TECHNICAL REPORT
R-346
June 2009

JMLR Workshop and Conference Proceedings 6:39–58 NIPS 2008 workshop on causality

Causal Inference

Judea Pearl JUDEA@CS.UCLA.EDU
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA

Editor: Isabelle Guyon, Dominik Janzing, and Bernhard Schölkopf

Abstract

This paper reviews a theory of causal inference based on the Structural Causal Model (SCM) described in (Pearl, 2000a). The theory unifies the graphical, potential-outcome (Neyman-Rubin), decision analytical, and structural equation approaches to causation, and provides both a mathematical foundation and a friendly calculus for the analysis of causes and counterfactuals. In particular, the paper establishes a methodology for inferring (from a combination of data and assumptions) the answers to three types of causal queries: (1) queries about the effect of potential interventions, (2) queries about counterfactuals, and (3) queries about the direct (or indirect) effect of one event on another.

Keywords: Structural equation models, confounding, graphical methods, counterfactuals, causal effects, potential-outcome.

1. Introduction

The research questions that motivate most quantitative studies in the health, social and behavioral sciences are not statistical but causal in nature. For example, what is the efficacy of a given drug in a given population? Whether data can prove an employer guilty of hiring discrimination? What fraction of past crimes could have been avoided by a given policy? What was the cause of death of a given individual, in a specific incident? These are causal questions because they require some knowledge of the data-generating process; they cannot be computed from the data alone.

Remarkably, although much of the conceptual framework and 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 research community find it hard to appreciate and benefit from the many results that causal analysis has produced in the past two decades. These results rest on advances in three areas:

1. Nonparametric structural equations

2. Graphical models

3. Symbiosis between counterfactual and graphical methods.

This paper aims at making these advances more accessible to the general research community by, first, contrasting causal analysis with standard statistical analysis, second, comparing and unifying existing approaches to causal analysis, and finally, providing a friendly formalism for counterfactual analysis, within which most (if not all) causal questions can be formulated, analyzed and resolved.

We will see that, although full description of the data generating process cannot be inferred from data alone, many useful features of the process can be estimated from a combination of (1) data, (2) prior qualitative knowledge, and/or (3) experiments. Thus, the challenge of causal inference is to answer causal queries of practical interest with minimum number of assumptions and with minimal experimentation. Following an introductory section which defines the demarcation line between associational and causal analysis, the rest of the paper will deal with the estimation of three types of causal queries: (1) queries about the effect of potential interventions, (2) queries about counterfactuals (e.g., whether event x would occur had event y been different), and (3) queries about the direct and indirect effects.

2. From Associational to Causal Analysis: Distinctions and Barriers

2.1 The Basic Distinction: Coping With Change

The aim of standard statistical analysis, typified by regression, 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.[1]

2.2 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

1. The methodology of "causal discovery" (Spirtes, et al. 2000; Pearl 2000a, chapter 2) is likewise based on the causal assumption of "faithfullness" or "stability."

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, odd ratio, marginalization, conditionalization, "controlling for," and so on. Examples of causal concepts are: randomization, influence, effect, confounding, "holding constant," disturbance, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, attribution, and so on. 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 statistical associations alone.

2.3 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." That this definition and all its many variants must fail (Pearl 2000a, Section 6.2)[2] is obvious from the demarcation line above; if confounding were definable in terms of statistical associations, we would have been able to identify confounders from features of nonexperimental data, adjust for those confounders and obtain unbiased estimates of causal effects. This would have violated our golden rule: behind any causal conclusion there must be some causal assumption, untested in observational studies. Hence the definition must be false. Therefore, to the bitter disappointment of generations of epidemiologist and social science 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 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."

2.4 Two Mental Barriers: Untested Assumptions and New Notation

The preceding two requirements: (1) to commence causal analysis with untested,[3] theoretically or judgmentally based assumptions, and (2) to extend the syntax of probability calculus, consti-

2. Any intermediate variable U on a causal path from X to Y satisfies this definition, without confounding the effect of X on Y.

