Annals of Mathematics and Artificial Intelligence 28 (2000) 287–313

Probabilities of causation: Bounds and identification

Jin Tian and Judea Pearl

Cognitive Systems Laboratory, Computer Science Department, University of California, Los Angeles, CA 90024, USA
E-mail: {jtian,judea}@cs.ucla.edu

This paper deals with the problem of estimating the probability of causation, that is, the probability that one event was the real cause of another, in a given scenario. Starting from structural-semantical definitions of the probabilities of necessary or sufficient causation (or both), we show how to bound these quantities from data obtained in experimental and observational studies, under general assumptions concerning the data-generating process. In particular, we strengthen the results of Pearl [39] by presenting sharp bounds based on combined experimental and nonexperimental data under no process assumptions, as well as under the mild assumptions of exogeneity (no confounding) and monotonicity (no prevention). These results delineate more precisely the basic assumptions that must be made before statistical measures such as the excess-risk-ratio could be used for assessing attributional quantities such as the probability of causation.

1. Introduction

Assessing the likelihood that one event *was the cause* of another guides much of what we understand about (and how we act in) the world. For example, few of us would take aspirin to combat headache if it were not for our conviction that, with high probability, it was aspirin that "actually caused" relief in previous headache episodes. Likewise, according to common judicial standard, judgment in favor of plaintiff should be made if and only if it is "more probable than not" that the defendant's action was a *cause* for the plaintiff's injury (or death). This paper deals with the question of estimating the probability of causation from statistical data.

Causation has two faces, *necessary* and *sufficient*. The most common conception of causation – that the effect E would not have occurred in the absence of the cause C – captures the notion of "necessary causation". Competing notions such as "sufficient cause" and "necessary-and-sufficient cause" are also of interest in a number of applications, and this paper analyzes the relationships among the three notions. Although the distinction between necessary and sufficient causes goes back to J.S. Mill [35], it has received semi-formal explications only in the 1960s – via conditional probabilities [20] and logical implications [32]. These explications suffer from basic semantical difficulties [30;40, pp. 249–256, 313–316] and they do not yield effective procedures for computing probabilities of causes. This paper defines probabilities of causes in a

language of counterfactuals that is based on a simple model-theoretic semantics (to be formulated in section 2).

Robins and Greenland [44] gave a counterfactual definition for the probability of necessary causation taking counterfactuals as primitives, and assuming that one is in possession of a consistent joint probability function on both ordinary and counterfactual events. Pearl [39] gave definitions for the probabilities of necessary or sufficient causation (or both) based on structural model semantics, which defines counterfactuals as quantities derived from modifiable sets of functions [17,18,23,40]. The structural models semantics, as we shall see in section 2, leads to effective procedures for computing probabilities of counterfactual expressions from a given causal theory [4,5]. Additionally, this semantics can be characterized by a complete set of axioms [18,23], which we will use as inference rules in our analysis.

The central aim of this paper is to estimate probabilities of causation from frequency data, as obtained in experimental and observational statistical studies. In general, such probabilities are *non-identifiable*, that is, non-estimable from frequency data alone. One factor that hinders identifiability is confounding – the cause and the effect may both be influenced by a third factor. Moreover, even in the absence of confounding, probabilities of causation are sensitive to the data-generating process, namely, the functional relationships that connect causes and effects [4,44]. Nonetheless, useful information in the form of *bounds* on the probabilities of causation can be extracted from empirical data without actually knowing the data-generating process. These bounds improve when data from observational and experimental studies are combined. Additionally, under certain assumptions about the data-generating process (such as exogeneity and monotonicity), the bounds may collapse to point estimates, which means that the probabilities of causation are identifiable – they can be expressed in terms of probabilities of observed quantities. These estimates will be recognized as familiar expressions that often appear in the literature as measures of *attribution*. Our analysis thus explicates the assumptions about the data-generating process that must be ascertained before those measures can legitimately be interpreted as probabilities of causation.

The analysis of this paper leans heavily on results reported in [39;40, pp. 283–308]. Pearl derived bounds and identification conditions under certain assumptions of exogeneity and monotonicity, and this paper improves on Pearl's results by narrowing his bounds and weakening his assumptions. In particular, we show that for most of Pearl's results, the assumption of strong exogeneity can be replaced by weak exogeneity (to be defined in section 4.3). Additionally, we show that the point estimates that Pearl obtained under the assumption of monotonicity (definition 14) constitute valid lower bounds when monotonicity is not assumed. Finally, we prove that the bounds derived by Pearl, as well as those provided in this paper are *sharp*, that is, they cannot be improved without strengthening the assumptions.

The rest of the paper is organized as follows. Section 2 reviews the structural model semantics of actions, counterfactuals and probability of counterfactuals. In section 3 we present formal definitions for the probabilities of causation and briefly discuss their applicability in epidemiology, artificial intelligence, and legal reasoning. In sec-

tion 4 we systematically investigate the maximal information (about the probabilities of causation) that can be obtained under various assumptions and from various types of data. Section 5 illustrates, by example, how the results presented in this paper can be applied to resolve issues of attribution in legal settings. Section 6 concludes the paper.

2. Structural model semantics

This section presents a brief summary of the structural-equation semantics of counterfactuals as defined in [5,17,18,23]. Related approaches have been proposed in [49] (see footnote 5) and [42]. For detailed exposition of the structural account and its applications see [40].

Structural models are generalizations of the structural equations used in engineering, biology, economics and social science.[1] World knowledge is represented as a collection of stable and autonomous relationships called "mechanisms", each represented as a function, and changes due to interventions or hypothetical eventualities are treated as local modifications of these functions.

A causal model is a mathematical object that assigns truth values to sentences involving causal relationships, actions, and counterfactuals. We will first define causal models, then discuss how causal sentences are evaluated in such models. We will restrict our discussion to recursive (or feedback-free) models; extensions to non-recursive models can be found in [17,18,23].

Definition 1 (Causal model). A *causal model* is a triple

$$M = \langle U, V, F \rangle$$

where

(i) U is a set of variables, called *exogenous*. (These variables will represent background conditions, that is, variables whose values are determined outside the model.)

(ii) V is an ordered set $\{V_1, V_2, \ldots, V_n\}$ of variables, called *endogenous*. (These represent variables that are determined in the model, namely, by 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 $U \times (V_1 \times \cdots \times V_{i-1})$ to V_i. In other words, each f_i tells us the value of V_i given the values of U and all predecessors of V_i. Symbolically, the set of equations F can be represented by writing [2]

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

[1] Similar models, called "neuron diagrams" [22;31, p. 200] are used informally by philosophers to illustrate chains of causal processes.

[2] We use capital letters (e.g., X, Y) as names of variables and sets of variables, and lower-case letters (e.g., x, y) for specific values (called realizations) of the corresponding variables.

where pa_i is any realization of a minimal set of variables PA_i in V (connoting *parents*) sufficient for representing f_i.[3] Likewise, $U_i \subseteq U$ stands for a minimal set of variables in U that is sufficient for representing f_i.

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 (by convention, the exogenous variables are usually not shown explicitly in the graph). 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.

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. To evaluate such sentences we need the notion of "submodel".

Definition 2 (Submodel). Let M be a causal model, X be a set of variables in V, and x be a particular assignment of values to the variables in X. A *submodel* M_x of M is the causal model

$$M_x = \langle U, V, F_x \rangle,$$

where

$$F_x = \{f_i\colon\ V_i \notin X\} \cup \{X = x\}. \tag{1}$$

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 represent the effect of actions and hypothetical changes, including those dictated by counterfactual antecedents. 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 [18].[4]

Definition 3 (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.

[3] A set of variables X is *sufficient* for representing a given function $y = f(x, z)$ if f is trivial in Z – that is, if for every x, z, z' we have $f(x, z) = f(x, z')$.

