Probabilities of Causation: Bounds and Identification

Jin Tian and Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
jtian@cs.ucla.edu *judea@cs.ucla.edu*

Abstract

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 (1999) 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 the belief that, with high probability, aspirin was "the actual cause of 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 (1843), it has received semi-formal explications only in the 1960s – via conditional probabilities [Good, 1961] and logical implications [Mackie, 1965]. These explications suffer from basic semantical difficulties [Kim, 1971] [Pearl, 2000, 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, 1989] gave a counterfactual definition for the probability of necessary causation taking counterfactuals as undefined primitives, and assuming that one is in possession of a consistent joint probability function on both ordinary and counterfactual events. [Pearl, 1999] gave counterfactual 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 [Galles and Pearl, 1997, Galles and Pearl, 1998, Halpern, 1998, Pearl, 2000]. 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 [Balke and Pearl, 1994, 1995]. Additionally, this semantics can be characterized by a complete set of axioms [Galles and Pearl, 1998, Halpern, 1998], 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 [Robins and Greenland, 1989, Balke and Pearl, 1994]. 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 [Pearl, 1999] [Pearl, 2000, 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 13) 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 gives a review of 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 Section 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 Balke and Pearl (1995), Galles and Pearl (1997, 1998), and Halpern (1998). Related approaches have been proposed in Simon and Rescher (1966) (see footnote 4) and Robins (1986). For detailed exposition of the structural account and its applications see [Pearl, 2000].

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 an equation, and changes due to interventions or hypothetical eventualities are treated as local modifications of those equations.

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.

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

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

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 $U \times (V \setminus V_i)$ to V_i. In other words, each f_i tells us the value of V_i*

[1] Similar models, called "neuron diagrams" [Lewis, 1986, p. 200; Hall, 1998] are used informally by philosophers to illustrate chains of causal processes.

given the values of all other variables in $U \cup V$. Symbolically, the set of equations F can be represented by writing

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

where pa_i is any realization of the unique minimal set of variables PA_i in V/V_i (connoting parents) *that renders f_i nontrivial. Likewise, $U_i \subseteq U$ stands for the unique minimal set of variables in U that renders f_i nontrivial.*

(iv) *The set F of functions defines a mapping from (the respective domains of) U to V. Likewise, every subset F' of F defines a mapping from the exogenous to the endogenous variables of that subset.* [2]

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.

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 realization of X. A submodel *M_x of M is the causal model*

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

where

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

[2]This requirement, which ensures a unique solution for every subset of F, is satisfied whenever F is recursive (feedback free). The requirement was relaxed by Halpern (1998).

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 [Galles and Pearl, 1998].[3]

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.*

Definition 4 (*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.*

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 5 (*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 $Y_x(u)$.

[3]Structural modifications date back to Marschak (1950) and Simon (1953). 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). A similar notion of sub-model is introduced in Fine (1985), though not specifically for representing actions and counterfactuals.

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 [Balke and Pearl, 1994], as it permits x to differ from the current value of $X(u)$ without creating logical contradiction; it also suppresses abductive inferences (or backtracking) from the counterfactual antecedent $X = x$.[4]

It can easily be shown [Galles and Pearl, 1997] that the counterfactual relationship just defined, $Y_x(u)$, satisfies the following two properties:

Effectiveness:

For all variables Y and W,

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

Composition:

For any two singleton variables Y and W, and any set of variables X,

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

Furthermore, effectiveness and composition are *complete* whenever M is recursive (i.e., $G(M)$ is acyclic) [Galles and Pearl, 1998, Halpern, 1998].

A corollary of composition is a property called *consistency* by [Robins, 1987]:

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

Consistency states that if we intervene and set the experimental conditions $X = x$ equal to those prevailing before the intervention, we should not expect any change in the response variable Y. This property will be used in several derivations of Section 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*

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

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

[4]Simon and Rescher (1966, 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.

$P(u)$, 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 [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. (10)-(12)) 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. (7) allows the evaluation of this quantity as follows:

$$\begin{aligned} P(Y_{x'} = y' | X = x, Y = y) &= \frac{P(Y_{x'} = y', X = x, Y = y)}{P(X = x, Y = y)} \\ &= \sum_u P(Y_{x'}(u) = y') P(u|x, y) \end{aligned} \tag{9}$$

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 $Y_{x'}(u) = y'$.