3. By "untested" I mean untested using frequency data in nonexperimental studies.

tute 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 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 it 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, 1988), can recognize such expressions through the subscripts that are attached to counterfactual events and variables, e.g. $Y_x(u)$ or Z_{xy}. (Some authors use parenthetical expressions, e.g. $Y(0)$, $Y(1)$, $Y(x,u)$ 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 $P(Y_x = y)$. 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.[4] Still a third notation that distinguishes causal expressions is provided by graphical models, where the arrows convey causal directionality.[5]

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 governs 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 not be affected by a treatment, a necessary assumption for the control of confounding (Cox, 1958, p. 48), 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, can be traced to the unfriendly and ad-hoc way in which causal analysis has been presented to the research community, resting primarily on the restricted paradigm of controlled randomized trials advanced by Rubin (1974).

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.

4. 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.
5. 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.

3. Structural Causal Models (SCM) and The Language of Diagrams

3.1 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 relationship mathematically was made in the 1920's 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 possibly affect Y. In interpreting this equation one should think of a physical process whereby Nature *examines* the values of x and u and, accordingly, *assigns* variable Y the value $y = \beta x + u$. Similarly, to "explain" the occurrence of disease X, one could write $x = v$, where V stand for all factors affecting X.

To express the directionality inherent in this 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 the value Nature assigns to one variable is not determined by the value taken by another. In Figure 1, for example, the absence of arrow from Y to X represent the claim that symptom Y is not among the factors V which affect disease X.

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).

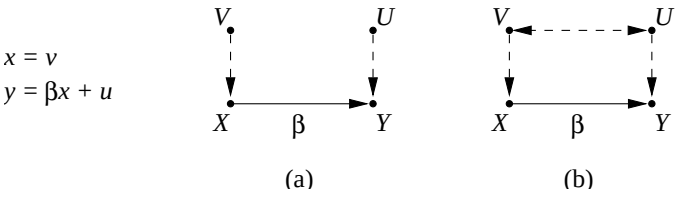

Figure 1: A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

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.

The generalization to nonlinear systems 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:

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

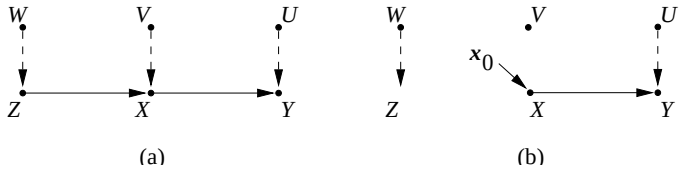

Figure 2: (a) The diagram associated with the structural model of equation (2). (b) The diagram associated with the modified model, M_{x_0}, of equation (3), representing the intervention $do(X = x_0)$.

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

Remarkably, unknown to most economists and pre-2000 philosophers,[6] 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 $\boldsymbol{U}$=$\boldsymbol{u}$," denoted $Y_x(u) = y$. Here $\boldsymbol{U}$ represents the vector of all exogenous variables.[7]

The key idea is to interpret the phrase "had X been x_0" as an instruction to modify the original model and replace the equation for X by a constant x_0, yielding the sub-model.

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

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 interpretation of counterfactuals in multi-stage models, where the dependent variable in one equation may be an independent variable in another (Balke and Pearl, 1994ab; Pearl, 2000b). For example, to compute $E(Y_{x_0})$, the expected effect of *setting* X to x_0, (also called the average causal effect of X on Y, denoted $E(Y|do(x_0))$ or, generically, $E(Y|do(x))$), 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. It is easy to show that in this simple system, the answer can be obtained without knowing the form of the function $f_Y(x,u)$ or the distribution $P(u)$. The answer is given by:

$$E(Y_{x_0}) = E(Y|do(X = x_0) = E(Y|x_0)$$

which is estimable from the observed distribution $P(x,y,z)$. This result hinges on the assumption that W, V, and U are mutually independent and on the topology of the graph (e.g., that there is no direct arrow from Z to Y.)

In general, it can be shown (Pearl 2000a, Chapter 3) that, whenever the graph is Markovian (i.e., acyclic with independent exogenous variables) the post-interventional distribution $P(Y = y|do(X = x))$ is given by the following expression:

$$P(Y = y|do(X = x)) = \sum_t P(y|t,x)P(t) \tag{4}$$

6. Connections between structural equations and a restricted class of counterfactuals were recognized by Simon and Rescher (1966). These were later generalized by Balke and Pearl (1995) who used modified models to permit counterfactual conditioning on dependent variables.

