Chapter 9

Probability of Causation: Interpretation and Identification

Come and let us cast lots to find out
who is to blame for this ordeal.
Jonah 1:7

Preface

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, according to common judicial standard, judgment in favor of the plaintiff should be made if and only if it is "more probable than not" that the defendant's action was the *cause* for the plaintiff's damage (or death). But causation has two faces, *necessary* and *sufficient*; which of the two have lawmakers meant us to consider? And how are we to evaluate their probabilities?

This chapter provides formal semantics for the probability that event x was a *necessary* or *sufficient* cause (or both) of another event y. We then explicate conditions under which the probability of necessary (or sufficient) causation can be learned from statistical data, and we show how data from both experimental and nonexperimental stud-

ies can be combined to yield information that neither study alone can provide.

9.1 Introduction

The standard counterfactual definition of causation (i.e., that E would not have occurred were it not for C) captures the notion of "necessary cause." Competing notions such as "sufficient cause" and "necessary and sufficient cause" are of interest in a number of applications, and these, too, can be given concise mathematical definitions in structural model semantics (Section 7.1). Although the distinction between necessary and sufficient causes goes back to J. S. Mill (1843), it has received semiformal explications only in the 1960s—via conditional probabilities (Good 1961) and logical implications (Mackie 1965). These explications suffer from basic semantical difficulties,[1] and they do not yield effective procedures for computing probabilities of causes as those provided by the structural account (Sections 7.1.3 and 8.3).

In this chapter we explore the counterfactual interpretation of necessary and sufficient causes, illustrate the application of structural model semantics to the problem of identifying probabilities of causes, and present, by way of examples, new ways of estimating probabilities of causes from statistical data. Additionally, we argue that necessity and sufficiency are two distinct facets of causation and that both facets should take part in the construction of causal explanations.

Our results have applications in epidemiology, legal reasoning, artificial intelligence (AI), and psychology. Epidemiologists have long been concerned with estimating the probability that a certain case of disease is "attributable" to a particular exposure, which is normally interpreted counterfactually as "the probability that disease would not have occurred in the absence of exposure, given that disease and exposure did in fact occur." This counterfactual notion, which Robins and Greenland (1989) called the "probability of causation," measures how *necessary* the cause is for the production of the effect.[2] It is used

[1] The limitations of the probabilistic account are discussed in Section 7.5; those of the logical account will be discussed in Section 10.1.4.

[2] Greenland and Robins (1988) further distinguish between two ways of mea-

frequently in lawsuits, where legal responsibility is at the center of contention (see e.g. Section 8.3). We shall denote this notion by the symbol PN, an acronym for probability of necessity.

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

As the examples illustrate, 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 is 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 in press).

It is known that PN is in general nonidentifiable, that is, it cannot be estimated from frequency data involving exposures and disease cases (Greenland and Robins 1988; Robins and Greenland 1989). The identification is hindered by two factors.

1. *Confounding*—Exposed and unexposed subjects may differ in several relevant factors or, more generally, the cause and the effect may both be influenced by a third factor. In this case we say

suring probabilities of causation: the first (called "excess fraction") concerns only *whether* the effect (e.g. disease) occurs by a particular time; the second (called "etiological fraction") requires consideration of *when* the effect occurs. We will confine our discussion here to events occurring within a specified time period, or to "all or none" outcomes (such as birth defects) for which the probability of occurrence but not the time to occurrence is important.

that the cause is not *exogenous* relative to the effect (see Section 7.4.5).

2. *Sensitivity to the generative process*—Even in the absence of confounding, probabilities of certain counterfactual relationships cannot be identified from frequency information unless we specify the functional relationships that connect causes and effects. Functional specification is needed whenever the facts at hand (e.g. disease) might be affected by the counterfactual antecedent (e.g. exposure) (see the examples in Sections 1.4, 7.5, and 8.3).

