

The Science and Ethics of Causal Modeling

Judea Pearl

University of California, Los Angeles

Computer Science Department

Los Angeles, CA, 90095-1596, USA

judaea@cs.ucla.edu

September 18, 2009

Abstract

The intrinsic schism between causal and associational relations presents profound ethical and methodological problems to researchers in the social and behavioral sciences, ranging from the statement of a problem, to the implementation of a study, to the reporting of finding. This paper describes a causal modeling framework that mitigates these problems and offers a simple, yet formal and principled methodology for empirical research. The framework is based on the Structural Causal Model (SCM) described in [Pearl, 2000a] – a nonparametric extension of structural equation models that provides 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) answers to three types of causal queries: (1) queries about the effects of potential interventions, (also called “causal effects” or “policy evaluation”), (2) queries about probabilities of counterfactuals, (including assessment of “regret,” “attribution,” or “causes of effects”), and (3) queries about direct and indirect effects (also known as “mediation” or “effect decomposition”). Finally, the paper defines the formal and conceptual relationships between the structural and potential-outcome frameworks

and demonstrates a symbiotic analysis that uses the strong features of both.

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

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 treatment or program 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.

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, a profound tension exists between the scientific questions that a researcher wishes to ask and the type of questions traditional analysis can accommodate, let alone answer. Bluntly, scientists speak causation and statistics delivers correlation. This tension has resulted in several ethical issues concerning the statement of a problem, the implementation of a study, and the reporting of finding. This paper describes a simple causal extension to the language of statistics, and shows how it leads to a coherent methodology that avoids the ethical problems mentioned, and permits researchers to benefit from the many results that causal analysis has produced in the past two decades.

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 probabilities of past and future events, as well as update probabilities 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 probabilities 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 whether 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 those 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 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, “Granger causality,” 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, mediation, 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.

¹The methodology of “causal discovery” (Spirtes et al. 2000; Pearl 2000a, Chapter 2) is likewise based on the causal assumption of “faithfulness” or “stability,” but will not be discussed in this paper.

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 $P(disease|symptom)$

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

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 in order to articulate such assumptions, constitute the two main sources of confusion in the ethics of formulating, conducting, and reporting empirical studies.

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 that representation that determine the reliability of the judgments upon which the analysis

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

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, may stand 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 statistical parameters (e.g., regression coefficients), 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],

⁴Clearly, $P(Y = y|do(X = x))$ 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 .

⁵These 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.

is expressed in plain English, not in a mathematical expression.

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

3 Structural causal models, diagrams, causal effects, and counterfactuals

3.1 Structural equations as oracles for causes 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 $y = \beta x + u$.

To express the directionality inherent in this assignment 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 indifferent to that taken by another. (See Fig. 1.)

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 Fig. 1(b).

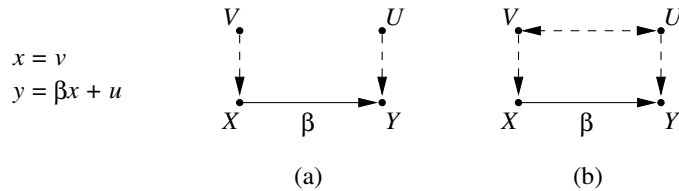


Figure 1: A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.

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 $Cov(U, V)=0$.

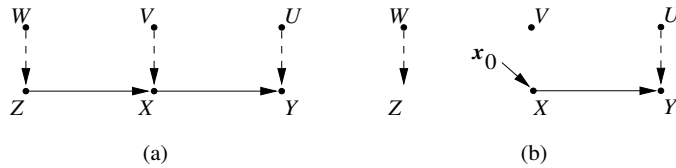


Figure 2: (a) The diagram associated with the structural model of Eq. (2). (b) The diagram associated with the modified model, M_{x_0} , of Eq. (3), representing the intervention $do(X = x_0)$.

The generalization to nonlinear system of equations is straightforward. For example, the non-parametric interpretation of the diagram of Fig. 2(a) corresponds to a set of three functions,

each corresponding to one of the observed variables:

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

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

Remarkably, unknown to most economists and 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 $\mathbf{U}=\mathbf{u}$,” denoted $Y_x(u) = y$. Here \mathbf{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, 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 Fig. 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

⁶Connections 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.

