In P.E. Shrout, K. Keyes, and K. Ornstein (Eds.), Causality and Psychopathology: Finding the Determinants of Disorders and their Cures, Oxford University Press, 47-65, 2010.

TECHNICAL REPORT
R-338
September 2010

The Mathematics of Causal Relations

Judea Pearl

Introduction

Almost two decades have passed since Paul Holland published his highly cited review paper on the Neyman-Rubin approach to causal inference [Holland, 1986]. Our understanding of causal inference has since increased severalfold, due primarily to advances in three areas:

1. Nonparametric structural equations
2. Graphical models
3. Symbiosis between counterfactual and graphical methods

These advances are central to the empirical sciences because the research questions that motivate most studies in the health, social, and behavioral sciences are not statistical but causal in nature. For example, what is the efficacy of a given drug in a given population? Can data prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual in a specific incident?

Remarkably, although much of the conceptual framework and many of the algorithmic tools needed for tackling such problems are now well established, they are hardly known to researchers in the field who could put them into practical use. Why?

Solving causal problems mathematically requires certain extensions in the standard mathematical language of statistics, and these extensions are not generally emphasized in the mainstream literature and education. As a result, large segments of the statistical research community find it hard to appreciate and benefit from the many results that causal analysis has produced in the past two decades.

This chapter aims at making these advances more accessible to the general research community by, first, contrasting causal analysis with standard statistical analysis and, second, comparing and unifying various approaches to causal analysis.

From Associational to Causal Analysis: Distinctions and Barriers

The Basic Distinction: Coping With Change

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 the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer not only the likelihood of events under static conditions but also the dynamics of events under *changing conditions*, for example, changes induced by treatments or 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 which identify relationships that remain invariant when external conditions change.

These considerations imply that the slogan "correlation does not imply causation" can be translated into a useful principle: One cannot substantiate causal claims from associations alone, even at the population level—behind every causal conclusion there must lie some causal assumption that is not testable in observational studies.

Formulating the Basic Distinction

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, risk ratio, odd ratio, propensity score, "Granger causality," marginalization, conditionalization, and "controlling for" Examples of causal concepts are randomization, influence, effect, confounding, "holding constant," disturbance, spurious correlation, instrumental variables, ignorability, exogeneity, exchangeability, 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 causal analysis for it helps investigators to trace 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 of statistical notions alone.

Ramifications of the Basic Distinction

This principle has far-reaching consequences that are not generally recognized in the standard statistical literature. Many researchers, for example, are still convinced that confounding is solidly founded in standard, frequentist statistics and that it can be given an associational definition, saying (roughly) "U is a potential confounder for examining the effect of treatment X on outcome Y when both U and X and U and Y are not independent" [Pearl, 2009b, p. 388]. That this definition and all of its many variants

must fail is obvious from the demarcation line above; "independence" is an associational concept while confounding is for a tool used in establishing causal relations. The two do not mix hence, the definition must be false. Therefore, to the bitter disappointment of generations of epidemiology researchers, confounding bias cannot be detected or corrected by statistical methods alone; one must make some judgmental assumptions regarding causal relationships in the problem before an adjustment (e.g., by stratification) can safely correct for confounding bias.

Another ramification of the sharp distinction between associational and causal concepts is that any mathematical approach to causal analysis must acquire new notation for expressing causal 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 to 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. Scientists seeking to express causal relationships must therefore supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the relation "symptoms cause disease" is distinct from the symbolic representation of "symptoms are associated with disease."

Two Mental Barriers: Untested Assumptions and New Notation

The preceding requirements—(1) to commence causal analysis with untested,[1] theoretically or judgmentally based assumptions, and (2) to extend the syntax of probability calculus, constitute the two main obstacles to the acceptance of causal analysis among statisticians and among professionals with traditional training in statistics.

Associational assumptions, even untested, are testable in principle, given a sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference 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 meaningful and unambiguous so that one can clearly judge the plausibility or inevitability of the assumptions articulated. Statisticians can no longer ignore the mental representation in which scientists store experiential knowledge since it is this representation, and the language used to access that representation that determine the reliability of the judgments upon which the analysis so crucially depends.

