Probabilities of causation: Three counterfactual interpretations and their identification

Judea Pearl

Cognitive Systems Laboratory

Computer Science Department

University of California, Los Angeles, CA 90024

judea@cs.ucla.edu

1 Introduction

In a recent report [Pearl, 1998] I have proposed a definition of "actual cause" and of the probability that an event C was an actual cause of event E in a given scenario. I have since come to realize that the standard counterfactual definition of causation[1] (i.e., that E would not have occurred if it were not for C), captures only the notion of "necessary cause." Competing notions such as "sufficient cause" and "necessary-and-sufficient cause" may be of great interest in a number of application,[2] and these, too, can be given concise counterfactual definitions. One advantage of casting aspects of causation in the language of counterfactuals is that the latter enjoys natural and formal semantics in terms of structural

[1]This definition dates back to Hume (1748, p. 115) and Mill (1843) and has been advocated in both the philosophical work of D. Lewis (1973) and the statistical analyses of J. Neyman (1923) and D. Rubin (1974).

[2]The importance of the distinction between between necessary and sufficient causes goes back to J.S. Mill (1843), and has received semi-formal explications in the 1960s using the syntax of conditional probabilities [Good, 1961] and logical implications [Mackie, 1965]. The basic limitations of the logical and probabilistic accounts are discussed in Pearl (1996, 1998) and stem primarily from lacking syntactic distinction between formulae that represent stable mechanisms and those that represent transitory logical or probabilistic relationships.

models [Galles and Pearl, 1997, 1998; Halpern, 1998], as well as effective procedures for computing probabilities of counterfactual expressions from a given causal theory [Balke and Pearl, 1994, 1995].

The purpose of this paper is to explore the counterfactual interpretation of necessary and sufficient causes, to illustrate the application of structural-model semantics (of counterfactuals) to the problem of identifying probabilities of causes, and to present, by way of examples, new ways of estimating probabilities of causes from statistical data.

The results have applications in epidemiology, legal reasoning, artificial intelligence (AI), and psychology. Epidemiologists have long been concerned with estimating the probability that a certain case of disease is *attributable* to a particular exposure, which is normally interpreted counterfactually as "the probability that disease would not have occurred in the absence of exposure, given that disease and exposure did in fact occur." This counterfactual notion, which Robins and Greenland (1989) called the "probability of causation" measures how *necessary* the cause is for the production of the effect.[3] It is used frequently in law suits, where legal responsibility is at the center of contention. We shall denote this notion by the symbol PN, an acronym for Probability of Necessity.

A parallel notion of causation, capturing how *sufficient* a cause is for the production of the effect, finds applications in policy analysis, AI, and psychology. A policy maker may well be interested in the dangers that a certain exposure may present to the healthy population [Khoury et al., 1989]. Counterfactually, this notion is expressed as the "probability that a healthy unexposed individual would have gotten the disease had he/she been exposed," and will be denoted by PS (Probability of Sufficiency). A natural extension would be to inquire for the probability PNS of necessary-and-sufficient causation, namely, how likely a given individual is to be affected both ways.

To summarize, PS assesses the presence of an active causal process capable of producing the effect, while PN emphasizes the absence of alternative processes, not involving the cause

[3] Greenland and Robins (1988) further distinguish between two ways of measuring probabilities of causation: the first (called "excess fraction") concerns only *whether* the effect (e.g., disease) occurs by a particular time, while the second, (called "etiological fraction") requires consideration of *when* the effect occurs. We will confine our discussion here to binary events occurring within a specified time period, hence, will not be concerned with the temporal aspects of etiological fractions.

in question, still capable of sustaining the effect. In legal settings, where the occurrence of the cause (x) and the effect (y) are fairly well established, PN is the measure that draws most attention, and the plaintiff must prove that y would not have occurred *but for* x [Robertson, 1997]. Still, lack of sufficiency may weaken arguments based on PN [Good, 1993; Michie, 1997].

It is known that PN is in general non-identifiable, namely, non-estimatable from frequency data involving exposures and disease cases [Greenland and Robins, 1988; Robins and Greenland, 1989]. The identification is hindered by two factors:

1. **Confounding**: exposed and unexposed subjects may differ in several relevant factors or, more generally, the cause and the effect may both be influenced by a third factor. In this case we say that the cause is not *exogenous* relative to the effect.

2. **Sensitivity to the generative process**: Even in the absence of confounding, probabilities of certain counterfactual relationships cannot be identified from probabilistic information unless we specify the functional relationships that connect causes and effects. Functional specification is needed whenever the evidence at hand involves variables that might be affected by the counterfactual antecedent [Balke and Pearl, 1994b]. In evaluating PN and PS, for example, identifiability is threatened because the evidence at hand consists of the current status of the outcome (e.g., disease), and that status is affected by the counterfactual antecedent (e.g., exposure).

Although PN is not identifiable in the general case, several formulas have nevertheless been proposed to estimate PN in terms of frequencies obtained in epidemiological studies [Bresslow and Day, 1980; Hennekens and Buring, 1987; Cole, 1997]. Naturally, any such formula must be predicated upon certain implicit assumptions about the data-generating process. This paper explicates some of those assumptions and explores conditions under which they can be relaxed.[4] It offers new formulas for PN and PS in cases where causes are confounded (with outcomes) but their effect can nevertheless be estimated (e.g., from

[4] A set of sufficient conditions for the identification of etiological fractions are given in Robins and Greenland (1989). These conditions, however, are too restrictive for the identification of PN, which is oblivious to the temporal aspects associated with etiological fractions.

clinical trials or from auxiliary measurements). We further provide a general condition for the identifiability of PN and PS when functional relationships are not known.

Clark Glymour (1998) has raised a number of issues concerning the identifiability of causal relationships when the functional relationships among the variables *are* known, but some variables are unobserved. These issues surfaced in connection with Cheng's model, according to which people assess the "causal power" between two events by estimating, using both frequency data and functional relationships, the probability that the effect will take place in a certain hypothetical model of the world [Cheng, 1997]. In the examples provided, Cheng's "causal power" coincides with PS and hence lends itself to counterfactual analysis. Accordingly we shall see that many of the issues raised by Glymour can be resolved and generalized using the counterfactual analysis of Balke and Pearl (1994).

The distinction between *necessary* and *sufficient* causes has important implications in AI, especially in systems that generate verbal explanations automatically. As can be seen from the epidemiological examples above, necessary causation is a concept tailored to a specific event under consideration, while sufficient causation is based on the general tendency of certain event *types* to produce other event types. Adequate explanations should respect both aspects. If we base explanations solely on generic tendencies (i.e., *sufficient* causation), we lose important specific information. For instance, aiming and shooting a person from 1000 meters away will not qualify as an explanation for that person's death, due to the very low tendency of typical shots fired from such long distances to hit their marks. If, on the other hand, we base explanations solely on singular-event considerations (i.e., *necessary* causation), then various background factors which are normally present in the world would awkwardly qualify as explanations. For example, the presence of oxygen in the room would qualify as an explanation for the fire that broke out, simply because the fire would not have occurred if it were not for the oxygen and, given that the fire did in fact occur, there is no question that a match has been struck. Clearly, some balance must be made between the necessary and the sufficient components of causal explanation, and the present paper illuminates this balance by formally explicating some of the basic relationships between the two components.

2 Necessary and Sufficient Causes: Conditions of Identification

2.1 Definitions, Notation, and Basic Relationships

The reader is assumed to be familiar with the counterfactual notation of Neyman-Rubin's model [Rubin 1990; Robins 1986] and with the structural-equation semantics of counterfactuals as defined in Balke and Pearl (1995), Galles and Pearl (1997, 1998), and Halpern (1998). A brief summary is presented in the Appendix to this paper. Using this notation and semantics, we give the following definitions for the three aspects of causation discussed in the introduction.

Definition 1 (*probability of necessity* (*PN*))
Let X and Y be two binary variables in a causal model M, let x and y stand for the propositions $X = \text{true}$ and $Y = \text{true}$, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$\begin{aligned} PN &\triangleq P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true}) \\ &\triangleq P(y'_{x'}|x, y) \end{aligned} \tag{1}$$

In words, PN stands for the probability that event y would not have occurred in the absence of event x, ($y'_{x'}$), given that x and y did in fact occur.[5]

Definition 2 (*Probability of sufficiency* (*PS*))

$$PS \triangleq P(y_x|y', x') \tag{2}$$

[5]Note the abbreviation y_x for $Y_x = true$ and y'_x for $Y_x = false$. These were proposed by Peyman Meshkat in class homework, and substantially simplify the derivations.

PS measures the capacity of x to *produce* y and, since "production" implies a transition from the absence to the presence of x and y, we condition the probability $P(y_x)$ on situations where x and y are both absent. It is a mirror image of the necessity of x, as measured by PN.

Definition 3 (*Probability of Necessity and Sufficiency (PNS)*)

$$PNS \triangleq P(y_x, y'_{x'}) \tag{3}$$

PNS stands for the probability that y would respond to x both ways, and therefore measures both the sufficiency and necessity of x to produce y.

Associated with these three basic notions, there are other counterfactual quantities that have attracted either practical or conceptual interest. We will mention two such quantities, but will not dwell on their analyses, since these can be easily inferred from our treatment of PN, PS, and PNS.

Definition 4 (*Probability of disablement (PD)*)

$$PD \triangleq P(y'_{x'}|y) \tag{4}$$

PD measures the probability that y would have been prevented if it were not for x; it is therefore of interest to policy makers who wish to assess the social effectiveness of various prevention programs.

Definition 5 (*Probability of enablement (PE)*)

$$PE \triangleq P(y_x|y')$$

PE is similar to PS, save for the fact that we do not condition on x'. It is applicable, for example, when we wish to assess the danger of an exposure on the entire population of healthy individuals, including those that were already exposed.

Although none of these quantities is sufficient for determining the others, they are not entirely independent, as shown in the next lemma.