7. Because $U = u$ may contain detailed information about a situation or an individual, $Y_x(u)$ is related to what philosophers called "token causation," while $P(Y_x = y|Z = z)$ characterizes "Type causation," that is, the tendency of X to influence Y in a sub-population characterized by $Z = z$.

where T is the set of direct causes of X (also called "parents") in the graph. Again, we see that all factors on the right hand side are estimable from the distribution P of observed variables and, hence, the counterfactual probability $P(Y_x = y)$ is estimable with mere partial knowledge of the generating process – the topology of the graph and independence of the exogenous variables is all that is needed.

When some variables in the graph (e.g., the parents of X) are unobserved, we may not be able to learn (or "identify" as it is called) the post-intervention distribution $P(y|do(x))$ by simple conditioning, and more sophisticated methods would be required. Likewise, when the query of interest involves several hypothetical worlds simultaneously, e.g., $P(Y_x = y, Y_{x'} = y')$[8], the Markovian assumption may not suffice for identification and additional assumptions, touching on the form of the data-generating functions (e.g., monotonicity) may need to be invoked. These issues will be discussed in Sections 3.2 and 5.

This interpretation 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 and the Neyman-Rubin potential-outcome framework to be discussed in Section 4. But first we discuss two long-standing problems that have been completely resolved in purely graphical terms, without delving into algebraic techniques.

3.2 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 be shaped by assumptions about how the data were generated, the relative roles of assumptions and data, and ways of using those assumptions to eliminate confounding bias have been a subject of much controversy. The structural framework of Section 3.1 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 and some are

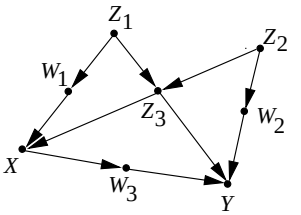

Figure 3: Graphical model illustrating the back-door criterion. Error terms are not shown explicitly.

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

8. Read: The probability that Y would be y if X were x and y' if X were x'.

we compare treated vs. untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of factors is called a "sufficient set" or a set "appropriate for adjustment". The problem of defining a sufficient set, let alone finding one, has baffled epidemiologists and social science for decades (see Greenland et al., 1999; Pearl, 1998, 2003 for review).

The following criterion, named "back-door" in Pearl (1993a), settles this this problem by providing a graphical method of selecting a sufficient set of factors for adjustment. It states that a set S is appropriate for adjustment if two conditions hold:

1. No element of S is a descendant of X
2. The elements of S "block" all "back-door" paths from X to Y, namely all paths that end with an arrow pointing to X.[9]

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

The implication of finding a sufficient set S is that, stratifying on S is guaranteed to remove all confounding bias relative the causal effect of X on Y. In other words, it renders the causal effect of X on Y estimable, via

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

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-outcome framework (Rosenbaum and Rubin, 1983). The criterion also enables the analyst to search for an optimal set of covariate—namely, a set S that minimizes measurement cost or sampling variability (Tian et al., 1998).

GENERAL CONTROL OF CONFOUNDING

Adjusting for covariates is only one of many methods that permits us to estimate causal effects in nonexperimental studies. 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.[10]

For example, if W_3 is the only observed covariate in the model of Figure 3, then there exists no sufficient set for adjustment (because no set of observed covariates can block the paths from

9. A set S of nodes is said to block a path p if either (i) p contains at least one arrow-emitting node that is in S, or (ii) p contains at least one collision node that is outside S and has no descendant in S. See (Pearl, 2000a, pp. 16-7). If S blocks *all* paths from X to Y it is said to "d-separate X and Y."
10. Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of Y.

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 $W_3 \rightarrow Y$, both emanating from a measured variable (W_3). In this example, the variable W_3 acts as a "mediating instrumental variable" (Pearl, 1993b; Chalak and White, 2006) and yields the estimand:

$$\begin{aligned} P(Y = y|do(X = x)) \\ &= \sum_{w_3} P(W_3 = w_3|do(X = x))P(Y = y|do(W_3 = w_3)) \\ &= \sum_{w_3} P(w_3|x) \sum_{x'} P(y|w_3, x')P(x') \end{aligned} \tag{6}$$

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 mathematical derivation of causal effect estimands, like equations (5) and (6) 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. Although the estimands derived in (5) and (6) are non-parametric, this does not mean that one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in (6) can be converted to the product $E(Y|do(x)) = r_{W_3X} r_{YW_3 \cdot X} x$, where $r_{YZ \cdot X}$ is the (standardized) coefficient of Z in the regression of Y on Z and X. More sophisticated estimation techniques can be found in Rosenbaum and Rubin (1983), and Robins (1999). For example, the "propensity score" method of Rosenbaum and Rubin (1983) was found to be quite useful when the dimensionality of the adjusted covariates is high and the data is sparse (See Pearl 2000a, 2nd edition, 2009a, pp. 348–52).