Although PN is not identifiable in the general case, several formulas have nevertheless been proposed to estimate attributions of various kinds in terms of frequencies obtained in epidemiological studies (Breslow and Day 1980; Hennekens and Buring 1987; Cole 1997). Naturally, any such formula must be predicated upon certain implicit assumptions about the data-generating process. Section 9.2 explicates some of those assumptions and explores conditions under which they can be relaxed.[3] It offers new formulas for PN and PS in cases where causes are confounded (with outcomes) but their effects can nevertheless be estimated (e.g., from clinical trials or from auxiliary measurements). Section 9.3 exemplifies the use of these formulas in legal and epidemiological settings, while Section 9.4 provides a general condition for the identifiability of PN and PS when functional relationships are only partially known.

The distinction between necessary and sufficient causes has important implications in AI, especially in systems that generate verbal explanations automatically (see Section 7.2.3). As can be seen from the epidemiological examples, 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.,

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

sufficient causation) then we lose important 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 actual cause 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. Clearly, some balance must be made between the necessary and the sufficient components of causal explanation, and the present chapter illuminates this balance by formally explicating the basic relationships between the two components.

9.2 Necessary and Sufficient Causes: Conditions of Identification

9.2.1 Definitions, Notation, and Basic Relationships

Using the counterfactual notation and the structural model semantics introduced in Section 7.1, we give the following definitions for the three aspects of causation discussed in the introduction.

Definition 9.2.1 (Probability of Necessity, PN)
Let X and Y be two binary variables in a causal model M. Let x and y stand (respectively) for the propositions $X = \text{true}$ and $Y = \text{true}$, and

let x' and y' denote their complements. The probability of necessity *is defined as the expression*

$$\begin{aligned} \text{PN} &\triangleq P(Y_{x'} = \mathit{false} \mid X = \mathit{true}, Y = \mathit{true}) \\ &\triangleq P(y'_{x'}|x, y). \end{aligned} \tag{9.1}$$

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

Observe the slight change in notation relative to that used Section 7.1. Lowercase letters (e.g., x and y) denoted values of variables in Section 7.1 but now stand for propositions (or events). Note also the abbreviations y_x for Y_x = true and y'_x for Y_x = false.[4] Readers accustomed to writing "$A > B$" for the counterfactual "B if it were A" can translate (9.1) to read $PN \triangleq P(x' > y'|x, y)$.[5]

Definition 9.2.2 (Probability of Sufficiency, PS)

$$PS \triangleq P(y_x|y', x'). \tag{9.2}$$

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.

[4] These were proposed by Peyman Meshkat (in class homework) and substantially simplify the derivations.

[5] Definition 9.2.1 generalizes naturally to cases where X and Y are multivalued, say $x \in \{x_1, x_2, \ldots, x_k\}$ and $y \in \{y_1, y_2, \ldots, y_l\}$. We say that event $C = \bigvee_{i \in I}(X = x_i)$ is "counterfactually necessary" for $E = \bigvee_{j \in J}(Y = y_j)$, written $\overline{C} > \overline{E}$, if Y_x falls outside E whenever $X = x$ is outside C. Accordingly, the probability that C was a necessary cause of E is defined as PN $\triangleq P(\overline{C} > \overline{E}|C, E)$. For simplicity, however, we will pursue the analysis in the binary case.

Definition 9.2.3 (Probability of Necessity and Sufficiency, PNS)

$$\text{PNS} \triangleq P(y_x, y'_{x'}). \tag{9.3}$$

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

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

Definition 9.2.4 (Probability of Disablement, PD)

$$\text{PD} \triangleq P(y'_{x'}|y). \tag{9.4}$$

PD measures the probability that y would have been prevented if it were not for x; it is therefore of interest to policy makers who wish to assess the social effectiveness of various prevention programs (Fleiss 1981, pp. 75–6).