How does one recognize causal expressions in the statistical literature? Those versed in the potential-outcome notation (Neyman, 1923; Rubin, 1974; Holland, 1986), can recognize such expressions through the subscripts that are attached to counterfactual events and variables, for example, $Y_x(u)$ or Z_{xy}—some authors use parenthetical expressions, such

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

as $Y(x,u)$ or $Z(x,y)$. The expression $Y_x(u)$, for example, stands for the value that outcome Y would take in individual u, had treatment X been at level x. If u is chosen at random, Y_x is a random variable, and one can talk about the probability that Y_x would attain a value y in the population, written $p(Y_x = y)$. Alternatively, Pearl (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.[2] Still a third notation that distinguishes causal expressions is provided by graphical models, where the arrows convey causal directionality.[3]

However, few have taken seriously the textbook requirement that any introduction of new notation must entail a systematic definition of the syntax and semantics that govern the notation. Moreover, in the bulk of the statistical literature before 2000, causal claims rarely appear in the mathematics. They surface only in the verbal interpretation that investigators occasionally attach to certain associations and in the verbal description with which investigators justify assumptions. For example, the assumption that a covariate is not affected by a treatment, a necessary assumption for the control of confounding [Cox, 1958], is expressed in plain English, not in a mathematical expression.

Remarkably, though the necessity of explicit causal notation is now recognized by most leaders in the field, the use of such notation has remained enigmatic to most rank-and-file researchers and its potentials still lay grossly underutilized in the statistics-based sciences. The reason for this, I am firmly convinced, can be traced to the way in which causal analysis has been presented to the research community, relying primarily on outdated paradigms of controlled randomized experiments and black-box "missing-data" models (Rubin, 1974; Holland, 1986).

The next section provides a conceptualization that overcomes these mental barriers; it offers both a friendly mathematical machinery for cause-effect analysis and a formal foundation for counterfactual analysis.

The Language of Diagrams and Structural Equations

Semantics: Causal Effects and Counterfactuals

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

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

where X stands for the level (or severity) of the disease, Y stands for the level (or severity) of the symptom, and u stands for all factors, other than the disease in question, that could

[2] 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.

[3] 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.

possibly affect Y.[4] In interpreting this equation one should think of a physical process whereby Nature *examines* the values of x and u and, accordingly, *assigns* variable Y the value $y = \beta x + u$.

To express the directionality inherent in this process, Wright augmented the equation with a diagram, later called a "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 the value nature assigns to one variable is not determined by the value taken by another.[5]

The variables V and U are called "exogenous"; they represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence, but are not influenced by, the other variables (called "endogenous") in the model.

If correlation is judged possible between two exogenous variables, U and V, it is customary to connect them by a dashed double arrow, as shown in Figure 1(b).

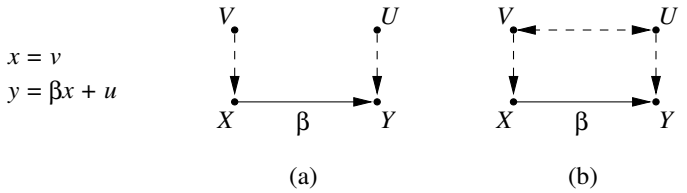

Figure 1: A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

To summarize, path diagrams encode causal assumptions via missing arrows, representing claims of zero influence, and missing double arrows (e.g., between V and U), representing the (causal) assumption $Cov(U, V)$=0.

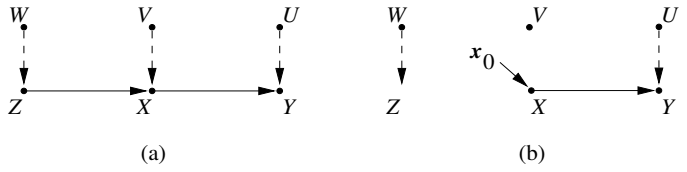

Figure 2: (a) The diagram associated with the structural model of Eq. (2). (b) The diagram associated with the modified model of Eq. (3), representing the intervention $do(X = x_0)$.

The generalization to a nonlinear system of equations is straightforward. For example, the non-parametric interpretation of the diagram of Figure 2(a) corresponds to a set of three functions, each corresponding to one of the observed variables:

$$z = f_Z(w)$$

[4]We use capital letters (e.g., X, Y, U) for variable names and lower case letters (e.g., x, y, u) for values taken by these variables.