It should be emphasized, however, that contrary to conventional wisdom (e.g., Rubin (2009)), propensity score methods are merely efficient estimators of the right hand side of (5); they cannot be expected to reduce bias in case the set S does not satisfy the back-door criterion (Pearl 2009abc).

3.3 Counterfactual Analysis in Structural Models

Not all questions of causal character can be encoded in $P(y|do(x))$ type expressions, in much the same way that not all causal questions can be answered from experimental studies. For example, questions of attribution (e.g., I took an aspirin and my headache is gone, was it *due* to the aspirin?) or of susceptibility (e.g., I am a healthy non-smoker, would I be as healthy had I been a smoker?) cannot be answered from experimental studies, and naturally, this kind of questions cannot be expressed in $P(y|do(x))$ notation.[11] 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 $\boldsymbol{U}$=$\boldsymbol{u}$," denoted $Y_x(\boldsymbol{u}) = y$.

As noted in Section 3.1, the structural definition of counterfactuals involves modified models, like M_{x_0} of equation (3), formed by the intervention $do(X = x_0)$ (Figure 2(b)). Call the solution of Y in model M_x the *potential response* of Y to x, and denote it by the symbol $Y_x(\boldsymbol{u})$.

11. 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.

In general, then, the formal definition of the counterfactual $Y_x(\boldsymbol{u})$ in SCM is given by (Pearl 2000a, p. 98):

$$Y_x(\boldsymbol{u}) = Y_{M_x}(\boldsymbol{u}).$$

The quantity $Y_x(\boldsymbol{u})$ can be given experimental interpretation; it stands for the way an individual with characteristics $(\boldsymbol{u})$ would respond, had the treatment been x, rather than the treatment $x = f_X(\boldsymbol{u})$ actually received by that individual. In our example, since Y does not depend on v and w, we can write:

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

Clearly, the distribution $P(u,v,w)$ induces a well defined probability on the counterfactual event $Y_{x_0} = y$, as well as on joint counterfactual events, such as '$Y_{x_0} = y$ AND $Y_{x_1} = y'$,' which are, in principle, unobservable if $x_0 \neq x_1$. Thus, to answer attributional questions, such as whether Y would be y_1 if X were x_1, given that in fact Y is y_0 and X is x_0, we need to compute the conditional probability $P(Y_{x_1} = y_1 | Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model. For example, assuming linear equations (as in Figure 1),

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

the conditions $Y = y_0$ and $X = x_0$ yield $v = x_0$ and $u = y_0 - \beta x_0$, and we can conclude that, with probability one, Y_{x_1} must take on the value: $Y_{x_1} = \beta x_1 + u = \beta(x_1 - x_0) + y_0$. In other words, if X were x_1 instead of x_0, Y would increase by β times the difference $(x_1 - x_0)$. In nonlinear systems, the result would also depend on the distribution of U and, for that reason, attributional queries are generally not identifiable in nonparametric models (Pearl, 2000a, Chapter 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 $X = x'$."[12] 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 structural interpretation of counterfactuals also 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 starting paradigm, assuming that nothing is known to the experimenter about the science behind the data. This "black-box" approach, which 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 2.4. The next section 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.

4. The Language of Potential Outcomes and Counterfactuals

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y_x(u)$, read: "the value that outcome Y would obtain in experimental unit u, had treatment X been x" (Neyman, 1923; Rubin, 1974). Here, *unit* may stand for an individual patient, an experimental subject, or an agricultural plot. In Section 3.3 we saw that this counterfactual entity has the natural interpretation as representing the solution for Y in a modified

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

system of equations, where *unit* is interpreted a vector $\boldsymbol{u}$ 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 some model. In this sense the structural interpretation of $Y_x(u)$ 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" could be realized, and what the logical consequence are of such a condition.

4.1 The "Black-Box" or "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

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

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

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

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:

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

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{9}$$

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

4.2 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 at hand. 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 our example of Figure 2(a), to communicate the understanding that the (Z) is randomized (hence independent of V and U), the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{X_z, Y_x\}$.[13] 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; 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

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

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 (5).