Definition 9.2.5 (Probability of Enablement, PE)

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

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

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

Lemma 9.2.6 *The probabilities of causation,* (PNS, PN *and* PS) *satisfy the following relationship:*

$$\text{PNS} = P(x, y)\text{PN} + P(x', y')\text{PS}. \tag{9.5}$$

Proof
The consistency conditions of (7.19), $X = x \Rightarrow Y_x = Y$, translate in our notation into

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

Hence we can write

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

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 9.2.6. □
To put into focus the aspects of causation captured by PN and PS, it is helpful to characterize those changes in the causal model that would leave each of the two measures invariant. The next two lemmas show that PN is insensitive to the introduction of potential inhibitors of y, while PS is insensitive to the introduction of alternative causes of y.

Lemma 9.2.7 *Let* $\text{PN}(x, y)$ *stand for the probability that* x *is a necessary cause of* y*. Let* $z = y \wedge q$ *be a consequence of* y*, that is potentially inhibited by* q'*. Then*

$$\text{PN}(x, z) \triangleq P(z'_{x'}|x, z) = P(y'_{x'}|x, y) \triangleq \text{PN}(x, y).$$

Cascading the process $Y_x(u)$ with the link $z = y \wedge q$ amounts to inhibiting the output of the process with probability $P(q')$. Lemma 9.2.7 asserts that we can add such a link without affecting PN. The reason is clear; conditioning on x and z implies that, in the scenario under consideration, the added link was not inhibited by q'.

Proof of Lemma 9.2.7 We have

$$\begin{aligned} \mathrm{PN}(x,z) &= P(z'_{x'}|x,z) = \frac{P(z'_{x'},x,z)}{P(x,z)} \\ &= \frac{P(z'_{x'},x,z|q)P(q) + P(z'_{x'},x,z|q')P(q')}{P(z,x,q) + P(z,x,q')}. \end{aligned} \tag{9.6}$$

Using $z = y \wedge q$, it follows that

$$q \Rightarrow (z = y),\ q \Rightarrow (z'_{x'} = y'_{x'}),\ \text{and } q' \Rightarrow z';$$

therefore,

$$\begin{aligned} \mathrm{PN}(x,z) &= \frac{P(y'_{x'},x,y|q)P(q) + 0}{P(y,x,q) + 0} \\ &= \frac{P(y'_{x'},x,y)}{P(y,x)} = P(y'_{x'}|xy) = \mathrm{PN}(x,y). \end{aligned}$$

□

Lemma 9.2.8 *Let* $\mathrm{PS}(x,y)$ *stand for the probability that* x *is a sufficient cause of* y*, and let* $z = y \vee r$ *be a consequence of* y *that may also be triggered by* r*. Then*

$$\mathrm{PS}(x,z) = P(z_x|x',z') = P(y_x|x',y') = \mathrm{PS}(x,y).$$

Lemma 9.2.8 asserts that we can add alternative causes (r), not involving x, without affecting PS. The reason again is clear; conditioning on the event x' and y' implies that the added causes (r) were not active. The proof of Lemma 9.2.8 is similar to that of Lemma 9.2.7.

Since all the causal measures defined above invoke conditionalization on y, and since y is presumed to be affected by x, we know that none of these quantities is identifiable from knowledge of the causal

diagram $G(M)$ and the data $P(v)$ alone, even under conditions of no-confounding. Moreover, none of these quantities determines the others in the general case. However, simple interrelationships and useful bounds can be derived for these quantities under the assumption of no-counfounding, and assumption that we call *exogeneity*.

9.2.2 Bounds and Basic Relationships under Exogeneity

Definition 9.2.9 (Exogeneity)
A variable X is said to be exogenous *relative to Y in model M if and only if*

$$P(y_x, y_{x'}|x) = P(y_x, y_{x'}). \tag{9.7}$$

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

Equation (9.7) is a strong version of those used in Chapter 5 (equation (5.31)) and in Chapter 6 (for no-confounding) in that it involves the joint event $\{y_x, y_{x'}\}$. This definition was named "strong ignorability" in Rosenbaum and Rubin (1983), and it corresponds to the classical econometric criterion for exogeneity that X be independent (jointly) of all the error terms in the equation for Y (Christ 1966, p. 156). A graphical criterion ensuring exogeneity is the absence of a common ancestor of X and Y in $G(M)$ or, more precisely, the absence of an active back-door path between X and Y (Section 3.3.1).