3 Probabilities of Causation: Definitions

In this section, we present the definitions for the three aspects of causation as defined in [Pearl, 1999]. We use the counterfactual language and the structural model semantics introduced in Section 2.

Definition 7 (*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 $X = true$ and $Y = true$, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$\begin{aligned} PN &\triangleq P(Y_{x'} = false \mid X = true, Y = true) \\ &\triangleq P(y'_{x'} | x, y) \end{aligned} \tag{10}$$

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.[5]

This quantity has applications in epidemiology, legal reasoning, and artificial intelligence (AI). Epidemiologists have long been concerned with estimating the probability that a certain case of disease is *attributable* to a

[5]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 Eq. (10) to read $PN \triangleq P(x' > y'|x, y)$.

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. It is used frequently in lawsuits, where legal responsibility is at the center of contention (see Section 5).

Definition 8 (*Probability of sufficiency (PS)*)

$$PS \triangleq P(y_x|y', x') \tag{11}$$

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 [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." In psychology, PS serves as the basis for Cheng's (1997) 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 [Pearl, 2000, pp. 221-223].

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

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

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 still 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 [Robertson, 1997]. Still, lack of sufficiency may weaken arguments based on PN [Good, 1993, Michie, 2000].

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:*

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

Proof of Lemma 1
Using the consistency condition of Eq. (4),

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

we can write

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

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

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

which proves Lemma 1. □

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 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 $P(y_{x'})$ is also available[6]. Moreover, under some general assumptions about the data-generating process, these quantities may even be identified.

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: $P_{XY} = \{P(x,y), P(x',y), P(x,y'), P(x',y')\}$.[7]

[6]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 [Pearl, 1995].

[7]By "experimental data" we mean data gathered under controlled randomized study on a large, randomly selected sample from a population characterizing events x and y. Likewise, by "nonexperimental data" we mean frequency counts obtained in uncontrolled study conducted on large, randomly selected sample from the population characterizing events x and y, and under the conditions prevailing during the occurrence of x and y. 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 the population which is deemed to be governed by the same causal model M that governed the behavior of I.

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:

$$
\begin{aligned}
p_{111} = &\quad P(y_x, y_{x'}, x) &= P(x, y, y_{x'}) \\
p_{110} = &\quad P(y_x, y_{x'}, x') &= P(x', y, y_x) \\
p_{101} = &\quad P(y_x, y'_{x'}, x) &= P(x, y, y'_{x'}) \\
p_{100} = &\quad P(y_x, y'_{x'}, x') &= P(x', y', y_x) \\
p_{011} = &\quad P(y'_x, y_{x'}, x) &= P(x, y', y_{x'}) \\
p_{010} = &\quad P(y'_x, y_{x'}, x') &= P(x', y, y'_x) \\
p_{001} = &\quad P(y'_x, y'_{x'}, x) &= P(x, y', y'_{x'}) \\
p_{000} = &\quad P(y'_x, y'_{x'}, x') &= P(x', y', y'_x)
\end{aligned}
$$

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

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

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

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

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:

$$
\begin{aligned}
PNS &= p_{101} + p_{100} && (18)\\
PN &= p_{101}/P(x,y) && (19)\\
PS &= p_{100}/P(x',y') && (20)
\end{aligned}
$$

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), (16) and (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 (1995, Appendix B) describes a computer program that takes symbolic description 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 [Balke, 1995, 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:

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

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

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

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

4.2.2 Given causal effects

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

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

with no constraints on PN and PS.