[5]A weaker class of causal diagrams, known as "causal Bayesian networks," encode interventional rather than functional dependencies; it can be used to predict outcomes of randomized experiments but not probabilities of counterfactuals (for formal definitions, see Pearl (2000a, pp. 22–24) for formal definition).

$$\begin{aligned} x &= f_X(z, v) \\ y &= f_Y(x, u) \end{aligned} \tag{2}$$

where W, V, and U are here assumed to be jointly independent but, otherwise, arbitrarily distributed.

Remarkably, unknown to most economists and philosophers, structural equation models provide a formal interpretation and symbolic machinery for analyzing counterfactual relationships of the type "Y would be y had X been x in situation $U = u$," denoted $Y_x(u) = y$. Here U stands for the vector of all exogenous variables, and represents all relevant features of an experimental *unit* (i.e., a patient or a subject).

The key idea is to interpret the phrase "had X been x_0" as an instruction to modify the original model M and replace the equation for X by a constant x_0, yielding a modified model, M_{x_0}:

$$\begin{aligned} z &= f_Z(w) \\ x &= x_0 \\ y &= f_Y(x, u) \end{aligned} \tag{3}$$

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

This replacement permits the constant x_0 to differ from the actual value of X—namely, $f_X(z, v)$—without rendering the system of equations inconsistent, thus yielding a formal definition of counterfactuals in multistage models, where the dependent variable in one equation may be an independent variable in another (Balke & Pearl, 1994a, 1994b; Pearl, 2000b). The general definition reads as follows:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{4}$$

In words, the counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the modified submodel M_x, in which the equation for X is replaced by $X = x$. For example, to compute the average causal effect of X on Y, that is, $E(Y_{x_0})$ we solve equation 3 for Y in terms of the exogenous variables, yielding $Y_{x_0} = f_Y(x_0, u)$, and average over U and V. To answer more sophisticated 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.

This formalization of counterfactuals, cast as solutions to modified systems of equations, provides the conceptual and formal link between structural equation models used in economics and social science, the potential-outcome framework, to be discussed later under The Language of Potential Outcomes Lewis (1973) "closest-world" counterfactuals, Woodward's (2003) "interventionalism" approach, Mackie's (1965) "insufficient but necessary components of unnecessary but sufficient" (INUS) condition; and Rothman's (1976) "sufficient component" framework (see VanderWeele and Robins's 2007). The next section discusses two long-standing problems that have been completely resolved in purely graphical terms, without delving into algebraic techniques.

Confounding and Causal Effect Estimation

The central target of most studies in the social and health sciences is the elucidation of cause-effect relationships among variables of interests, for example, treatments, policies, preconditions, and outcomes. While good statisticians have always known that the elucidation of causal relationships from observational studies must rest on assumptions about how the data were generated, the relative roles of assumptions and data and the ways of using those assumptions to eliminate confounding bias have been a subject of much controversy. The preceding structural framework puts these controversies to rest.

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 3; some are affecting the response, some are affecting the treatment,

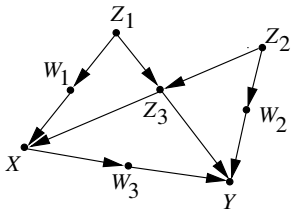

Figure 3: Graphical model illustrating the back-door criterion. Error terms are not shown explicitly.

and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or lifestyle, while others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment so that if we compare treated vs. untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a "sufficient set," "admissible" or a set "appropriate for adjustment." The problem of defining a sufficient set, let alone finding one, has baffled epidemiologists and social scientists for decades (for review, see Greenland, Pearl & Robins, 1999; Pearl, 2000a; 2009a).

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

1. No element of S is a descendant of X.
2. The elements of S "block" all back-door paths from X to Y, that is, all paths that end with an arrow pointing to X.[6]

[6]In this criterion, a set S of nodes is said to *block* a path P if either (i) P contains at least one arrow-emitting node that is in S, or (ii) P contains at least one collision node (e.g., $\rightarrow Z \leftarrow$) that is outside S and has no descendant in S (see Pearl, 2000a, 2009b, pp. 16–17, 335–337).