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

$$P(y_x) = P(y_x|x) = P(y|x), \tag{9.8}$$

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

Theorem 9.2.10 *Under condition of exogeneity,* PNS *is bounded as follows:*

$$\max[0, P(y|x) - P(y|x')] \leq \text{PNS} \leq \min[P(y|x), P(y'|x')]. \tag{9.9}$$

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

Proof
For any two events A and B, we have the sharp bounds

$$\max[0, P(A) + P(B) - 1] \leq P(A, B) \leq \min[P(A), P(B)]. \quad (9.10)$$

Equation (9.9) follows from (9.10) using $A = y_x$, $B = y'_{x'}$, $P(y_x) = P(y|x)$, and $P(y'_{x'}) = P(y'|x')$. □
Clearly, if exogeneity cannot be ascertained, then PNS is bound by inequalities similar to those of (9.9), with $P(y_x)$ and $P(y'_{x'})$ replacing $P(y|x)$ and $P(y'|x')$, respectively.

Theorem 9.2.11 *Under condition of exogeneity, the probabilities* PN, PS, *and* PNS *are related to each other as follows:*

$$\text{PN} = \frac{\text{PNS}}{P(y|x)}, \quad (9.11)$$

$$\text{PS} = \frac{\text{PNS}}{P(y'|x')}. \quad (9.12)$$

Thus, the bounds for PNS *in* (9.9) *provide corresponding bounds for* PN *and* PS.

The resulting bounds for PN,

$$\frac{\max[0, P(y|x) - P(y|x')]}{P(y|x)} \leq \text{PN} \leq \frac{\min[P(y|x), P(y'|x')]}{P(y|x)}, \quad (9.13)$$

place limits on our ability to identify PN in experimental studies, where exogeneity holds.

Corollary 9.2.12 *If x and y occur in an experimental study, and $P(y_x)$ and $P(y_{x'})$ are the causal effects measured in that study, then, for any point p in the range*

$$\frac{\max[0, P(y_x) - P(y_{x'})]}{P(y_x)} \leq p \leq \frac{\min[P(y_x), P(y'_{x'})]}{P(y_x)}, \quad (9.14)$$

there exists a causal model M that agrees with $P(y_x)$, $P(y_{x'})$ and for which PN $= p$.

Other bounds can be established for nonexperimental events, if we have data from both experimental and observational studies (as in Section 9.3.4). The non-zero widths of these bounds imply that probabilities of causation cannot be defined uniquely in stochastic (non-Laplacian) models where, for each u, $Y_x(u)$ is specified in probability $P(Y_x(u) = y)$ instead of a single number.[6]

Proof of Theorem 9.2.11:
Using $x \Rightarrow (y_x = y)$, we can write $x \wedge y_x = x \wedge y$ and so obtain

$$\begin{aligned} \text{PN} = P(y'_{x'}|x,y) &= P(y'_{x'},x,y)/P(x,y) && (9.15)\\ &= P(y'_{x'},x,y_x)/P(x,y) && (9.16)\\ &= P(y'_{x'},y_x)P(x)/P(x,y) && (9.17)\\ &= \frac{\text{PNS}}{P(y|x)}, && (9.18) \end{aligned}$$

which establishes (9.11). Equation (9.12) follows by identical steps. □
For completeness, we write the relationship between PNS and the probabilities of enablement and disablement:

$$\text{PD} = \frac{P(x)\ \text{PNS}}{P(y)}, \quad \text{PE} = \frac{P(x')\ \text{PNS}}{P(y')}. \tag{9.19}$$

9.2.3 Identifiability under Monotonicity and Exogeneity

Before attacking the general problem of identifying the counterfactual quantities in (9.1)–(9.3), it is instructive to treat a special condition, called *monotonicity*, which is often assumed in practice and which renders these quantities identifiable. The resulting probabilistic expressions will be recognized as familiar measures of causation that often appear in the literature.