[4] Structural modifications date back to Marschak [33] and Simon [48]. An explicit translation of interventions into "wiping out" equations from the model was first proposed by Strotz and Wold [52] and later used in [14,38,50,51]. A similar notion of sub-model is introduced in Fine [13], though not specifically for representing actions and counterfactuals.

Definition 4 (Potential response). Let Y be a variable in V, let X be a subset of V, and let u be a particular value of U. The *potential response* of Y to action *do*$(X = x)$ in situation u, denoted $Y_x(u)$, is the (unique) solution for Y of the set of equations F_x.

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 [37]. We will not consider disjunctive actions, of the form "*do*$(X = x$ or $X = x')$", since these complicate the probabilistic treatment of counterfactuals.

Definition 5 (Counterfactual). Let Y be a variable in V, and let X be a subset of V. The counterfactual expression "The value that Y would have obtained, had X been x" is interpreted as denoting the potential response $Y_x(u)$.

Definition 5 thus interprets the counterfactual phrase "had X been x" in terms of a hypothetical external action that modifies the actual course of history and enforces the condition "$X = x$" with minimal change of mechanisms. This is a crucial step in the semantics of counterfactuals [4], as it permits x to differ from the actual value $X(u)$ of X without creating logical contradiction; it also suppresses abductive inferences (or backtracking) from the counterfactual antecedent $X = x$.[5]

It can easily be shown [17] that the counterfactual relationship just defined, $Y_x(u)$, satisfies the following two properties:

- *Effectiveness*: For any two disjoint sets of variables, Y and W, we have

$$Y_{yw}(u) = y. \tag{2}$$

  In words, setting the variables in W to w has no effect on Y, once we set the value of Y to y.

- *Composition*: For any two disjoint sets of variables X and W, and any set of variables Y,

$$W_x(u) = w \Longrightarrow Y_{xw}(u) = Y_x(u). \tag{3}$$

  In words, once we set X to x, setting the variables in W to the same values, w, that they would attain (under x) should have no effect on Y.

Furthermore, effectiveness and composition are *complete* whenever M is recursive (i.e., $G(M)$ is acyclic) [18,23], that is, every property of counterfactuals that follows from the structural model semantics can be derived by repeated application of effectiveness and composition.

A corollary of composition is a property called *consistency* by Robins [43]:

$$\big(X(u) = x\big) \Longrightarrow \big(Y_x(u) = Y(u)\big). \tag{4}$$

[5] Simon and Rescher [49, p. 339] did not include this step in their account of counterfactuals and noted that backward inferences triggered by the antecedents can lead to ambiguous interpretations.

Consistency states that, if in a certain context u we find variable X at value x, and we intervene and set X to that same value, x, we should not expect any change in the response variable Y. This property will be used in several derivations of sections 3 and 4.

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

Definition 6 (Probabilistic causal model). A *probabilistic causal model* is a pair

$$\langle M, P(u) \rangle,$$

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

$$P(y) \triangleq P(Y = y) = \sum_{\{u|Y(u)=y\}} P(u). \tag{5}$$

The probability of counterfactual statements is defined in the same manner, through the function $Y_x(u)$ induced by the submodel M_x. For example, the *causal effect* of x on y is defined as

$$P(Y_x = y) = \sum_{\{u|Y_x(u)=y\}} P(u). \tag{6}$$

Likewise, a probabilistic 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

$$P(Y_x = y, X = x') = \sum_{\{u|Y_x(u)=y \,\&\, X(u)=x'\}} P(u) \tag{7}$$

and

$$P(Y_x = y, Y_{x'} = y') = \sum_{\{u|Y_x(u)=y \,\&\, Y_{x'}(u)=y'\}} P(u). \tag{8}$$

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 [10]. The definition of Y_x and $Y_{x'}$ in terms of two distinct submodels, driven by a standard probability space over U, demonstrates that joint probabilities of counterfactuals have

solid mathematical and conceptual underpinning and, moreover, these probabilities can be encoded rather parsimoniously using $P(u)$ and F.

In particular, the probabilities of causation analyzed in this paper (see equations (10)–(12)) require the evaluation of expressions of the form $P(Y_{x'} = y' \mid X = x, Y = y)$ with x and y incompatible with x' and y', respectively. Equation (7) allows the evaluation of this quantity as follows:

$$P\big(Y_{x'} = y' \mid X = x, Y = y\big) = \frac{P(Y_{x'} = y', X = x, Y = y)}{P(X = x, Y = y)} = \sum_u P\big(Y_{x'}(u) = y'\big) P(u \mid x, y). \tag{9}$$

In other words, we first update $P(u)$ to obtain $P(u \mid x, y)$, then we use the updated distribution $P(u \mid x, y)$ to compute the expectation of the propositional variable $Y_{x'}(u) = y'$.[6]

3. Probabilities of causation: Definitions

In this section, we present the definitions for the three aspects of causation as defined in [39]. We use the counterfactual language and the structural model semantics introduced in section 2. For notational simplicity, we limit the discussion to binary variables; extension to multi-valued variables are straightforward (see [40, p. 286, footnote 5]).

Definition 7 (Probability of necessity (PN)). Let X and Y be two binary variables in a causal mode M, let x and y stand for the propositions $X =$ true and $Y =$ true, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$PN \triangleq P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true}) \triangleq P\big(y'_{x'} \mid x, y\big). \tag{10}$$

In other 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.[7]

This quantity has applications in epidemiology, legal reasoning, and artificial intelligence (AI). Epidemiologists have long been concerned with estimating the prob-

[6] In our deterministic model, $P(Y_{x'}(u) = y')$ takes on the values zero and one, but in models involving intrinsic nondeterminism (see section 6), or memoryless stochastic fluctuations, $P(Y_{x'}(u) = y')$ expresses the residual uncertainty in Y, under the setting $X = x'$, in situation $U = u$. Equation (9) then captures the uncertainty associated with the effect of action *do*$(X = x')$, conditioned on the pre-action evidence $X = x$ and $Y = y$.

[7] Note a slight change in notation relative to that used section 2. Lower case letters (e.g., x, y) denoted arbitrary values of variables in section 2, and now stand for propositions (or events). Note also the abbreviations y_x for $Y_x =$ true and y'_x for $Y_x =$ false. Readers accustomed to writing "$A > B$" for the counterfactual "B if it were A" can translate equation (10) to read PN $\triangleq P(x' > y' \mid x, y)$.

ability 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 [44] called the "probability of causation", measures how *necessary* the cause is for the production of the effect. It is used frequently in lawsuits, where legal responsibility is at the center of contention (see section 5).

Definition 8 (Probability of sufficiency (PS)).

$$\text{PS} \triangleq P\big(y_x \mid y', x'\big). \tag{11}$$

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. Thus, mirroring the necessity of x (as measured by PN), PS gives the probability that setting x would produce y in a situation where x and y are in fact absent.

PS 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 [29]. Counterfactually, this notion is expressed as the "probability that a healthy unexposed individual would have gotten the disease had he/she been exposed". In psychology, PS serves as the basis for Cheng's [8] causal power theory, which attempts to explain how humans judge causal strength among events. In AI, PS plays a major role in the generation of explanations [40, pp. 221–223].

Definition 9 (Probability of necessity and sufficiency (PNS)).

$$\text{PNS} \triangleq P\big(y_x, y'_{x'}\big). \tag{12}$$

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.

As illustrated above, 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 in question, that are capable of explaining 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 [41]. Still, lack of sufficiency may weaken arguments based on PN [21,34].