Based on this criterion we see, for example, that each of the sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$, and $\{W_2, Z_3\}$ is sufficient for adjustment because each blocks all back-door paths between X and Y. The set $\{Z_3\}$, however, is not sufficient for adjustment because 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 implication of finding a sufficient set, S, is that stratifying on S is guaranteed to remove all confounding bias relative to the causal effect of X on Y. In other words, it renders the causal effect of X on Y identifiable, via

$$P(Y = y|do(X = x)) = \sum_s P(Y = y|X = x, S = s)P(S = s) \quad (5)$$

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

The back-door criterion allows us to write equation 5 directly, after selecting a sufficient set, S, from the diagram, without resorting to any algebraic manipulation. 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 & 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, Paz, & Pearl, 1998).

LEFT OFF HERE

General Control of Confounding

Adjusting for covariates is only one of many methods that permit us to estimate causal effects in nonexperimental studies. A much more general identification criterion is provided by the following theorem:

Theorem 1 [*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.[7]

For example, if W_3 is the only observed covariate in the model of Fig. 3, 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 nevertheless 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 $W_3 \rightarrow Y$, each emanating from a measured variable. In this example, the variable W_3 acts as a "mediating instrumental variable" [Pearl, 1993b; Chalak and White, 2006] and yields the estimand:

$$P(Y = y|do(X = x))$$

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

$$= \sum_{w_3} P(W_3 = w|do(X = x))P(Y = y|do(W_3 = w))$$
$$= \sum_{w} P(w|x) \sum_{x'} P(y|w, x')P(x') \quad (6)$$

More recent results extend this theorem by (1) presenting a necessary and sufficient condition for identification [Shpitser and Pearl, 2006], and (2) extending the condition from causal effects to any counterfactual expression [Shpitser and Pearl, 2007]. The corresponding unbiased estimands for these causal quantities, are readable directly from the diagram.

The Language of Potential Outcomes

The elementary object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y_x(u)$, read: "the value that Y would obtain in unit u, had treatment X been x" [Neyman, 1923; Rubin, 1974]. These subscripted variables are treated as undefined quantities, useful for expressing the causal quantities we seek, but are not derived from other quantities in the model. In contrast, in the previous section counterfactual entities were *derived* from a set of meaningful physical processes, each represented by an equation, and *unit* was interpreted a vector u of background factors that characterize an experimental unit. Each structural equation model thus provides a compact representation for a huge number of counterfactual claims, guaranteed to be consistent.

In view of these features, the structural definition of $Y_x(u)$ (Eq. (4)) can be regarded as the formal basis for the potential outcome approach. It interprets the opaque English phrase "the value that Y would obtain in unit u, had X been x" in terms of a scientifically-based mathematical model that allows such values to be computed unambiguously. Consequently, important concepts in potential response analysis that researchers find ill-defined or esoteric often obtain meaningful and natural interpretation in the structural semantics. Examples are: "unit" ("exogenous variables" in structural semantics), "principal stratification" ("equivalence classes" in structural semantics [Balke and Pearl, 1994b] and [Pearl, 2000b] "conditional ignorability" ("back-door condition" in [Pearl, 1993a]"assignment mechanism" ($P(x|$*direct-causes of* $X)$ in structural semantics) and so on. The next two subsections examine how assumptions and inferences are handled in the potential outcome approach vis a vis the graphical-structural approach.

0.1 Formulating Assumptions

The distinct characteristic of the potential outcome approach is that, although its primitive objects are undefined, hypothetical quantities, 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, treating the former as "missing data". In other words, 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 are phrased as queries about the probability distribution of the counterfactual variable of interest, written $P^*(Y_x = y)$. The new hypothetical entities Y_x are treated as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence. Moreover, these hypothetical entities are not entirely whimsy, but are assumed to be connected to observed variables via consistency constraints [Robins, 1986] such as

$$X = x \implies Y_x = Y, \tag{7}$$

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

The main conceptual difference between the two approaches is that, whereas the structural approach views the subscript x as an operation that changes the distribution but keeps the variables the same, the potential-outcome approach views the variable Y_x, to be a different variable, unobserved, loosely connected to Y through relations such as (7).