Lemma 1 *The probabilities of causation, PNS, PN and PS satisfy the following relationship:*

$$PNS = P(x, y)PN + P(x', y')PS \tag{5}$$

Proof of Lemma 1

Using the consistency conditions [Robins, 1986]

$$x \Rightarrow (y_x = y), \quad x' \Rightarrow (y_{x'} = y)$$

we expand $P(y_x, y'_{x'})$ and obtain:

$$\begin{aligned} P(y_x, y'_{x'}) &= P(y_x, y'_{x'}|x)P(x) + P(y_x, y'_{x'}|x')P(x') \\ &= P(y'_{x'}, y|x)P(x) + P(y_x, y'|x')P(x') \\ &= P(y'_{x'}|y, x)P(y|x)P(x) + P(y_x|y', x')P(y', x') \end{aligned}$$

which proves Lemma 1. □

Definition 6 (*Identifiability*)

Let $Q(M)$ be any quantity defined on a causal model M. Q is identifiable *in a class $\boldsymbol{M}$ of models if any two models M_1 and M_2 from $\boldsymbol{M}$ that satisfy $P_{M_1}(v) = P_{M_2}(v)$ also satisfy $Q(M_1) = Q(M_2)$. In words, Q is identifiable if it can be determined uniquely from the probability distribution $P(v)$ of the endogenous variables V.*

The class $\boldsymbol{M}$ that we will consider when discussing identifiability will be determined by assumptions that one is willing to make about the model under study. For example, if our assumptions consist of the structure of a causal graph G_0, $\boldsymbol{M}$ will consist of all models M for which $G(M) = G_0$. If, in addition to G_0, we are also willing to make assumptions about the functional form of some mechanisms in M, $\boldsymbol{M}$ will consist of all models M that incorporate those mechanisms, and so on.

Since all the causal measures defined above invoke conditionalization on y, and since y is presumed affected by x, the antecedent of the the counterfactual y_x, we know that none of these quantities is identifiable from knowledge of the structure $G(M)$ and the data $P(v)$ alone, even under condition of no confounding. Moreover, none of these quantities determines the others in the most general case. However, simple interrelationships and useful bounds can be derived for these quantities under the assumption of no-confounding, an assumption that we call *exogeneity*.

2.2 Bounds and basic relationships under exogeneity

Definition 7 (*Exogeneity*)
A variable X is said to be exogenous relative to Y in model M iff

$$P(y_x, y_{x'}|x) = P(y_x, y_{x'}) \tag{6}$$

namely, the way Y would potentially respond to conditions x or x' is independent of the actual value of X.

Eq. (6) has been given a variety of (equivalent) definitions and interpretations. Epidemiologists refer to this condition as "no-confounding" [Robins and Greenland, 1989], statisticians call it "as if randomized," and Rosenbaum and Rubin (1983) call it "ignorability." A graphical criterion ensuring exogeneity is the absence of a common ancestor of X and Y in $G(M)$. The classical econometric criterion for exogeneity (e.g., Dhrymes (1970, p. 169) states that X be independent of the error term in the equation for Y.[6]

The importance of exogeneity lies in permitting the identification of $P(y_x)$, the *causal effect* of X on Y, since (using $x \Rightarrow (y_x = y)$)

$$P(y_x) = P(y_x|x) = P(y|x) \tag{7}$$

with similar reduction for $P(y_{x'})$.

[6]This criterion has been the subject of relentless objections by modern econometricians [Engle et al., 1983; Hendry, 1995; Imbens, 1997], but see Aldrich (1993) and Galles and Pearl (1998), for a reconciliatory perspective on this controversy.

Theorem 1 *Under condition of exogeneity, PNS is bounded as follows:*

$$\max[0, P(y|x) + P(y'|x') - 1] \leq PNS \leq \min[P(y|x), P(y'|x')] \tag{8}$$

Both bounds are tight in the sense that for every joint distribution $P(x, y)$ there exists a model $y = f(x, u)$, with u independent of x, that realizes any value of PNS permitted by the bounds.

Proof of Theorem 1:

For any two events A and B we have the tight bounds:

$$\max[0, P(A) + P(B) - 1] \leq P(A, B) \leq \min[P(A), P(B)] \tag{9}$$

Eq. (8) follows from (9) using $A = y_x$, $B = y'_{x'}$, $P(y_x) = P(y|x)$ and $P(y'_{x'}) = P(y'|x')$ $\Box$

Clearly, if exogeneity cannot be ascertained, then PNS is bound by inequalities similar to those of Eq. (8), with $P(y_x)$ and $P(y'_{x'})$ replacing $P(y|x)$ and $P(y'|x')$, respectively.

Theorem 2 *Under condition of exogeneity, the probabilities PN, PS, and PNS are related to each other as follows:*

$$PN = \frac{PNS}{P(y|x)} \tag{10}$$

$$PS = \frac{PNS}{1 - P(y|x')} \tag{11}$$

Thus, the bounds for PNS in Eq. (8) provide corresponding bounds for PN and PS.

The resulting bounds for PN

$$\frac{\max[0, P(y|x) + P(y'|x') - 1]}{P(y|x)} \leq PN \leq \frac{\min[P(y|x), P(y'|x')]}{P(y|x)} \tag{12}$$

were established by Robins and Greenland (1989) using a stochastic model of $Y_x(u)$. Theorem 1 reaffirms the validity of these bounds for deterministic $Y_x(u)$, provided the functional relationships in the model, $x_i = f_i(pa_i, u)$, are not known.

Proof of Theorem 2:

Using $x \Rightarrow (y_x = y)$, we can write $x \wedge y_x = x \wedge y$, and obtain

$$\begin{aligned} PN = P(y'_{x'}|x, y) &= P(y'_{x'}, x, y)/P(x, y) && (13)\\ &= P(y'_{x'}, x, y_x)/P(x, y) && (14)\\ &= P(y'_{x'}, y_x)P(x)/P(x, y) && (15)\\ &= \frac{PNS}{P(y|x)} && (16) \end{aligned}$$

which establishes Eq. (10). Eq. (11) follows by identical steps. □

For completion, we note the relationship between PNS and the probabilities of enablement and disablement:

$$PD = \frac{P(x)\ PNS}{P(y)}, \quad PE = \frac{P(x')\ PNS}{P(y')} \tag{17}$$

2.3 Identifiability under monotonicity and exogeneity

Before attacking the general problem of identifying the counterfactual quantities in Eqs. (1)–(3) it is instructive to treat a special condition, called *monotonicity*, which is often assumed in practice, and which renders these quantities identifiable. The resulting probabilistic expressions will be recognized as familiar measures of causation that often appear in the literature.

Definition 8 (*Monotonicity*)
A variable Y is said to be monotonic relative to variable X in a causal model M iff the function $Y_x(u)$ is monotonic in x for all u. Equivalently, Y is monotonic relative to X iff

$$y'_x \wedge y_{x'} = \mathit{false} \tag{18}$$

Monotonicity expresses the assumption that a change from $X = \mathit{false}$ to $X = \mathit{true}$ cannot, under any circumstance make Y change from *true* to *false*.[7] In epidemiology, this assumption is often expressed as "no prevention," that is, no individual in the population

[7]Our analysis remains invariant to complementing x or y (or both), hence, the general condition of monotonicity should read: either $y'_x \wedge y_{x'} = \mathit{false}$ or $y'_{x'} \wedge y_x = \mathit{false}$. For simplicity, however, we will adhere to the definition in Eq. (18).

can be helped by the exposure. Angrist, Imbens, and Rubin (1996) used this assumption to identify treatment effects from studies involving non-compliance (see also Balke and Pearl (1997)). Glymour (1998) and Cheng (1997) resort to this assumption in using disjunctive or conjunctive relationships between causes and effects, excluding functions such as exclusive-or, or parity.

Theorem 3 (*Identifiability under exogeneity and monotonicity*)
If X is exogenous and Y is monotonic relative to X, then the probabilities PN, PS, and PNS are all identifiable, and are given by Eqs. (10)–(11) with

$$PNS = P(y|x) - P(y|x') \tag{19}$$

The r.h.s. of (19) is called "risk-difference" in epidemiology, also misnomered "attributable risk" [Hennekens and Buring, 1987, p. 87].

From (10) we see that the probability of necessity, PN, is identifiable and given by the *excess-risk-ratio*

$$PN = [P(y|x) - P(y|x')]/P(y|x) \tag{20}$$

also misnomered as the *attributable fraction* [Schlesselman, 1982], *attributable-rate percent* [Hennekens and Buring, 1987, p. 88], or *attributable proportion* [Cole, 1997]. Taken literally, the ratio presented in (20) has nothing to do with attribution, since it is made up of statistical terms and not of causal or counterfactual relationships. However, the assumptions of exogeneity and monotonicity together enable us to translate the notion of attribution embedded in the definition of PN (Eq. (1)) into a ratio of purely statistical associations. This suggests that exogeneity and monotonicity were tacitly assumed by authors who proposed or derived Eq. (20) as a measure for the "fraction of exposed cases that are attributable to the exposure."

Robins and Greenland (1989) have analyzed the identification of PN under the assumption of stochastic monotonicity (i.e., $P(Y_x(u) = y) > P(Y_{x'}(u) = y)$) and have shown that this assumption is too weak to permit such identification; in fact, it yields the same bounds as in Eq. (12). This indicates that stochastic monotonicity imposes no constraints whatsoever on the functional mechanisms that mediate between X and Y.

The expression for PS (Eq. (11)), likewise, is quite revealing:

$$PS = [P(y|x) - P(y|x')]/[1 - P(y|x')], \tag{21}$$

as it coincide with what epidemiologists call the "relative difference" [Shep, 1958], which is used to measure the *susceptibility* of a population to a certain risk factor x. Susceptibility is defined as the proportion of persons who possess "an underlying factor sufficient to make a person contract a disease following exposure" [Khoury et al., 1989]. PS offers a formal counterfactual interpretation of susceptibility, which sharpens this definition and renders susceptibility amenable to systematic analysis. Khoury et al. (1989) have recognized that susceptibility in general is not identifiable, and have derived Eq. (21) by making three assumptions: no confounding, monotonicity[8] and independence (i.e., assuming that susceptibility to exposure is independent of susceptibility to background not involving exposure). This last assumption is often criticized as untenable, and Theorem 3 assures us that independence is in fact unnecessary; Eq. (21) attains its validity through exogeneity and monotonicity alone.