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

Lemma 1. The probabilities of causation satisfy the following relationship:

$$\text{PNS} = P(x, y)\text{PN} + P\big(x', y'\big)\text{PS}. \tag{13}$$

Proof. Using the consistency condition of equation (4),

$$x \Rightarrow (y_x = y), \qquad x' \Rightarrow (y_{x'} = y), \tag{14}$$

we can write

$$\begin{aligned} y_x \wedge y'_{x'} &= \left(y_x \wedge y'_{x'}\right) \wedge \left(x \vee x'\right) \\ &= \left(y_x \wedge x \wedge y'_{x'}\right) \vee \left(y_x \wedge y'_{x'} \wedge x'\right) \\ &= \left(y \wedge x \wedge y'_{x'}\right) \vee \left(y_x \wedge y' \wedge x'\right). \end{aligned}$$

Taking probabilities on both sides, and using the disjointness of x and x', we obtain:

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

which proves the lemma. □

Definition 10 (Identifiability). Let $Q(M)$ be any quantity defined on a causal model M. Q is identifiable in a class $\boldsymbol{M}$ of models iff 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 other 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 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. However, useful information in the form of bounds may be derived for these quantities from $P(v)$, especially when knowledge about causal effects $P(y_x)$ and $P(y_{x'})$ are also available.[8] Moreover, under some general assumptions about the data-generating process, these quantities may even be identified.

To formulate precisely what it means to identify a counterfactual quantity from various types of data, we now generalize definition 10 to capture the notion of "identification from experiments". By *experiment* we mean a prescribed modification of the underlying causal model, together with the probability distribution that the modified model induces on the variables observed in the experiment.

[8] The causal effects $P(y_x)$ and $P(y_{x'})$ can be estimated reliably from controlled experimental studies, and from certain observational (i.e., nonexperimental) studies which permit the control of confounding through adjustment of covariates [38].

Definition 11 (Identifiability from experiments). Let $Q(M)$ be any quantity defined on a causal model M, let M^{exp} be a modification of M induced by some experiment, *exp*, and let Y be a set of variables observed under *exp*. We say that Q is *identifiable from experiment exp* in a class $\boldsymbol{M}$ of models iff any two models M_1 and M_2 from $\boldsymbol{M}$ that satisfy $P_{M_1^{\text{exp}}}(y) = P_{M_2^{\text{exp}}}(y)$ also satisfy $Q(M_1) = Q(M_2)$. In other words, Q is identifiable from *exp* if it can be determined uniquely from the probability distribution that the observed variables Y attain under the experimental conditions created by *exp*.

In the sequel, we will consider standard controlled experiments, in which the values of the control variable X are assigned at random. The outcomes of such experiments are the causal effects probabilities, $P(y_x)$ and $P(y_{x'})$, which are also induced by the submodels M_x and $M_{x'}$, respectively. However, definition 11 is applicable to a much broader class of experimental designs, corresponding to both deletion and replacement of the model equations. Note that standard identifiability (definition 10) is a special case of identifiability from experiments, where $Y = V$ and $M^{\text{exp}} = M$.

4. Bounds and conditions of identification

In this section we estimate the three probabilities of causation defined in section 3 when given experimental or nonexperimental data (or both) and additional assumptions about the data-generating process. We will assume that experimental data will be summarized in the form of the causal effects $P(y_x)$ and $P(y_{x'})$, and nonexperimental data will be summarized in the form of the joint probability function: $P_{XY} = \{P(x, y), P(x', y), P(x, y'), P(x', y')\}$.[9]

4.1. Linear programming formulation

In principle, in order to compute the probability of any counterfactual sentence involving variables X and Y we need to specify a causal model, namely, the functional relation between X and Y and the probability distribution on U. However, since every such model induces a joint probability distribution on the four binary variables: X, Y, Y_x and $Y_{x'}$, specifying the sixteen parameters of this distribution would suffice. Moreover, since Y is a deterministic function of the other three variables, the problem is fully specified by the following set of eight parameters:

$$
\begin{aligned}
p_{111} &= P\big(y_x, y_{x'}, x\big) = P\big(x, y, y_{x'}\big),\\
p_{110} &= P\big(y_x, y_{x'}, x'\big) = P\big(x', y, y_x\big),\\
p_{101} &= P\big(y_x, y'_{x'}, x\big) = P\big(x, y, y'_{x'}\big),
\end{aligned}
$$

[9] For example, if x represents a specific exposure and y represents the outcome of a specific individual I, then P_{XY} is estimated from sampled frequency counts in a population that is deemed representative of the relevant characteristics of I. The choice of an appropriate reference population is usually based on causal consideration (often suppressed), and involves matching the characteristics of I against the causal model (M, $P(u)$) judged to govern the population.

$$
\begin{aligned}
p_{100} &= P\left(y_x, y'_{x'}, x'\right) = P\left(x', y', y_x\right),\\
p_{011} &= P\left(y'_x, y_{x'}, x\right) = P\left(x, y', y_{x'}\right),\\
p_{010} &= P\left(y'_x, y_{x'}, x'\right) = P\left(x', y, y'_x\right),\\
p_{001} &= P\left(y'_x, y'_{x'}, x\right) = P\left(x, y', y'_{x'}\right),\\
p_{000} &= P\left(y'_x, y'_{x'}, x'\right) = P\left(x', y', y'_x\right),
\end{aligned}
$$

where we have used the consistency condition (14). These parameters are constrained by the probabilistic constraints

$$
\begin{aligned}
&\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} p_{ijk} = 1,\\
&p_{ijk} \geqslant 0 \quad \text{for } i, j, k \in \{0, 1\}.
\end{aligned}
\tag{15}
$$

In addition, the nonexperimental probabilities P_{XY} impose the constraints

$$
\begin{aligned}
p_{111} + p_{101} &= P(x, y),\\
p_{011} + p_{001} &= P\left(x, y'\right),\\
p_{110} + p_{010} &= P\left(x', y\right)
\end{aligned}
\tag{16}
$$

and the causal effects, $P(y_x)$ and $P(y_{x'})$, impose the constraints:

$$
\begin{aligned}
P(y_x) &= p_{111} + p_{110} + p_{101} + p_{100},\\
P(y_{x'}) &= p_{111} + p_{110} + p_{011} + p_{010}.
\end{aligned}
\tag{17}
$$

The quantities we wish to bound are:

$$\text{PNS} = p_{101} + p_{100}, \tag{18}$$

$$\text{PN} = \frac{p_{101}}{P(x, y)}, \tag{19}$$

$$\text{PS} = \frac{p_{100}}{P(x', y')}. \tag{20}$$

In the following sections we obtain bounds for these quantities by solving various linear programming problems. For example, given both experimental and nonexperimental data, the lower (and upper) bounds for PNS are obtained by minimizing (or maximizing, respectively) $p_{101} + p_{100}$ subject to the constraints (15)–(17). The bounds obtained are guaranteed to be sharp because the optimization is global.

Optimizing the functions in (18)–(20), subject to equality constraints, defines a linear programming (LP) problem that lends itself to closed-form solution. Balke [3, appendix B] describes a computer program that takes symbolic descriptions of LP problems and returns symbolic expressions for the desired bounds. The program works by systematically enumerating the vertices of the constraint polygon of the dual problem. The bounds reported in this paper were produced (or tested) using Balke's

program, and will be stated here without proofs; their correctness can be verified by manually enumerating the vertices as described in [3, appendix B].

4.2. Bounds with no assumptions

4.2.1. Given nonexperimental data

Given P_{XY}, constraints (15) and (16) induce the following upper bound on PNS:

$$0 \leqslant \text{PNS} \leqslant P(x, y) + P\big(x', y'\big). \tag{21}$$

However, PN and PS are not constrained by P_{XY}.