We see that the assumption of conditional ignorability (10) qualifies Z as a sufficient covariate for adjustment; it is entailed indeed by the "back-door" criterion of Section 3.2, which qualifies such covariates by tracing paths in the causal diagram.

The derivation above may explain why the potential-outcome approach appeals to mathematical statisticians; instead of constructing new vocabulary (e.g., arrows), new operators $(do(x))$ and new logic for causal analysis, almost all mathematical operations in this framework are conducted within the safe confines of probability calculus. Save for an occasional application of rule (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.

13. 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).

However, this mathematical orthodoxy exacts a very high cost: 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 (10), the key to the derivation of (11), 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 3 would satisfy this conditional independence condition.[14] Likewise, to derive equation (6) in the language of potential-outcome (see Pearl 2000a, page 233), one 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 graph-less potential-outcome camp rarely use "conditional ignorability" (10) 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.[15]

Replacing "ignorability" with a simple 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 whether *all* relevant counterfactual independence judgments have been articulated, whether the judgments articulated are *redundant*, or whether those judgments are *self-consistent.* The need to express, defend, and manage formidable counterfactual relationships of this type explain the slow acceptance of causal analysis among health scientists and statisticians, and why economists and social scientists continue to use structural equation models instead of the potential-outcome alternatives advocated in Angrist et al. (1996); Holland (1988); Sobel (1998).

On the other hand, the algebraic machinery offered by the counterfactual notation, $Y_x(u)$, once a problem is properly formalized, can be extremely powerful in refining assumptions (Angrist et al., 1996), deriving consistent estimands (Robins, 1986), bounding probabilities of necessary and sufficient causation (Tian and Pearl, 2000), and combining data from experimental and nonexperimental studies (Pearl, 2000a). 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 counterfactual notation, performing the mathematics in the algebraic language of counterfactuals (using (7), (8), and (9)) and, finally, interpreting the result in plain causal language. The next section illustrates such symbiosis.

14. Inquisitive readers are invited to guess whether $X_z \perp\!\!\!\perp Z|Y$ holds in Figure 2(a).

15. 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 3) may actually *increase* confounding bias. Paul Rosenbaum, for example, writes: "there is no reason to avoid adjustment for a variable describing subjects before treatment" Rosenbaum (2002), p. 76. Don 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 2009bc for a discussion of this fallacy).

5. Mediation: Direct and Indirect Effects

5.1 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.

Another example concerns the identification of neural pathways in the brain or the structural features of protein-signaling networks in molecular biology (Brent and Lok, 2005). Here, the decomposition of effects into their direct and indirect components carries theoretical scientific importance, for it predicts behavior under a rich variety of hypothetical interventions.

In all such examples, the requirement of holding the mediating variables fixed must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, conditioning, or adjustment. For example, it will not be sufficient to measure the association between gender (X) and hiring (Y) for a given level of qualification Z, because, by conditioning on the mediator Z, we may create spurious associations between X and Y even when there is no direct effect of X on Y. This can easily be illustrated in the model $X \rightarrow Z \leftarrow U \rightarrow Y$, where X has no direct effect on Y. Physically holding Z constant would permit no association between X and Y, as can be seen by deleting all arrows entering Z. But if we were to condition on Z, a spurious association would be created through U (unobserved) that might be construed as a direct effect of X on Y.

Using the $do(x)$ notation, and focusing on expectations, this leads to a simple definition of *controlled direct effect*:

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

or, equivalently, using counterfactual notation:

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

where Z is any set of mediating variables that intercept all indirect paths between X and Y. Graphical identification conditions for expressions of the type $E(Y|do(x),do(z_1),do(z_2),\ldots,do(z_k))$ were derived by Pearl and Robins (1995) (see Pearl 2000a, Chapter 4) and invoke sequential application of the back-door conditions discussed in Section 3.2.

5.2 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 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 meaningful to average the direct effect over those levels. Conceptually, we can define the average direct effect $DE_{x,x'}(Y)$ 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 (1991) called "pure" and Pearl (2001) called "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)).

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

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \tag{12}$$

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, even with the help of ideal, controlled experiments (see footnote 11 for intuitive explanation). Pearl (2001) has nevertheless shown that, if certain assumptions of "no confounding" are deemed valid,[16] the natural direct effect can be reduced to

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

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.

In particular, expression (13) is both valid and identifiable in Markovian models, where each term on the right can be reduced to a "do-free" expression using equation (4).

