**Statistics Surveys**
Vol. 3 (2009) 96–146
ISSN: 1935-7516
DOI: 10.1214/09-SS057

# Causal inference in statistics: An overview*†‡

**Judea Pearl**

*Computer Science Department*
*University of California, Los Angeles, CA 90095 USA*
*e-mail:* judea@cs.ucla.edu

**Abstract:** This review presents empirical researchers with recent advances in causal inference, and stresses the paradigmatic shifts that must be undertaken in moving from traditional statistical analysis to causal analysis of multivariate data. Special emphasis is placed on the assumptions that underly all causal inferences, the languages used in formulating those assumptions, the conditional nature of all causal and counterfactual claims, and the methods that have been developed for the assessment of such claims. These advances are illustrated using a general theory of causation based on the Structural Causal Model (SCM) described in Pearl (2000a), which subsumes and unifies other approaches to causation, and provides a coherent mathematical foundation for the analysis of causes and counterfactuals. In particular, the paper surveys the development of mathematical tools 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"). Finally, the paper defines the formal and conceptual relationships between the structural and potential-outcome frameworks and presents tools for a symbiotic analysis that uses the strong features of both.

**Keywords and phrases:** Structural equation models, confounding, graphical methods, counterfactuals, causal effects, potential-outcome, mediation, policy evaluation, causes of effects.

Received September 2009.

## Contents



*Portions of this paper are based on my book *Causality* (Pearl, 2000, 2nd edition 2009), and have benefited appreciably from conversations with readers, students, and colleagues.
†This research was supported in parts by an ONR grant #N000-14-09-1-0665.
‡This paper was accepted by Elja Arjas, Executive Editor for the Bernoulli.

## 1. Introduction

The questions that motivate most studies in the health, social and behavioral sciences are not associational 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, nor from the distributions that govern the data.

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 who could put them into practical use. The main reason is educational. Solving causal problems systematically requires certain extensions

in the standard mathematical language of statistics, and these extensions are not generally emphasized in the mainstream literature and education. As a result, large segments of the statistical research community find it hard to appreciate and benefit from the many results that causal analysis has produced in the past two decades. These results rest on contemporary advances in four areas:

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

This survey aims at making these advances more accessible to the general research community by, first, contrasting causal analysis with standard statistical analysis, second, presenting a unifying theory, called "structural," within which most (if not all) aspects of causation can be formulated, analyzed and compared, thirdly, presenting a set of simple yet effective tools, spawned by the structural theory, for solving a wide variety of causal problems and, finally, demonstrating how former approaches to causal analysis emerge as special cases of the general structural theory.

To this end, Section 2 begins by illuminating two conceptual barriers that impede the transition from statistical to causal analysis: (i) coping with untested assumptions and (ii) acquiring new mathematical notation. Crossing these barriers, Section 3.1 then introduces the fundamentals of the structural theory of causation, with emphasis on the formal representation of causal assumptions, and formal definitions of causal effects, counterfactuals and joint probabilities of counterfactuals. Section 3.2 uses these modeling fundamentals to represent interventions and develop mathematical tools for estimating causal effects (Section 3.3) and counterfactual quantities (Section 3.4). These tools are demonstrated by attending to the analysis of instrumental variables and their role in bounding treatment effects in experiments marred by noncompliance (Section 3.5).

The tools described in this section permit investigators to communicate causal assumptions formally using diagrams, then inspect the diagram and

1. Decide whether the assumptions made are sufficient for obtaining consistent estimates of the target quantity;
2. Derive (if the answer to item 1 is affirmative) a closed-form expression for the target quantity in terms of distributions of observed quantities; and
3. Suggest (if the answer to item 1 is negative) a set of observations and experiments that, if performed, would render a consistent estimate feasible.

Section 4 relates these tools to those used in the potential-outcome framework, and offers a formal mapping between the two frameworks and a symbiosis (Section 4.3) that exploits the best features of both. Finally, the benefit of this symbiosis is demonstrated in Section 5, in which the structure-based logic of counterfactuals is harnessed to estimate causal quantities that cannot be defined within the paradigm of controlled randomized experiments. These include direct and indirect effects, the effect of treatment on the treated, and ques-

tions of attribution, i.e., whether one event can be deemed "responsible" for another.

## 2. From association to causation

### 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 beliefs or probabilities of past and future events, as well as update those probabilities in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer not only beliefs or probabilities under static conditions, but also the dynamics of beliefs under *changing conditions*, for example, changes induced by treatments 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 concept is any relationship that cannot be defined from the distribution alone. Examples of associational concepts are: correlation, regression, dependence, conditional independence, likelihood, collapsibility, propensity score, risk ratio, odds ratio, marginalization,

[1] The methodology of "causal discovery" (Spirtes et al. 2000; Pearl 2000a, Chapter 2) is likewise based on the causal assumption of "faithfulness" or "stability," a problem-independent assumption that concerns relationships between the structure of a model and the data it generates.

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

[2]For example, any intermediate variable $U$ on a causal path from $X$ to $Y$ satisfies this definition, without confounding the effect of $X$ on $Y$.

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, constitute the two main obstacles to the acceptance of causal analysis among statisticians and among professionals with traditional training in statistics.

Associational assumptions, even untested, are testable in principle, given 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 $P(Y_x = y)$ (see Section 4 for semantics). Alternatively, Pearl (1995a) 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]

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

[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

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 many academic scholars, 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 semi-formal way in which causal analysis has been presented to the research community, resting primarily on the restricted paradigm of controlled randomized trials.

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

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

Any conception of causation worthy of the title "theory" must be able to (1) represent causal questions in some mathematical language, (2) provide a precise language for communicating assumptions under which the questions need to be answered, (3) provide a systematic way of answering at least some of these questions and labeling others "unanswerable," and (4) provide a method of determining what assumptions or new measurements would be needed to answer the "unanswerable" questions.

A "general theory" should do more. In addition to embracing *all* questions judged to have causal character, a general theory must also *subsume* any other theory or method that scientists have found useful in exploring the various aspects of causation. In other words, any alternative theory needs to evolve as a special case of the "general theory" when restrictions are imposed on either the model, the type of assumptions admitted, or the language in which those assumptions are cast.

The structural theory that we use in this survey satisfies the criteria above. It is based on the Structural Causal Model (SCM) developed in (Pearl, 1995a, 2000a) which combines features of the structural equation models (SEM) used in economics and social science (Goldberger, 1973; Duncan, 1975), the potential-outcome framework of Neyman (1923) and Rubin (1974), and the graphical models developed for probabilistic reasoning and causal analysis (Pearl, 1988; Lauritzen, 1996; Spirtes et al., 2000; Pearl, 2000a).

standard probability expressions, void of graphs, counterfactual subscripts or $do(*)$ operators, can safely be discarded as inadequate.

Although the basic elements of SCM were introduced in the mid 1990's (Pearl, 1995a), and have been adapted widely by epidemiologists (Greenland et al., 1999; Glymour and Greenland, 2008), statisticians (Cox and Wermuth, 2004; Lauritzen, 2001), and social scientists (Morgan and Winship, 2007), its potentials as a comprehensive theory of causation are yet to be fully utilized. Its ramifications thus far include:

1. The unification of the graphical, potential outcome, structural equations, decision analytical (Dawid, 2002), interventional (Woodward, 2003), sufficient component (Rothman, 1976) and probabilistic (Suppes, 1970) approaches to causation; with each approach viewed as a restricted version of the SCM.
2. The definition, axiomatization and algorithmization of counterfactuals and joint probabilities of counterfactuals
3. Reducing the evaluation of "effects of causes," "mediated effects," and "causes of effects" to an algorithmic level of analysis.
4. Solidifying the mathematical foundations of the potential-outcome model, and formulating the counterfactual foundations of structural equation models.
5. Demystifying enigmatic notions such as "confounding," "mediation," "ignorability," "comparability," "exchangeability (of populations)," "superexogeneity" and others within a single and familiar conceptual framework.
6. Weeding out myths and misconceptions from outdated traditions (Meek and Glymour, 1994; Greenland et al., 1999; Cole and Hernán, 2002; Arah, 2008; Shrier, 2009; Pearl, 2009b).

This section provides a gentle introduction to the structural framework and uses it to present the main advances in causal inference that have emerged in the past two decades.

### *3.1. Introduction to structural equation models*

How can one express mathematically the common understanding that symptoms do not cause diseases? The earliest attempt to formulate such relationship mathematically was made in the 1920's by the geneticist Sewall Wright (1921). Wright used a combination of equations and graphs to communicate causal relationships. For example, if $X$ stands for a disease variable and $Y$ stands for a certain symptom of the disease, Wright would write a linear equation:[6]

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

where $x$ stands for the level (or severity) of the disease, $y$ stands for the level (or severity) of the symptom, and $u_Y$ stands for all factors, other than the disease in question, that could possibly affect $Y$ when $X$ is held constant. In interpreting

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

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 $y = \beta x + u_Y$. Similarly, to "explain" the occurrence of disease $X$, one could write $x = u_X$, where $U_X$ stands for all factors affecting $X$.

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

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

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

The variables $U_X$ and $U_Y$ are called "exogenous;" they represent observed or unobserved background factors that the modeler decides to keep unexplained, that is, factors that influence but are not influenced by the other variables (called "endogenous") in the model. Unobserved exogenous variables are sometimes called "disturbances" or "errors", they represent factors omitted from the model but judged to be relevant for explaining the behavior of variables in the model. Variable $U_X$, for example, represents factors that contribute to the disease $X$, which may or may not be correlated with $U_Y$ (the factors that influence the symptom $Y$). Thus, background factors in structural equations differ fundamentally from residual terms in regression equations. The latters are artifacts of analysis which, by definition, are uncorrelated with the regressors. The formers are part of physical reality (e.g., genetic factors, socio-economic conditions) which are responsible for variations observed in the data; they are treated as any other variable, though we often cannot measure their values precisely and must resign to merely acknowledging their existence and assessing qualitatively how they relate to other variables in the system.