4.2.3 Given both nonexperimental data and causal effects

Given the constraints (15), (16) and (17), the following bounds are induced on the three probabilities of causation:

$$\max \left\{ \begin{array}{c} 0 \\ P(y_x) - P(y_{x'}) \\ P(y) - P(y_{x'}) \\ P(y_x) - P(y) \end{array} \right\} \leq PNS \leq \min \left\{ \begin{array}{c} P(y_x) \\ P(y'_{x'}) \\ P(x,y) + P(x',y') \\ P(y_x) - P(y_{x'}) + P(x,y') + P(x',y) \end{array} \right\} \tag{24}$$

$$\max \left\{ \begin{array}{c} 0 \\ \frac{P(y)-P(y_{x'})}{P(x,y)} \end{array} \right\} \leq PN \leq \min \left\{ \begin{array}{c} 1 \\ \frac{P(y'_{x'})-P(x',y')}{P(x,y)} \end{array} \right\} \tag{25}$$

$$\max \left\{ \begin{array}{c} 0 \\ \frac{P(y_x)-P(y)}{P(x',y')} \end{array} \right\} \leq PS \leq \min \left\{ \begin{array}{c} 1 \\ \frac{P(y_x)-P(x,y)}{P(x',y')} \end{array} \right\} \tag{26}$$

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

Definition 11 (*Exogeneity*)
A variable X is said to be exogenous for Y in model M iff

$$\begin{aligned} P(y_x) &= P(y|x) \\ P(y_{x'}) &= P(y|x') \end{aligned} \tag{27}$$

or, equivalently,

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

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

Eq. (27) 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 "weak ignorability." A graphical criterion ensuring exogeneity is the absence of a common ancestor of X and Y in $G(M)$ (including latent ancestors which represent dependencies among variables in U). The classical econometric criterion for exogeneity (e.g., Dhrymes 1970, p. 169) states that X be independent of the error term (u) in the equation for Y.[8] 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 Eq. (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:*

$$\max[0, P(y|x) - P(y|x')] \leq PNS \leq \min[P(y|x), P(y'|x')] \tag{29}$$
$$\frac{\max[0, P(y|x) - P(y|x')]}{P(y|x)} \leq PN \leq \frac{\min[P(y|x), P(y'|x')]}{P(y|x)} \tag{30}$$
$$\frac{\max[0, P(y|x) - P(y|x')]}{P(y'|x')} \leq PS \leq \frac{\min[P(y|x), P(y'|x')]}{P(y'|x')} \tag{31}$$

The bounds expressed in Eq. (30) were first derived by Robins and Greenland (1989) ; a more elaborate proof can be found in [Freedman and Stark, 1999]. [Pearl, 1999] derived Eqs. (29)-(31) under a stronger condition of exogeneity (see Definition 12). We see that under the condition of no-confounding the lower bound for PN can be expressed as

$$PN \geq 1 - \frac{1}{P(y|x)/P(y|x')} \triangleq 1 - \frac{1}{RR} \tag{32}$$

where $RR = P(y|x)/P(y|x')$ is called *relative risk* in epidemiology. Courts have often used the condition $RR > 2$ as a criterion for legal responsibility

[8] 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 Pearl (2000, pp. 169-170; 245-247) for a reconciliatory perspective on this controversy.

[Bailey *et al.*, 1994]. Eq. (32) shows that this practice represents a conservative interpretation of the "more probable than not" standard (assuming no confounding); PN must be higher than 0.5 if RR exceeds 2. Freedman and Stark (1999) argue that, in general, epidemiological evidence may not be applicable as proof for specific causation [Freedman and Stark, 1999] because such evidence cannot account for all characteristics specific to the plaintiff. This argument represents an overly narrow interpretation of the concept "probability of causation," for it insists on characterizing the plaintiff to minute detail and reduces to zero or one when all relevant details are accounted for. We doubt that this interpretation underlies the intent of judicial standards. By using the probabilistic 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. PN further ensures us that two obvious features of the plaintiff will not be ignored: the exposure (x) and the injury (y); 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 Eq. (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 (1983). Though untestable, such joint independence is assumed to hold when we assert the absence of factors that simultaneously affect exposure and outcome.

Definition 12 (*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,

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

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

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

Remarkably, the added constraints introduced by strong exogeneity do not alter the bounds of Eqs. (29)–(31):

Theorem 2 *Under condition of strong exogeneity, the probabilities PN, PS, and PNS are constrained by the bounds of Eqs. (29)–(31). Moreover, PN, PS, and PNS are related to each other as follows* [Pearl, 1999] *:*

$$PN = \frac{PNS}{P(y|x)} \tag{35}$$

$$PS = \frac{PNS}{1 - P(y|x')} \tag{36}$$

4.4 Identifiability under monotonicity

Definition 13 (*Monotonicity*)
A variable Y is said to be monotonic relative to variable X in a causal model M iff

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

Monotonicity expresses the assumption that a change from $X =$ false to $X =$ 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, 1997] used this assumption to tighten bounds of treatment effects from studies involving non-compliance. 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.

