpern, eds.). AUAI Press, Arlington, VA, 446–453.

PEARL, J. (2009a). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.

PEARL, J. (2009b). Causal inference in statistics: An overview. *Statistics Surveys* **3** 96–146, <http://ftp.cs.ucla.edu/pub/stat_ser/r350.pdf>.

PEARL, J. (2009c). Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine* **28** 1415–1416. <http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf>.

PEARL, J. (2009d). Myth, confusion, and science in causal analysis. Tech. Rep. R-348, <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>, University of California, Los Angeles, CA.

PEARL, J. (2010a). An introduction to causal inference. *The International Journal of Biostatistics* **6** DOI: 10.2202/1557–4679.1203, <http://www.bepress.com/ijb/vol6/iss2/7/>.

PEARL, J. (2010b). The mediation formula: A guide to the assessment of causal pathways in non-linear models. Tech. Rep. R-363, <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>, Department of Computer Science, University of California, Los Angeles, CA.

PEARL, J. (2010c). On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 425–432. <http://ftp.cs.ucla.edu/pub/stat_ser/r357.pdf>.

Pearl, J. (2010d). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 425–432. <http://ftp.cs.ucla.edu/pub/stat_ser/r356.pdf>.

Pearl, J. (2010e). On the consistency rule in causal inference: An axiom, definition, assumption, or a theorem? Tech. Rep. R-358, <http://ftp.cs.ucla.edu/pub/stat_ser/r358.pdf>, Department of Computer Science, University of California, Los Angeles, CA. Forthcoming, *Epidemiology*.

Pearl, J. and Paz, A. (2010). Confounding equivalence in causal equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. AUAI, Corvallis, OR, 433–441.

Pearl, J. and Robins, J. (1995). 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, 444–453.

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

Peikes, D., Moreno, L. and Orzol, S. (2008). Propensity scores matching: A note of caution for evaluators of social programs. *The American Statistician* **62** 222–231.

Petersen, M., Sinisi, S. and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology* **17** 276–284.

Richard, J. (1980). Models with several regimes and changes in exogeneity. *Review of Economic Studies* **47** 1–20.

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

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

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

Robins, J. (1999). Testing and estimation of directed effects by reparameterizing directed acyclic with structural nested models. In *Computation, Causation, and Discovery* (C. Glymour and G. Cooper, eds.). AAAI/MIT Press, Cambridge, MA, 349–405.

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

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

Rosenbaum, P. (2002). *Observational Studies*. 2nd ed. Springer-Verlag, New York.

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

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

Rubin, D. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* **31** 161–170.

Rubin, D. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322–331.

Rubin, D. (2007). The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine* **26** 20–36.

RUBIN, D. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine* **28** 1420–1423.

SHADISH, W. and COOK, T. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology* **60** 607–629.

SHPITSER, I. and PEARL, J. (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence* (R. Dechter and T. Richardson, eds.). AUAI Press, Corvallis, OR, 437–444.

SHPITSER, I. and PEARL, J. (2008). Dormant independence. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence*. AAAI Press, Menlo Park, CA, 1081–1087.

SHRIER, I. (2009). Letter to the editor: Propensity scores. *Statistics in Medicine* **28** 1317–1318. See also Pearl 2009 <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>.

SHROUT, P. and BOLGER, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods* **7** 422–445.

SIMON, H. (1953). Causal ordering and identifiability. In *Studies in Econometric Method* (W. C. Hood and T. Koopmans, eds.). Wiley and Sons, Inc., New York, NY, 49–74.

SIMON, H. and RESCHER, N. (1966). Cause and counterfactual. *Philosophy and Science* **33** 323–340.

SJÖLANDER, A. (2009). Letter to the editor: Propensity scores and M-structures. *Statistics in Medicine* **28** 1416–1423.

SMITH, J. and TODD, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* **125** 305–353.

SOBEL, M. (1987). Direct and indirect effects in linear structural equation models. *Sociological Methods & Research* **16** 1155–176.

SOBEL, M. (1996). An introduction to causal inference. *Sociological Methods & Research* **24** 353–379.

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

SOBEL, M. (2008). Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics* **33** 230–231.

SØRENSEN, A. (1998). Theoretical methanisms and the empirical study of social processes. In *Social Mechanisms: An Analytical Approach to Social Theory, Studies in Rationality and Social Change* (P. Hedström and R. Swedberg, eds.). Cambridge University Press, Cambridge, 238–266.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*. 2nd ed. MIT Press, Cambridge, MA.

STOCK, J. and WATSON, M. (2003). *Introduction to Econometrics*. Addison Wesley, New York.

STROTZ, R. and WOLD, H. (1960). Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28** 417–427.

TIAN, J., PAZ, A. and PEARL, J. (1998). Finding minimal separating sets. Tech. Rep. R-254, <http://ftp.cs.ucla.edu/pub/stat_ser/r254.pdf>, Computer Science Department, University of California, Los Angeles, CA.

TIAN, J. and PEARL, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence* **28** 287–313.

TIAN, J. and PEARL, J. (2002). A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*. AAAI Press/The MIT Press, Menlo Park, CA, 567–573.

VANDERWEELE, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* **20** 18–26.

VanderWeele, T. and Robins, J. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology* **18** 561–568.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Uncertainty in Artificial Intelligence, Proceedings of the Sixth Conference*. Cambridge, MA. Also in P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 6*, Elsevier Science Publishers, B.V., 255–268, 1991.

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

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

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

Wilkinson, L., the Task Force on Statistical Inference and *APA Board of Scientific Affairs* (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* **54** 594–604.

Winship, C. and Mare, R. (1983). Structural equations and path analysis for discrete data. *The American Journal of Sociology* **89** 54–110.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge and London.

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

Wright, S. (1923). The theory of path coefficients: A reply to Niles' criticism. *Genetics* **8** 239–255.

Xie, Y. (2007). Review: Otis Dudley Duncan's legacy: The demographic approach to quantitative reasoning in social science. *Research in Social Stratification and Mobility* **25** 141–156.