If correlation is presumed possible, it is customary to connect the two variables, $U_Y$ and $U_X$, by a dashed double arrow, as shown in Fig. 1(b).

In reading path diagrams, it is common to use kinship relations such as parent, child, ancestor, and descendent, the interpretation of which is usually

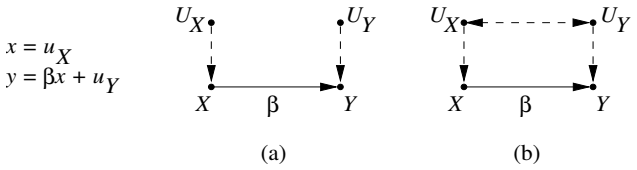


FIG 1. *A simple structural equation model, and its associated diagrams. Unobserved exogenous variables are connected by dashed arrows.*

self evident. For example, an arrow $X \rightarrow Y$ designates $X$ as a parent of $Y$ and $Y$ as a child of $X$. A "path" is any consecutive sequence of edges, solid or dashed. For example, there are two paths between $X$ and $Y$ in Fig. 1(b), one consisting of the direct arrow $X \rightarrow Y$ while the other tracing the nodes $X, U_X, U_Y$ and $Y$.

Wright's major contribution to causal analysis, aside from introducing the language of path diagrams, has been the development of graphical rules for writing down the covariance of any pair of observed variables in terms of path coefficients and of covariances among the error terms. In our simple example, one can immediately write the relations

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

for Fig. 1(a), and

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

for Fig. 1(b) (These can be derived of course from the equations, but, for large models, algebraic methods tend to obscure the origin of the derived quantities). Under certain conditions, (e.g. if $Cov(U_Y, U_X) = 0$), such relationships may allow one to solve for the path coefficients in term of observed covariance terms only, and this amounts to inferring the magnitude of (direct) causal effects from observed, nonexperimental associations, assuming of course that one is prepared to defend the causal assumptions encoded in the diagram.

It is important to note that, in path diagrams, causal assumptions are encoded not in the links but, rather, in the missing links. An arrow merely indicates the possibility of causal connection, the strength of which remains to be determined (from data); a missing arrow represents a claim of zero influence, while a missing double arrow represents a claim of zero covariance. In Fig. 1(a), for example, the assumptions that permits us to identify the direct effect $\beta$ are encoded by the missing double arrow between $U_X$ and $U_Y$, indicating $Cov(U_Y, U_X)$=0, together with the missing arrow from $Y$ to $X$. Had any of these two links been added to the diagram, we would not have been able to identify the direct effect $\beta$. Such additions would amount to relaxing the assumption $Cov(U_Y, U_X) = 0$, or the assumption that $Y$ does not effect $X$, respectively. Note also that both assumptions are causal, not associational, since none can be determined from the joint density of the observed variables, $X$ and $Y$; the association between the unobserved terms, $U_Y$ and $U_X$, can only be uncovered in an experimental setting; or (in more intricate models, as in Fig. 5) from other causal assumptions.

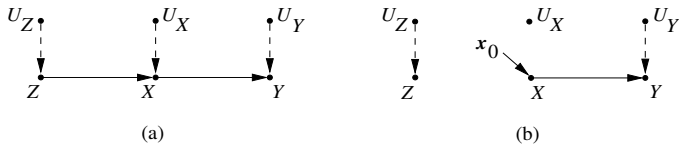


FIG 2. *(a) The diagram associated with the structural model of Eq. (5). (b) The diagram associated with the modified model of Eq. (6), representing the intervention* $do(X = x_0)$.

Although each causal assumption in isolation cannot be tested, the sum total of all causal assumptions in a model often has testable implications. The chain model of Fig. 2(a), for example, encodes seven causal assumptions, each corresponding to a missing arrow or a missing double-arrow between a pair of variables. None of those assumptions is testable in isolation, yet the totality of all those assumptions implies that $Z$ is unassociated with $Y$ in every stratum of $X$. Such testable implications can be read off the diagrams using a graphical criterion known as *d-separation* (Pearl, 1988).

**Definition 1** ($d$-separation)**.** 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 $S$. If $S$ blocks *all* paths from $X$ to $Y$, it is said to "$d$-separate $X$ and $Y$," and then, $X$ and $Y$ are independent given $S$, written $X \perp\!\!\!\perp Y|S$.

To illustrate, the path $U_Z \to Z \to X \to Y$ is blocked by $S = \{Z\}$ and by $S = \{X\}$, since each emits an arrow along that path. Consequently we can infer that the conditional independencies $U_X \perp\!\!\!\perp Y|Z$ and $U_Z \perp\!\!\!\perp Y|X$ will be satisfied in any probability function that this model can generate, regardless of how we parametrize the arrows. Likewise, the path $U_Z \to Z \to X \leftarrow U_X$ is blocked by the null set $\{\emptyset\}$ but is not blocked by $S = \{Y\}$, since $Y$ is a descendant of the collider $X$. Consequently, the marginal independence $U_Z \perp\!\!\!\perp U_X$ will hold in the distribution, but $U_Z \perp\!\!\!\perp U_X|Y$ may or may not hold. This special handling of colliders (e.g., $Z \to X \leftarrow U_X$)) reflects a general phenomenon known as *Berkson's paradox* (Berkson, 1946), whereby observations on a common consequence of two independent causes render those causes dependent. For example, the outcomes of two independent coins are rendered dependent by the testimony that at least one of them is a tail.

The conditional independencies induced by $d$-separation constitute the main opening through which the assumptions embodied in structural equation models can confront the scrutiny of nonexperimental data. In other words, almost all statistical tests capable of invalidating the model are entailed by those implications.[7]

[7]Additional implications called "dormant independence" (Shpitser and Pearl, 2008) may be deduced from some graphs with correlated errors.

### 3.2. From linear to nonparametric models and graphs

Structural equation modeling (SEM) has been the main vehicle for effect analysis in economics and the behavioral and social sciences (Goldberger, 1972; Duncan, 1975; Bollen, 1989). However, the bulk of SEM methodology was developed for linear analysis and, until recently, no comparable methodology has been devised to extend its capabilities to models involving dichotomous variables or nonlinear dependencies. A central requirement for any such extension is to detach the notion of "effect" from its algebraic representation as a coefficient in an equation, and redefine "effect" as a general capacity to transmit *changes* among variables. Such an extension, based on simulating hypothetical interventions in the model, was proposed in (Haavelmo, 1943; Strotz and Wold, 1960; Spirtes et al., 1993; Pearl, 1993a, 2000a; Lindley, 2002) and has led to new ways of defining and estimating causal effects in nonlinear and nonparametric models (that is, models in which the functional form of the equations is unknown).

The central idea is to exploit the invariant characteristics of structural equations without committing to a specific functional form. For example, the nonparametric interpretation of the diagram of Fig. 2(a) corresponds to a set of three functions, each corresponding to one of the observed variables:

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

where $U_Z, U_X$ and $U_Y$ are assumed to be jointly independent but, otherwise, arbitrarily distributed. Each of these functions represents a causal process (or mechanism) that determines the value of the left variable (output) from those on the right variables (inputs). The absence of a variable from the right hand side of an equation encodes the assumption that Nature ignores that variable in the process of determining the value of the output variable. For example, the absence of variable $Z$ from the arguments of $f_Y$ conveys the empirical claim that variations in $Z$ will leave $Y$ unchanged, as long as variables $U_Y$, and $X$ remain constant. A system of such functions are said to be *structural* if they are assumed to be autonomous, that is, each function is invariant to possible changes in the form of the other functions (Simon, 1953; Koopmans, 1953).

#### 3.2.1. Representing interventions

This feature of invariance permits us to use structural equations as a basis for modeling causal effects and counterfactuals. This is done through a mathematical operator called $do(x)$ which simulates physical interventions by deleting certain functions from the model, replacing them by a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $do(x_0)$ that holds $X$ constant (at $X = x_0$) in model $M$ of Fig. 2(a), we

replace the equation for $x$ in Eq. (5) with $x = x_0$, and obtain a new model, $M_{x_0}$,

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

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

The joint distribution associated with the modified model, denoted $P(z, y|do(x_0))$ describes the post-intervention distribution of variables $Y$ and $Z$ (also called "controlled" or "experimental" distribution), to be distinguished from the pre-intervention distribution, $P(x, y, z)$, associated with the original model of Eq. (5). For example, if $X$ represents a treatment variable, $Y$ a response variable, and $Z$ some covariate that affects the amount of treatment received, then the distribution $P(z, y|do(x_0))$ gives the proportion of individuals that would attain response level $Y = y$ and covariate level $Z = z$ under the hypothetical situation in which treatment $X = x_0$ is administered uniformly to the population.

In general, we can formally define the post-intervention distribution by the equation:

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

In words: In the framework of model $M$, the 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 average difference

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

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

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

The variance $Var(Y|do(x_0))$, or any other distributional parameter, may also enter the comparison; all these measures can be obtained from the controlled distribution function $P(Y = y|do(x)) = \sum_z P(z, y|do(x))$ which was called "causal effect" in Pearl (2000a, 1995a) (see footnote 4). The central question in the analysis of causal effects is the question of *identification*: Can the controlled (post-intervention) distribution, $P(Y = y|do(x))$, be estimated from data governed by the pre-intervention distribution, $P(z, x, y)$?