⁷Because $U = u$ 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 $Z = z$.

independent variable in another [Balke and Pearl, 1994, Pearl, 2000b]. In general, we can formally define the post-intervention distribution by the equation:

$$P_M(y|do(x)) \triangleq P_{M_x}(y) \quad (4)$$

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

From this distribution, one is able to assess treatment efficacy by comparing aspects of this distribution at different levels of x_0 . A common measure of treatment efficacy is the difference

$$E(Y|do(x'_0)) - E(Y|do(x_0)) \quad (5)$$

where x'_0 and x_0 are two levels (or types) of treatment selected for comparison. 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 Eq. (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:

$$P(Y = y|do(X = x)) = \sum_t P(y|t, x)P(t) \quad (6)$$

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 variable 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 estimate (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., $P(Y_x = y|Y_{x'} = y')$, 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 3.4. 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

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.⁸ The structural framework of Section

⁸A recent flare-up of this controversy can be found in Pearl [2009a,b] and Rubin [2009] which demonstrates the difficulties statisticians encounter in articulating causal assumptions and typical mistakes that arise from pursuing

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 Fig. 3; some are affecting the response, some are affecting the treatment and some are affecting

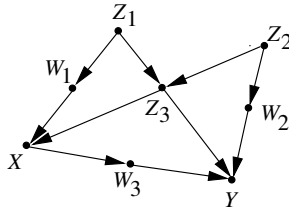


Figure 3: Graphical model illustrating the back-door criterion. Error terms are not shown explicitly.

both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or life style, others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment, namely, that if we compare treated vs. untreated subjects having the same values of the selected factors, we get the correct treatment effect in that subpopulation of subjects. Such a set of “deconfounding” factors is called a “sufficient set” or a set “admissible 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 [1993], settles this problem by providing a graphical method of selecting a sufficient set of factors for adjustment. It states that a set S is admissible for adjustment if two conditions hold:

1. No element of S is a descendant of X

causal analysis within the statistical paradigm of clinical trials or “missing data.”

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 in footnote 9, 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

$$\begin{aligned} P(Y = y|do(X = x)) \\ = \sum_s P(Y = y|X = x, S = s)P(S = s) \end{aligned} \tag{7}$$

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 Eq. (7) 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]. A complete identification condition, including models with no sufficient sets (e.g., Fig. 3, assuming that X, Y , and W_3 are the only measured

⁹In this criterion, a set S of nodes is said to *block* a path p if either (i) p contains at least one arrow-emitting node that is in S , or (ii) p contains at least one collision node that is outside S and has no descendant in S . See [Pearl, 2000a, pp. 16–7, 335–7].

variables) are given in [Shpitser and Pearl, 2006].

Another problem that has a simple graphical solution is to determine whether adjustment for two sets of covariates would result in the same confounding bias [Pearl and Paz, 2009]. This criterion allows one to assess, prior to taking any measurement, whether two candidate sets of covariates, differing substantially in dimensionality, measurement error, cost, or sample variability are equally valuable in their bias-reduction potential.

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.¹⁰ 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 $\mathbf{U}=\mathbf{u}$,” denoted $Y_x(\mathbf{u}) = y$.

As noted in Section 3.1, the structural definition of counterfactuals involves modified models, like M_{x_0} of Eq. (3), formed by the intervention $do(X = x_0)$ (Fig. 2(b)). Denote the solution of Y in model M_x by the symbol $Y_{M_x}(\mathbf{u})$, the formal definition of the counterfactual $Y_x(\mathbf{u})$ in SCM is

¹⁰The 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.

given by [Pearl, 2000a, p. 98]:

$$Y_x(\mathbf{u}) \triangleq Y_{M_x}(\mathbf{u}). \quad (8)$$

The quantity $Y_x(\mathbf{u})$ can be given experimental interpretation; it stands for the way an individual with characteristics (\mathbf{u}) would respond, had the treatment been x , rather than the treatment $x = f_X(\mathbf{u})$ actually received by that individual. In our example, since Y does not depend on v and w , we can write: $Y_{x_0}(\mathbf{u}) = f_Y(x_0, u)$. Clearly, the distribution $P(u, v, w)$ induces a well defined probability on the counterfactual event $Y_{x_0} = y$, as well as on joint counterfactual events, such as ‘ $Y_{x_0} = y$ AND $Y_{x_1} = y'$,’ which are, in principle, unobservable if $x_0 \neq x_1$. Thus, to answer attributional questions, such as whether Y would be y_1 if X were x_1 , given that in fact Y is y_0 and X is x_0 , we need to compute the conditional probability $P(Y_{x_1} = y_1 | Y = y_0, X = x_0)$ which is well defined once we know the forms of the structural equations and the distribution of the exogenous variables in the model. For example, assuming a linear equation for Y (as in Fig. 1),

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

the conditions $Y = y_0$ and $X = x_0$ yield $V = x_0$ and $U = y_0 - \beta x_0$, and we can conclude that, with probability one, Y_{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 $X = x'$.” 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, 2000a], since the

contradictory joint statement is mapped into an ordinary event (among the background variables) that satisfies both statements simultaneously, each in its own distinct submodel; such events have well defined probabilities.

The structural interpretation of counterfactuals (8) also provides the conceptual and formal basis for the Neyman-Rubin potential-outcome framework, an approach 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 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.

3.4 Relation to potential outcomes and the demystification of “ignorability”

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 (Eq. (8)) that this counterfactual entity has a natural interpretation in structural equations as the solution for Y in a modified system of equation, where *unit* is interpreted a vector \mathbf{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.