Pearl [2000a, Chapter 7] shows, using the structural interpretation of $Y_x(u)$, that it is indeed legitimate to treat counterfactuals as jointly distributed random variables in all respects, that consistency constraints like (7) are automatically satisfied in the structural interpretation and, moreover, that investigators need not be concerned about any additional constraints except the following two:[8]

$$Y_{yz} = y \quad \text{for all } y \text{ and } z \tag{8}$$

$$X_z = x \implies Y_{xz} = Y_z \quad \text{for all } x \text{ and } z \tag{9}$$

Eq. (8) ensures that the interventions $do(Y = y)$ results in the condition $Y = y$, regardless of concurrent interventions, say $do(Z = z)$, that are applied to variables other than Y. Equation (9) generalizes (7) to cases where Z is held fixed, at z.

To communicate substantive causal knowledge, the potential-outcome analyst must express causal assumptions as constraints on P^*, usually in the form of conditional independence assertions involving counterfactual variables. In Fig. 2(a) for instance, to communicate the understanding that a treatment assignment (Z) is randomized (hence independent of both U and V), the potential-outcome analyst needs to use the independence constraint $Z \perp\!\!\!\perp \{X_z, Y_x\}$. To further formulate the understanding that Z does not affect Y directly, except through X, the analyst would write a, so called, "exclusion restriction": $Y_{xz} = Y_x$. Clearly, no mortal can judge the validity of such assumptions in any real life problem without resorting to graphs.[9]

[8] This *completeness* result is due to Halpern [1998], who noted that an additional axiom

$$\{Y_{xz} = y\} \ \& \ \{Z_{xy} = z\} \implies Y_x = y$$

must hold in non-recursive models. This fundamental axiom may come to haunt economists and social scientists who blindly apply NR analysis in their fields.

[9] Even with the use of graphs the task is not easy, for example, the reader should try to verify whether $\{Z \perp\!\!\!\perp X_z | Y\}$ holds in the simple model of Fig. 2(a). The answer is given in Pearl [2000a, p. 214].

0.2 Performing Inferences

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

$$Y_x \perp\!\!\!\perp X | Z \tag{10}$$

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

$$\begin{aligned} P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\ &= \sum_z P^*(Y_x = y|x,z)P(z) \quad \text{(using (10))} \\ &= \sum_z P^*(Y = y|x,z)P(z) \quad \text{(using (7))} \\ &= \sum_z P(y|x,z)P(z). \end{aligned} \tag{11}$$

which is the usual covariate-adjustment formula, as in Eq. (5).

Note that almost all mathematical operations in this derivation are conducted within the safe confines of probability calculus. Save for an occasional application of rule (9) or (7)), the analyst may forget that Y_x stands for a counterfactual quantity—it is treated as any other random variable, and the entire derivation follows the course of routine probability exercises.

However, this mathematical illusion comes at the expense of conceptual clarity, especially at a stage where causal assumptions need be formulated. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability Eq. (10), the key to the derivation of Eq. (11), holds in any familiar situation, say in the experimental setup of Fig. 2(a). This assumption reads: "the value that Y would obtain had X been x, is independent of X, given Z" (see footnote 5). Such assumptions of conditional independence among counterfactual variables are not straightforward to comprehend or ascertain, for they are cast in a language far removed from ordinary understanding of cause and effect. When counterfactual variables are not viewed as byproducts of a deeper, process-based model, it is also hard to ascertain whether *all* relevant counterfactual independence judgments have been articulated, whether the judgments articulated are redundant, or whether those judgments are self-consistent.

The need to express, defend, and manage formidable counterfactual relationships of this type explains the slow acceptance of causal analysis among epidemiologists and statisticians, and why economists and social scientists continue to use structural equation models instead of the potential-outcome alternatives advocated in Holland [1988], Angrist et al. [1996], and Sobel [1998].

On the other hand, the algebraic machinery offered by the potential-outcome notation, once a problem is properly formalized, can be powerful in refining assumptions [Angrist et al., 1996], deriving consistent estimands [Robins, 1986], analyzing mediation [Pearl, 2001], bounding probabilities of causation [Tian and Pearl, 2000], and combining data from experimental and nonexperimental studies [Pearl, 2000a, pp. 302–303].

0.3 Combining Graphs and Counterfactuals – The Mediation Formula

Pearl [2000a, p. 232] presents a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams, translating these assumptions into potential outcome notation, performing the mathematics in the algebraic language of counterfactuals and, finally, interpreting the result in plain causal language. Often, the answer desired can be obtained directly from the diagram, and no translation is necessary (as demonstrated in Section).