The problem of *identification* has received considerable attention in econometrics (Hurwicz, 1950; Marschak, 1950; Koopmans, 1953) and social science (Duncan, 1975; Bollen, 1989), usually in linear parametric settings, were it reduces to asking whether some model parameter, $\beta$, has a unique solution in terms of the parameters of $P$ (the distribution of the observed variables). In the nonparametric formulation, identification is more involved, since the notion

of "has a unique solution" does not directly apply to causal quantities such as $Q(M) = P(y|do(x))$ which have no distinct parametric signature, and are defined procedurally by simulating an intervention in a causal model $M$ (7). The following definition overcomes these difficulties:

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

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

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

### *3.2.2. Estimating the effect of interventions*

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

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

whereas the pre-intervention model of Eq. (5) gives

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

which is identical to (11). Therefore,

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

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

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

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

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

This equality endows $\beta$ with its causal meaning as "effect coefficient." It is extremely important to keep in mind that in structural (as opposed to regressional) models, $\beta$ is not "interpreted" as an effect coefficient but is "proven" to be one by the derivation above. $\beta$ will retain this causal interpretation regardless of how $X$ is actually selected (through the function $f_X$, Fig. 2(a)) and regardless of whether $U_X$ and $U_Y$ are correlated (as in Fig. 1(b)) or uncorrelated (as in Fig. 1(a)). Correlations may only impede our ability to estimate $\beta$ from nonexperimental data, but will not change its definition as given in (14). Accordingly, and contrary to endless confusions in the literature (see footnote 15) structural equations say absolutely nothing about the conditional expectation $E(Y|X = x)$. Such connection may be exist under special circumstances, e.g., if $cov(X, U_Y) = 0$, as in Eq. (13), but is otherwise irrelevant to the definition or interpretation of $\beta$ as effect coefficient, or to the empirical claims of Eq. (1).

The next subsection will circumvent these derivations altogether by reducing the identification problem to a graphical procedure. Indeed, since graphs encode all the information that non-parametric structural equations represent, they should permit us to solve the identification problem without resorting to algebraic analysis.

### *3.2.3. Causal effects from data and graphs*

Causal analysis in graphical models begins with the realization that all causal effects are identifiable whenever the model is *Markovian*, that is, the graph is acyclic (i.e., containing no directed cycles) and all the error terms are jointly independent. Non-Markovian models, such as those involving correlated errors (resulting from unmeasured confounders), permit identification only under certain conditions, and these conditions too can be determined from the graph structure (Section 3.3). The key to these results rests with the following basic theorem.

**Theorem 1** (The Causal Markov Condition)**.** *Any distribution generated by a Markovian model $M$ can be factorized as:*

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

*where $V_1, V_2, \ldots, V_n$ are the endogenous variables in $M$, and $pa_i$ are (values of) the endogenous "parents" of $V_i$ in the causal diagram associated with $M$.*

For example, the distribution associated with the model in Fig. 2(a) can be factorized as

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

since $X$ is the (endogenous) parent of $Y, Z$ is the parent of $X$, and $Z$ has no parents.

**Corollary 1** (Truncated factorization)**.** *For any Markovian model, the distribution generated by an intervention $do(X = x_0)$ on a set $X$ of endogenous variables is given by the truncated factorization*

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

*where $P(v_i|pa_i)$ are the pre-intervention conditional probabilities.*[8]

Corollary 1 instructs us to remove from the product of Eq. (15) all factors associated with the intervened variables (members of set $X$). This follows from the fact that the post-intervention model is Markovian as well, hence, following Theorem 1, it must generate a distribution that is factorized according to the modified graph, yielding the truncated product of Corollary 1. In our example of Fig. 2(b), the distribution $P(z, y|do(x_0))$ associated with the modified model is given by

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

where $P(z)$ and $P(y|x_0)$ are identical to those associated with the pre-intervention distribution of Eq. (16). As expected, the distribution of $Z$ is not affected by the intervention, since

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

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

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

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

[8] A simple proof of the Causal Markov Theorem is given in Pearl (2000a, p. 30). This theorem was first presented in Pearl and Verma (1991), but it is implicit in the works of Kiiveri et al. (1984) and others. Corollary 1 was named "Manipulation Theorem" in Spirtes et al. (1993), and is also implicit in Robins' (1987) *G*-computation formula. See Lauritzen (2001).

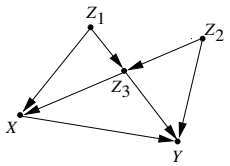


FIG 3. *Markovian model illustrating the derivation of the causal effect of $X$ on $Y$, Eq. (20). Error terms are not shown explicitly.*

The truncated factorization formula enables us to derive causal quantities directly, without dealing with equations or equation modification as in Eqs. (11)–(13). Consider, for example, the model shown in Fig. 3, in which the error variables are kept implicit. Instead of writing down the corresponding five nonparametric equations, we can write the joint distribution directly as

$$P(x, z_1, z_2, z_3, y) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(x|z_1, z_3)P(y|z_2, z_3, x) \tag{18}$$

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

$$P(z_1, z_2, z_3, y|do(x_0)) = P(z_1)P(z_2)P(z_3|z_1, z_2)P(y|z_2, z_3, x_0) \tag{19}$$

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

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

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

Note also that the truncated factorization formula is not restricted to interventions on a single variable; it is applicable to simultaneous or sequential interventions such as those invoked in the analysis of time varying treatment with time varying confounders (Robins, 1986; Arjas and Parner, 2004). For example, if $X$ and $Z_2$ are both treatment variables, and $Z_1$ and $Z_3$ are measured covariates, then the post-intervention distribution would be

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

and the causal effect of the treatment sequence $do(X = x), do(Z_2 = z_2)$[9] would be

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

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

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

### *3.3. Coping with unmeasured confounders*

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

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

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

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

where $T$ is the set of direct causes of $X$ (also called "parents") in the graph. This allows us to write (23) directly from the graph, thus skipping the algebra that led to (23). It further implies that, no matter how complicated the model, the parents of $X$ are the only variables that need to be measured to estimate the causal effects of $X$.

It is not immediately clear however whether other sets of variables beside $X$'s parents suffice for estimating the effect of $X$, whether some algebraic manipulation can further reduce Eq. (23), or that measurement of $Z_3$ (unlike $Z_1$, or $Z_2$) is necessary in any estimation of $P(y|do(x_0))$. Such considerations become transparent from a graphical criterion to be discussed next.

#### *3.3.1. Covariate selection – the back-door criterion*

Consider an observational study where we wish to find the effect of $X$ on $Y$, for example, treatment on response, and assume that the factors deemed relevant to the problem are structured as in Fig. 4; some are affecting the response, some are affecting the treatment and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetic trait or life style, others are measurable, such as gender, age, and salary level. Our problem is to select a subset of these factors for measurement and adjustment, namely, that if we compare 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 "admissible set" for

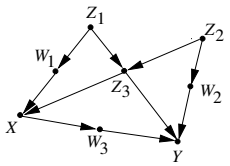


FIG 4. *Markovian model illustrating the back-door criterion. Error terms are not shown explicitly.*

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

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

**Definition 3** (Admissible sets – the back-door criterion)**.** A set $S$ is admissible (or "sufficient") for adjustment if two conditions hold:

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

In this criterion, "blocking" is interpreted as in Definition 1. For example, the set $S = \{Z_3\}$ blocks the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$, because the arrow-emitting node $Z_3$ is in $S$. However, the set $S = \{Z_3\}$ does not block the path $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$, because none of the arrow-emitting nodes, $Z_1$ and $Z_2$, is in $S$, and the collision node $Z_3$ is not outside $S$.

Based on this criterion we see, for example, that the sets $\{Z_1, Z_2, Z_3\}, \{Z_1, Z_3\}$, $\{W_1, Z_3\}$, and $\{W_2, Z_3\}$, 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 intuition behind the back-door criterion is as follows. The back-door paths in the diagram carry spurious associations from $X$ to $Y$, while the paths directed along the arrows from $X$ to $Y$ carry causative associations. Blocking the former paths (by conditioning on $S$) ensures that the measured association between $X$ and $Y$ is purely causative, namely, it correctly represents the target quantity: the causal effect of $X$ on $Y$. The reason for excluding descendants of $X$ (e.g., $W_3$ or any of its descendants) is given in (Pearl, 2009a, p. 338–41).

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

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

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

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

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

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

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

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

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

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

and

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

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

Interestingly, it can be shown that any irreducible sufficient set, $S$, taken as a unit, satisfies the associational criterion that epidemiologists have been using to define "confounders". In other words, $S$ must be associated with $X$ and, simultaneously, associated with $Y$, given $X$. This need not hold for any specific members of $S$. For example, the variable $Z_3$ in Fig. 4, though it is a member of every sufficient set and hence a confounder, can be unassociated with both $Y$ and $X$ (Pearl, 2000a, p. 195). Conversely, a pre-treatment variable $Z$ that is associated with both $Y$ and $X$ may need to be excluded from entering a sufficient set.

The back-door criterion allows us to write Eq. (25) directly, by selecting a sufficient set $S$ directly from the diagram, without manipulating the truncated factorization formula. The selection criterion can be applied systematically to diagrams of any size and shape, thus freeing analysts from judging whether "$X$ is conditionally ignorable given $S$," a formidable mental task required in the potential-response framework (Rosenbaum and Rubin, 1983). The criterion also enables the analyst to search for an optimal set of covariate—namely, a set $S$ that minimizes measurement cost or sampling variability (Tian et al., 1998).