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 treating the new hypothetical entities Y_x 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

$$X = x \implies Y_x = Y, \tag{9}$$

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 (9), 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 $Y_x(u)$ (Eq. (8)), that it is indeed legitimate to treat counterfactuals as jointly distributed random variables in all respects, that consistency constraints like (9) are automatically satisfied in the structural interpretation and, moreover, that investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \quad \text{for all } y, \text{ subsets } Z, \text{ and values } z \text{ for } Z \quad (10)$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x, \text{ subsets } Z, \text{ and values } z \text{ for } Z \quad (11)$$

Equation (10) 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 (11) generalizes (9) to cases where Z is held fixed, at z .

3.5 Problem Formulation and the Demystification of “Ignorability”

The main drawback of this black-box approach surfaces in 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 causal assumptions in the form of assertions involving counterfactual variables. For instance, in our example of Fig. 2(a)), to communicate the understanding that Z is randomized (hence independent of V and U), the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{X_z, Y_x\}$.¹¹

To further formulate the understanding that Z does not affect Y directly, except through X , the

¹¹The 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].

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 relation

$$Y_x \perp\!\!\!\perp X | Z \tag{12}$$

(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

$$\begin{aligned} P(Y_x = y) &= \sum_z P(Y_x = y|z)P(z) \\ &= \sum_z P(Y_x = y|x, z)P(z) \quad (\text{using (12)}) \\ &= \sum_z P(Y = y|x, z)P(z) \quad (\text{using (9)}) \\ &= \sum_z P(y|x, z)P(z). \end{aligned} \tag{13}$$

The last expression contains no counterfactual quantities and coincides precisely with the standard covariate-adjustment formula of Eq. (7).

We see that the assumption of conditional ignorability (12) qualifies Z as a sufficient covariate for adjustment; indeed, one can show formally [Pearl, 2000a, pp. 98–102, 341–43] that (12) is entailed by the “back-door” criterion of Section 3.2.

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 (11) or (9), 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 orthodoxy exacts a very high cost at the inevitable stage where causal assumptions are formulated. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability (12), the key to the derivation of (13), 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 ”. Even the most experienced potential-outcome expert would be unable to discern whether any subset Z of covariates in Fig. 3 would satisfy this conditional independence condition.¹² Likewise, to convey the structure of the chain $X \rightarrow W_3 \rightarrow Y$ (Fig. 3) in the language of potential-outcome, one would need to write 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 cause and effect 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” (12) 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.¹³

Having translated “ignorability” into 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

¹²Inquisitive readers are invited to guess whether $X_z \perp\!\!\!\perp Z|Y$ holds in Fig. 2(a).

¹³The 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 Fig. 3) may actually *increase* confounding bias. Paul Rosenbaum, for example, writes: “there is little or no reason to avoid adjustment for a variable describing subjects before treatment” [Rosenbaum, 2002, p. 76]. 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 [2009a,b] for a discussion of this fallacy).

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 formulated, 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 (9), (10), and (11)) and, finally, interpreting the result in plain causal language. Section 5 illustrates such symbiosis.

4 Methodological Dictates and Ethical Considerations

The structural theory described in the previous sections dictates a principled methodology that eliminates the confusion between causal and statistical interpretations of study results as well as the ethical dilemmas that this confusion tends to spawn. The methodology dictates that every

investigation involving causal relationships (and this entails the vast majority of empirical studies in the social and behavioral sciences) should be structured along the following four-step process:

1. **Define:** Express the target quantity Q as a function $Q(M)$ that can be computed from any model M .
2. **Assume:** Formulate causal assumptions using ordinary scientific language and represent their structural part in graphical form.
3. **Identify:** Determine if the target quantity is identifiable (i.e., expressible as distributions).
4. **Estimate:** Estimate the target quantity if it is identifiable, or approximate it, if it is not.

4.1 Defining the target quantity

The definitional phase is the most neglected step in current practice of quantitative analysis (Section 5). The structural modeling approach insists on defining the target quantity, be it “causal effect,” “mediated effect,” “effect on the treated,” or “probability of causation” before specifying any aspect of the model, without making functional or distributional assumptions and prior to choosing a method of estimation.

The investigator should view this definition as an *algorithm* that receives a model M as an input and delivers the desired quantity $Q(M)$ as the output. Surely, such algorithm should not be tailored to any aspect of the input M ; it should be general, and ready to accommodate any conceivable model M whatsoever. Moreover, the investigator should imagine that the input M is a completely specified model, with all the functions f_X, f_Y, \dots and all the U variables (or their associated probabilities) given precisely. This is the hardest step for statistically trained investigators to make; knowing in advance that such model details will never be estimable from the data, the definition of $Q(M)$ appears like a futile exercise in fantasy land – it is not.