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 (Eq. (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 $P(y_{x'})$:

$$\begin{array}{rcl} P(y) & \leq P(y_x) \leq & 1 - P(x, y') \\ P(x', y) & \leq P(y_{x'}) \leq & P(y) \end{array} \tag{39}$$

Compared with Eq. (22), the lower bound for $P(y_x)$ and the upper bound for $P(y_{x'})$ are tightened. The importance of Eq. (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 Eq. (39) 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_{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:

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

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 identifiable whenever the causal effects $P(y_x)$ and $P(y_{x'})$ are identifiable, and are given by*

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

$$PN = P(y'_{x'}|x, y) = \frac{P(y) - P(y_{x'})}{P(x, y)} \tag{41}$$

$$PS = P(y_x|x', y') = \frac{P(y_x) - P(y)}{P(x', y')} \tag{42}$$

Eqs. (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 14 (*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 Pearl (1993, 1995) 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

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

where pa_X are (realizations of) the *parents* of X in the causal graph associate with M (see also Spirtes et al. (1993), Robins (1986), and Pearl (2000, p. 73)). Thus, we can combine Eq. (43) with Theorem 3 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. (40)–(42), with $P(y_x)$ given in Eq. (43). If monotonicity cannot be ascertained, then PNS, PN and PS are bounded by Eqs. (24)–(26), with $P(y_x)$ given in Eq. (43).*

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. (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 Eq. (27), and from theorem 3 we conclude that PNS, PN, and PS are all identifiable.

[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 $P(y_x)$. See also [Pearl, 2000, 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

$$PNS = P(y|x) - P(y|x') \tag{44}$$

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

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

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 misnomered "attributable risk" [Hennekens and Buring, 1987, p. 87]. The probability of necessity, PN, is given by the *excess-risk-ratio* (ERR)

$$PN = [P(y|x) - P(y|x')]/P(y|x) \tag{47}$$

often misnomered as the *attributable fraction* [Schlesselman, 1982], *attributable-rate percent* [Hennekens and Buring, 1987, p. 88], *attributed fraction for the exposed* [Kelsey *et al.*, 1996, p. 38], or *attributable proportion* [Cole, 1997]. The reason we consider these labels to be misnomers 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 which 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

$$PS = [P(y|x) - P(y|x')]/[1 - P(y|x')], \tag{48}$$

as it coincides with what epidemiologists call the "relative difference" [Shep, 1958], which is used to measure the *susceptibility* of a population to a risk factor x. It also coincides with what Cheng calls "causal power" (1997), namely, the effect of x on y after suppressing "all other causes of y." See Pearl (1999) for additional discussions of these expressions.

To appreciate the difference between Eqs. (41) and (47) we can rewrite Eq. (41) as

$$\begin{aligned} PN &= \frac{P(y|x)P(x) + P(y|x')P(x') - P(y_{x'})}{P(y|x)P(x)} \\ &= \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y_{x'})}{P(x,y)} \end{aligned} \quad (49)$$

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. $P(y_{x'}) \neq P(y|x')$. We will call the r.h.s. of (49) by corrected excess-risk-ratio (CERR).

¿From Eqs. (44)–(46) we see that the three notions of causation satisfy the simple relationships given by Eqs. (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 of Theorem 5:
¿From monotonicity condition, we have

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

Thus we can write

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

and

$$P(y|x') = P(y_{x'}|x') = P(y_x, y_{x'}|x') \quad (52)$$

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

$$P(y_x, y_{x'}|x') = P(y_x, y_{x'}), \quad (53)$$

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

$$P(y_x, y_{x'}|x) = P(y_x, y_{x'}). \quad (54)$$

Combining Eq. (54) with

$$\begin{aligned} P(y_x) &= P(y_x, y_{x'}) + P(y_x, y'_{x'}), & (55) \\ P(y|x) &= P(y_x|x) = P(y_x, y_{x'}|x) + P(y_x, y'_{x'}|x), & (56) \end{aligned}$$

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

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

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

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

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 estimand 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 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 Eq. (30). (The upper bound is usually unity.) In the presence of confounding, ERR must be corrected by the additive term $[P(y|x') - P(y_{x'})]/P(x, y)$, 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 $P(y|x') - P(y_{x'})$. 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 [Khoury *et al.*, 1989, Glymour, 1998].