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

### 3.3.2. General control of confounding

Adjusting for covariates is only one of many methods that permits us to estimate causal effects in nonexperimental studies. Pearl (1995a) has presented examples in which there exists no set of variables that is sufficient for adjustment and where the causal effect can nevertheless be estimated consistently. The estimation, in such cases, employs multi-stage adjustments. For example, if $W_3$ is the only observed covariate in the model of Fig. 4, then there exists no sufficient set for adjustment (because no set of observed covariates can block the paths from $X$ to $Y$ through $Z_3$), yet $P(y|do(x))$ can be estimated in two steps; first we estimate $P(w_3|do(x)) = P(w_3|x)$ (by virtue of the fact that there exists no unblocked back-door path from $X$ to $W_3$), second we estimate $P(y|do(w_3))$ (since $X$ constitutes a sufficient set for the effect of $W_3$ on $Y$) and, finally, we combine the two effects together and obtain

$$P(y|do(x)) = \sum_{w_3} P(w_3|do(x))P(y|do(w_3)) \tag{26}$$

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

The analysis used in the derivation and validation of such results invokes mathematical rules of transforming causal quantities, represented by expressions such as $P(Y = y|do(x))$, into *do*-free expressions derivable from $P(z, x, y)$, since only *do*-free expressions are estimable from non-experimental data. When such a transformation is feasible, we are ensured that the causal quantity is identifiable.

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

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

**Theorem 2** (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 Fig. 4, $P(y|do(x))$ can be estimated since every path from $X$ to $W_3$ (the only child of $X$) traces either the arrow $X \to W_3$, or the arrow $W_3 \to Y$, both emanating from a measured variable ($W_3$).

More recent results extend this theorem by (1) presenting a *necessary* and sufficient condition for identification (Shpitser and Pearl, 2006), and (2) extending

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

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.

### *3.3.3. From identification to estimation*

The mathematical derivation of causal effect estimands, like Eqs. (25) and (26) is merely a first step toward computing quantitative estimates of those effects from finite samples, using the rich traditions of statistical estimation and machine learning Bayesian as well as non-Bayesian. Although the estimands derived in (25) and (26) are non-parametric, this does not mean that one should refrain from using parametric forms in the estimation phase of the study. Parametrization is in fact necessary when the dimensionality of a problem is high. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in (26) 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 are the "marginal structural models" of (Robins, 1999), and the "propensity score" method of (Rosenbaum and Rubin, 1983) which were found to be particularly useful when dimensionality is high and data are sparse (see Pearl (2009a, pp. 348–52)).

It should be emphasized, however, that contrary to conventional wisdom (e.g., (Rubin, 2007, 2009)), propensity score methods are merely efficient estimators of the right hand side of (25); they cannot be expected to reduce bias in case the set $S$ does not satisfy the back-door criterion (Pearl, 2009a,b,c). Consequently, the prevailing practice of conditioning on as many pre-treatment measurements as possible should be approached with great caution; some covariates (e.g., $Z_3$ in Fig. 3) may actually increase bias if included in the analysis (see footnote 20). Using simulation and parametric analysis, Heckman and Navarro-Lozano (2004) and Wooldridge (2009) indeed confirmed the bias-raising potential of certain covariates in propensity-score methods. The graphical tools presented in this section unveil the character of these covariates and show precisely what covariates should, and should not be included in the conditioning set for propensity-score matching (see also (Pearl and Paz, 2009)).

### *3.3.4. Bayesianism and causality, or where do the probabilities come from?*

Looking back at the derivation of causal effects in Sections 3.2 and 3.3, the reader should note that at no time did the analysis require numerical assessment of probabilities. True, we assumed that the causal model $M$ is loaded with a probability function $P(u)$ over the exogenous variables in $U$, and we likewise assumed that the functions $v_i = f_i(pa_i, u)$ map $P(u)$ into a probability $P(v_1, v_2, \ldots, v_n)$ over the endogenous observed variables. But we never used or required any numerical assessment of $P(u)$ nor any assumption on the

form of the structural equations $f_i$. The question naturally arises: Where do the numerical values of the post-intervention probabilities $P(y|do(x))$ come from?

The answer is, of course, that they come from the data together with standard estimation techniques that turn data into numerical estimates of statistical parameters (i.e., aspects of a probability distribution). Subjective judgments were required only in *qualitative* form, to jump start the identification process, the purpose of which was to determine what statistical parameters need be estimated. Moreover, even the qualitative judgments were not about properties of probability distributions but about cause-effect relationships, the latter being more transparent, communicable and meaningful. For example, judgments about potential correlations between two $U$ variables were essentially judgments about whether the two have a latent common cause or not.

Naturally, the influx of traditional estimation techniques into causal analysis carries with it traditional debates between Bayesians and frequentists, subjectivists and objectivists. However, this debate is orthogonal to the distinct problems confronted by causal analysis, as delineated by the demarcation line between causal and statistical analysis (Section 2).

As is well known, many estimation methods in statistics invoke subjective judgment at some level or another; for example, what parametric family of functions one should select, what type of prior one should assign to the model parameters, and more. However, these judgments all refer to properties or parameters of a static distribution function and, accordingly, they are expressible in the language of probability theory. The new ingredient that causal analysis brings to this tradition is the necessity of obtaining explicit judgments not about properties of distributions but about the invariants of a distribution, namely, judgment about cause-effect relationships, and those, as we discussed in Section 2, cannot be expressed in the language of probability.

Causal judgments are tacitly being used at many levels of traditional statistical estimation. For example, most judgments about conditional independence emanate from our understanding of cause effect relationships. Likewise, the standard decision to assume independence among certain statistical parameters and not others (in a Bayesian prior) rely on causal information (see discussions with Joseph Kadane and Serafin Moral (Pearl, 2003)). However the causal rationale for these judgments has remained implicit for many decades, for lack of adequate language; only their probabilistic ramifications received formal representation. Causal analysis now requires explicit articulation of the underlying causal assumptions, a vocabulary that differs substantially from the one Bayesian statisticians have been accustomed to articulate.

The classical example demonstrating the obstacle of causal vocabulary is Simpson's paradox (Simpson, 1951) – a reversal phenomenon that earns its claim to fame only through a causal interpretation of the data (Pearl, 2000a, Chapter 6). The phenomenon was discovered by statisticians a century ago (Pearson et al., 1899; Yule, 1903) analyzed by statisticians for half a century (Simpson, 1951; Blyth, 1972; Cox and Wermuth, 2003) lamented by statisticians (Good and Mittal, 1987; Bishop et al., 1975) and wrestled with by statisticians till this very day (Chen et al., 2009; Pavlides and Perlman, 2009). Still, to the

best of my knowledge, Wasserman (2004) is the first statistics textbook to treat Simpson's paradox in its correct causal context (Pearl, 2000a, p. 200).

Lindley and Novick (1981) explained this century-long impediment to the understanding of Simpson's paradox as a case of linguistic handicap: "We have not chosen to do this; nor to discuss causation, because the concept, although widely used, does not seem to be well-defined" (p. 51). Instead, they attribute the paradox to another untestable relationship in the story—exchangeability (DeFinetti, 1974) which is cognitively formidable yet, at least formally, can be cast as a property of some imaginary probability function.

The same reluctance to extending the boundaries of probability language can be found among some scholars in the potential-outcome framework (Section 4), where judgments about conditional independence of counterfactual variables, however incomprehensible, are preferred to plain causal talk: "Mud does not cause rain."

This reluctance however is diminishing among Bayesians primarily due to recognition that, orthogonal to the traditional debate between frequentists and subjectivists, causal analysis is about change, and change demands a new vocabulary that distinguishes "seeing" from "doing" (Lindley, 2002) (see discussion with Dennis Lindley (Pearl, 2009a, 2nd Edition, Chapter 11).

Indeed, whether the conditional probabilities that enter Eqs. (15)–(25) originate from frequency data or subjective assessment matters not in causal analysis. Likewise, whether the causal effect $P(y|do(x))$ is interpreted as one's degree of belief in the effect of action $do(x)$, or as the fraction of the population that will be affected by the action matters not in causal analysis. What matters is one's readiness to accept and formulate qualitative judgments about cause-effect relationship with the same seriousness that one accepts and formulates subjective judgment about prior distributions in Bayesian analysis.

Trained to accept the human mind as a reliable transducer of experience, and human experience as a faithful mirror of reality, Bayesian statisticians are beginning to accept the language chosen by the mind to communicate experience – the language of cause and effect.

### *3.4. Counterfactual analysis in structural models*

Not all questions of causal character can be encoded in $P(y|do(x))$ type expressions, thus implying that not all causal questions can be answered from experimental studies. For example, questions of attribution (e.g., what fraction of death cases are *due to* specific exposure?) or of susceptibility (what fraction of the healthy unexposed population would have gotten the disease had they been exposed?) cannot be answered from experimental studies, and naturally, this kind of questions cannot be expressed in $P(y|do(x))$ notation.[11] To answer

[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

such questions, a probabilistic analysis of counterfactuals is required, one dedicated to the relation "$Y$ would be $y$ had $X$ been $x$ in situation $U = u$," denoted $Y_x(u) = y$. Remarkably, unknown to most economists and philosophers, structural equation models provide the formal interpretation and symbolic machinery for analyzing such counterfactual relationships.[12]

The key idea is to interpret the phrase "had $X$ been $x$" as an instruction to make a minimal modification in the current model, which may have assigned $X$ a different value, say $X = x'$, so as to ensure the specified condition $X = x$. Such a minimal modification amounts to replacing the equation for $X$ by a constant $x$, as we have done in Eq. (6). This replacement permits the constant $x$ to differ from the actual value of $X$ (namely $f_X(z, u_X)$) without rendering the system of equations inconsistent, thus yielding a formal interpretation of counterfactuals in multi-stage models, where the dependent variable in one equation may be an independent variable in another.