These constraints also induce bounds on the causal effects $P(y_x)$ and $P(y_{x'})$:

$$\begin{aligned} P(x, y) &\leqslant P(y_x) \leqslant 1 - P\big(x, y'\big), \\ P(x', y) &\leqslant P(y_{x'}) \leqslant 1 - P\big(x', y'\big). \end{aligned} \tag{22}$$

4.2.2. Given causal effects

Given constraints (15) and (17), the bounds induced on PNS are:

$$\max\big[0, P(y_x) - P(y_{x'})\big] \leqslant \text{PNS} \leqslant \min\big[P(y_x), P\big(y'_{x'}\big)\big] \tag{23}$$

with no constraints on PN and PS.

4.2.3. Given both nonexperimental data and causal effects

Given the constraints (15)–(17), the following bounds are induced on the three probabilities of causation:

$$\max\left\{\begin{matrix} 0 \\ P(y_x) - P(y_{x'}) \\ P(y) - P(y_{x'}) \\ P(y_x) - P(y) \end{matrix}\right\} \leqslant \text{PNS} \leqslant \min\left\{\begin{matrix} P(y_x) \\ P\big(y'_{x'}\big) \\ P(x, y) + P\big(x', y'\big) \\ P(y_x) - P(y_{x'}) + P\big(x, y'\big) + P\big(x', y\big) \end{matrix}\right\}, \tag{24}$$

$$\max\left\{\begin{matrix} 0 \\ \dfrac{P(y) - P(y_{x'})}{P(x, y)} \end{matrix}\right\} \leqslant \text{PN} \leqslant \min\left\{\begin{matrix} 1 \\ \dfrac{P(y'_{x'}) - P(x', y')}{P(x, y)} \end{matrix}\right\}, \tag{25}$$

$$\max\left\{\begin{matrix} 0 \\ \dfrac{P(y_x) - P(y)}{P(x', y')} \end{matrix}\right\} \leqslant \text{PS} \leqslant \min\left\{\begin{matrix} 1 \\ \dfrac{P(y_x) - P(x, y)}{P(x', y')} \end{matrix}\right\}. \tag{26}$$

Thus we see that some information about PN and PS can be extracted without making any assumptions about the data-generating process. Furthermore, combined data from both experimental and nonexperimental studies yield information that neither study alone can provide.

4.3. Bounds under exogeneity (no confounding)

Definition 12 (Exogeneity). A variable X is said to be exogenous for Y in model M iff

$$P(y_x) = P(y \mid x) \quad \text{and} \quad P(y_{x'}) = P\big(y \mid x'\big), \tag{27}$$

or, equivalently,

$$Y_x \perp\!\!\!\perp X \quad \text{and} \quad Y_{x'} \perp\!\!\!\perp X. \tag{28}$$

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

Equation (27) has been given a variety of (equivalent) definitions and interpretations. Epidemiologists refer to this condition as "no-confounding" [44], statisticians call it "as if randomized", and Rosenbaum and Rubin [45] call it "weak ignorability". A graphical criterion ensuring exogeneity is the absence of a common ancestor of X and Y in $G(M)$ (more precisely, a common ancestor that is connected to Y through a path not containing X, including latent ancestors, which represent dependencies among variables in U). The classical econometric criterion for exogeneity (e.g., [11, p. 169]) states that X be independent of the error term (u) in the equation for Y.[10] We will use the term "exogeneity", since it was under this term that the relations given in (27) first received their precise definition (by economists).

Combining equation (27) with the constraints of (15)–(17), the linear programming optimization (section 4.1) yields the following results:

Theorem 1. Under condition of exogeneity, the three probabilities of causation are bounded as follows:

$$\max\big[0, P(y \mid x) - P\big(y \mid x'\big)\big] \leqslant \text{PNS} \leqslant \min\big[P(y \mid x), P\big(y' \mid x'\big)\big], \tag{29}$$

$$\frac{\max[0, P(y \mid x) - P(y \mid x')]}{P(y \mid x)} \leqslant \text{PN} \leqslant \frac{\min[P(y \mid x), P(y' \mid x')]}{P(y \mid x)}, \tag{30}$$

$$\frac{\max[0, P(y \mid x) - P(y \mid x')]}{P(y' \mid x')} \leqslant \text{PS} \leqslant \frac{\min[P(y \mid x), P(y' \mid x')]}{P(y' \mid x')}. \tag{31}$$

The bounds expressed in equation (30) were first derived by Robins and Greenland [44]; a more elaborate proof can be found in [15]. Pearl [39] derived equations (29)–(31) under a stronger condition of exogeneity (see definition 13). We see that under the condition of no-confounding the lower bound for PN can be expressed as

$$\text{PN} \geqslant 1 - \frac{1}{P(y \mid x)/P(y \mid x')} \triangleq 1 - \frac{1}{\text{RR}}, \tag{32}$$

[10] This criterion has been the subject of relentless objections by modern econometricians [12,25,27], but see [1;40, pp. 169–170; 245–247] for a reconciliatory perspective on this controversy.

where RR $= P(y \mid x)/P(y \mid x')$ is the *risk ratio* (also called *relative risk*) in epidemiology. Courts have often used the condition RR > 2 as a criterion for legal responsibility [2]. Equation (32) shows that this practice represents a conservative interpretation of the "more probable than not" standard (assuming no confounding); PN must indeed be higher than 0.5 if RR exceeds 2. Freedman and Stark [15] argue that, in general, epidemiological evidence may not be applicable as proof for specific causation [15] because such evidence cannot account for all characteristics specific to the plaintiff. Freedman and Stark further imply that the appropriate way of interpreting the "more probable than not" criterion would be to consider the probability of causation in a restricted subpopulation, one that shares the plaintiff characteristics. Taken to extreme, such restrictive interpretation would insist on characterizing the plaintiff to minute detail, and would reduce PN to zero or one when all relevant details are accounted for. We doubt that this interpretation underlies the intent of judicial standards. We believe that, by using the wording "more probable than not", law makers have instructed us to ignore specific features for which data is not available, and to base our determination on the most specific features for which reliable data is available (see footnote 9).[11] PN ensures us that two obvious features of the plaintiff will not be ignored: the exposure, x, and the injury, y. In contrast, these two features are ignored in the causal effect measure $P(y_x)$ which is a quantity averaged over the entire population, including unexposed and uninjured.

4.3.1. Bounds under strong exogeneity

The condition of exogeneity, as defined in equation (27) is testable by comparing experimental and nonexperimental data. A stronger version of exogeneity can be defined as the joint independence $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$ which was called "strong ignorability" by Rosenbaum and Rubin [45]. Though untestable, such joint independence is assumed to hold when we assert the absence of factors that simultaneously affect exposure and outcome.

Definition 13 (Strong Exogeneity). A variable X is said to be strongly exogenous for Y in model M iff $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$, that is,

$$\begin{aligned} P(y_x, y_{x'} \mid x) &= P(y_x, y_{x'}), \\ P(y_x, y'_{x'} \mid x) &= P(y_x, y'_{x'}), \\ P(y'_x, y_{x'} \mid x) &= P(y'_x, y_{x'}), \\ P(y'_x, y'_{x'} \mid x) &= P(y'_x, y'_{x'}). \end{aligned} \tag{33}$$

The four conditions in (33) are sufficient to represent $\{Y_x, Y_{x'}\} \perp\!\!\!\perp X$, because for every event E we have

$$P(E \mid x) = P(E) \Longrightarrow P(E \mid x') = P(E). \tag{34}$$

[11] Our results remain valid when we condition P_{XY} on a set of covariates that characterize the specific case at hand.