For example, the formal definition of the causal effect $P(y|do(x))$, as given in Eq. (4), is universally applicable to all models, parametric as well as nonparametric, through the formation of a submodel M_x . By defining causal effect procedurally, thus divorcing it from its traditional parametric representation, the structural theory avoids the many pitfalls and confusions that have plagued the interpretation of structural and regressional parameters for the past half century.¹⁴

4.2 Explicating Causal Assumptions

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

Structural equation models, in their counterfactual reading, have settled this difficulty. Today we understand that the versatility and natural appeal of structural equations stem from the fact that they permit investigators to communicate causal assumptions formally and in the very same vocabulary that scientific knowledge is stored.

Unfortunately, however, this understanding is not shared by all causal analysts; some analysts vehemently resist the resurrection of structural models and insist, instead, on articulating causal assumptions exclusively in the unnatural (though formally equivalent) language of potential outcomes, ignorability, treatment assignment, and other metaphors borrowed from clinical trials. This assault on structural modeling is perhaps more dangerous than the causal-associational

¹⁴Note that β in Eq. (1), the incremental causal effect of X on Y , is defined procedurally by

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

Naturally, all attempts to give β statistical interpretation have ended in frustrations [Holland, 1988, Whittaker, 1990, Wermuth, 1992, Wermuth and Cox, 1993], some persisting well into the 21st century [Sobel, 2008].

confusion, because it is riding on a halo of exclusive ownership to scientific principles and, instead of prohibiting causation, secludes it away from its natural habitat.

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

The methodology expounded in this paper testifies against such restrictions. It demonstrates a viable and principled formalism based on traditional structural equations paradigm, which stands diametrically opposed to the "missing data" paradigm. It renders the vocabulary of "treatment assignment" stifling and irrelevant (e.g., there is no "treatment assignment" in sex discrimination cases). Most importantly, it strongly prefers the use of "uncorrelated error terms," (or "omitted factors") over its "strong ignorability" alternative, which even experts admit cannot be used (and has not been used) to reason about underlying assumptions.

In short, APA's guidelines should be vastly more inclusive, and borrow strength from multiple approaches. The next section demonstrates the benefit of a symbiotic, graphical-structural-counterfactual approach to deal with the problem of mediation, or effect decomposition.

5 An Example: 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.

From a policy making viewpoint, an investigator may be interested in decomposing effects to quantify the extent to which racial salary disparity is due to educational disparity. 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 tells us “how nature works” and, therefore, enables us to predict behavior under a rich variety of conditions.

Yet despite its ubiquity, the analysis of mediation has long been a thorny issue in the social and behavioral sciences [Judd and Kenny, 1981, Baron and Kenny, 1986, Muller et al., 2005, Shrout

and Bolger, 2002, MacKinnon et al., 2007a] primarily because structural equation modeling in those sciences were deeply entrenched in linear analysis, where the distinction between causal and regressional parameters can easily be conflated. As demands grew to tackle problems involving binary and categorical variables, researchers could no longer define direct and indirect effects in terms of structural or regressional coefficients, and all attempts to extend the linear paradigms of effect decomposition to non-linear systems produced distorted results [MacKinnon et al., 2007b]. These difficulties have accentuated the need to redefine and derive causal effects from first principles, uncommitted to distributional assumptions or a particular parametric form of the equations. The structural methodology presented in this paper adheres to this philosophy and it has produced indeed a principled solution to the mediation problem, based on the counterfactual reading of structural equations (8). The following subsections summarize the method and its solution.

5.2 Controlled Direct-Effects

A major impediment to progress in mediation analysis has been the lack of notational facility for expressing the key notion of “holding the mediating variables fixed” in the definition of direct effect. Clearly, this notion must be interpreted as (hypothetically) setting the intermediate variables to constants by physical intervention, not by analytical means such as selection, conditioning, matching or adjustment. For example, consider the simple mediation models of Fig. 4, where the error terms (not shown explicitly) are assumed to be independent. It will not be sufficient to measure the association between gender (X) and hiring (Y) for a given level of qualification (Z), (see Fig. 4(b)) because, by conditioning on the mediator Z , we create spurious associations between X and Y through W_2 , even when there is no direct effect of X on Y [Pearl, 1998].

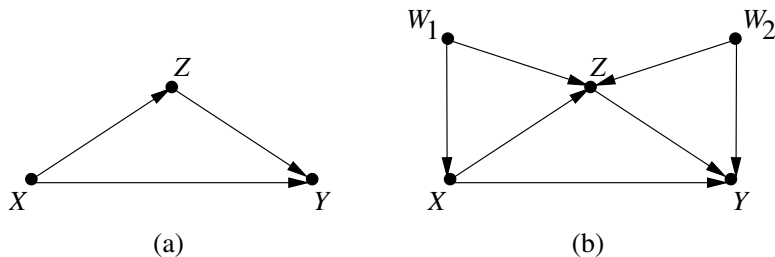


Figure 4: (a) A generic model depicting mediation through Z with no confounders (b) A mediation model with two confounders, W_1 and W_2 .