**Definition 4** (Unit-level Counterfactuals, Pearl (2000a, p. 98))**.** Let $M$ be a structural model and $M_x$ a modified version of $M$, with the equation(s) of $X$ replaced by $X = x$. Denote the solution for $Y$ in the equations of $M_x$ by the symbol $Y_{M_x}(u)$. The counterfactual $Y_x(u)$ (Read: "The value of $Y$ in unit $u$, had $X$ been $x$" is given by:

$$Y_x(u) \triangleq Y_{M_x}(u). \tag{27}$$

We see that the unit-level counterfactual $Y_x(u)$, which in the Neyman-Rubin approach is treated as a primitive, undefined quantity, is actually a derived quantity in the structural framework. The fact that we equate the experimental unit $u$ with a vector of background conditions, $U = u$, in $M$, reflects the understanding that the name of a unit or its identity do not matter; it is only the vector $U = u$ of attributes characterizing a unit which determines its behavior or response. As we go from one unit to another, the laws of nature, as they are reflected in the functions $f_X, f_Y$, etc. remain invariant; only the attributes $U = u$ vary from individual to individual.[13]

To illustrate, consider the solution of $Y$ in the modified model $M_{x_0}$ of Eq. (6), which Definition 4 endows with the symbol $Y_{x_0}(u_X, u_Y, u_Z)$. This entity has a

be alive if untreated and, simultaneously, many of those who survived with treatment would have died if not treated.

[12] Connections between structural equations and a restricted class of counterfactuals were first recognized by Simon and Rescher (1966). These were later generalized by Balke and Pearl (1995) to permit endogenous variables to serve as counterfactual antecedents.

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

clear counterfactual interpretation, for it stands for the way an individual with characteristics $(u_X, u_Y, u_Z)$ would respond, had the treatment been $x_0$, rather than the treatment $x = f_X(z, u_X)$ actually received by that individual. In our example, since $Y$ does not depend on $u_X$ and $u_Z$, we can write:

$$Y_{x_0}(u) = Y_{x_0}(u_Y, u_X, u_Z) = f_Y(x_0, u_Y). \tag{28}$$

In a similar fashion, we can derive

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

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

and so on. These examples reveal the counterfactual reading of each individual structural equation in the model of Eq. (5). The equation $x = f_X(z, u_X)$, for example, advertises the empirical claim that, regardless of the values taken by other variables in the system, had $Z$ been $z_0$, $X$ would take on no other value but $x = f_X(z_0, u_X)$.

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

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

the conditioning events $Y = y_0$ and $X = x_0$ yield $U_X = x_0$ and $U_Y = y_0 - \beta x_0$, and we can conclude that, with probability one, $Y_{x_1}$ must take on the value: $Y_{x_1} = \beta x_1 + U_Y = \beta(x_1 - x_0) + y_0$. In other words, if $X$ were $x_1$ instead of $x_0$, $Y$ would increase by $\beta$ times the difference $(x_1 - x_0)$. In nonlinear systems, the result would also depend on the distribution of $\{U_X, U_Y\}$ and, for that reason, attributional queries are generally not identifiable in nonparametric models (see Section 5.2 and 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 $X = x'$."[14] 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.

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

The structural definition 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 ruling 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. Section 4 establishes the precise relationship between the structural and potential-outcome paradigms, and outlines how the latter can benefit from the richer representational power of the former.

### *3.5. An example: Non-compliance in clinical trials*

To illustrate the methodology of the structural approach to causation, let us consider the practical problem of estimating treatment effect in a typical clinical trial with partial compliance. Treatment effect in such a setting is in general nonidentifiable, yet this example is well suited for illustrating the four major steps that should be part of every exercise in causal inference:

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

#### *3.5.1. Defining the target quantity*

The definition phase in our example is not altered by the specifics of the experimental setup under discussion. The structural modeling approach insists on defining the target quantity, in our case "causal effect," before specifying the process of treatment selection, and without making functional form or distributional assumptions. The formal definition of the causal effect $P(y|do(x))$, as given in Eq. (7), is universally applicable to all models, and invokes 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 confusions and controversies that have plagued the interpretation of structural equations and econometric parameters for the past half century (see footnote 15).

#### *3.5.2. Formulating the assumptions – Instrumental variables*

The experimental setup in a typical clinical trial with partial compliance can be represented by the model of Fig. 5(a) and Eq. (5) where $Z$ represents a randomized treatment assignment, $X$ is the treatment actually received, and $Y$ is

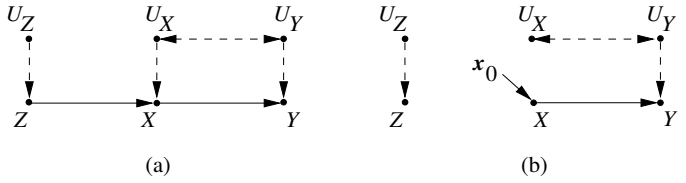


FIG 5. *(a) Causal diagram representing a clinical trial with imperfect compliance. (b) A diagram representing interventional treatment control.*

the observed response. The $U_Y$ term represents all factors (unobserved) that influence the way a subject responds to treatments; hence, an arrow is drawn from $U_Y$ to $Y$. Similarly, $U_X$ denotes all factors that influence the subject's compliance with the assignment, and $U_Z$ represents the random device used in deciding assignment. The dependence between $U_X$ and $U_Y$ allows for certain factors (e.g., socio economic status or predisposition to disease and complications) to influence both compliance and response. In Eq. (5), $f_X$ represents the process by which subjects select treatment level and $f_Y$ represents th process that determines the outcome $Y$. Clearly, perfect compliance would amount to setting $f_X(z, u_X) = z$ while any dependence on $u_X$ represents imperfect compliance.

The graphical model of Fig. 5(a) reflects two assumptions.

1. The assignment $Z$ does not influence $Y$ directly but rather through the actual treatment taken, $X$. This type of assumption is called "exclusion" restriction, for it excludes a variable ($Z$) from being a determining argument of the function $f_Y$, as in Eq. (5).
2. The variable $Z$ is independent of $U_Y$ and $U_X$; this is ensured through the randomization of $Z$, which rules out a common cause for both $Z$ and $U_Y$ (as well as for $Z$ and $U_X$).

By drawing the diagram of Fig. 5(a) an investigator encodes an unambiguous specification of these two assumptions, and permits the technical part of the analysis to commence, under the interpretation provided by Eq. (5).

The target of causal analysis in this setting is to estimate the causal effect of the treatment ($X$) on the outcome ($Y$), as defined by the modified model of Eq. (6) and the corresponding distribution $P(y|do(x_0))$. In words, this distribution describes the response of the population to a hypothetical experiment in which we administer treatment at level $X = x_0$ uniformly to the entire population and let $x_0$ take different values on hypothetical copies of the population. An inspection of the diagram in Fig. 5(a) reveals immediately that this distribution is not identifiable by adjusting for confounders. The graphical criterion for such identification (Definition 3) requires the existence of observed covariates on the "back-door" path $X \leftarrow U_X \leftrightarrow U_Y \rightarrow Y$, that blocks the spurious associations created by that path. Had $U_X$ (or $U_Y$) been observable, the treatment effect

would have been obtained by stratification on the levels of $U_X$.

$$P(Y = y|do(x_0)) = \sum_{u_X} P(Y = y|X = x_0, U_X = u_X)P(U_X = u_X) \tag{29}$$

thus yielding an estimable expression that requires no measurement of $U_Y$ and no assumptions relative the dependence between $U_Y$ and $U_X$. However, since $U_X$ (and $U_Y$) are assumed to be unobserved, and since no other blocking covariates exist, the investigator can conclude that confounding bias cannot be removed by adjustment. Moreover, it can be shown that, in the absence of additional assumptions, the treatment effect in such graphs cannot be identified by any method whatsoever (Balke and Pearl, 1997); one must therefore resort to approximate methods of assessment.

It is interesting to note that it is our insistence on allowing arbitrary functions in Eq. (5) that curtails our ability to infer the treatment effect from nonexperimental data (when $U_X$ and $U_Y$ are unobserved). In linear systems, for example, the causal effect of $X$ on $Y$ is identifiable, as can be seen by writing:[15]

$$y = f_Y(x, u) = \beta x + u_Y; \tag{30}$$

multiplying this equation by $z$ and taking expectations, gives

$$\beta = Cov(Z, Y)/(Cov(Z, X) \tag{31}$$

which reduces $\beta$ to correlations among observed measurements. Eq. (31) is known as the *instrumental variable* estimand (Bowden and Turkington, 1984). Similarly, Angrist and Imbens (1991) have shown that a broader class of nonlinear functions $f_X$ and $f_Y$ may render the causal effect identifiable. Angrist et al. (1996) and Heckman and Vytlacil (2005) further refined this analysis by considering a variety of causal effect measures, each applicable to a special (albeit non-identifiable and transient) segment of the population.

### *3.5.3. Bounding causal effects*

When conditions for identification are not met, the best one can do is derive *bounds* for the quantities of interest—namely, a range of possible values that represents our ignorance about the data-generating process and that cannot be improved with increasing sample size. In our example, this amounts to bounding the average difference of Eq. (8) subject to the constraint provided by the

[15] Note that $\beta$ represents the incremental causal effect of $X$ on $Y$, defined 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 $\beta$ 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).

observed distribution

$$P(x,y|z) = \sum_{u_X,u_Y} P(x,y,u_X,u_Y|z)$$
$$= \sum_{u_X,u_Y} P(y|x,u_Y,u_X)P(x|z,u_X)P(u_Y,u_X) \quad (32)$$

where the product decomposition is licensed by the conditional independencies shown in Fig. 5(a). Likewise, since the causal effect is governed by the modified model of Fig. 5(b), it can be written

$$P(y|do(x')) - P(y|do(x'')) = \sum_u [P(y|x',u_Y) - P(y|x'',u_Y)]P(u_Y) \quad (33)$$

Our task is then to bound the expression in Eq. (33) given the observed probabilities $P(y,x|z)$ as expressed in Eq. (32). This task amounts to a constrained optimization exercise of finding the highest and lowest values of Eq. (33) subject to the equality constraints in Eq. (32), where the maximization ranges over all possible functions $P(u_Y,u_X)$, $P(y|x,u_Y,u_X)$ and $P(x|z,u_Y)$ that satisfy those constraints.