Remarkably, the added constraints introduced by strong exogeneity do not alter the bounds of equations (29)–(31). They do, however, strengthen lemma 1:

Theorem 2. If strong exogeneity holds, the probabilities PN, PS, and PNS are constrained by the bounds of equations (29)–(31), and, moreover, PN, PS, and PNS are related to each other as follows [39]:

$$\text{PN} = \frac{\text{PNS}}{P(y \mid x)}, \tag{35}$$

$$\text{PS} = \frac{\text{PNS}}{P(y' \mid x')}. \tag{36}$$

4.4. Identifiability under monotonicity

Definition 14 (Monotonicity). A variable Y is said to be monotonic relative to variable X in a causal model M iff

$$y'_x \wedge y_{x'} = \text{false}. \tag{37}$$

Monotonicity expresses the assumption that a change from $X = \text{false}$ to $X = \text{true}$ cannot, under any circumstance make Y change from true to false. In epidemiology, this assumption is often expressed as "no prevention", that is, no individual in the population can be helped by exposure to the risk factor. Balke and Pearl [6] used this assumption to tighten bounds of treatment effects from studies involving noncompliance. Glymour [19] and Cheng [8] resort to this assumption in using disjunctive or conjunctive relationships between causes and effects, excluding functions such as exclusive-or, or parity.

In the linear programming formulation of section 4.1, monotonicity narrows the feasible space to the manifold:

$$\begin{aligned} p_{011} &= 0, \\ p_{010} &= 0. \end{aligned} \tag{38}$$

4.4.1. Given nonexperimental data

Under the constraints (15), (16), and (38), we find the same bounds for PNS as the ones obtained under no assumptions (equation (21)). Moreover, there are still no constraints on PN and PS. Thus, with nonexperimental data alone, the monotonicity assumption does not provide new information.

However, the monotonicity assumption induces sharper bounds on the causal effects $P(y_x)$ and $P(y_{x'})$:

$$\begin{aligned} P(y) \leqslant P(y_x) &\leqslant 1 - P\big(x, y'\big), \\ P\big(x', y\big) \leqslant P(y_{x'}) &\leqslant P(y). \end{aligned} \tag{39}$$

Compared with equation (22), the lower bound for $P(y_x)$ and the upper bound for $P(y_{x'})$ are tightened. The importance of equation (39) lies in providing a simple necessary test for the assumption of monotonicity. These inequalities are sharp, in the sense that every combination of experimental and non-experimental 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 equation (39) can be used for testing the compatibility of the experimental and non-experimental data, namely, whether subjects used in clinical trials were sampled from the same target population, characterized by the joint distribution P_{XY}.

4.4.2. Given causal effects

Constraints (15), (17), and (38) induce no constraints on PN and PS, while the value of PNS is fully determined:

$$\text{PNS} = P\left(y_x, y'_{x'}\right) = P(y_x) - P(y_{x'}).$$

That is, under the assumption of monotonicity, PNS can be determined by experimental data alone, despite the fact that the joint event $y_x \wedge y'_{x'}$ can never be observed.

4.4.3. Given both nonexperimental data and causal effects

Under the constraints (15)–(17) and (38), the values of PN, PS, and PNS are all determined precisely.

Theorem 3. If Y is monotonic relative to X, then PNS, PN, and PS are given by

$$\text{PNS} = P\left(y_x, y'_{x'}\right) = P(y_x) - P(y_{x'}), \tag{40}$$

$$\text{PN} = P\left(y'_{x'} \mid x, y\right) = \frac{P(y) - P(y_{x'})}{P(x, y)}, \tag{41}$$

$$\text{PS} = P\left(y_x \mid x', y'\right) = \frac{P(y_x) - P(y)}{P(x', y')}. \tag{42}$$

Corollary 1. If Y is monotonic relative to X, then PNS, PN, and PS are identifiable whenever the causal effects $P(y_x)$ and $P(y_{x'})$ are identifiable,

Equations (40)–(42) are applicable to situations where, in addition to observational probabilities, we also have information about the causal effects $P(y_x)$ and $P(y_{x'})$. Such information may be obtained either directly, through separate experimental studies, or indirectly, from observational studies in which certain identifying assumptions are deemed plausible (e.g., assumptions that permits identification through adjustment of covariates). Note that the identification of PN requires only $P(y_{x'})$ while that of PS requires $P(y_x)$. In practice, however, any method that yields the former also yields the latter.

One common class of models which permits the identification of $P(y_x)$ is called *Markovian*.

Definition 15 (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 iff $G(M)$ 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 [36,38] that for every two variables, X and Y, in a positive-Markovian model M, the causal effects $P(y_x)$ and $P(y_{x'})$ are identifiable and are given by

$$\begin{aligned} P(y_x) &= \sum_{pa_X} P(y \mid pa_X, x) P(pa_X), \\ P(y_{x'}) &= \sum_{pa_X} P\bigl(y \mid pa_X, x'\bigr) P(pa_X), \end{aligned} \tag{43}$$

where pa_X are (values of) the *parents* of X in the causal graph associate with M (see also [40, p. 73;42,51]. Thus, we can combine equation (43) with theorem 3 and obtain a concrete condition for the identification of the probability of causation.

Corollary 2. 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 equations (40)–(42), with $P(y_x)$ given in equation (43). If monotonicity cannot be ascertained, then PNS, PN and PS are bounded by equations (24)–(26), with $P(y_x)$ given in equation (43).

A broader identification condition can be obtained through the use of the back-door and front-door criteria [38], which are applicable to semi-Markovian models. These were further generalized in [16][12] and lead to the following corollary:

Corollary 3. Let $\boldsymbol{GP}$ be the class of semi-Markovian models that satisfy the graphical criterion of Galles and Pearl [16]. If $Y_x(u)$ is monotonic, then the probabilities of causation PNS, PS and PN are identifiable in $\boldsymbol{GP}$ and are given by equations (40)–(42), with $P(y_x)$ determined by the topology of $G(M)$ through the GP criterion.

4.5. Identifiability under monotonicity and exogeneity

Under the assumption of monotonicity, if we further assume exogeneity, then $P(y_x)$ and $P(y_{x'})$ are identified through equation (27), and from theorem 3 we conclude that PNS, PN, and PS are all identifiable.

[12] Galles and Pearl [16] 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 $P(y_x)$. See also [40, pp. 114–118].

Theorem 4 (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

$$\text{PNS} = P(y \mid x) - P\big(y \mid x'\big), \tag{44}$$

$$\text{PN} = \frac{P(y) - P(y \mid x')}{P(x, y)} = \frac{P(y \mid x) - P(y \mid x')}{P(y \mid x)}, \tag{45}$$

$$\text{PS} = \frac{P(y \mid x) - P(y)}{P(x', y')} = \frac{P(y \mid x) - P(y \mid x')}{P(y' \mid x')}. \tag{46}$$

These expressions are to be recognized as familiar measures of attribution that often appear in the literature. The r.h.s. of (44) is called "risk-difference" in epidemiology, and is also misnamed "attributable risk" [26, p. 87]. The probability of necessity, PN, is given by the *excess-risk-ratio* (ERR)

$$\text{PN} = \frac{P(y \mid x) - P(y \mid x')}{P(y \mid x)} = 1 - \frac{1}{\text{RR}} \tag{47}$$

often misnamed as the *attributable fraction* [46], *attributable-rate percent* [26, p. 88], *attributed fraction for the exposed* [28, p. 38], or *attributable proportion* [9]. The reason we consider these labels to be misnamed is that ERR invokes purely statistical relationships, hence it cannot in itself serve to measure attribution, unless fortified with some causal assumptions. Exogeneity and monotonicity are the causal assumptions that endow ERR with attributional interpretation, and these assumptions are rarely made explicit in the literature on attribution.

The expression for PS is likewise quite revealing

$$\text{PS} = \frac{P(y \mid x) - P(y \mid x')}{1 - P(y \mid x')}, \tag{48}$$

as it coincides with what epidemiologists call the "relative difference" [47], which is used to measure the *susceptibility* of a population to a risk factor x. It also coincides with what Cheng calls "causal power" [8], namely, the effect of x on y after suppressing "all other causes of y". See [39] for additional discussions of these expressions.