Using the $do(x)$ notation, we avoid this confusion and obtain a simple definition of the *controlled direct effect* of the transition from $X = x$ to $X = x'$:

$$CDE \triangleq E(Y|do(x), do(z)) - E(Y|do(x'), do(z))$$

or, equivalently, using counterfactual notation:

$$CDE \triangleq E(Y_{xz}) - E(Y_{x'z})$$

where Z is the set of all mediating variables. The readers can easily verify that, in linear systems, the controlled direct effect reduces to the path coefficient of the link $X \rightarrow Y$ (see footnote 14) regardless of whether confounders are present (as in Fig. 4(b)) and regardless of whether the error terms are correlated or not.

This separates the task of definition from that of identification, as demanded by Section 4.1. The identification of CDE would depend, of course, on whether confounders are present and whether they can be neutralized by adjustment, but these do not alter its definition. Graphical identification conditions for expressions of the type $E(Y|do(x), do(z_1), do(z_2), \dots, do(z_k))$ in the presence of unmeasured confounders 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.

5.3 Natural Direct Effects

In linear systems, the direct effect is fully specified by the path coefficient attached to the link from X to Y ; therefore, the direct effect is independent of the values at which we hold Z . In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, it is not uncommon to find employers who prefer males for the high-paying jobs (i.e., high z) and females for low-paying jobs (low z).

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

Extending the subscript notation to express nested counterfactuals, Pearl [2001] gave a formal definition for the “natural direct effect”:

$$DE_{x,x'}(Y) = E(Y_{x',Z_x}) - E(Y_x). \quad (14)$$

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 10 for intuitive explanation). However, aided by Eq. (8) and the notational power of nested counterfactuals, Pearl [2001] was nevertheless able to show that, if certain assumptions of “no confounding” are deemed valid, the natural direct effect can be reduced to

$$DE_{x,x'}(Y) = \sum_z [E(Y|do(x', z)) - E(Y|do(x, z))]P(z|do(x)). \quad (15)$$

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

One condition for the validity of (15) is that $Z_x \perp\!\!\!\perp Y_{x',z} | W$ holds for some set W of measured covariates. This technical condition in itself, like the ignorability condition of (12), is close to meaningless for most investigators, as it is not phrased in terms of realized variables. The structural interpretation of counterfactuals (8) can be invoked at this point to unveil the graphical interpretation of this condition. It states that W should be admissible (i.e., satisfy the back-door condition) relative the path(s) from Z to Y . This condition, satisfied by W_2 in Fig. 4(b), is readily comprehended by empirical researchers, and the task of selecting such measurements, W , can then be guided by the available scientific knowledge. Additional graphical and counterfactual conditions for identification are derived in Pearl [2001] Petersen et al. [2006] and Imai et al. [2008].

In particular, it can be shown (Pearl 2001) that expression (15) is both valid and identifiable in Markovian models (i.e., no unobserved confounders) where each term on the right can be reduced to a “*do-free*” expression using Eq. (6) and then estimated by regression.

For example, for the model in Fig. 4(b), Eq. (15) reads:

$$DE_{x,x'}(Y) = \sum_z \sum_{w_1} P(w_1)[E(Y|x', z, w_1) - E(Y|x, z, w_1)] \sum_{w_2} P(z|x, w_2)P(w_2). \quad (16)$$

while for the confounding-free model of Fig. 4(a) we have:

$$DE_{x,x'}(Y) = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \quad (17)$$

Both (16) and (17) can easily be estimated by a two-step regression.

5.4 Natural Indirect Effects

Remarkably, the definition of the natural direct effect (14) can be turned around and provide an operational definition for the *indirect effect* – a concept shrouded in mystery and controversy, because it is impossible, using 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]:

$$IE_{x,x'}(Y) \triangleq E[(Y_{x,Z_{x'}}) - E(Y_x)], \quad (18)$$

which is almost identical to the direct effect (Eq. (14)) save for exchanging x and x' in the first term.

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

$$TE_{x,x'}(Y) \triangleq E(Y_x - Y_{x'}) = DE_{x,x'}(Y) - IE_{x',x}(Y). \quad (19)$$

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

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \quad (20)$$

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].

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 IE 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, and serve to guide policies. Awareness of this potential should embolden researchers to go through the definitional step of the study and freely

articulate the target quantity $Q(M)$ in the language of science, i.e., counterfactuals [Pearl, 2000b].

6 The Mediation Formula: a simple solution to a thorny problem

This subsection demonstrates how the solution provided in equations (15) and (20) can be applied to practical problems of assessing mediation effects in non-linear models. We will use the simple mediation model of Fig. 4(a), where all error terms (not shown explicitly) are assumed to be mutually independent, with the understanding that adjustment for appropriate sets of covariates W may be necessary to achieve this independence and that integrals should replace summations when dealing with continuous variables [Imai et al., 2008].