Realizing that units in this example fall into 16 equivalent classes, each representing a binary function $X = f(z)$ paired with a binary function $y = g(x)$, Balke and Pearl (1997) were able to derive closed-form solutions for these bounds.[16] They showed that despite the imperfection of the experiments, the derived bounds can yield significant and sometimes accurate information on the treatment efficacy. Chickering and Pearl (1997) further used Bayesian techniques (with Gibbs sampling) to investigate the sharpness of these bounds as a function of sample size.

### *3.5.4. Testable implications of instrumental variables*

The two assumptions embodied in the model of Fig. 5(a), that $Z$ is randomized and has no direct effect on $Y$, are untestable in general (Bonet, 2001). However, if the treatment variable may take only a finite number of values, the combination of these two assumptions yields testable implications, and these can be used to alert investigators to possible violations of these assumptions. The testable implications take the form of inequalities which restrict aspects of the observed conditional distribution $P(x,y|z)$ from exceeding certain bounds (Pearl, 1995b).

One specially convenient form that these restrictions assume is given by the inequality

$$\max_x \sum_y [\max_z P(x,y|z)] \leq 1 \quad (34)$$

Pearl (1995b) called this restriction an *instrumental inequality*, because it constitutes a necessary condition for any variable $Z$ to qualify as an instrument

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

relative to the pair $(X, Y)$. This inequality is sharp for binary valued $X$, but becomes loose when the cardinality of $X$ increases.[17]

If all observed variables are binary, Eq. (34) reduces to the four inequalities

$$\begin{aligned} P(Y=0, X=0|Z=0) &+ P(Y=1, X=0|Z=1) \leq 1 \\ P(Y=0, X=1|Z=0) &+ P(Y=1, X=1|Z=1) \leq 1 \\ P(Y=1, X=0|Z=0) &+ P(Y=0, X=0|Z=1) \leq 1 \\ P(Y=1, X=1|Z=0) &+ P(Y=0, X=1|Z=1) \leq 1 \end{aligned} \tag{35}$$

We see that the instrumental inequality is violated when the controlling instrument $Z$ manages to produce significant changes in the response variable $Y$ while the direct cause, $X$, remains constant.

The instrumental inequality can be used in the detection of undesirable side-effects. Violations of this inequality can be attributed to one of two possibilities: either there is a direct causal effect of the assignment $(Z)$ on the response $(Y)$, unmediated by the treatment $(X)$, or there is a common causal factor influencing both variables. If the assignment is carefully randomized, then the latter possibility is ruled out and any violation of the instrumental inequality (even under conditions of imperfect compliance) can safely be attributed to some direct influence of the assignment process on subjects' response (e.g., psychological aversion to being treated). Alternatively, if one can rule out any direct effects of $Z$ on $Y$, say through effective use of a placebo, then any observed violation of the instrumental inequality can safely be attributed to spurious dependence between $Z$ and $U_Y$, namely, to selection bias.

## 4. The potential outcome framework

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

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y_x(u)$, read: "the value that outcome $Y$ would obtain in experimental unit $u$, had treatment $X$ been $x$." Here, *unit* may stand for an individual patient, an experimental subject, or an agricultural plot. In Section 3.4 (Eq. (27) we saw that this counterfactual entity has a natural interpretation in the SCM; it is the solution for $Y$ in a modified system

[17] The inequality is sharp in the sense that every distribution $P(x, y, z)$ satisfying Eq. (34) can be generated by the model defined in Fig. 5(a).

of equations, where *unit* is interpreted a vector $u$ of background factors that characterize an experimental unit. Each structural equation model thus carries a collection of assumptions about the behavior of hypothetical units, and these assumptions permit us to derive the counterfactual quantities of interest. In the potential-outcome framework, however, no equations are available for guidance and $Y_x(u)$ is taken as primitive, that is, an undefined quantity in terms of which other quantities are defined; not a quantity that can be derived *from* the model. In this sense the structural interpretation of $Y_x(u)$ given in (27) provides the formal basis for the potential-outcome approach; the formation of the submodel $M_x$ explicates mathematically how the hypothetical condition "had $X$ been $x$" is realized, and what the logical consequences are of such a condition.

### *4.1. The "Black-Box" missing-data paradigm*

The distinct characteristic of the potential-outcome approach is that, although investigators must think and communicate in terms of undefined, hypothetical quantities such as $Y_x(u)$, the analysis itself is conducted almost entirely within the axiomatic framework of probability theory. This is accomplished, by postulating a "super" probability function on both hypothetical and real events. If $U$ is treated as a random variable then the value of the counterfactual $Y_x(u)$ becomes a random variable as well, denoted as $Y_x$. The potential-outcome analysis proceeds by treating the observed distribution $P(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function $P^*$ defined over both observed and counterfactual variables. Queries about causal effects (written $P(y|do(x))$ in the structural analysis) are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y_x = y)$. The new hypothetical entities $Y_x$ are treated as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence.

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

$$X = x \implies Y_x = Y, \tag{36}$$

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. When $X$ is binary, it is sometimes more convenient to write (36) as:

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

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

The main conceptual difference between the two approaches is that, whereas the structural approach views the intervention $do(x)$ as an operation that changes

a distribution but keeps the variables the same, the potential-outcome approach views the variable $Y$ under $do(x)$ to be a different variable, $Y_x$, loosely connected to $Y$ through relations such as (36), 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 (36) are automatically satisfied in the structural interpretation and, moreover, that investigators need not be concerned about any additional constraints except the following two:

$$Y_{yz} = y \quad \text{for all } y, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{37}$$

$$X_z = x \Rightarrow Y_{xz} = Y_z \quad \text{for all } x, \text{ subsets } Z, \text{ and values } z \text{ for } Z \tag{38}$$

Equation (37) 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 (38) generalizes (36) 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 Fig. 5(a), to communicate the understanding that $Z$ is randomized (hence independent of $U_X$ and $U_Y$), the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp \{Y_{z_1}, Y_{z_2}, \ldots, Y_{z_k}\}$.[18] To further formulate the understanding that $Z$ does not affect $Y$ directly, except through $X$, the analyst would write a, so called, "exclusion restriction": $Y_{xz} = Y_x$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest. For example, if one can plausibly assume that, in Fig. 4, a set $Z$ of covariates satisfies the conditional independence

$$Y_x \perp\!\!\!\perp X | Z \tag{39}$$

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

[18] 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).

to yield

$$\begin{aligned} P^*(Y_x = y) &= \sum_z P^*(Y_x = y|z)P(z) \\ &= \sum_z P^*(Y_x = y|x,z)P(z) \quad \text{(using (39))} \\ &= \sum_z P^*(Y = y|x,z)P(z) \quad \text{(using (36))} \\ &= \sum_z P(y|x,z)P(z). \end{aligned} \tag{40}$$

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 Eq. (25).

We see that the assumption of conditional ignorability (39) qualifies $Z$ as an admissible covariate for adjustment; it mirrors therefore the "back-door" criterion of Definition 3, which bases the admissibility of $Z$ on an explicit causal structure encoded in the diagram.

The derivation above may explain why the potential-outcome approach appeals to mathematical statisticians; instead of constructing new vocabulary (e.g., arrows), new operators ($do(x)$) and new logic for causal analysis, almost all mathematical operations in this framework are conducted within the safe confines of probability calculus. Save for an occasional application of rule (38) or (36)), the analyst may forget that $Y_x$ stands for a counterfactual quantity—it is treated as any other random variable, and the entire derivation follows the course of routine probability exercises.

This orthodoxy exacts a high cost: Instead of bringing the theory to the problem, the problem must be reformulated to fit the theory; all background knowledge pertaining to a given problem must first be translated into the language of counterfactuals (e.g., ignorability conditions) before analysis can commence. This translation may in fact be the hardest part of the problem. The reader may appreciate this aspect by attempting to judge whether the assumption of conditional ignorability (39), the key to the derivation of (40), 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. 4 would satisfy this conditional independence condition.[19] Likewise, to derive Eq. (39) in the language of potential-outcome (see (Pearl, 2000a, p. 223)), one would need to convey the structure of the chain $X \to W_3 \to Y$ 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

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

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" (39) 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.[20]

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

Aside from offering no guidance in covariate selection, formulating a problem in the potential-outcome language encounters three additional hurdles. When counterfactual variables are not viewed as byproducts of a deeper, process-based model, it is hard to ascertain whether *all* relevant 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 most economists and social scientists continue to use structural equation models (Wooldridge, 2002; Stock and Watson, 2003; Heckman, 2008) instead of the potential-outcome alternatives advocated in Angrist et al. (1996); Holland (1988); Sobel (1998, 2008).

On the other hand, the algebraic machinery offered by the counterfactual notation, $Y_x(u)$, once a problem is properly formalized, can be extremely powerful in refining assumptions (Angrist et al., 1996; Heckman and Vytlacil, 2005), deriving consistent estimands (Robins, 1986), bounding probabilities of necessary and sufficient causation (Tian and Pearl, 2000), and combining data from experimental and nonexperimental studies (Pearl, 2000a). The next subsection (4.3) presents a way of combining the best features of the two approaches. It is based on encoding causal assumptions in the language of diagrams, translating these assumptions into counterfactual notation, performing the mathematics in the algebraic language of counterfactuals (using (36), (37), and (38)) and, finally, interpreting the result in graphical terms or plain causal language. The mediation problem of Section 5.1 illustrates how such symbiosis clarifies the definition and identification of direct and indirect effects.

In contrast, when the mediation problem is approached from an orthodox potential-outcome viewpoint, void of the structural guidance of Eq. (27), 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

[20]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 Fig. 4, $Z$ in Fig. 5a) may actually *increase* confounding bias in propensity-score matching. Paul Rosenbaum, for example, writes: "there is little or no reason to avoid adjustment for a true covariate, 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, 2009b,c; Heckman and Navarro-Lozano, 2004) for a discussion of this fallacy).