Eq. (21) also coincides with what Cheng calls "causal power" (1997), namely, the effect of x on y after suppressing "all other causes of y." The counterfactual definition of PS, $P(y_x|x', y')$, suggests another interpretation of this quantity. It measures the probability that setting x would produce y in a situation where x and y are in fact absent. Conditioning on y' amounts to selecting (or hypothesizing) only those worlds in which "all other causes of y" are indeed suppressed.

It is important to note, however, that the simple relationships among the three notions of causation (Eqs. 10–11) only hold under the assumption of exogeneity; the weaker relationship of Eq. (5) prevails in the more general, non-exogenous case. Additionally, all these notions of causation are defined in terms of the global relationships $Y_x(u)$ and $Y_{x'}(u)$ which, as it is argued in Pearl (1998), is too crude to fully characterize the many nuances of causation;

[8]Monotonicity is not mentioned in [Khoury et al., 1989], but it must have been assumed implicitly to make their derivations valid.

the detailed structure of the causal model leading from X to Y is often needed to explicate more refined notions, such as "actual cause."

Proof of Theorem 3:

Writing $y_{x'} \vee y'_{x'} = true$, we have

$$y_x = y_x \wedge (y_{x'} \vee y'_{x'}) = (y_x \wedge y_{x'}) \vee (y_x \wedge y'_{x'}) \tag{22}$$

and

$$y_{x'} = y_{x'} \wedge (y_x \vee y'_x) = (y_{x'} \wedge y_x) \vee (y_{x'} \wedge y'_x) = y_{x'} \wedge y_x \tag{23}$$

since monotonicity entails $y_{x'} \wedge y'_x = false$. Substituting (23) into (22) yields

$$y_x = y_{x'} \vee (y_x \wedge y'_{x'}) \tag{24}$$

Taking the probability of (24), and using the disjointness of $y_{x'}$ and $y'_{x'}$, we obtain:

$$P(y_x) = P(y_{x'}) + P(y_x, y'_{x'})$$

or:

$$P(y_x, y'_{x'}) = P(y_x) - P(y_{x'}) \tag{25}$$

Eq. (25) together with the assumption of exogeneity (Eq. (7)) establish Eq. (19). $\square$

2.4 Identifiability under monotonicity and non-exogeneity

The relations established in Theorems 1–3 were based on the assumption of exogeneity. In this section, we relax this assumption and consider cases where the effect of X on Y is confounded, i.e., $P(y_x) \neq P(y|x)$. In such cases $P(y_x)$ may still be estimated by auxiliary means (e.g., through adjustment of certain covariates, or through experimental studies) and the question is whether this added estimate can render the probability of causation identifiable. The answer is affirmative.

Theorem 4 *If Y is monotonic relative to X, then PNS, PN, PS are identifiable whenever the causal effect $P(y_x)$ is identifiable and are given by:*

$$PNS = P(y_x, y'_{x'}) = P(y_x) - P(y_{x'}) \tag{26}$$

$$PN = P(y'_{x'}|x,y) = \frac{P(y) - P(y_{x'})}{P(x,y)} \tag{27}$$

$$PS = P(y_x|x',y') = \frac{P(y_x) - P(y)}{P(x',y')} \tag{28}$$

To appreciate the difference between Eqs. (27) and (20) we can expand $P(y)$ and write

$$\begin{aligned} PN &= \frac{P(y|x)P(x) + P(y|x')P(x') - P(y_{x'})}{P(y|x)P(x)} \\ &= \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y_{x'})}{P(x,y)} \end{aligned} \tag{29}$$

The first term on the r.h.s. of (29) is the familiar excess-risk-ratio as in (20), and represents the value of PN under exogeneity. The second term represents the correction needed to account for X's non-exogeneity, i.e. $P(y_{x'}) \neq P(y|x')$.

Eqs. (26)–(28) thus provide more refined measures of causation, which can be used in situations where the causal effect $P(y_x)$ can be identified through auxiliary means (see Example-4, Section 3.4). Note however that these measures are no longer governed by the simple relationships given in Eqs. (10)–(11). Instead, the governing relation is Eq. (5).

Remarkably, since PS and PN must be non-negative, Eqs. (27)-(28) provide a simple test for the assumption of monotonicity:

$$P(y_x) \leq P(y) \leq P(y_{x'}) \tag{30}$$

to which one should join the standard inequalities

$$P(y_x) \geq P(x,y), \quad P(y_{x'}) \geq P(x',y)$$

It can be shown that these inequalities are in fact tight, that is, every combination of experimental and nonexperimental data that satisfy these inequalities can be generated from some causal model in which Y is monotonic in X. That the commonly made assumption of "no-prevention" is not entirely exempt from empirical scrutiny should come as a relief to many epidemiologists. Alternatively, if the no-prevention assumption is theoretically unassailable, the inequalities of Eq. (30) can be used for testing the compatibility of the experimental and non-experimental data, namely, whether subjects used in clinical trials are representative of the target population, characterized by the joint distribution $P(x, y)$.

Proof of Theorem 4:

Eq. (26) was established in (25). To prove (28), we write

$$P(y_x|x', y') = \frac{P(y_x, x', y')}{P(x', y')} = \frac{P(y_x, x', y'_{x'})}{P(x', y')} \tag{31}$$

because $x' \wedge y' = x' \wedge y'_{x'}$ (by consistency). To calculate the numerator of (31), we conjoin (24) with x'

$$x' \wedge y_x = (x' \wedge y_{x'}) \vee (y_x \wedge y'_{x'} \wedge x')$$

and take the probability on both sides, which gives (since $y_{x'}$ and $y'_{x'}$ are disjoint):

$$\begin{aligned} P(y_x, y'_{x'}, x') &= P(x', y_x) - P(x', y_{x'}) \\ &= P(x', y_x) - P(x', y) \\ &= P(y_x) - P(x, y_x) - P(x', y) \\ &= P(y_x) - P(x, y) - P(x', y) \\ &= P(y_x) - P(y) \end{aligned}$$

Substituting in (31), we finally obtain

$$P(y_x|x', y') = \frac{P(y_x) - P(y)}{P(x', y')}$$

which establishes (28). Eq. (27) follows through identical steps. $\square$

One common class of models which permits the identification of $P(y_x)$ under conditions of non-exogeneity is called *Markovian.*

Definition 9 (*Markovian models*)
A causal model M is said to be Markovian if the graph $G(M)$ associated with M is acyclic, and if the exogenous factors u_i are mutually independent. A model is semi-Markovian if $G(M)$ is acyclic and the exogenous variables are not necessarily independent. A causal model is said to be positive-Markovian if it is Markovian and $P(v) > 0$ for every v.

It is shown in Pearl (1993, 1995) that for every two variables, X and Y, in a positive-Markovian model M, the causal effect $P(y_x)$ is identifiable and is given by

$$P(y_x) = \sum_{pa_X} P(y|pa_X, x)P(pa_X) \tag{32}$$

where pa_X are (realizations of) the *parents* of X in the causal graph associate with M (see also Spirtes et al. (1993) and Robins (1986)). Thus, we can combine Eq. (32) with Theorem 4 and obtain a concrete condition for the identification of the probability of causation:

Corollary 1 *If in a positive-Markovian model M, the function $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable and are given by Eqs. (26)–(28), with $P(y_x)$ given in Eq. (32).*

A broader identification condition can be obtained through the use of the back-door and front-door criteria [Pearl, 1995], which are applicable to semi-Markovian models. These were further generalized in Galles and Pearl (1995)[9] and lead to the following corollary:

Corollary 2 *Let **GP** be the class of semi-Markovian models that satisfy the graphical criterion of Galles and Pearl (1995). If $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable in **GP** and are given by Eqs. (26)–(28), with $P(y_x)$ determined by the topology of $G(M)$ through the GP criterion.*

3 Examples and Applications

3.1 Example-1: Betting against a Fair Coin

We must bet heads or tails on the outcome of a fair coin toss; we win a dollar if we guess correctly, lose if we don't. Suppose we bet heads and we win a dollar, without glancing at the outcome of the coin, was our bet a necessary cause (respectively, sufficient cause, or both) for winning?

[9]Galles and Pearl (1995) provide an efficient method of deciding, from the graph $G(M)$, whether $P(y_x)$ is identifiable and, if the answer is affirmative, deriving the expression for $P(y_x)$.

Let x stand for "we bet on heads," y for "we win a dollar" and u for "the coin turned up head." The functional relationship between y, x and u is

$$y = (x \wedge u) \vee (x' \wedge u') \tag{33}$$

which is not monotonic but, nevertheless, permits us to compute the probabilities of causation from the basic definitions of Eqs. (1)–(3). To exemplify,

$$PN = P(y'_{x'}|x, y) = P(y'_{x'}|u) = 1$$

because $x \wedge y \Rightarrow u$, and $Y_{x'}(u) =$ *false*. In words, knowing the current bet (x) and current win (y) permits us to infer that the coin outcome must have been a head (u), from which we can further deduce that betting tails (x') instead of heads, would have resulted in a loss. Similarly,

$$PS = P(y_x|x', y') = P(y_x|u) = 1$$

because $x' \wedge y' \Rightarrow u$, and

$$\begin{aligned} PNS &= P(y_x, y'_{x'}) \\ &= P(y_x, y'_{x'}|u)P(u) + P(y_x, y'_{x'}|u')P(u') \\ &= 1\frac{1}{2} + 0\frac{1}{2} = \frac{1}{2} \end{aligned}$$

We see that betting heads has 50% chance of being a necessary-and-sufficient cause of winning. Still, once we win, we can be 100% sure that our bet was necessary for our win, and once we lose (say on betting tails) we can be 100% sure that betting head would have been sufficient for producing a win.