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 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 non-experimental studies,

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

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

$$\begin{aligned} P(y_x) &= 16/1000 &= 0.016 \\ P(y_{x'}) &= 14/1000 &= 0.014 \\ P(y'_{x'}) &= 1 - P(y_{x'}) &= 0.986 \end{aligned}$$

The non-experimental data provide the estimates

$$\begin{aligned} P(y) &= 30/2000 &= 0.015 \\ P(x, y) &= 2/2000 &= 0.001 \\ P(x', y') &= 972/2000 &= 0.486 \end{aligned}$$

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

$$0.002 \leq PNS \leq 0.016 \tag{58}$$

$$1.0 \leq PN \leq 1.0 \tag{59}$$

$$0.002 \leq PS \leq 0.031 \tag{60}$$

These figures show that although surviving patients who didn't 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 Eqs. (40)–(42) yield

$$\begin{aligned} PNS &= 0.002 \quad (61)\\ PN &= 1.0 \quad (62)\\ PS &= 0.002 \quad (63) \end{aligned}$$

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:

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

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 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 contrived to show the usefulness of bounds Eqs. (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 Eq. (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 population at large. The invariant quantities are simply the causal effects probabilities, $P(y_{x'})$ and $P(y_x)$. Although these counterfactual probabilities were not measured in the nonexperimental group, they must (by definition) 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. The invariance of these quantities, 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 clarify 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. We further illustrate the applicability of these results to problems in epidemiology and legal reasoning.

The main application of this analysis to artificial intelligence lies in the automatic generation of verbal explanations, where the distinction between necessary and sufficient 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 various background 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 the singular event at hand (where each factor alone is both necessary and sufficient) and consider situations of the same general type – where oxygen alone is obviously insufficient to start a fire.

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) $= p_o$ 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.

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 Pearl (2000, 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.

References

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

[Bailey *et al.*, 1994] L. A. Bailey, L. Gordis, and M. Green. Reference guide on epidemiology. *Reference Manual on Scientific Evidence*, 1994. Federal Judicial Center. Available online at http://www.fjc.gov/EVIDENCE/science/sc_ev_sec.html.

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

[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. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176, September 1997.

[Balke, 1995] A. Balke. *Probabilistic Counterfactuals: Semantics, Computation, and Applications.* PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 1995.

[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. To appear [with discussion] in *Journal American Statistical Association*, 2000.

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

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

[Fine, 1985] K. Fine. *Reasoning with Arbitrary Objects.* B. Blackwell, New York, 1985.

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

[Freedman and Stark, 1999] 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):619–647, Dec. 1999.

[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.

[Hall, 1998] N. Hall. Two concepts of causation, 1998. 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.

[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.

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

[Kelsey *et al.*, 1996] J.L. Kelsey, A.S. Whittemore, A.S. Evans, and W.D Thompson. *Methods in Observational Epidemiology.* Oxford University Press, New York, 1996.

[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.

[Kim, 1971] J. Kim. Causes and events: Mackie on causation. *Journal of Philosophy*, 68:426–471, 1971. Reprinted in E. Sosa and M. Tooley (Eds.), *Causation*, Oxford University Press, 1993.

[Lewis, 1986] D. Lewis. *Philosophical Papers.* Oxford University Press, New York, 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.

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

[Michie, 2000] D. Michie. Adapting Good's q theory to the causation of individual events. *Machine Intelligence*, 15, 2000. Forthcoming.

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

[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 empirical research. *Biometrika*, 82(4):669–710, December 1995.

[Pearl, 1999] J. Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 1999.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[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.

[Robins, 1987] 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):139S–161S, 1987.

[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.

[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.

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

[Simon, 1953] H.A. Simon. Causal ordering and identifiability. In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. Wiley and Sons, Inc., 1953.

[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.