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 a 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. (27).[21]

### *4.3. Combining graphs and potential outcomes*

The formulation of causal assumptions using graphs was discussed in Section 3. In this subsection we will systematize the translation of these assumptions from graphs to counterfactual notation.

Structural equation models embody causal information in both the equations and the probability function $P(u)$ assigned to the exogenous variables; the former is encoded as missing arrows in the diagrams the latter as missing (double arrows) dashed arcs. Each parent-child family $(PA_i, X_i)$ in a causal diagram $G$ corresponds to an equation in the model $M$. Hence, missing arrows encode exclusion assumptions, that is, claims that manipulating variables that are excluded from an equation will not change the outcome of the hypothetical experiment described by that equation. Missing dashed arcs encode independencies among error terms in two or more equations. For example, the absence of dashed arcs between a node $Y$ and a set of nodes $\{Z_1, \ldots, Z_k\}$ implies that the corresponding background variables, $U_Y$ and $\{U_{Z_1}, \ldots, U_{Z_k}\}$, are independent in $P(u)$.

These assumptions can be translated into the potential-outcome notation using two simple rules (Pearl, 2000a, p. 232); the first interprets the missing arrows in the graph, the second, the missing dashed arcs.

1. *Exclusion restrictions:* For every variable $Y$ having parents $PA_Y$ and for every set of endogenous variables $S$ disjoint of $PA_Y$, we have

$$Y_{pa_Y} = Y_{pa_Y, s}. \tag{41}$$

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

$$Y_{pa_Y} \perp\!\!\!\perp \{Z_{1\ pa_1}, \ldots, Z_{k\ pa_k}\}. \tag{42}$$

The exclusion restrictions expresses the fact that each parent set includes *all* direct causes of the child variable, hence, fixing the parents of $Y$, determines the value of $Y$ uniquely, and intervention on any other set $S$ of (endogenous) variables can no longer affect $Y$. The independence restriction translates the

[21] Such symbiosis is now standard in epidemiology research (Robins, 2001; Petersen et al., 2006; VanderWeele and Robins, 2007; Hafeman and Schwartz, 2009; VanderWeele, 2009) yet still lacking in econometrics (Heckman, 2008; Imbens and Wooldridge, 2009).

independence between $U_Y$ and $\{U_{Z_1}, \ldots, U_{Z_k}\}$ into independence between the corresponding potential-outcome variables. This follows from the observation that, once we set their parents, the variables in $\{Y, Z_1, \ldots, Z_k\}$ stand in functional relationships to the $U$ terms in their corresponding equations.

As an example, the model shown in Fig. 5(a) displays the following parent sets:

$$PA_Z = \{\emptyset\},\ PA_X = \{Z\},\ PA_Y = \{X\}. \tag{43}$$

Consequently, the exclusion restrictions translate into:

$$\begin{aligned} X_z &= X_{yz} \\ Z_y &= Z_{xy} = Z_x = Z \\ Y_x &= Y_{xz} \end{aligned} \tag{44}$$

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

$$Z \perp\!\!\!\perp \{Y_x, X_z\}. \tag{45}$$

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

The role of graphs is not ended with the formulation of causal assumptions. Throughout an algebraic derivation, like the one shown in Eq. (40), the analyst may need to employ additional assumptions that are entailed by the original exclusion and independence assumptions, yet are not shown explicitly in their respective algebraic expressions. For example, it is hardly straightforward to show that the assumptions of Eqs. (44)–(45) imply the conditional independence $(Y_x \perp\!\!\!\perp Z | \{X_z, X\})$ but do not imply the conditional independence $(Y_x \perp\!\!\!\perp Z | X)$. These are not easily derived by algebraic means alone. Such implications can, however, easily be tested in the graph of Fig. 5(a) using the graphical reading for conditional independence (Definition 1). (See (Pearl, 2000a, pp. 16–17, 213–215).) Thus, when the need arises to employ independencies in the course of a derivation, the graph may assist the procedure by vividly displaying the independencies that logically follow from our assumptions.

## 5. Counterfactuals at work

### *5.1. Mediation: Direct and indirect effects*

#### *5.1.1. 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 $Y$ (Pearl, 1998; Cole and Hernán, 2002). 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 sustain the independence 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 $Y$.[22]

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

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

or, equivalently, using counterfactual notation:

$$CDE \triangleq E(Y_{x'z}) - E(Y_{xz}) \tag{46}$$

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)) using sequential application of the back-door condition (Definition 3).

[22] According to Rubin (2004, 2005), R.A. Fisher made this mistake in the context of agriculture experiments. Fisher, in fairness, did not have graphs for guidance.

*5.1.2. Natural direct effects*

In linear systems, Eq. (46) yields the path coefficient of the link from $X$ to $Y$; independent of the values at which we hold $Z$, independent of the distribution of the error terms, and regardless of whether those coefficients are identifiable or not. In nonlinear systems, the values at which we hold $Z$ 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 define the direct effect relative to a "natural representative" of those levels or, more specifically, 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) 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":

$$DE_{x,x'}(Y) \triangleq E(Y_{x',Z_x}) - E(Y_x). \tag{47}$$

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 original 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 "unconfoundedness" are deemed valid, the natural direct effect can be reduced to

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

The intuition is simple; the natural direct effect is the weighted average of the controlled direct effect (46), using the causal effect $P(z|do(x))$ as a weighing function.

One sufficient condition for the identification of (47) is that $Z_x \perp\!\!\!\perp Y_{x',z}|W$ holds for some set $W$ of measured covariates. However, this condition in itself, like the ignorability condition of (42), is close to meaningless for most investigators, as it is not phrased in terms of realized variables. The symbiotic analysis

of Section 4.3 can be invoked at this point to unveil the graphical interpretation of this condition (through Eq. (45).) It states that $W$ should be admissible (i.e., satisfy the back-door condition) relative the path(s) from $Z$ to $Y$. This condition is readily comprehended by empirical researchers, and the task of selecting such measurements, $W$, can then be guided by the available scientific knowledge. See details and graphical criteria in Pearl (2001, 2005) and in Petersen et al. (2006).

In particular, expression (48) is both valid and identifiable in Markovian models, where each term on the right can be reduced to a "*do*-free" expression using Eq. (24).

### *5.1.3. Indirect effects and the Mediation Formula*

Remarkably, the definition of the natural direct effect (47) 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):

$$IE_{x,x'}(Y) \triangleq E((Y_{x,Z_{x'}}) - E(Y_x)), \tag{49}$$

which is almost identical to the direct effect (Eq. (47)) save for exchanging $x$ and $x'$.

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

$$TE_{x,x'}(Y) \triangleq E(Y_{x'} - Y_x) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{50}$$

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

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{51}$$

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.

For completeness, we explicate (from (48) and (51)) the expression for indirect effects under conditions of nonconfoundedness:

$$IE_{x,x'}(Y) = \sum_z E(Y|x,z)[P(z|x') - P(z|x)] \tag{52}$$

This expression deserves the label *Mediation Formula*, due to its pivotal role in mediation analysis (Imai et al., 2008), which has been a thorny issue in several

sciences (Shrout and Bolger, 2002; MacKinnon et al., 2007; Mortensen et al., 2009). 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. In addition to providing researchers with a principled, parametric-free target quantity that is valid in both linear and nonlinear models, the formula can also serve as an analytical laboratory for testing the effectiveness of various estimation techniques under various types of model mispecification (VanderWeele, 2009).

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 testing the former in experimental setup. The mantra "No causation without manipulation" must be rejected. (See (Pearl, 2000a, Section 11.4.5.).)

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 the next section and 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).

### 5.2. Causes of effects and probabilities of causation

The likelihood that one event *was the cause* of another guides much of what we understand about the world (and how we act in it). For example, knowing whether it was the aspirin that cured my headache or the TV program I was watching would surely affect my future use of aspirin. Likewise, to take an example from common judicial standard, judgment in favor of a plaintiff should be made if and only if it is "more probable than not" that the damage would not have occurred *but for* the defendant's action (Robertson, 1997).

These two examples fall under the category of "causes of effects" because they concern situations in which we observe both the effect, $Y = y$, and the putative cause $X = x$ and we are asked to assess, counterfactually, whether the former would have occurred absent the latter.

We have remarked earlier (footnote 11) that counterfactual probabilities conditioned on the outcome cannot in general be identified from observational or even experimental studies. This does not mean however that such probabilities are useless or void of empirical content; the structural perspective may guide us in fact toward discovering the conditions under which they can be assessed from data, thus defining the empirical content of these counterfactuals.

Following the 4-step process of structural methodology – define, assume, identify, and estimate – our first step is to express the target quantity in counterfactual notation and verify that it is well defined, namely, that it can be computed unambiguously from any fully-specified causal model.

In our case, this step is simple. Assuming binary events, with $X = x$ and $Y = y$ representing treatment and outcome, respectively, and $X = x'$, $Y = y'$ their negations, our target quantity can be formulated directly from the English sentence:

> "Find the probability that $Y$ would be $y'$ had $X$ been $x'$, given that, in reality, $Y$ is actually $y$ and $X$ is $x$,"

to give:

$$PN(x, y) = P(Y_{x'} = y' | X = x, Y = y) \tag{53}$$

This counterfactual quantity, which Robins and Greenland (1989a) named "probability of causation" and Pearl (2000a, p. 296) named "probability of necessity" (PN), to be distinguished from other nuances of "causation," is certainly computable from any fully specified structural model, i.e., one in which $P(u)$ and all functional relationships are given. This follows from the fact that every structural model defines a joint distribution of counterfactuals, through Eq. (27).

Having written a formal expression for PN, Eq. (53), we can move on to the formulation and identification phases and ask what assumptions would permit us to identify PN from empirical studies, be they observational, experimental or a combination thereof.