Combining (15) and (20), the expression for the indirect effect, IE , becomes:

$$IE_{x,x'}(Y) = \sum_z E(Y|x, z)[P(z|x') - P(z|x)] \quad (21)$$

which provides a general formula for mediation effects, applicable to any nonlinear system, any distribution (of U), and any type of variables. Moreover, the formula is readily estimable by regression. Owing to its generality and ubiquity, I will refer to this expression as the “Mediation Formula.”

The Mediation Formula represents the average increase in the outcome Y that the transition from $X = x$ to $X = x'$ is expected to produce absent any direct effect of X on Y . Though based on solid causal principles, it embodies no causal assumption other than the generic mediation structure of Fig. 4(a). When the outcome Y is binary (e.g., recovery, or hiring) the ratio $(1 - IE)/TE$ represents the fraction of responding individuals who owe their response to direct paths, while $(1 - DE)/TE$ represents the fraction who owe their response to Z -mediated paths.

The Mediation Formula tells us that IE depends only on the expectation of the counterfactual Y_{xz} , not on its functional form $f_Y(x, z, u_Y)$ or its distribution $P(Y_{xz} = y)$. It calls therefore for a

two-step regression which, in principle, can be performed non-parametrically. In the first step we regress Y on X and Z , and obtain the estimate

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

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

$$IE_{x,x'}(Y) = E_z(g(x, z)|x') - E_z(g(x, z)|x)$$

Nonparametric estimation is not always practical. When Z consists of a vector of several mediators, the dimensionality of the problem would prohibit the estimation of $E(Y|x, z)$ for every (x, z) cell, and the need arises to use parametric approximation. We can then choose any convenient parametric form for $E(Y|x, z)$ (e.g., linear, logit, probit), estimate the parameters separately (e.g., by regression or maximum likelihood methods), insert the parametric approximation into (21) and estimate its two conditional expectations (over z) to get the mediated effect [VanderWeele, 2009].

Let us examine what the Mediation Formula yields when applied to both linear and non-linear versions of model 4(a). In the linear case, the structural model reads:

$$\begin{aligned} x &= u_X \\ z &= b_x x + u_Z \\ y &= c_x x + c_z z + u_Y \end{aligned} \tag{22}$$

Computing the conditional expectation in (21) gives

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

and yields

$$IE_{x,x'}(Y) = \sum_z (c_x x + c_z z) [P(z|x') - P(z|x)].$$

$$= c_z [E(Z|x') - E(Z|x)] \quad (23)$$

$$= (x' - x)(c_z b_x) \quad (24)$$

$$= (x' - x)(b - c_x) \quad (25)$$

where b is the total effect coefficient, $b = (E(Y|x') - E(Y|x))/(x' - x) = c_x + c_z b_x$.

We thus obtained the standard expressions for indirect effects in linear systems, which can be estimated either as a difference in two regression coefficients (Eq. 25) or a product of two regression coefficients (Eq. 24), with Y regressed on both X and Z . (see [MacKinnon et al., 2007b]). These two strategies do not generalize to non-linear system as we shall see next.

Suppose we apply (21) to a non-linear process (Fig. 5) in which X, Y , and Z are binary

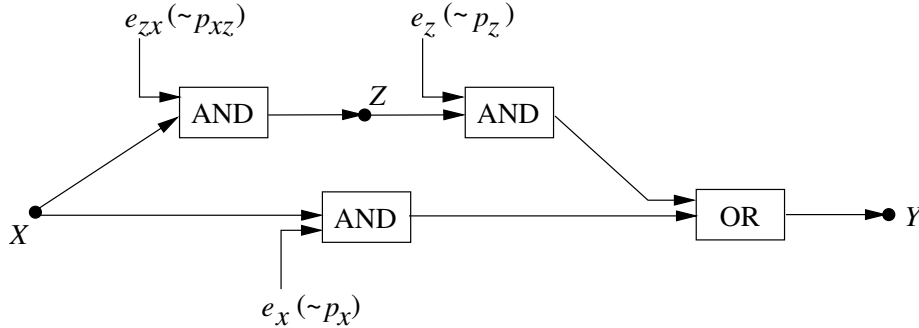


Figure 5: Stochastic non-linear model of mediation. All variables are binary.

variables, and Y and Z are given by the Boolean formula

$$Y = \text{AND}(x, e_x) \vee \text{AND}(z, e_z) \quad x, z, e_x, e_z = 0, 1$$

$$z = \text{AND}(x, e_{xz}) \quad z, e_{xz} = 0, 1$$

Such disjunctive interaction would describe, for example, a disease Y that would be triggered

either by X directly, if enabled by e_x , or by Z , if enabled by e_z . Let us further assume that e_x, e_z and e_{xz} are three independent Bernoulli variables with probabilities p_x, p_z , and p_{xz} , respectively.