Note that these counterfactual quantities cannot be computed from the joint probability of X and Y without knowledge of the functional relationship in Eq. (33) which tells us the policy by which a win or a loss is determined. This can be seen from the conditional probabilities and causal effects associated with this example

$$P(y|x) = P(y|x') = P(y_x) = P(y_{x'}) = P(y) = \frac{1}{2}$$

and noting that identical probabilities would be generated by a random policy in which y is functionally independent of x, say by a bookie who hands us a dollar at random, without even looking at our bet. In such a random policy, the probabilities of causation PN, PS and PNS are all zero. Indeed, the bounds delineated in Theorem 1 (Eq. (8)) read $0 \leq PNS \leq \frac{1}{2}$, meaning that the three probabilities of causation cannot be determined from statistical data on X and Y alone, not even in a randomized experiment; knowledge of the functional mechanism is required, as in Eq. (33).

It is interesting to note that whether the coin is tossed before or after the bet has no bearing on the probabilities of causation as defined above. This stands in contrast with some theories of probabilistic causality which attempt to avoid deterministic mechanisms by conditioning all probabilities on "the state of the world just before" the occurrence of the cause in question (x) (e.g., [Good, 1961]). In the betting story above, the intention is to condition all probabilities on the state of the coin (u) which is not fulfilled if the coin is tossed after the bet is placed. Attempts to enrich the conditioning set with events occurring after the cause in question have led back to deterministic relationships involving counterfactual variables (see [Cartwright, 1989; Eells, 1991]).

One may argue, of course, that if the coin is tossed after the bet, then it is not at all clear what my winning would be had I bet differently; merely uttering my bet could conceivably affect the trajectory of the coin [Dawid, 1997]. This objection can be diffused by placing x and u in two remote locations and tossing the coin a split second after the bet is placed, but before any light ray could arrive from the betting room to the coin-tossing room. In such hypothetical situation the counterfactual statement: "my winning would be different had I bet differently" is rather compelling, even though the conditioning event (u) occurs after the cause in question (x). We conclude that temporal descriptions such as "the state of the world just before x" cannot be used to properly identify the appropriate conditioning events (u) in a problem; a deterministic model of the mechanisms involved is needed for such identification.

3.2 Example-2: The Firing Squad

Consider a 2-man firing squad (see Figure 1) in which A and B are riflemen, C is the squad's Captain who is waiting for the court order, U, and T is a condemned prisoner. Let u be the proposition that the court has ordered an execution, x the proposition stating that A

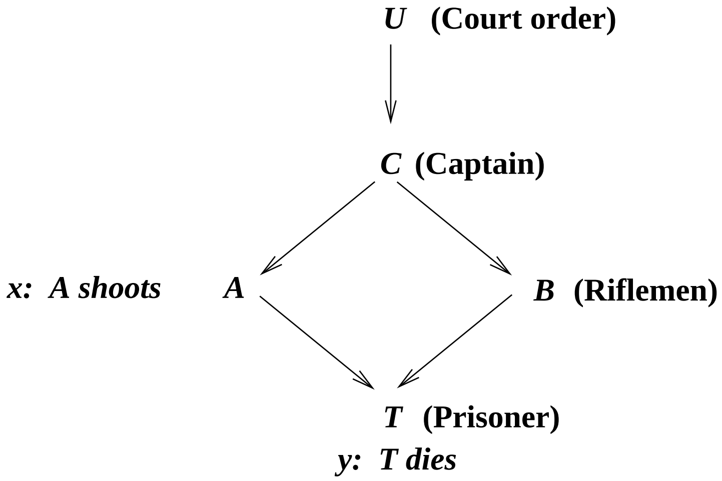

Figure 1:

pulled the trigger and y that T is dead. Assume that $P(u) = \frac{1}{2}$, that A and B are expert marksmen who are alert, sober and law abiding, and that T is not likely to die from fright or other extraneous causes. We wish to compute the probability that x was a necessary (or sufficient, or both) cause for y (i.e., PN, PS and PNS).

Definitions (1)–(3) permit us to compute these probabilities directly from the given causal model, since all functions and all probabilities are specified, with the truth value of each variable following that of U. Accordingly, we can write[10]

$$\begin{aligned} P(y_x) &= P(Y_x(u) = true)P(u) + P(Y_x(u') = true)P(u') \\ &= \frac{1}{2}(1+1) = 1 \end{aligned} \tag{34}$$

Similarly, we have,

$$\begin{aligned} P(y_{x'}) &= P(Y_{x'}(u) = true)P(u) + P(Y_{x'}(u') = true)P(u') \\ &= \frac{1}{2}(1+0) = \frac{1}{2} \end{aligned} \tag{35}$$

To compute PNS, we need to evaluate the probability of the joint event $y_{x'} \wedge y_x$. Considering that these two events are jointly true only when $U = true$, we have:

[10]Recall that $P(Y_x(u') = true)$ involves the submodel M_x, in which X is set to $true$ independently of U. Thus, although under condition u' the captain has not given a signal, the potential outcome $Y_x(u')$ is evaluated by hypothesizing rifleman-A pulling the trigger (x) despite a court order to stay the execution.

$$\begin{aligned} PNS &= P(y_x, y_{x'}) \\ &= P(y_x, y_{x'}|u)P(u) + P(y_x, y_{x'}|u')P(u') \\ &= \frac{1}{2}(1+0) = \frac{1}{2} \end{aligned} \tag{36}$$

The calculation of PS and PN, likewise, are simplified by the fact that the conditioning events, $x \wedge y$ for PN and $x' \wedge y'$ for PS, are true in only one state of U. We thus have:

$$PN = P(y'_{x'}|x, y) = P(y'_{x'}|u) = 0$$

reflecting the fact that, once the court orders an execution (u), T will die (y) from the shot of rifleman B, even if A refrains from shooting (x'). Indeed, upon learning of T's death, we can categorically state that rifleman-A's shot was *not* a necessary cause of the death.

Similarly,

$$PS = P(y_x|x', y') = P(y_x|u') = 1$$

matching our intuition that the a shot fired by an expert rifleman would be sufficient for causing the death of T, regardless of the court decision.

Note that Theorems 1 and 2 are not applicable to this example, because x is not exogenous; events x and y have a common cause (the Captain's signal C) which renders $P(y|x) = 1 \neq P(y_x) = \frac{1}{2}$. However, the monotonicity of Y (in x) permits us to compute PNS, PS and PN from the joint distribution $P(x, y)$ (using Eq. (26)–(28)), instead of consulting the basic model. Indeed, writing

$$P(x, y) = P(x', y') = \frac{1}{2} \tag{37}$$

$$P(x, y') = P(x', y) = 0 \tag{38}$$

we obtain

$$PN = \frac{P(y) - P(y_{x'})}{P(x, y)} = \frac{\frac{1}{2} - \frac{1}{2}}{\frac{1}{2}} = 0 \tag{39}$$

$$PS = \frac{P(y_x) - P(y)}{P(x', y')} = \frac{1 - \frac{1}{2}}{\frac{1}{2}} = 1 \tag{40}$$

as expected.

3.3 Example-3: The Effect of Radiation on Leukemia

Consider the following data (adapted from Finkelstein and Levin[11] (1990)) comparing leukemia deaths in children in Southern Utah with high and low exposure to radiation from fallout from nuclear tests in Nevada: Given these data, we wish to estimate the probabilities

		Exposure	
		High	**Low**
		x	x'
Deaths	y	30	16
Non-deaths	y'	69,130	59,010

Table 1:

that high exposure to radiation was a necessary (or sufficient or both) cause of death due to leukemia.

Assuming that exposure to nuclear radiation had no remedial effect on any individual in the study (i.e., monotonicity), the process can be modeled by a simple disjunctive mechanism represented by the equation

$$y = f(x, u, q) = (x \wedge q) \vee u \tag{41}$$

where u represents "all other causes" of y, and q represents all "enabling" mechanisms that must be present for x to trigger y. Assuming q and u are both unobserved, the question we ask is under what conditions we can identify the probability of causation, PNS, PN and PS, from the joint distribution of X and Y.

Since Eq. (41) is monotonic in x, Theorem 3 states that all three quantities would be identifiable provided X is exogenous, namely, x should be independent of q and u. Under this assumption, Eqs. (19)–(21) further permit us to compute the probabilities of causation

[11] The data in Finkelstein and Levin (1990) are given in person-year units. For the purpose of illustration we have converted the data to absolute numbers (of deaths and non-deaths) assuming a 10-year observation period.

from frequency data. Taking fractions to represent probabilities, the data in Table 1 imply the following numerical results:

$$PNS = P(y|x) - P(y|x') = \frac{30}{30 + 69,130} - \frac{16}{16 + 59,010} = .0001625 \tag{42}$$

$$PN = \frac{PNS}{P(y|x)} = \frac{PNS}{30/(30 + 69,130)} = .37535 \tag{43}$$

$$PS = \frac{PNS}{1 - P(y|x')} = \frac{PNS}{1 - 16/(16 + 59,010)} = .0001625 \tag{44}$$

Statistically, these figures mean: There is a 1.625 in ten-thousand chance that a randomly chosen child would both die of leukemia if exposed and survive if not exposed. There is a 37.544% chance that a child who died from leukemia after exposure would have survived had he/she not been exposed. There is a 1.625 in ten-thousand chance that any unexposed surviving child would have died of leukemia had he/she been exposed.

Glymour (1998) analyzes this example with the aim of identifying the probability $P(q)$ (Cheng's "causal power") which coincides with PS. Glymour concludes that $P(q)$ is identifiable and is given by Eq. (21), provided x, u, and q are mutually independent. Our analysis shows that Glymour's result can be generalized in several ways. First, since Y is monotonic in X, the validity of Eq. (21) is assured even when q and u are dependent, because exogeneity requires merely independence between x and $\{u, q\}$ jointly. This is important in epidemiological settings, because an individual's susceptibility to nuclear radiation is likely to be associated with his/her susceptibility to other potential causes of leukemia (e.g., natural kinds of radiation).