To appreciate the difference between equations (41) and (47) we can rewrite equation (41) as

$$\begin{aligned}\text{PN} &= \frac{P(y \mid x)P(x) + P(y \mid x')P(x') - P(y_{x'})}{P(y \mid x)P(x)} \\ &= \frac{P(y \mid x) - P(y \mid x')}{P(y \mid x)} + \frac{P(y \mid x') - P(y_{x'})}{P(x, y)}.\end{aligned} \tag{49}$$

The first term on the r.h.s. of (49) is the familiar ERR as in (47), and represents the value of PN under exogeneity. The second term represents the correction needed to account for X's non-exogeneity, i.e., $P(y_{x'}) \neq P(y \mid x')$. We will call the r.h.s. of (49) by corrected excess-risk-ratio (CERR).

From equations (44)–(46) we see that the three notions of causation satisfy the simple relationships given by equations (35) and (36) which we obtained under the strong exogeneity condition. In fact, we have the following theorem.

Theorem 5. Monotonicity (37) and exogeneity (27) together imply strong exogeneity (33).

Proof. From the monotonicity condition, we have

$$y_{x'} = y_{x'} \wedge \left(y_x \vee y'_x\right) = (y_{x'} \wedge y_x) \vee \left(y_{x'} \vee y'_x\right) = y_{x'} \wedge y_x. \tag{50}$$

Thus we can write

$$P(y_{x'}) = P(y_x, y_{x'}), \tag{51}$$

and

$$P\left(y \mid x'\right) = P\left(y_{x'} \mid x'\right) = P\left(y_x, y_{x'} \mid x'\right), \tag{52}$$

where consistency condition (14) is used. The exogeneity condition (27) allows us to equate (51) and (52), and we obtain

$$P\left(y_x, y_{x'} \mid x'\right) = P(y_x, y_{x'}), \tag{53}$$

which implies the first of the four conditions in (33):

$$P(y_x, y_{x'} \mid x) = P(y_x, y_{x'}). \tag{54}$$

Combining equation (54) with

$$P(y_x) = P(y_x, y_{x'}) + P\left(y_x, y'_{x'}\right), \tag{55}$$

$$P(y \mid x) = P(y_x \mid x) = P(y_x, y_{x'} \mid x) + P\left(y_x, y'_{x'} \mid x\right), \tag{56}$$

and the exogeneity condition (27), we obtain the second equation in (33):

$$P\left(y_x, y'_{x'} \mid x\right) = P\left(y_x, y'_{x'}\right). \tag{57}$$

Both sides of the third equation in (33) are equal to zero from monotonicity condition and the last equation in (33) follows because the four quantities sum up to 1 on both sides of the four equations. □

4.6. Summary of results

We now summarize the results from section 4 that should be of value to practicing epidemiologists and policy makers. These results are shown in table 1, which lists the best estimate of PN under various assumptions and various types of data—the stronger the assumptions, the more informative the estimates.

We see that the excess-risk-ratio (ERR), which epidemiologists commonly identify with the probability of causation, is a valid measure of PN only when two

Table 1
PN (the probability of necessary causation) as a function of assumptions and available data. ERR stands of the excess-risk-ratio $1 - P(y \mid x')/P(y \mid x)$ and CERR is given in equation (49). The non-entries (—) represent vacuous bounds, that is, $0 \leqslant \text{PN} \leqslant 1$.

Assumptions		Data available		
Exogeneity	Monotonicity	Experimental	Nonexperimental	Combined
+	+	ERR	ERR	ERR
+	−	bounds	bounds	bounds
−	+	—	—	CERR
−	−	—	—	bounds

assumptions can be ascertained: exogeneity (i.e., no confounding) and monotonicity (i.e., no prevention). When monotonicity does not hold, ERR provides merely a lower bound for PN, as shown in equation (30). (The upper bound is usually unity.) In the presence of confounding, ERR must be corrected by the additive term $[P(y \mid x') - P(y_{x'})]/P(x, y)$, as stated in (49). In other words, when confounding bias (of the causal effect) is positive, PN is higher than ERR by the amount of this additive term. Clearly, owing to the division by $P(x, y)$, the PN bias can be many times higher than the causal effect bias $P(y \mid x') - P(y_{x'})$. However, confounding results only from association between exposure and other factors that affect the outcome; one need not be concerned with associations between such factors and susceptibility to exposure, as is often assumed in the literature [19,29].

The last two rows in table 1 correspond to no assumptions about exogeneity, and they yield vacuous bounds for PN when data come from either experimental or observational study. In contrast, informative bounds (25) or point estimates (49) are obtained when data from experimental and observational studies are combined. Concrete use of such combination will be illustrated in section 5.

5. An example: 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 minor 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

Table 2
Frequency data (hypothetical) obtained in experimental and nonexperimental studies, comparing deaths (in thousands) among drug users, x, and non-users, x'.

	Experimental		Nonexperimental	
	x	x'	x	x'
Deaths (y)	16	14	2	28
Survivals (y')	984	986	998	972

stating that: (1) counterfactual speculations regarding whether patients would or would not have died are purely metaphysical and should be avoided, 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 non-experimental studies, what the probability is that drug x was in fact the cause of Mr. A's death.

The (hypothetical) data associated with the two studies are shown in table 2. The experimental data provide the estimates

$$P(y_x) = 16/1000 = 0.016,$$
$$P(y_{x'}) = 14/1000 = 0.014,$$
$$P\left(y'_{x'}\right) = 1 - P(y_{x'}) = 0.986.$$

The non-experimental data provide the estimates

$$P(y) = 30/2000 = 0.015,$$
$$P(x, y) = 2/2000 = 0.001,$$
$$P\left(x', y'\right) = 972/2000 = 0.486.$$

Since both the experimental and nonexperimental data are available, we can obtain bounds on all three probabilities of causation through equations (24)–(26) without making any assumptions about the underlying mechanisms. The data in table 2 imply the following numerical results:

$$0.002 \leqslant \text{PNS} \leqslant 0.016, \tag{58}$$
$$1.0 \leqslant \text{PN} \leqslant 1.0, \tag{59}$$
$$0.002 \leqslant \text{PS} \leqslant 0.031. \tag{60}$$

These figures show that although surviving patients who did not take drug x have only less than 3.1% chance to die had they taken the drug, there is 100% assurance (barring sample errors) that those who took the drug and died would have survived had they not taken the drug. Thus the plaintiff was correct; drug x was in fact responsible for the death of Mr. A.

If we assume that drug x can only cause, but never prevent, death, theorem 3 is applicable and equations (40)–(42) yield

$$\text{PNS} = 0.002, \tag{61}$$

$$\text{PN} = 1.0, \tag{62}$$

$$\text{PS} = 0.002. \tag{63}$$

Thus, we conclude that drug x was responsible for the death of Mr. A, with or without the no-prevention assumption.

Note that a straightforward use of the experimental excess-risk-ratio would yield a much lower (and incorrect) result:

$$\frac{P(y_x) - P(y_{x'})}{P(y_x)} = \frac{0.016 - 0.014}{0.016} = 0.125. \tag{64}$$

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 so the proportion of deaths among the control group $P(y_{x'})$ would 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 present 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 contrived to show the usefulness of the bounds in equations (24)–(26). Nevertheless, it is instructive to note that a combination of experimental and non-experimental studies may unravel what experimental studies alone will not reveal. In addition, such combination may provide a test for the assumption of no-prevention, as outlined in section 4.4.1. For example, if the frequencies in table 2 were slightly different, they could easily violate the inequalities of equation (39). Such violation may be due either to nonmonotonicity or to incompatibility of the experimental and nonexperimental groups.