As investigators, we are not aware, of course, of these underlying mechanisms; all we know is that X, Y , and Z are binary, that Z is hypothesized to be a mediator, and that the assumption of nonconfoundedness permits us to use the Mediation Formula (21) for estimating the Z -mediated effect of X on Y . Assume that our plan is to conduct a nonparametric estimation of the terms in (21) over a very large sample drawn from $P(x, y, z)$; it is interesting to ask what the asymptotic value of the Mediation Formula would be, as a function of the model parameters: p_x, p_z , and p_{xz} .

From knowledge of the underlying mechanism, we have:

$$\begin{aligned} P(Z = 1|x) &= p_{xz}x & x = 0, 1 \\ P(Y = 1|x, z) &= p_x x + p_z z - p_x p_z x z & x, z = 0, 1 \end{aligned}$$

Therefore,

$$\begin{aligned} E(Z|x) &= p_{xz}x & x = 0, 1 \\ E(Y|x, z) &= x p_x + z p_z - x z p_x p_z & x, z = 0, 1 \\ E(Y|x) &= \sum_z E(Y|x, z) P(z|x) \\ &= x p_x + (p_z - x p_x p_z) E(Z|x) \\ &= x(p_x + p_{xz} p_z - x p_x p_z p_{xz}) & x = 0, 1 \end{aligned}$$

Taking $x = 0, x' = 1$ and substituting these expressions in (15), (20), and (21) yields

$$IE(Y) = p_z p_{xz} \tag{26}$$

$$DE(Y) = p_x \tag{27}$$

$$TE(Y) = p_z p_{xz} + p_x + p_x p_z p_{xz} \tag{28}$$

Two observations are worth noting. First, we see that, despite the non-linear interaction between the two causal paths, the parameters of one do not influence on the causal effect mediated

by the other. Second, the total effect is not the sum of the direct and indirect effects. Instead, we have:

$$TE = DE + IE - DE * IE$$

which means that a fraction $DE \cdot ID/TE$ of outcome cases triggered by the transition from $X = 0$ to $X = 1$ are triggered simultaneously, by both causal paths, and would have been triggered even if one of the paths was disabled.

Now assume that we choose to approximate $E(Y|x, z)$ by the linear expression

$$g(x, z) = a_0 + a_1x + a_2z. \quad (29)$$

After fitting the a 's parameters to the data (e.g., by OLS) and substituting in (21) one would obtain

$$\begin{aligned} IE_{x,x'}(Y) &= \sum_z (a_0 + a_1x + a_2z)[P(z|x') - P(z|x)] \\ &= a_2[E(Z|x') - E(Z|x)] \end{aligned} \quad (30)$$

which holds whenever we use the approximation in (29), regardless of the underlying mechanism.

If the correct data-generating process was the linear model of (22), we would obtain the expected estimates $a_2 = c_z$, $E(z|x') - E(z|x) = b_x(x' - x)$ and

$$IE_{x,x'}(Y) = b_x c_z (x' - x).$$

If however we were to apply the approximation in (29) to data generated by the nonlinear model of Fig. 5, a distorted solution would ensue; a_2 would evaluate to

$$\begin{aligned} a_2 &= \sum_x [E(Y|x, z = 1) - E(Y|x, z = 0)]P(x) \\ &= P(x = 1)[E(Y|x = 1, z = 1) - E(Y|x = 1, z = 0)] \\ &= P(x = 1)[(p_x + p_z - p_x p_z) - p_x] \\ &= P(x = 1)p_z(1 - p_x), \end{aligned}$$

$E(z|x') - E(z|x)$ would evaluate to $p_{xz}(x' - x)$, and (30) would yield the approximation

$$\begin{aligned}\hat{I}E_{x,x'}(Y) &= a_2[E(Z|x') - E(Z|x)] \\ &= p_{xz}P(x=1)p_z(1 - p_x)\end{aligned}\tag{31}$$

We see immediately that the result differs from the correct value $p_z p_{xz}$ derived in (26).

Whereas the approximate value depends on $P(x=1)$, the correct value shows no such dependence, and rightly so; no causal effect should depend on the probability of the causal variable.

Fortunately, the analysis permits us to examine under what condition the distortion would be significant. Comparing (31) and (26) reveals that the approximate method always underestimates the indirect effect and the distortion is minimal for high values of $P(x=1)$ and $(1 - p_x)$.

Had we chosen to include an interaction term in the approximation of $E(Y|x, z)$, the correct result would obtain. To witness, writing

$$E(Y|x, z) = a_0 + a_1x + a_2z + a_3xz,$$

a_2 would evaluate to p_z , a_3 to $p_x p_z$, and the correct result obtains through:

$$\begin{aligned}IE_{x,x'}(Y) &= \sum_z (a_0 + a_1x + a_2z + a_3xz)[P(z|x') - P(z|x)] \\ &= (a_2 + a_3x)[E(Z|x') - E(Z|x)] \\ &= (a_2 + a_3x)p_{xz}(x' - x) \\ &= (p_z - p_x p_z)x p_{xz}(x' - x)\end{aligned}$$

We see that, in addition to providing causally-sound estimates for mediation effects, the Mediation Formula also enables researchers to evaluate analytically the effectiveness of various parametric specifications relative to any assumed model. This type of analytical “sensitivity analysis” has been used extensively in statistics for parameter estimation, but could not be applied to mediation analysis, owed to the absence of an objective target quantity that captures

the notion of indirect effect in both linear and non-linear systems, free of parametric assumptions. The Mediation Formula of Eq. (21) explicates this target quantity formally, and casts it in terms of estimable quantities.