Second, Theorem 2 assures us that the relationships between PN, PS and PNS (Eqs. (10)–(11)), which Glymour derives for independent q and u, should remain valid even when u and q are dependent.

Finally, Theorem 4 assures us that PN and PS are identifiable even when x is not independent of $\{u, q\}$, provided only that the mechanism of Eq. (41) is embedded in a larger causal structure which permits the identification of $P(y_x)$. For example, assume that exposure to nuclear radiation (x) is suspect of being associated with terrain and altitude, which are also factors in determining exposure to cosmic radiation. A model reflecting such consideration is depicted in Figure 2, where W represents factors affecting both X and U. A natural way

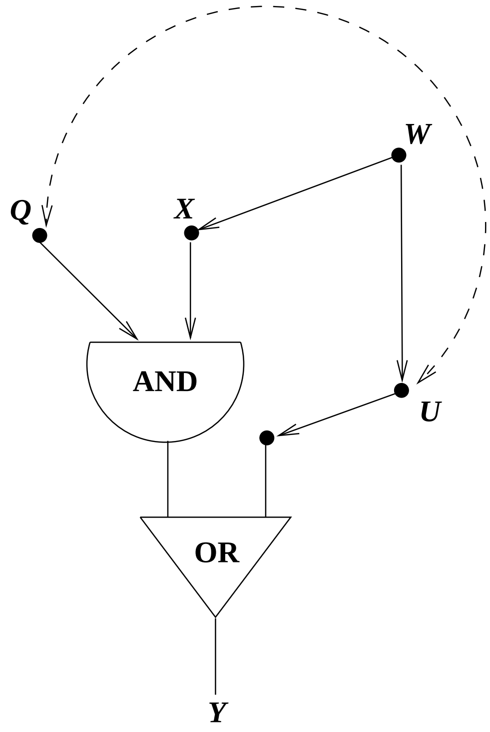

Figure 2:

to correct for possible confounding bias in the causal effect of X on Y would be to adjust for W, that is, to calculate $P(y_x)$ using the adjustment formula

$$P(y_x) = \sum_w P(y|x,w)P(w) \tag{45}$$

(instead of $P(y|x)$) where the summation runs over levels of W. This adjustment formula, which follows from Eq. (32), is correct regardless of the mechanisms mediating X and Y, provided only that W represents *all* common factors affecting X and Y [Pearl, 1995]. Theorem 4 instructs us to evaluate PN and PS by substituting (45) into Eqs. (27) and (28), respectively, and it assures us that the resulting expressions constitute consistent estimates of PN and PS. This consistency is guaranteed jointly by the assumption of monotonicity and by the (assumed) topology of the causal graph.

Note that monotonicity as defined in Eq. (18) is a global property of all pathways between x and y. The causal model may include several nonmonotonic mechanisms along these pathways without affecting the validity of (18). Arguments for the validity of monotonicity, however, must be based on substantive information, as it is not testable in general. For example, Robins and Greenland (1989) argue that, exposure to nuclear radiation may conceivably be of benefit to some individuals, since such radiation is routinely used clinically in treating cancer patients.

3.4 Example-4: Legal Responsibility from Experimental and Nonexperimental Data

A lawsuit is filed against the manufacturer of drug x, charging that the drug is likely to have caused the death of Mr. A, who took the drug to relieve symptom S associated with disease D. The manufacturer claims that experimental data on patients with symptom S show conclusively that drug x may cause only negligible increase in death rates. The plaintiff argues, however, that the experimental study is of little relevance to this case, because it represents the effect of the drug on *all* patients, not on patients like Mr. A who actually died while using drug x. Moreover, argues the plaintiff, Mr. A is unique in that he used the drug on his own volition, unlike subjects in the experimental study who took the drug to comply with experimental protocols. To support this argument, the plaintiff furnishes nonexperimental data indicating that most patients who chose drug x would have been alive if it were not for the drug. The manufacturer counter argues by stating that: (1) counterfactual speculations regarding whether patients would or would not have died, are purely metaphysical and should be avoided [Dawid, 1997], and (2) nonexperimental data should be dismissed a priori, on the ground that such data may be highly biased; for example, incurable terminal patients might be more inclined to use drug x if it provides them greater symptomatic relief. The court must now decide, based on both the experimental and nonexperimental studies, what the probability is that drug x was in fact the cause of Mr. A's death.

The data associated with the two studies are shown in Table 2 below.

The experimental data provide the estimates

$$P(y_x) = 16/1000 = 0.016 \tag{46}$$

$$P(y_{x'}) = 14/1000 = 0.014 \tag{47}$$

Experimental		x	x'
Deaths	y	16	14
non-deaths	y'	984	986

Non-Experimental		x	x'
Deaths	y	2	28
non-deaths	y'	998	972

Table 2:

The nonexperimental data provide the estimates

$$P(y) = 30/2000 = 0.015 \tag{48}$$

$$P(y, x) = 2/2000 = 0.001 \tag{49}$$

Assuming that drug x can only cause, never prevent death, Theorem 4 is applicable and Eq. (27) gives

$$PN = \frac{P(y) - P(y_{x'})}{P(y, x)} = \frac{0.015 - 0.014}{0.001} = 1.00 \tag{50}$$

Thus, the plaintiff was correct; barring sampling errors, the data provide us with 100% assurance that drug x was in fact responsible for the death of Mr. A. Note that a straightforward use of the experimental excess-risk-ratio would yield a much lower (and incorrect) result:

$$\frac{P(y_x) - P(y_{x'})}{P(y_x)} = \frac{0.016 - 0.014}{0.016} = 0.125 \tag{51}$$

Evidently, what the experimental study does not reveal is that, given a choice, terminal patients stay away from drug x. Indeed, if there were any terminal patients who would choose x (given the choice) then the control group (x') would have included some such patients (due to randomization) and then the proportion of deaths among the control group $P(y_{x'})$ should have been higher than $P(x', y)$, the population proportion of terminal patients avoiding x. However, the equality $P(y_{x'}) = P(y, x')$ tells us that no such patients were included in the control group, hence, (by randomization) no such patients exist in the population at large and, therefore, none of the patients who freely chose drug x was a terminal case; all were susceptible to x.

The numbers in Table 2 were obviously chosen to represent an extreme case, so as to facilitate a qualitative explanation of the validity of Eq. (27). Nevertheless, it is instructive to note that a combination of experimental and nonexperimental studies may unravel what experimental studies alone will not reveal and, in addition, that such combination may provide a test for the assumption of no-prevention, as outlined in Section 2.4 (Eq. (30)).

4 Identification in Non-monotonic Models

In this section we discuss the identification of probabilities of causation without making the monotonicity assumption. We will assume that we are given a causal model M in which all functional relationships are known, but since the exogenous variables U are not observed, their distributions are not known.

A straightforward way to identify any causal or counterfactual quantity (including PN, PS and PNS) would be to infer the probability distribution of the exogenous variables – that would amount to inferring the entire model, from which all quantities can be computed. Thus, our first step would be to study under what conditions the function $P(u)$ can be identified.

If M is Markovian, the problem can be analyzed by considering each parents-child family separately. Consider any arbitrary equation in M:

$$\begin{aligned} y &= f(pa_Y, u_Y) \\ &= f(x_1, x_2, \ldots, x_k, u_1, \ldots, u_m) \end{aligned} \quad (52)$$

where $U_Y = \{U_1, \ldots, U_m\}$ is the set of exogenous, possibly dependent variables that appear in the equation for Y. In general, the domain of U_Y can be arbitrary, discrete, or continuous, since these variables represent unobserved factors that were omitted from the model. However, since the observed variables are binary, there is only a finite number $(2^{(2^k)})$ of functions from PA_Y to Y and, for any point $U_Y = u$, only one of those function is realized. This defines a partition of the domain of U_Y into a set S of equivalence classes, where each equivalence class $s \in S$ induces the same function $f^{(s)}$ from PA_Y to Y. Thus, as u varies over its domain, a set S of such functions is realized, and we can regard S as a new exogenous variable, whose values are the set $\{f^{(s)} : s \in S\}$ of functions from PA_Y to Y that are realizable in U_Y. The number of such functions will usually be smaller than $2^{(2^k)}$.[12]

[12]Balke and Pearl (1994) called these S variables "response variables," and Heckerman and Shachter (1995)

For example, consider the model described in Figure 2. As the exogenous variables (q, u) vary over their respective domains, the relation between X and Y spans three distinct functions:

$$Y = \ true, \quad Y = \ false, \text{ and } Y = X$$

The fourth possible function, $Y = not\text{-}X$, is never realized because $f_Y(\cdot)$ is monotonic. The cells (q, u) and (q', u) induce the same function between X and Y, hence they belong to the same equivalence class.

If we are given the distribution $P(u_Y)$, we can compute the distribution $P(s)$ and this will determine the conditional probabilities $P(y|pa_Y)$ by summing $P(s)$ over all those functions $f^{(s)}$ that map pa_Y into the value *true*,

$$P(y|pa_Y) = \sum_{s: f^{(s)}(pa_Y) \ = \ true} P(s) \tag{53}$$

To insure model identifiability it is sufficient that we can invert the process and determine $P(s)$ from $P(y|pa_Y)$. If we let the set of conditional probabilities $P(y|pa_Y)$ be represented by a vector $\boldsymbol{p}$ (of dimensionality 2^k), and $P(s)$ by a vector $\boldsymbol{q}$, then the relation between $\boldsymbol{q}$ is $\boldsymbol{p}$ is linear and can be represented as a matrix multiplication [Balke and Pearl, 1994b]

$$\boldsymbol{p} = \boldsymbol{R}\boldsymbol{q} \tag{54}$$

where $\boldsymbol{R}$ is a 0-1 matrix, with dimension $2^k \times |S|$. Thus, a sufficient condition for identification is simply that $\boldsymbol{R}$, together with the normalizing equation $\sum_j \boldsymbol{q}_j = 1$, be invertible.