This last point may warrant a word of explanation, lest the reader wonders why two data sets, taken from two separate groups under different experimental conditions, should constrain one another. The explanation is that certain quantities in the two subpopulations are expected to remain invariant to all these differences, provided that the two subpopulations were sampled properly from the same general population. In fact, every quantity of the form $P(Q)$, where Q is computable from a causal model M, enjoys this invariance property, because the two subpopulations are assumed to be governed by the same causal model. Thus, the question whether two data sets, obtained under different experimental conditions, should constrain one another reduces to a purely mathematical question of whether the quantities that represent the two experimental conditions, $P(Q)$ and $P(Q')$, necessarily constrain one another in the same causal model considered. In our case, the quantities in question are simply the causal effects probabilities, $P(y_{x'})$ and $P(y_x)$. Although these probabilities were not

measured in the nonexperimental group, they must nevertheless be the same as those measured in the experimental group. The invariance of these quantities is the basic axiom of controlled experimentation, without which *no* inference would be possible from experimental studies to general behavior of the population. This invariance, together with monotonicity, imply the inequalities of (39).

6. Conclusion

This paper shows how useful information about probabilities of causation can be obtained from experimental and observational studies, with weak or no assumptions about the data-generating process. We have shown that, in general, bounds for the probabilities of causation can be obtained from combined experimental and nonexperimental data. These bounds were proven to be sharp and, therefore, they represent the ultimate information that can be extracted from statistical methods. We have further illustrated the applicability of these results to problems in epidemiology and legal reasoning, and we have clarified the two basic assumptions – exogeneity and monotonicity – that must be ascertained before statistical measures such as excess-risk-ratio could represent attributional quantities such as probability of causation.

It is appropriate at this point to discuss the relation between the assumptions in the example of section 5 (where we have population probabilities and available experiments) and the general framework with which the paper begins (where we have exogenous variables that determine everything and the probabilities enter as an add-on feature). Traditional statisticians might judge the deterministic model incompatible with the stochastic nature of the data, and would be tempted to start the analysis at section 3 (see [15,44]), without the counterfactual model expounded in section 2. However, traditional statistical analysis cannot commence without explicating the quantity we wish to estimate (that is, PN), for which we have no empirical data and for which we have no statistical definition. Instead, our target quantity is defined verbally by law makers as a mixture of probabilistic and deterministic components: "it is more *probable* than not, that the plaintiff injury would not have occurred *but for* the defender action". The "more probable than not" criterion is probabilistic while the "but for" criterion is deterministic, implying counterfactual necessity.

The structural approach expounded in this paper gives a clear semantics to this mixture, typical of counterfactual expressions, and relates it in a natural way to empirical data. The stochastic nature of the data is viewed as emerging from our ignorance of the detailed experimental conditions that prevailed in the study. The exogenous variables in U represent these missing details, and include the physiology and previous history of each person, his/her mental and spiritual attitude, as well as the time and manner in which the exposure occurred. In short, U summarizes all the factors which "determine" in the classical physical sense the outcome of the study. $P(u)$ summarizes our ignorance of those factors.

The main application of our analysis to artificial intelligence lies in the automatic generation of causal explanations, where the distinction between necessary and suf-

ficient causes has important ramifications. As can be seen from the definitions and examples discussed in this paper, necessary causation is a concept tailored to a specific event under consideration (singular causation), whereas 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) then we lose important scenario-specific information. For instance, aiming a gun at and shooting a person from 1,000 meters away will not qualify as an explanation for that person's death, owing to the very low tendency of shots fired from such long distances to hit their marks. This stands contrary to common sense, for when the shot does hit its mark on that singular day, regardless of the reason, the shooter is an obvious culprit for the consequence. If, on the other hand, we base explanations solely on singular-event considerations (i.e., necessary causation), then ambient factors that 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 were it not for the oxygen. That we judge the match struck, not the oxygen, to be the more adequate explanation of the fire indicates that we go beyond necessity considerations.

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, the PN is actually higher for the oxygen if we allow for alternative ways of igniting a spark). Thus, it must be the sufficiency component 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 denoted p_{m} and p_{o}, respectively, then the PS measures associated with these explanations evaluate to PS(match) $= p_{\mathrm{o}}$ and PS(oxygen) $= p_{\mathrm{m}}$, clearly favoring the match when $p_{\mathrm{o}} \gg p_{\mathrm{m}}$. Thus, a robot instructed to explain why a fire broke out has no choice but to consider both PN and PS in its deliberations.

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 the basic relationships between the two components. In [40, chapter 10] it is further shown that PN and PS are too crude for capturing probabilities of causation in multi-stage scenarios, and that the structure of the intermediate process leading from cause to effect must enter the definitions of causation and explanation. Such considerations will be the subject of future investigation (see [24]).

Another important application of probabilities of causation is found in decision making problems, such as those encountered in medicine, system maintenance, and planning under uncertainty. As was pointed out in [40, pp. 217–219], the counterfactual "y would have been true if x were true" can often be translated into a conditional action claim "given that currently x and y are false, y will be true if we do x". The evaluation of such conditional predictions, and the probabilities of such predictions, are commonplace in decision making situations, where actions are brought into focus by certain eventualities that demand remedial correction. In troubleshooting, for example, we observe undesirable effects $Y = y$ that are potentially caused by other conditions $X = x$ and we wish to predict whether an action that brings about a change in X would

remedy the situation. The information provided by the evidence y and x is extremely valuable, and it must be processed (using the updated distribution $P(u \mid x, y)$, as in equation (9)) before we can predict the effect of any action.[13] Thus, the expressions developed in this paper constitute bounds on the effectiveness of pending policies, when full knowledge of the current state of affairs (u) is not available, yet the current states of the decision variable (X) and the outcome variable (Y) are measured.

For these bounds to be valid in policy making, the context u must be time-invariant, that is, the probability $P(u)$ should represent epistemic uncertainty about a static, albeit unknown context $U = u$. The constancy of u is well justified in the control and diagnosis of physical systems, where u represents fixed, but unknown physical characteristic of devices or subsystems. The constancy approximation is also justified in the health sciences where patients' genetic attributes and physical characteristics can be assumed relatively constant between observation and treatment. For instance, if a patient in the example of section 5 wishes to assess the risk of switching from x' (no drug) to x (drug), it is reasonable to assume that this patient's susceptibility to the drug remains constant through the period of decision. The risk of death associated with this patient's decision to start using the drug will then be given by PS $= P(y_x \mid x', y')$, and may be assessed by the bounds in equation (60).

The constancy assumption is less justified in economic systems, where agents are bombarded by rapidly fluctuating stream of external forces ("shocks" in econometric terminology) as well as by inter-agents communication messages. These exogenous factors may vary substantially during the policy making interval and they require, therefore, time-dependent analysis. The canonical violation of the constancy assumption occurs, of course, in quantum mechanical systems, where the indeterminism associated with U is "intrinsic", and the existence of a deterministic relationship between U and V is no longer a good approximation. A method of incorporating such intrinsic indeterminism into counterfactual analysis is outlined in [40, p. 220], and leads to equation (9), where $P(Y_{x'}(u) = y')$ represents the intrinsic uncertainty in Y associated with the macroscopic state $U = u$, under the action $do(X = x)$ (see footnote 6).

Acknowledgements

We thank two anonymous referees for making useful suggestions on the first draft of this paper. Sander Greenland has provided valuable insight from epidemiological perspectives. This research was supported in parts by grants from NSF, ONR and AFOSR and by a Microsoft Fellowship to the first author.

References

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

[13] Such processing have been applied indeed to the evaluation of economic policies [5] and to repair-test strategies in troubleshooting [7].