5.3 Natural Indirect Effects

Remarkably, the definition of the natural direct effect (12) can easily be turned around and provide an operational definition for the *indirect effect* – a concept shrouded in mystery and controversy, because it is impossible, using the $do(x)$ operator, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

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 (Pearl, 2001):

$$IE_{x,x'}(Y) \triangleq E[(Y_{x,Z_{x'}}) - E(Y_x)], \tag{14}$$

which is almost identical to the direct effect (equation (12)) save for exchanging x and x'.

16. One sufficient condition is that $Z_x \perp\!\!\!\perp Y_{x',z} | W$ holds for some set W of measured covariates. See details and graphical criteria in Pearl (2001, 2005) and in Petersen et al. (2006).

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,

$$TE_{x,x'}(Y) \stackrel{\Delta}{=} E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{15}$$

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

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{16}$$

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.

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

Note that in all these cases, the policy intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Pearl (2001) has suggested therefore 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 experimental setup. The mantra "No causation without manipulation" must be rejected. (See Pearl 2000a, Section 11.4.5, 2nd Ed.)

It is remarkable that counterfactual quantities like DE and ID that could not be expressed in terms of $do(x)$ operators, and appear therefore void of empirical content, can, under certain conditions be estimated from empirical studies. A general characterization of those conditions is given in Shpitser and Pearl (2007).

Additional examples of this "marvel of formal analysis" are given in (Pearl, 2000a, Chapters 7, 9, 11). It constitutes an unassailable argument in defense of counterfactual analysis, as expressed in Pearl (2000b) against the stance of Dawid (2000) and Geneletti (2007).

6. 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 phenomenon. This paper introduces nonparametric structural causal models (SCM) 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 empirical investigators a powerful methodology for causal inference which resolves long-standing problems in the empirical sciences. These include the control of confounding, the evaluation of policies, the analysis of mediation and the algorithmization of counterfactuals.

Acknowledgments

Portions of this paper are based on my book *Causality* (Pearl, 2000, 2nd edition forthcoming 2009a). This research was supported in parts by grants from NSF #IIS-0535223 and ONR #N000-14-09-1-0665.

References

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

C. Avin, I. Shpitser, and J. Pearl. Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence IJCAI-05*, pages 357–363, Edinburgh, UK, 2005. Morgan-Kaufmann Publishers.

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

A. Balke and J. Pearl. 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, 1994b.

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

R. Brent and L. Lok. A fishing buddy for hypothesis generators. *Science*, 308(5721):523–529, 2005.

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

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

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

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

S. Geneletti. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society, Series B (Methodological)*, 69(2):199–215, 2007.

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

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

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

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

J. Pearl. Mediating instrumental variables. Technical Report Technical Report R-210, Computer Science Department, UCLA, 1993b. http://ftp.cs.ucla.edu/pub/stat_ser/r210.pdf.

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

J. Pearl. Graphs, causality, and structural equation models. *Sociological Methods and Research*, 27(2):226–284, 1998.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000a. Second Edition forthcoming 2009.

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

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

J. Pearl. Statistics and causal inference: A review. *Test Journal*, 12(2):281–345, December 2003.

J. Pearl. Direct and indirect effects. In *Proceedings of the American Statistical Association, Joint Statistical Meetings*, pages 1572–1581. MIRA Digital Publishing, Minn., MN, 2005.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, second edition, 2009a. Forthcoming.

J. Pearl. Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine*, 28:1420–1423, 2009b. http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf.

J. Pearl. Myth, confusion, and science in causal analysis. Technical Report R-348, University of California, Los Angeles, CA, 2009c. http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf.

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

M.L. Petersen, S.E. Sinisi, and M.J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.

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

J.M. Robins. Testing and estimation of direct effects by reparameterizing directed acyclic with structural nested models. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, pages 349–405. AAAI Press/The MIT Press, Menlo Park, CA, 1999.

J.M. Robins and S. Greenland. Estimability and estimation of expected years of life lost due to a hazardous exposure. *Statistics in Medicine*, 10:79–93, 1991.

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

P.R. Rosenbaum. *Observational Studies*. Springer-Verlag, New York, second edition, 2002.

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

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

I. Shpitser and J Pearl. Identification of conditional interventional distributions. In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444. AUAI Press, Corvallis, OR, 2006.

I. Shpitser and J Pearl. 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, 2007.

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

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

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

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

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

J. Tian and J. Pearl. 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, 2002.

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