In general, $\boldsymbol{R}$ will not be invertible because the dimensionality of $\boldsymbol{q}$ can be much larger than that of $\boldsymbol{p}$. However, in many cases, such as the Noisy-OR mechanism

$$Y = U_0 \bigvee_{i=1,\ldots,k} (X_i \wedge U_i), \tag{55}$$

symmetry permits $\boldsymbol{q}$ to be identified from $P(y|pa_Y)$ even when the exogenous variables $U_0, U_1, \ldots, U_k$ are not independent. This can be seen by noting that every point u for which $U_0 = \ false$ defines a unique function $f^{(s)}$ because, if T is the set of indices i for which U_i is true, the relationship between PA_Y and Y becomes

$$Y = U_0 \bigvee_{i \in T} X_i \tag{56}$$

called them "mapping variables."

and, for $U_0 =$ *false*, this equation defines a distinct function for each T. The number of induced functions is $2^k + 1$, which (subtracting 1 for normalization) is exactly the number of distinct realizations of PA_Y. Moreover, it is easy to show that the matrix connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ is invertible. We thus conclude that the probability of every counterfactual sentence can be identified in any Markovian model composed of Noisy-OR mechanisms, regardless of whether the exogenous variables in each family are mutually independent. The same holds of course for Noisy-AND mechanisms or any combination thereof, including negating mechanisms, provided that each family consists of one type of mechanism.

To generalize this results to mechanisms other than Noisy-OR and Noisy-AND, we note that although $f_Y(\cdot)$ in this example was monotonic (in each X_i), it was the redundancy of $f_Y(\cdot)$, not its monotonicity, that ensured identifiability. The following is an example of a monotonic function for which the $\boldsymbol{R}$ matrix is not invertible

$$Y = (X_1 \wedge U_1) \vee (X_2 \wedge U_1) \vee (X_1 \wedge X_2 \wedge U_3)$$

It represents a Noisy-OR gate for $U_3 =$ *false*, and becomes a Noisy-AND gate for $U_3 = \text{ } true, U_1 = U_2 =$ *false*. The number of equivalence-classes induced is six, which would require five independent equations to determine their probabilities; the data $P(y|pa_Y)$ provide only four such equations.

In contrast, the mechanism governed by the equation below, although non-monotonic, is invertible:

$$Y = XOR(X_1, XOR(U_2, \ldots, XOR(U_{k-1}, XOR(X_k, U_k))))),$$

where $XOR(*)$ stands for Exclusive-OR. This equation induces only two functions from PA_Y to Y;

$$Y = \begin{cases} XOR(X_1, ..., X_k) & if \quad XOR(U_1, \ldots, U_k) = \text{ } false \\ \neg XOR(X_1, \ldots, X_k) & if \quad XOR(U_1, \ldots, U_k) = \text{ } true \end{cases}$$

A single conditional probability, say $P(y|x_1, \ldots, x_k)$, would suffice therefore for computing the one parameter needed for identification: $P[XOR(U_1, \ldots, U_k) = \text{ } true]$.

We summarize these considerations with a theorem.

Definition 10 (*Local invertability*)

A model M is said to be locally invertible if for every variable $V_i \in V$ the set of $2^k + 1$

equations

$$P(y|pa_i) = \sum_{s:f^{(s)}(pa_i)= \ true} q_i(s) \tag{57}$$

$$\sum_s q_i(s) = 1 \tag{58}$$

has a unique solution for $q_i(s)$*, where each* $f_i^{(s)}(pa_i)$ *corresponds to the function* $f_i(pa_i, u_i)$ *induced by* u_i *in equivalence-class* s*.*

Theorem 5 *Given a Markovian model* $M =< U, V, \{f_i\} >$ *in which the functions* $\{f_i\}$ *are known and the exogenous variables* U *are unobserved, if* M *is locally invertible, then the probability of every counterfactual sentence is identifiable from the joint probability* $P(v)$*.*

Proof:

If Eq. (57) has a unique solution for $q_i(s)$, we can replace U with S and obtain an equivalent model

$$M' =< S, V, \{f_i'\} > \text{ where } f_i' = f_i^{(s)}(pa_i).$$

M' together with $q_i(s)$ completely specifies a probabilistic model $< M', P(s) >$ (due to the Markov property) from which probabilities of counterfactuals are derivable by definition. $\square$

Theorem 5 provides a sufficient condition for identifying probabilities of causation, but of course does not exhaust the spectrum of assumptions that are helpful in achieving identification. In many cases we might be justified in hypothesizing additional structure on the model, for example, that the U variables entering each family are themselves independent. In such cases, additional constraints are imposed on the probabilities $P(s)$ and Eq. (57) may be solved even when the cardinality of S far exceeds the number of conditional probabilities $P(y|pa_Y)$.

5 Conclusion

The paper explicates and analyzes the necessary and sufficient components of causation. Using counterfactual interpretations that rest on structural-model semantics, the paper demonstrates how simple techniques of computing probabilities of counterfactuals can be used in computing probabilities of causes, deciding questions of identification, defining conditions under which probabilities of causes can be estimated from statistical data, and uncovering tests for assumptions that are routinely made (often unknowingly) by analysts and investigators.

On the practical side, the paper offers several useful tools to epidemiologists and health scientists. It formulates and calls attention to basic assumptions that must be ascertained before statistical measures such as excess-risk-ratio could represent causal quantities such as attributable-risk or probability of causes. It shows how data from both experimental and non-experimental studies can be combined to yield information that neither study alone can reveal. Finally, it provides tests for the commonly made assumption of "no prevention," and for the often asked question of whether a clinical study is representative of its target population.

On the conceptual side, we have seen that both the probability of necessity (PN) and probability of sufficiency (PS) play a role in our understanding of causation, and that both components have their logics and computational rules. Although the counterfactual concept of necessary cause (i.e., that an outcome would not have occurred 'but for' the action) is predominant in legal settings [Robertson, 1997] and ordinary discourse, the sufficient component has a definite influence on causal thoughts.

The sufficiency component of causation plays a major role in scientific and legal explanations, as can be seen from examples where the necessary component is dormant. Why do we consider striking a match to be a more adequate explanation (of a fire) than the presence of oxygen? Recasting the question in the language of PN and PS, we note that, since both explanations are necessary for the fire, each will command a PN of unity. (In fact PN is higher for the oxygen, if we allow for alternative ways of igniting a spark). It is the sufficiency component alone that endows the match with greater explanatory power than the oxygen. If the probabilities associated with striking a match and the presence of oxygen are p_m and p_o, respectively, the PS measures associated with these explanations evaluate to $PS(match) = p_o$ and $PS(oxygen) = p_m$, clearly favoring the match when $p_o >> p_m$. Thus, a robot instructed to explain why a fire broke out has no choice but to consider both PN and PS in its deliberations.

Should PS enter legal considerations in criminal and tort law? I believe (as does I.J. Good (1993)) that it should, because attention to sufficiency implies attention to the consequences of one's action. The person who lighted the match ought to have anticipated the presence of oxygen, while the person who supplied (or failed to remove) the oxygen is not generally expected to have anticipated match-striking episodes.

However, what weight should the law assign to the necessary vis-a-vis the sufficient component of causation? This question obviously lies beyond the scope of this investigation, and it is not at all clear who would be qualified to tackle the question or whether our legal system would be prepared to implement the recommendation. I am hopeful, however, that whoever undertakes to consider such questions will find the analysis of this paper to be of some use.

Acknowledgments

I am indebted to Sander Greenland for many suggestions and discussions concerning the treatment of causation in the epidemiological literature and potential applications of this analysis in practical epidemiological studies. Donald Michie and Jack Good are responsible for swaying my attention from PN to PS and PNS. Clark Glymour and Patricia Cheng have helped unravel some of the mysteries of causal power theory, and Michelle Pearl has provided useful pointers to the epidemiological literature.

References

[Aldrich, 1993] J. Aldrich. Cowles' exogeneity and core exogeneity. Technical Report Discussion Paper 9308, Department of Economics, University of Southampton, England, 1993.

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

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

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

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

[Balke and Pearl, 1997] A. Balke and J. Pearl. Nonparametric bounds on causal effects from partial compliance data. *Journal of the American Statistical Association*, 92(439):1–6, September 1997.

[Breslow and Day, 1980] N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research; Vol. 1, The Analysis of Case-Control Studies*. IARC, Lyon, 1980.

[Cartwright, 1989] N. Cartwright. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford, 1989.

[Cheng, 1997] P.W. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367–405, 1997.

[Cole, 1997] P. Cole. Causality in epidemiology, health policy, and law. *Journal of Marketing Research*, 27:10279–10285, 1997.

[Dawid, 1997] A.P. Dawid. Causal inference without counterfactuals. Technical report, Department of Statistical Science, University College London, UK, 1997.

[Dhrymes, 1970] P.J. Dhrymes. *Econometrics*. Springer-Verlag, New York, 1970.

[Eells, 1991] E. Eells. *Probabilistic Causality*. Cambridge University Press, Cambridge, MA, 1991.

[Engle et al., 1983] R.F. Engle, D.F. Hendry, and J.F. Richard. Exogeneity. *Econometrica*, 51:277–304, 1983.

[Finkelstein and Levin, 1990] M.O. Finkelstein and B. Levin. *Statistics for Lawyers*. Springer-Verlag, New York, 1990.

[Fisher, 1970] F.M. Fisher. A correspondence principle for simultaneous equations models. *Econometrica*, 38(1):73–92, January 1970.

[Galles and Pearl, 1995] D. Galles and J. Pearl. Testing identifiability of causal effects. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 185–195. Morgan Kaufmann, San Francisco, 1995.

[Galles and Pearl, 1997] D. Galles and J. Pearl. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

[Galles and Pearl, 1998] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3(1):151–182, 1998.