One area that has benefited substantially from this symbiosis is the analysis of direct and indirect effects, also known as "mediation analysis" [Shrout and Bolger, 2002], which has resisted generalizations to discrete variables and non-linear interactions for several decades [Robins and Greenland, 1992; Mackinnon et al., 2007]. The obstacles were definitional; the direct effect is sensitive to the level at which we condition the intermediate variable, while the indirect effect cannot be defined by conditioning on a third variable, or taking the difference between the total and direct effects.

The structural definition of counterfactuals (Eq. (4)) and the graphical analysis of Section combined to produce formal definitions of, and graphical conditions under which direct and indirect effects can be estimated from data [Pearl, 2001; Petersen et al., 2006]. In particular, under conditions of no unmeasured (or uncontrolled for) confounders, this symbiosis has produced the following "Mediation Formulas" for the expected direct (DE) and indirect (IE) effects of the transition from $X = x$ to $X = x'$ (with outcome Y, and mediating set Z):

$$DE = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \tag{12}$$

$$IE = \sum_z E(Y|x, z)[P(z|x') - P(z|x)] \tag{13}$$

These general formulas are applicable to any type of variables,[10] any nonlinear interactions, any distribution and, moreover, are readily estimable by regression. IE (respectively, DE) represents the average increase in the outcome Y that the transition from $X = x$ to $X = x'$ is expected to produce absent any direct (respectively indirect) effect of X on Y. 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. TE stands for the total effect $TE = E(Y|x') - E(Y|x)$ which, in nonlinear systems may or may not be the sum of the direct and indirect effects.

Additional results spawned by the structural-graphical-counterfactual symbiosis include: effect estimation under non compliance [Balke and Pearl, 1997; Chickering and Pearl, 1997], mediating instrumental variables [Pearl, 1993b; Brito and Pearl, 2006], robustness analysis [Pearl, 2004], selecting predictors for propensity scores [Pearl, 2009c;

[10] Integrals should replace summations when Z is continuous. Generalizations to cases involving observed or unobserved confounders are given in [Pearl, 2001] and exemplified in [Pearl, 2010a, Pearl, 2010b]. Conceptually, IE measures the average change in Y under the operation of setting X to x and, simultaneously, setting Z to whatever value it would have obtained under $X = x'$ [Robins and Greenland, 1992].

2010c], and estimating the effect of treatment on the treated Shpitser and Pearl [2009]. Detailed descriptions of these results are given in the corresponding articles which are available on ⟨http://bayes.cs.ucla.edu/csl_papers.html⟩.

Conclusions

Statistics is strong in devising ways of describing data and inferring distributional parameters from sample. Causal inference require two addition ingredients: a science-friendly language for articulating causal knowledge, and a mathematical machinery for processing that knowledge, combining it with data and drawing new causal conclusions about a phenomena. This paper introduces nonparametric structural equations models as a formal and meaningful language for formulating causal knowledge and for explicating causal concepts used in scientific discourse. These include: randomization, intervention, direct and indirect effects, confounding, counterfactuals, and attribution. The algebraic component of the structural language coincides with the potential-outcome framework, and its graphical component embraces Wright's method of path diagrams (in its nonparametric version.) When unified and synthesized, the two components offer investigators a powerful methodology for empirical research [e.g., Morgan and Winship, 2007; Greenland et al., 1999; Glymour and Greenland, 2008; Chalak and White, 2006; Pearl, 2009a].

Perhaps the most important message of the discussion and methods presented in this paper would be a widespread awareness that (1) all studies concerning causal relations must begin with causal assumptions of some sort and (2) that a friendly and formal language is currently available for articulating such assumptions. This means that scientific articles concerning questions of causation must contain a section in which causal assumptions are articulated using either graphs or subscripted formulas. Authors who wish their assumptions to be understood, scrutinized and discussed by readers and colleagues would do well to use graphs. Authors who refrain from using graphs would be risking a suspicion of attempting to avoid transparency of their working assumptions.