[2] L.A. Bailey, L. Gordis and M. Green, Reference guide on epidemiology, Reference Manual on Scientific Evidence, Federal Judicial Center (1994) Available online at http://www.fjc.gov/EVIDENCE/science/sc_ev_sec.html.
[3] A. Balke, Probabilistic counterfactuals: Semantics, computation, and applications, Ph.D. thesis, Computer Science Department, University of California, Los Angeles, CA (November 1995).
[4] A. Balke and J. Pearl, Probabilistic evaluation of counterfactual queries, in: *Proceedings of the Twelfth National Conference on Artificial Intelligence*, Vol. I (MIT Press, Menlo Park, CA, 1994) pp. 230–237.
[5] A. Balke and J. Pearl, Counterfactuals and policy analysis in structural models, in: *Uncertainty in Artificial Intelligence*, Vol. 11, eds. P. Besnard and S. Hanks (Morgan Kaufmann, San Francisco, CA, 1995) pp. 11–18.
[6] A. Balke and J. Pearl, Nonparametric bounds on causal effects from partial compliance data, Journal of the American Statistical Association 92(439) (1997) 1–6.
[7] J.S. Breese and D. Heckerman, Decision-theoretic troubleshooting: A framework for repair and experiment, in: *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, eds. E. Horvitz and F. Jensen (Morgan Kaufmann, San Francisco, CA, 1996) pp. 124–132.
[8] P.W. Cheng, From covariation to causation: A causal power theory, Psychological Review 104(2) (1997) 367–405.
[9] P. Cole, Causality in epidemiology, health policy, and law, Journal of Marketing Research 27 (1997) 10279–10285.
[10] A.P. Dawid, Causal inference without counterfactuals, Technical Report, Department of Statistical Science, University College London, UK, 1997. Forthcoming, with discussion, Journal of the American Statistical Association, 2000.
[11] P.J. Dhrymes, *Econometrics* (Springer, New York, 1970).
[12] R.F. Engle, D.F. Hendry and J.F. Richard, Exogeneity, Econometrica 51 (1983) 277–304.
[13] K. Fine, *Reasoning with Arbitrary Objects* (B. Blackwell, New York, 1985).
[14] F.M. Fisher, A correspondence principle for simultaneous equations models, Econometrica 38(1) (1970) 73–92.
[15] D.A. Freedman and P.B. Stark, The swine flu vaccine and Guillain–Barré syndrome: A case study in relative risk and specific causation, *Evaluation Review* 23(6) (1999) 619–647.
[16] D. Galles and J. Pearl, Testing identifiability of causal effects, in: *Uncertainty in Artificial Intelligence*, Vol. 11, eds. P. Besnard and S. Hanks (Morgan Kaufmann, San Francisco, CA, 1995) pp. 185–195.
[17] D. Galles and J. Pearl, Axioms of causal relevance, *Artificial Intelligence* 97(1–2) (1997) 9–43.
[18] D. Galles and J. Pearl, An axiomatic characterization of causal counterfactuals, Foundations of Science 3(1) (1998) 151–182.
[19] C. Glymour, Psychological and normative theories of causal power and the probabilities of causes, in: *Uncertainty in Artificial Intelligence*, eds. G.F. Cooper and S. Moral (Morgan Kaufmann, San Francisco, CA, 1998) pp. 166–172.
[20] I.J. Good, A causal calculus, I, British Journal for the Philosophy of Science 11 (1961) 305–318.
[21] I.J. Good, A tentative measure of probabilistic causation relevant to the philosophy of the law, Journal of Statistical Computation and Simulation 47 (1993) 99–105.
[22] N. Hall, Two concepts of causation (1998) in press.
[23] J.Y. Halpern, Axiomatizing causal reasoning, in: *Uncertainty in Artificial Intelligence*, eds. G.F. Cooper and S. Moral (Morgan Kaufmann, San Francisco, CA, 1998) pp. 202–210.
[24] J.Y. Halpern and J. Pearl, Causes and explanations: A structural-model approach, Technical Report R-266, Cognitive System Laboratory, Department of Computer Science, University of California, Los Angeles (March 2000).
[25] D.F. Hendry, *Dynamic Econometrics* (Oxford University Press, New York, 1995).
[26] C.H. Hennekens and J.E. Buring, *Epidemiology in Medicine* (Brown, Little, Boston, 1987).
[27] G.W. Imbens, Book reviews, Journal of Applied Econometrics 12 (1997).

[28] J.L. Kelsey, A.S. Whittemore, A.S. Evans and W.D. Thompson, *Methods in Observational Epidemiology* (Oxford University Press, New York, 1996).
[29] 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) (1989) 183–190.
[30] J. Kim, Causes and events: Mackie on causation, Journal of Philosophy 68 (1971) 426–471. Reprinted in: *Causation*, eds. E. Sosa and M. Tooley (Oxford University Press, 1993).
[31] D. Lewis, *Philosophical Papers* (Oxford University Press, New York, 1986).
[32] J.L. Mackie, Causes and conditions, American Philosophical Quarterly 2/4 (1965) 261–264. Reprinted in: *Causation*, eds. E. Sosa and M. Tooley (Oxford University Press, 1993).
[33] J. Marschak, Statistical inference in economics, in: *Statistical Inference in Dynamic Economic Models*, ed. T. Koopmans (Wiley, New York, 1950) pp. 1–50. Cowles Commission for Research in Economics, Monograph 10.
[34] D. Michie, Adapting Good's q theory to the causation of individual events, Machine Intelligence 15 (2000).
[35] J.S. Mill, *System of Logic*, Vol. 1 (John W. Parker, London, 1843).
[36] J. Pearl, Comment: Graphical models, causality, and intervention, Statistical Science 8 (1993) 266–269.
[37] J. Pearl, A probabilistic calculus of actions, in: *Uncertainty in Artificial Intelligence*, Vol. 10, eds. R. Lopez de Mantaras and D. Poole (Morgan Kaufmann, San Mateo, CA, 1994) pp. 454–462.
[38] J. Pearl, Causal diagrams for experimental research, Biometrika 82 (1995) 669–710.
[39] J. Pearl, Probabilities of causation: three counterfactual interpretations and their identification, Synthese 121(1–2) (1999) 93–149.
[40] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, NY, 2000).
[41] D.W. Robertson, The common sense of cause in fact, Texas Law Review 75(7) (1997) 1765–1800.
[42] 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 (1986) 1393–1512.
[43] J.M. Robins, A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods, Journal of Chronic Diseases 40(Suppl 2) (1987) 139S–161S.
[44] J.M. Robins and S. Greenland, The probability of causation under a stochastic model for individual risk, Biometrics 45 (1989) 1125–1138.
[45] P. Rosenbaum and D. Rubin, The central role of propensity score in observational studies for causal effects, Biometrica 70 (1983) 41–55.
[46] J.J. Schlesselman, *Case-Control Studies: Design Conduct Analysis* (Oxford University Press, New York, 1982).
[47] M.C. Shep, Shall we count the living or the dead? New England Journal of Medicine 259 (1958) 1210–1214.
[48] H.A. Simon, Causal ordering and identifiability, in: *Studies in Econometric Method*, eds. Wm.C. Hood and T.C. Koopmans (Wiley, New York, 1953) pp. 49–74.
[49] H.A. Simon and N. Rescher, Cause and counterfactual, Philosophy and Science 33 (1966) 323–340.
[50] M.E. Sobel, Effect analysis and causation in linear structural equation models, Psychometrika 55(3) (1990) 495–515.
[51] P. Spirtes, C. Glymour and R. Scheines, *Causation, Prediction, and Search* (Springer, New York, 1993).
[52] R.H. Strotz and H.O.A. Wold, Recursive versus nonrecursive systems: An attempt at synthesis, Econometrica 28 (1960) 417–427.