This problem was analyzed by Pearl (2000a, Chapter 9) and yielded the following results:

**Theorem 3.** *If $Y$ is monotonic relative to $X$, i.e., $Y_1(u) \geq Y_0(u)$, then* PN *is identifiable whenever the causal effect $P(y|do(x))$ is identifiable and, moreover,*

$$\text{PN} = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x, y)}. \tag{54}$$

The first term on the r.h.s. of (54) is the familiar excess risk ratio (ERR) that epidemiologists have been using as a surrogate for PN in court cases (Cole, 1997; Robins and Greenland, 1989a). The second term represents the *correction* needed to account for confounding bias, that is, $P(y|do(x')) \neq P(y|x')$.

This suggests that monotonicity and unconfoundedness were tacitly assumed by the many authors who proposed or derived ERR as a measure for the "fraction of exposed cases that are attributable to the exposure" (Greenland, 1999).

Equation (54) thus provides a more refined measure of causation, which can be used in situations where the causal effect $P(y|do(x))$ can be estimated from either randomized trials or graph-assisted observational studies (e.g., through Theorem 2 or Eq. (25)). It can also be shown (Tian and Pearl, 2000) that the expression in (54) provides a lower bound for PN in the general, nonmonotonic case. (See also (Robins and Greenland, 1989b).) In particular, the tight upper and lower bounds on PN are given by:

$$\max\left\{0, \frac{P(y)-P(y|do(x'))}{P(x,y)}\right\} \leq PN \leq \min\left\{1, \frac{P(y'|do(x'))-P(x',y')}{P(x,y)}\right\} \quad (55)$$

It is worth noting that, in drug related litigation, it is not uncommon to obtain data from both experimental and observational studies. The former is usually available at the manufacturer or the agency that approved the drug for distribution (e.g., FDA), while the latter is easy to obtain by random surveys of the population. In such cases, the standard lower bound used by epidemiologists to establish legal responsibility, the Excess Risk Ratio, can be substantially improved using the lower bound of Eq. (55). Likewise, the upper bound of Eq. (55) can be used to exonerate drug-makers from legal responsibility. Cai and Kuroki (2006) analyzed the statistical properties of PN.

Pearl (2000a, p. 302) shows that combining data from experimental and observational studies which, taken separately, may indicate no causal relations between $X$ and $Y$, can nevertheless bring the lower bound of Eq. (55) to unity, thus implying causation *with probability one*.

Such extreme results dispel all fears and trepidations concerning the empirical content of counterfactuals (Dawid, 2000; Pearl, 2000b). They demonstrate that a quantity PN which at first glance appears to be hypothetical, ill-defined, untestable and, hence, unworthy of scientific analysis is nevertheless definable, testable and, in certain cases, even identifiable. Moreover, the fact that, under certain combination of data, and making no assumptions whatsoever, an important legal claim such as "the plaintiff would be alive had he not taken the drug" can be ascertained with probability one, is a remarkable tribute to formal analysis.

Another counterfactual quantity that has been fully characterized recently is the Effect of Treatment on the Treated (ETT):

$$ETT = P(Y_x = y|X = x')$$

ETT has been used in econometrics to evaluate the effectiveness of social programs on their participants (Heckman, 1992) and has long been the target of research in epidemiology, where it came to be known as "the effect of exposure on the exposed," or "standardized morbidity" (Miettinen, 1974; Greenland and Robins, 1986).

Shpitser and Pearl (2009) have derived a complete characterization of those models in which ETT can be identified from either experimental or observa-

tional studies. They have shown that, despite its blatant counterfactual character, (e.g., "I just took an aspirin, perhaps I shouldn't have?") ETT can be evaluated from experimental studies in many, though not all cases. It can also be evaluated from observational studies whenever a sufficient set of covariates can be measured that satisfies the back-door criterion and, more generally, in a wide class of graphs that permit the identification of conditional interventions.

These results further illuminate the empirical content of counterfactuals and their essential role in causal analysis. They prove once again the triumph of logic and analysis over traditions that a-priori exclude from the analysis quantities that are not testable in isolation. Most of all, they demonstrate the effectiveness and viability of the *scientific* approach to causation whereby the dominant paradigm is to model the activities of Nature, rather than those of the experimenter. In contrast to the ruling paradigm of conservative statistics, we begin with relationships that we know in advance will never be estimated, tested or falsified. Only after assembling a host of such relationships and judging them to faithfully represent our theory about how Nature operates, we ask whether the parameter of interest, crisply defined in terms of those theoretical relationships, can be estimated consistently from empirical data and how. It often does, to the credit of progressive statistics.

## 6. Conclusions

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

## References

Angrist, J. and Imbens, G. (1991). Source of identifying information in evaluation models. Tech. Rep. Discussion Paper 1568, Department of Economics, Harvard University, Cambridge, MA.

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

Arah, O. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: Covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology* **4** doi:10.1186/1742–7622–5–5. Online at <http://www.ete-online.com/content/5/1/5>.

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

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

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

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

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

Bishop, Y., Fienberg, S. and Holland, P. (1975). *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, MA.

Blyth, C. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* **67** 364–366.

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

Bonet, B. (2001). Instrumentality tests revisited. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, 48–55.

Bowden, R. and Turkington, D. (1984). *Instrumental Variables*. Cambridge University Press, Cambridge, England.

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

Cai, Z. and Kuroki, M. (2006). Variance estimators for three 'probabilities of causation'. *Risk Analysis* **25** 1611–1620.

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

Chen, A., Bengtsson, T. and Ho, T. (2009). A regression paradox for linear models: Sufficient conditions and relation to Simpson's paradox. *The American Statistician* **63** 218–225.

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

Cole, P. (1997). Causality in epidemiology, health policy, and law. *Journal of*

*Marketing Research* **27** 10279–10285.

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

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

Cox, D. and Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **65** 937–941.

Cox, D. and Wermuth, N. (2004). Causality: A statistical view. *International Statistical Review* **72** 285–305.

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

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

Dawid, A. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review* **70** 161–189.

DeFinetti, B. (1974). *Theory of Probability: A Critical Introductory Treatment*. Wiley, London. 2 volumes. Translated by A. Machi and A. Smith.

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

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

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

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

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

Goldberger, A. (1973). Structural equation models: An overview. In *Structural Equation Models in the Social Sciences* (A. Goldberger and O. Duncan, eds.). Seminar Press, New York, NY, 1–18.

Good, I. and Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics* **15** 694–711.

Greenland, S. (1999). Relation of probability of causation, relative risk, and doubling dose: A methodologic error that has become a social problem. *American Journal of Public Health* **89** 1166–1169.

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

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

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

Hafeman, D. and Schwartz, S. (2009). Opening the black box: A motivation

for the assessment of mediation. *International Journal of Epidemiology* **3** 838–845.

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

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

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

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

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

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

Imai, K., Keele, L. and Yamamoto, T. (2008). Identification, inference, and sensitivity analysis for causal mediation effects. Tech. rep., Department of Politics, Princton University.

Imbens, G. and Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**.

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

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

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

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press, Oxford.

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

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

Lindley, D. and Novick, M. (1981). The role of exchangeability in inference. *The Annals of Statistics* **9** 45–58.

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

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

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

Meek, C. and Glymour, C. (1994). Conditioning and intervening. *British*

*Journal of Philosophy Science* **45** 1001–1021.

Miettinen, O. (1974). Proportion of disease caused or prevented by a given exposure, trait, or intervention. *Journal of Epidemiology* **99** 325–332.

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

Mortensen, L., Diderichsen, F., Smith, G. and Andersen, A. (2009). The social gradient in birthweight at term: quantification of the mediating role of maternal smoking and body mass index. *Human Reproduction* To appear, doi:10.1093/humrep/dep211.

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

Pavlides, M. and Perlman, M. (2009). How likely is Simpson's paradox? *The American Statistician* **63** 226–233.

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

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

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

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

Pearl, J. (1995b). On the testability of causal models with latent and instrumental variables. In *Uncertainty in Artificial Intelligence 11* (P. Besnard and S. Hanks, eds.). Morgan Kaufmann, San Francisco, CA, 435–443.

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

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

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

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

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

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

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

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

Pearl, J. (2009c). Myth, confusion, and science in causal analy-

sis. Tech. Rep. R-348, University of California, Los Angeles, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r348.pdf>.

Pearl, J. and Paz, A. (2009). Confounding equivalence in observational studies. Tech. Rep. TR-343, University of California, Los Angeles, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r343.pdf>.

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

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

Pearson, K., Lee, A. and Bramley-Moore, L. (1899). Genetic (reproductive) selection: Inheritance of fertility in man. *Philosophical Transactions of the Royal Society A* **73** 534–539.

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

Robertson, D. (1997). The common sense of cause in fact. *Texas Law Review* **75** 1765–1800.

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

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

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

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

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

Robins, J. and Greenland, S. (1989a). The probability of causation under a stochastic model for individual risk. *Biometrics* **45** 1125–1138.

Robins, J. and Greenland, S. (1989b). Estimability and estimation of excess and etiologic fractions. *Statistics in Medicine* **8** 845–859.

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

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

Rosenbaum, P. and Rubin, D. (1983). The central role of propensity score in

observational studies for causal effects. *Biometrika* **70** 41–55.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **13** 238–241.

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

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

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag, New York.

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

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

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

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

Tian, J., Paz, A. and Pearl, J. (1998). Finding minimal separating sets. Tech. Rep. R-254, University of California, Los Angeles, CA.

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

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

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

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

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Science+Business Media, Inc., New York, NY.

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

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

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

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

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

Wooldridge, J. (2009). Should instrumental variables be used as matching variables? Tech. Rep. <https://www.msu.edu/∼ec/faculty/wooldridge/current%20research/treat1r6.pdf>, Michigan State University, MI.

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

Yule, G. (1903). Notes on the theory of association of attributes in statistics. *Biometrika* **2** 121–134.