[Glymour, 1998] C. Glymour. Psychological and normative theories of causal power and the probabilities of causes. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 166–172. Morgan Kaufmann, San Francisco, CA, 1998.

[Good, 1961] I.J. Good. A causal calculus, I. *British Journal for the Philosophy of Science*, 11:305–318, 1961.

[Good, 1993] I.J. Good. A tentative measure of probabilistic causation relevant to the philosophy of the law. *J. Statist. Comput. and Simulation*, 47:99–105, 1993.

[Greenland and Robins, 1988] S. Greenland and J. Robins. Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology*, 128:1185–1197, 1988.

[Hall,] N. Hall. Two concepts of causation. In press.

[Halpern, 1998] J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence*, pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998.

[Heckerman and Shachter, 1995a] D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

[Heckerman and Shachter, 1995b] D. Heckerman and R. Shachter. A definition and graphical representation for causality. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 262–273, San Mateo, CA, 1995. Morgan Kaufmann.

[Hendry, 1995] David F. Hendry. *Dynamic Econometrics.* Oxford University Press, New York, 1995.

[Hennekens and Buring, 1987] C.H. Hennekens and J.E. Buring. *Epidemiology in Medicine.* Brown, Little, Boston, 1987.

[Hume, 1748 reprinted 1988] D. Hume. *An Enquiry concerning Human Understanding.* Open Court Press, LaSalle, 1748, reprinted 1988.

[Imbens, 1997] G.W. Imbens. Book reviews. *Journal of Applied Econometrics*, 12, 1997.

[Khoury et al., 1989] M.J. Khoury, W.D Flanders, S. Greenland, and M.J. Adams. On the measurement of susceptibility in epidemiologic studies. *American Journal of Epidemiology*, 129(1):183–190, 1989.

[Lewis, 1973] D. Lewis. Causation. *The Journal of Philosophy*, 70:556–567, 1973. Reprinted with postscript in D. Lewis, *Philosophical Papers*, vol. II. New York: Oxford, 1986.

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

[Michie, 1997] D. Michie. Adapting Good's q theory to the causation of individual events. Technical report, University of Edinburgh, UK, 1997. Submitted for publication in *Machine Intelligence 15.*

[Mill, 1843] J.S. Mill. *System of Logic*, volume 1. John W. Parker, London, 1843.

[Neyman, 1923] J. Neyman. Sur les applications de la thar des probabilities aux experiences Agaricales: Essay des principle. *Statistical Science*, 5:463–472, 1923. English translation of excerpts by Dabrowska, D. and Speed, T. (1990).

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

[Pearl, 1994] J. Pearl. A probabilistic calculus of actions. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 454–462. Morgan Kaufmann, San Mateo, CA, 1994.

[Pearl, 1995] J. Pearl. Causal diagrams for experimental research. *Biometrika*, 82:669–710, December 1995.

[Pearl, 1996a] J. Pearl. Causation, action, and counterfactuals. In Y. Shoham, editor, *Theoretical Aspects of Rationality and Knowledge, Proceedings of the Sixth Conference*, pages 51–73. Morgan Kaufmann, San Francisco, CA, 1996.

[Pearl, 1996b] J. Pearl. Structural and probabilistic causality. In D.R. Shanks, K.J. Holyoak, and D.L. Medin, editors, *The Psychology of Learning and Motivation*, volume 34, pages 393–435. Academic Press, San Diego, CA, 1996.

[Pearl, 1998] J. Pearl. On the definition of actual cause. Technical Report R-259, Department of Computer Science, University of California, Los Angeles, CA, 1998.

[Robertson, 1997] D.W. Robertson. The common sense of cause in fact. *Texas Law Review*, 75(7):1765–1800, 1997.

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

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

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

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

[Rubin, 1990] D.B. Rubin. Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25:279–292, 1990.

[Schlesselman, 1982] J.J. Schlesselman. *Case-Control Studies: Design Conduct Analysis.* Oxford University Press, New York, 1982.

[Shep, 1958] M.C. Shep. Shall we count the living or the dead? *New England Journal of Medicine*, 259:1210–1214, 1958.

[Sobel, 1990] M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55(3):495–515, 1990.

[Spirtes et al., 1993] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer-Verlag, New York, 1993.

[Strotz and Wold, 1960] R.H. Strotz and H.O.A. Wold. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.

A APPENDIX: Structural Model Semantics

This appendix presents a brief review of the basic definitions and assumptions behind structural model semantics, as formulated in [Galles and Pearl, 1998].

A.1 Definitions: Causal models, actions and counterfactuals

A causal model is a mathematical object that assigns truth values to sentences involving causal and counterfactual relationships. Basic of our analysis are sentences involving actions or external interventions, such as, "p will be true if we do q" where q is any elementary proposition. Structural models are generalizations of the structural equations used in engineering, biology, economics and social science.[13] World knowledge is represented as a collection of stable and autonomous relationships called "mechanisms," each represented as an equation, and changes due to interventions or hypothetical novel eventualities are treated as local modifications of those equations.

Definition 11 (causal model) A *causal model* is a triple

$$M = \ < U, V, F >$$

where

(i) U is a set of variables, called *exogenous*, that are determined by factors outside the model.

(ii) V is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called *endogenous*, that are determined by variables in the model, namely, variables in $U \cup V$.

(iii) F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ where each f_i is a mapping from $U \cup (V \setminus V_i)$ to V_i. In other words, each f_i tells us the value of V_i given the values of all other variables in $U \cup V$. Symbolically, the set of equations F can be represented by writing

$$v_i = f_i(pa_i, u) \quad i = 1, \ldots, n$$

[13] Similar models, called "neuron diagrams" [Lewis, 1973; Hall, 1998] are used informally by philosophers to illustrate chains of causal processors.

where pa_i is any realization of the minimal set of variables PA_i in V/V_i (connoting *parents*) that renders f_i nontrivial.

Every causal model M can be associated with a directed graph, $G(M)$, in which each node corresponds to a variable in V and the directed edges point from members of PA_i toward V_i. We call such a graph the *causal graph* associated with M. This graph merely identifies the endogenous variables PA_i that have direct influence on each V_i but it does not specify the functional form of f_i.

Definition 12 (submodel) Let M be a causal model, X be a set of variables in V, and x be a particular realization of X. A submodel M_x of M is the causal model

$$M_x = \;\; < U, V, F_x >$$

where

$$F_x = \{f_i : V_i \notin X\} \cup \{X = x\} \tag{59}$$

In words, F_x is formed by deleting from F all functions f_i corresponding to members of set X and replacing them with the set of constant functions $X = x$.

Submodels are useful for representing the effect of local actions and hypothetical changes. If we interpret each function f_i in F as an independent physical mechanism and define the action $do(X = x)$ as the minimal change in M required to make $X = x$ hold true under any u, then M_x represents the model that results from such a minimal change, since it differs from M by only those mechanisms that directly determine the variables in X. The transformation from M to M_x modifies the algebraic content of F, which is the reason for the name *modifiable structural equations* used in [Galles and Pearl, 1998].

Definition 13 (effect of action) Let M be a causal model, X be a set of variables in V, and x be a particular realization of X. The *effect of action* $do(X = x)$ on M is given by the submodel M_x.[14]

[14]An explicit translation of interventions into "wiping out" equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970), Sobel (1990), Spirtes et al. (1993), and Pearl (1995).

Definition 14 (potential response) Let Y be a variable in V, and let X be a subset of V. The *potential response* of Y to action $do(X = x)$, denoted $Y_x(u)$, is the solution for Y of the set of equations F_x.[15]

We will confine our attention to actions in the form of $do(X = x)$. Conditional actions, of the form "$do(X = x)$ if $Z = z$" can be formalized using the replacement of equations by functions of Z, rather than by constants [Pearl, 1994]. We will not consider disjunctive actions, of the form "$do(X = x$ or $X = x')$", since these complicate the probabilistic treatment of counterfactuals.

Definition 15 (counterfactual) Let Y be a variable in V, and let X a subset of V. The counterfactual sentence "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.[16]

This formulation generalizes naturally to probabilistic systems, as is seen below.

Definition 16 (probabilistic causal model) A probabilistic causal model is a pair

$$< M, P(u) >$$

where M is a causal model and $P(u)$ is a probability function defined over the domain of U.

$P(u)$, together with the fact that each endogenous variable is a function of U, defines a probability distribution over the endogenous variables. That is, for every set of variables $Y \subseteq V$, we have

$$P(y) \triangleq P(Y = y) = \sum_{\{u \mid Y(u)=y\}} P(u) \tag{60}$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x:

$$P(Y_x = y) = \sum_{\{u \mid Y_x(u)=y\}} P(u) \tag{61}$$

[15] Galles and Pearl (1998) required that F_x has a unique solution, a requirement later relaxed by Halpern (1998). In this paper we are dealing with recursive systems (i.e., $G(M)$ is a cyclic) where uniqueness of solution is ensured.

[16] The connection between counterfactuals and local actions (sometimes resembling "miracles") is made in Lewis (1973) and is further elaborated in Balke and Pearl (1994) and Heckerman and Shachter (1995).

Likewise a causal model defines a joint distribution on counterfactual statements, i.e., $P(Y_x = y, Z_w = z)$ is defined for any sets of variables Y, X, Z, W, not necessarily disjoint. In particular, $P(Y_x = y, X = x')$ and $P(Y_x = y, Y_{x'} = y')$ are well defined for $x \neq x'$, and are given by

$$P(Y_x = y, X = x') = \sum_{\{u | Y_x(u)=y \ \& \ X(u)=x'\}} P(u) \tag{62}$$

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u \ | \ Y_x(u)=y \ \& \ Y_{x'}(u)=y'\}} P(u). \tag{63}$$

When x and x' are incompatible, Y_x and $Y_{x'}$ cannot be measured simultaneously, and it may seem meaningless to attribute probability to the joint statement "Y would be y if $X = x$ *and* Y would be y' if $X = x'$." Such concerns have been a source of recent objections to treating counterfactuals as jointly distributed random variables [Dawid, 1997]. The definition of Y_x and Y'_x in terms of two distinct submodels, driven by a standard probability space over U, explains away these objections and further illustrates that joint probabilities of counterfactuals can be encoded rather parsimoniously using $P(u)$ and F.