Another important implication of this paper is that every causal inquiry can be mathematized. In other words, mechanical procedures can now be invoked to determine what assumptions investigators must be willing to make in order for desired quantities to be estimable consistently from the data. This is not to say that the needed assumptions would be reasonable, or that the resulting estimation method would be easy. It means that the needed causal assumptions can be made transparent, brought up for discussion and refinement and, once consistency is assured, causal quantities can be estimated from data through ordinary statistical methods, free of the mystical aura that has shrouded causal analysis in the past.

Acknowledgments

This research was supported in parts by grants from NSF IIS-0535223 and NLM T15 LM07356.

References

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

[Balke and Pearl, 1994a] Balke, A. and Pearl, J. (1994a). Counterfactual probabilities: Computational methods, bounds, and applications. In de Mantaras, R. L. and Poole, D., editors, *Uncertainty in Artificial Intelligence 10*, pages 46–54. Morgan Kaufmann, San Mateo, CA.

[Balke and Pearl, 1994b] Balke, A. and Pearl, J. (1994b). Probabilistic evaluation of counterfactual queries. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, volume I, pages 230–237. MIT Press, Menlo Park, CA.

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

[Brito and Pearl, 2006] Brito, C. and Pearl, J. (2006). Graphical condition for identification in recursive SEM. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 47–54. AUAI Press, Corvallis, OR.

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

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

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

[Glymour and Greenland, 2008] Glymour, M. and Greenland, S. (2008). Causal diagrams. In Rothman, K., Greenland, S., and Lash, T., editors, *Modern Epidemiology*, pages 183–209. Lippincott Williams & Wilkins, Philadelphia, PA, 3rd edition.

[Greenland et al., 1999] Greenland, S., Pearl, J., and Robins, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

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

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

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

[Lewis, 1973] Lewis, D. (1973). *Counterfactuals*. Harvard University Press, Cambridge, MA.

[Mackie, 1965] Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly*, 2/4:261–264. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

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

[Morgan and Winship, 2007] 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.

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

[Pearl, 1993a] Pearl, J. (1993a). Comment: Graphical models, causality, and intervention. *Statistical Science*, 8(3):266–269.

[Pearl, 1993b] Pearl, J. (1993b). Mediating instrumental variables. Technical Report TR-210, <http://ftp.cs.ucla.edu/pub/stat_ser/R210.pdf>, Department of Computer Science, University of California, Los Angeles.

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

[Pearl, 2000a] Pearl, J. (2000a). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. Second ed., 2009.

[Pearl, 2000b] Pearl, J. (2000b). Comment on A.P. Dawid's, Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):428–431.

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

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

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

[Pearl, 2009b] Pearl, J. (2009b). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, second edition.

[Pearl, 2009c] Pearl, J. (2009c). 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, 2010a] Pearl, J. (2010a). The foundation of causal inference. Technical Report R-355, University of California, Los Angeles, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r355.pdf>. Forthcoming *Sociological Methodology*.

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

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

[Petersen et al., 2006] Petersen, M., Sinisi, S., and van der Laan, M. (2006). Estimation of direct causal effects. *Epidemiology*, 17(3):276–284.

[Robins, 1986] 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 and Greenland, 1992] Robins, J. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

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

[Rothman, 1976] Rothman, K. (1976). Causes. *American Journal of Epidemiology*, 104:587–592.

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

[Shpitser and Pearl, 2006] Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226. AAAI Press, Menlo Park, CA.

[Shpitser and Pearl, 2007] Shpitser, I. and Pearl, J. (2007). What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 352–359. AUAI Press, Vancouver, BC, Canada. Also, *Journal of Machine Learning Research*, 9:1941–1979, 2008.

[Shpitser and Pearl, 2009] Shpitser, I. and Pearl, J. (2009). Effects of treatment on the treated: Identification and generalization. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Montreal, Quebec.

[Shrout and Bolger, 2002] Shrout, P. and Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7(4):422–445.

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

[Tian et al., 1998] Tian, J., Paz, A., and Pearl, J. (1998). Finding minimal separating sets. Technical Report R-254, University of California, Los Angeles, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r254.pdf>.

[Tian and Pearl, 2000] Tian, J. and Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313.

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

[VanderWeele and Robins, 2007] VanderWeele, T. and Robins, J. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568.

[Woodward, 2003] Woodward, J. (2003). *Making Things Happen*. Oxford University Press, New York, NY.

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