The derivation of the Mediation Formula was facilitated by taking seriously the four steps of the structural methodology (Section 4) together with the graph-counterfactual-structural symbiosis spawned by the structural interpretation of counterfactuals (Eq. (8)).

In contrast, when the mediation problem is approached from an exclusivist potential-outcome viewpoint, void of the structural guidance of Eq. (8), paradoxical results ensue. For example, the direct effect is definable only in units absent of indirect effects [Rubin, 2004, 2005]. This means that a grandfather would be deemed to have no direct effect on his grandson's behavior in families where he has had some effect on the father. This precludes from the analysis all typical families, in which a father and a grandfather have simultaneous, complementary influences on children's upbringing. In linear systems, to take an even sharper example, the direct effect would be undefined whenever indirect paths exist from the cause to its effect. The emergence of such paradoxical conclusions underscores the wisdom, if not necessity of a symbiotic analysis, in which the counterfactual notation $Y_x(u)$ is governed by its structural definition, Eq. (8).¹⁵

7 Conclusions

Statistics is strong in inferring distributional parameters from sample data. 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

¹⁵Such symbiosis is now standard in epidemiology research [Robins, 2001, Petersen et al., 2006, VanderWeele and Robins, 2007, Hafeman and Schwartz, 2009, VanderWeele, 2009] and is making its way slowly toward the social and behavioral sciences.

new causal conclusions about a phenomena. This paper presents nonparametric structural causal models (SCM) as a formal and meaningful language for meeting these challenges, thus resolving ethical tensions that follow the disparity between causal quantities sought by scientists and associational quantities inferred from observational studies. 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 2000a; 2nd edition, 2009). 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, 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. 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, 1994.
- 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.M. Baron and D.A. Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.
- R. Brent and L. Lok. A fishing buddy for hypothesis generators. *Science*, 308(5721):523–529, 2005.
- 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, 2000.
- S. Greenland, J. Pearl, and J.M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48, 1999.
- D.M. Hafeman and S. Schwartz. Opening the black box: A motivation for the assessment of mediation. *International Journal of Epidemiology*, 3:838–845, 2009.
- 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.

- K. Imai, L. Keele, and T. Yamamoto. Identification, inference, and sensitivity analysis for causal mediation effects. Technical report, Department of Politics, Princeton University, December 2008.
- C.M. Judd and D.A. Kenny. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5):602–619, 1981.
- D.P. MacKinnon, A.J. Fairchild, and M.S. Fritz. Mediation analysis. *Annual Review of Psychology*, 58:593–614, 2007a.
- D.P. MacKinnon, C.M. Lockwood, C.H. Brown, W. Wang, and J.M Hoffman. The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, 4:499–513, 2007b.
- D. Muller, C.M. Judd, and V.Y. Yzerbyt. When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89(6):852–863, 2005.
- 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, 1993.
- J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 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. 2nd edition, 2009.
- J. Pearl. Comment on A.P. Dawid’s, Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):428–431, 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. Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine*, 28:1415–1416, 2009a. <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, 2009b. <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>.
- J. Pearl and A. Paz. Confounding equivalence in observational studies. Technical Report TR-343, University of California, Los Angeles, CA, July 2009. <http://ftp.cs.ucla.edu/pub/stat_ser/r343.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. Data, design, and background knowledge in etiologic inference. *Epidemiology*, 12(3):313–320, 2001.
- J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

- 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. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31:161–170, 2004.
- D.B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
- D.B. Rubin. Author’s reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine*, 28:1420–1423, 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.
- P.E. Shrout and N. Bolger. Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, 7(4):422–445, 2002.
- 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, 1998.
- M.E. Sobel. Identification of causal parameters in randomized studies with mediating variables. *Journal of Educational and Behavioral Statistics*, 33(2):230–231, 2008.

- P. Spirtes, C.N. 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.
- T.J. VanderWeele. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20(1):18–26, 2009.
- T.J. VanderWeele and J.M. Robins. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568, 2007.
- N. Wermuth. On block-recursive regression equations. *Brazilian Journal of Probability and Statistics* (with discussion), 6:1–56, 1992.
- N. Wermuth and D. Cox. Linear dependencies represented by chain graphs. *Statistical Science*, 8(3):204–218, 1993.
- J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester, England, 1990.
- L. Wilkinson, the Task Force on Statistical Inference, and APA Board of Scientific Affairs. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8):594–604, 1999.
- S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.