In particular, the probabilities of causation analyzed in this paper (see Eqs. (1)-(3)) require the evaluation of expressions of the form $P(Y_{x'} = y'|X = x, Y = y)$ with x and y incompatible with x' and y', respectively. Eq. (62) allows the evaluation of this quantity as follows:

$$\begin{aligned} P(Y_{x'} = y'|X = x, Y = y) &= \frac{P(Y_{x'} = y', X = x, Y = y)}{P(X = x, Y = y)} \\ &= \sum_u P(Y_{x'}(u) = y')P(u|x, y) \end{aligned} \tag{64}$$

In other words, we first update $P(u)$ to obtain $P(u|x, y)$, then we use the updated distribution $P(u|x, y)$ to compute the expectation of the index function $Y_{x'}(u) = y'$.

A.2 Examples

Figure A-1 describes the causal relationships among the season of the year (X_1), whether rain falls (X_2) during the season, whether the sprinkler is on (X_3) during the season, whether the pavement is wet (X_4), and whether the pavement is slippery (X_5). All variables in this graph except the root variable X_1 take a value of either "True" or "False" (encoded "1"

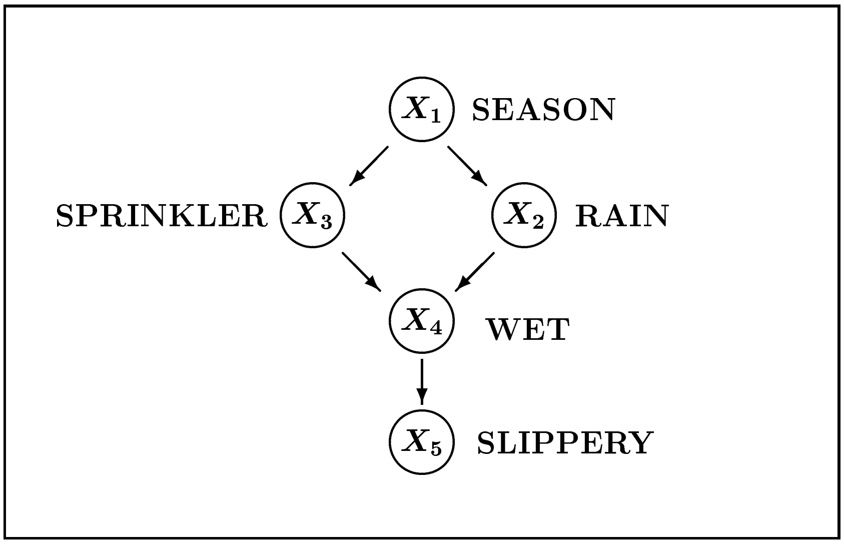

Figure A-1: Causal graph illustrating causal relationships among five variables.

and “0” for convenience.) X_1 takes one of four values: “Spring,” “Summer,” “Fall,” or “Winter.” Here, the absence of a direct link between, for example, X_1 and X_5, captures our understanding that the influence of the season on the slipperiness of the pavement is mediated by other conditions (e.g., the wetness of the pavement). The corresponding model consists of five functions, each representing an autonomous mechanism:

$$
\begin{aligned}
x_1 &= u_1 \\
x_2 &= f_2(x_1, u_2) \\
x_3 &= f_3(x_1, u_3) \\
x_4 &= f_4(x_3, x_2, u_4) \\
x_5 &= f_5(x_4, u_5)
\end{aligned}
\tag{65}
$$

The exogenous variables $U_1, \ldots, U_5$, represent factors omitted from the analysis. For example, U_4 may stand for (unspecified) events that would cause the pavement to get wet ($x_4 = 1$) when the sprinkler is off ($x_2 = 0$) and it does not rain ($x_3 = 0$) (e.g., a leaking water pipe). These factors are not shown explicitly in Figure A-1 to communicate, by convention, that the U's are assumed independent of one another. When some of these factors are judged to be dependent, it is customary to encode such dependencies by augmenting the graph with double-headed arrows [Pearl, 1995].

To represent the action "turning the sprinkler ON," or $do(X_3 = \text{ON})$, we replace the equation $x_3 = f_3(x_1, u_3)$ in the model of Eq. (65) with the equation $x_3 = 1$. The resulting submodel, $M_{X_3=\text{ON}}$, contains all the information needed for computing the effect of the action on the other variables. Note that the operation $do(X_3 = \text{ON})$ stands in marked contrast to that of *finding* the sprinkler ON; the latter involves making the substitution *without* removing the equation for X_3, and therefore may potentially influence (the belief in) every variable in the network. In contrast, the only variables affected by the action $do(X_3 = \text{ON})$ are X_4 and X_5, that is, the descendants of the manipulated variable X_3. This mirrors the difference between *seeing* and *doing*: after observing that the sprinkler is ON, we wish to infer that the season is dry, that it probably did not rain, and so on; no such inferences should be drawn in evaluating the effects of the action "turning the sprinkler ON" that a person may consider taking.

This distinction obtains a vivid symbolic representation in cases where the U_i's are assumed independent, because the joint distribution of the endogenous variables then admits the product decomposition

$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2, x_3)P(x_5|x_4) \tag{66}$$

Similarly, the joint distribution associated with the submodel M_x representing the action $do(X_3 = \text{ON})$ is obtained from the product above by deleting the factor $P(x_3|x_1)$ and substituting $x_3 = 1$.

$$P(x_1, x_2, x_4, x_5|do(X_3 = \text{ON})) = P(x_1)\ P(x_2|x_1)\ P(x_4|x_2, x_3 = 1)\ P(x_5|x_4) \tag{67}$$

The difference between the action $do(X_3 = \text{ON})$ and the observation $X_3 = \text{ON}$ is thus seen from the corresponding distributions. The former is represented by Eq. (67), while the latter by *conditioning* Eq. (66) on the observation, i.e.,

$$P(x_1, x_2, x_4, x_5|X_3 = \text{ON}) = \frac{P(x_1)\ P(x_2|x_1)\ P(x_3 = 1|x_1)P(x_4|x_2, x_3 = 1)P(x_5|x_4)}{P(x_3 = 1)}$$

Note that the conditional probabilities on the r.h.s. of Eq. (67) are the same as those in Eq. (66), and can therefore be estimated from pre-action observations, provided $G(M)$ is available. However, the pre-action distribution P together with the causal graph $G(M)$ is generally not sufficient for evaluating all counterfactuals sentences. For example, the probability that "the pavement would be slippery if the sprinkler were off, given that currently the pavement *is* slippery," cannot be evaluated from the conditional probabilities $P(x_i|pa_i)$ alone; the functional forms of the f_i's (Eq. 65) are necessary for evaluating such queries [Balke and Pearl 1994; Pearl 1996].

To illustrate the evaluation of counterfactuals, consider a deterministic version of the model given by Eq. (7) assuming that the only uncertainty in the model lies in the identity of the season, summarized by a probability distribution $P(u_1)$ (or $P(x_1)$.) We observe the ground slippery and the sprinkler on and we wish to assess the probability that the ground would be slippery had the sprinkler been off. Formally, the quantity desired is given by

$$P(X_{5_{x_3=0}} = 1|X_5 = 1, X_3 = 1)$$

According to Eq. (64), the expression above is evaluated by summing over all states of U that are compatible with the information at hand. In our example, the only state compatible with the evidence $X_5 = 1$ and $X_3 = 1$ is that which yields $X_1 =$ Summer $\vee$ Spring, and in this state $X_2 =$ no-rain, hence $X_{5_{x_3=0}} = 0$. Thus, matching intuition, we obtain

$$P(X_{5_{x_3=0}} = 1|X_5 = 1, X_3 = 1) = 0.$$

In general, the conditional probability of a counterfactual sentence "If it were A then B", given evidence e, can be computed in three steps:

1. **Abduction** – update $P(u)$ by the evidence e, to obtain $P(u|e)$.

2. **Action** – Modify M by the action $do(A)$, where A is the antecedent of the counterfactual, to obtain the submodel M_A.

3. **Deduction** – Use the updated probability $P(u|e)$ in conjunction with M_A to compute the probability of the counterfactual consequence B.

Effective methods of computing probabilities of counterfactuals are presented in Balke and Pearl (1994, 1995).

A.3 Relation to Neyman-Rubin model

Many of the concepts defined in this appendix bear similarity to parallel concepts in the potential-outcome model used by Neyman (1923) and Rubin (1974) in the analysis treatment effects. In that model, $Y_x(u)$ stands for the outcome of experimental unit u (e.g., a subject, or an agricultural lot) under experimental condition $X = x$, and is taken as a primitive, that is, an undefined starting point. In the structural model framework, the quantity $Y_x(u)$ is not a primitive, but is derived mathematically from a set of equations that is defined by the operator $do(X = x)$. The variable U represents any set of exogenous factors relevant to the analysis, not necessarily the identity of a specific individual in the population. Using this semantics, it is possible to derive a complete axiomatic characterization of the constraints that govern the potential response function $Y_x(u)$ vis-a-vis those that govern directly observed variables, such as $X(u)$ and $Y(u)$ [Galles and Pearl, 1998; Halpern, 1998]. Among these basic axioms, we find the consistency condition [Robins, 1986]:

$$(X = x) \Rightarrow (Y_x = Y)$$

stating that if we intervene and enforce an experimental condition $X = x$ that equals precisely to the one prevailing before the intervention, we should not expect to see any difference in the response variable Y. (For example, a subject who selected treatment $X = x$ by choice and responded with $Y = y$ would respond in exactly the same way to treatment $X = x$ under controlled experiment.) This axiom is used in several of the derivations of Section 2.