[6]Robins and Greenland (1989), who used a stochastic model of $Y_x(u)$, defined the probability of causation as

$$\text{PN}(u) = [P(y|x,u) - P(y|x',u)]/P(y|x,u)$$

instead of the counterfactual definition in (9.1).

Definition 9.2.13 (Monotonicity)
A variable Y is said to be monotonic *relative to variable X in a causal model M if and only if the function $Y_x(u)$ is monotonic in x for all u. Equivalently, Y is monotonic relative to X if and only if*

$$y'_x \wedge y_{x'} = false. \tag{9.20}$$

Monotonicity expresses the assumption that a change from $X =$ false to $X =$ true cannot, under any circumstance make Y change from true to false.[7] In epidemiology, this assumption is often expressed as "no prevention," that is, no individual in the population can be helped by exposure to the risk factor.

Theorem 9.2.14 (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 (9.11)–(9.12), with*

$$\text{PNS} = P(y|x) - P(y|x'). \tag{9.21}$$

The r.h.s. of (9.21) is called "risk difference" in epidemiology, and is also misnomered "attributable risk" (Hennekens and Buring 1987, p. 87).

From (9.11) we see that the probability of necessity is identifiable and given by the *excess risk ratio*

$$\text{PN} = \frac{P(y|x) - P(y|x')}{P(y|x)}, \tag{9.22}$$

often misnomered as the "attributable fraction" (Schlesselman 1982), "attributable-rate percent" [Hennekens and Buring, 1987](Hennekens and Buring 1987, p. 88), or "attributable proportion" (Cole 1997). Taken literally, the ratio presented in (9.22) has nothing to do with attribution, since it is made up of statistical terms and not of causal or

[7] Our analysis remains invariant to complementing x or y (or both); hence, the general condition of monotonicity should read: Either $y'_x \wedge y_{x'} =$ false or $y'_{x'} \wedge y_x =$ false. For simplicity, however, we will adhere to the definition in (9.20).

counterfactual relationships. However, the assumptions of exogeneity and monotonicity together enable us to translate the notion of attribution embedded in the definition of PN (equation (9.1)) into a ratio of purely statistical associations. This suggests that exogeneity and monotonicity were tacitly assumed by the many authors who proposed or derived (9.22) as a measure for the "fraction of exposed cases that are attributable to the exposure."

Robins and Greenland (1989) analyzed the identification of PN under the assumption of stochastic monotonicity (i.e., $P(Y_x(u) = y) > P(Y_{x'}(u) = y)$) and showed that this assumption is too weak to permit such identification; in fact, it yields the same bounds as in (9.13). This indicates that stochastic monotonicity imposes no constraints whatsoever on the functional mechanisms that mediate between X and Y.

The expression for PS (equation (9.12)) is likewise quite revealing,

$$\text{PS} = \frac{P(y|x) - P(y|x')}{1 - P(y|x')}, \tag{9.23}$$

since 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. Susceptibility is defined as the proportion of persons who possess "an underlying factor sufficient to make a person contract a disease following exposure" (Khoury et al. 1989). PS offers a formal counterfactual interpretation of susceptibility, which sharpens this definition and renders susceptibility amenable to systematic analysis.

Khoury et al. (1989) recognized that susceptibility in general is not identifiable and derived (9.23) by making three assumptions: no-confounding, monotonicity,[8] and independence (i.e., assuming that susceptibility to exposure is independent of susceptibility to background not involving exposure). This last assumption is often criticized as untenable, and Theorem 9.2.14 assures us that independence is in fact unnecessary; (9.23) attains its validity through exogeneity and monotonicity alone.

Equation (9.23) also coincides with what Cheng (1997) calls "causal power," namely, the effect of x on y after suppressing "all other causes of

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

y." The counterfactual definition of PS, $P(y_x|x', y')$, suggests another interpretation of this quantity. It measures the probability that setting x would produce y in a situation where x and y are in fact absent. Conditioning on y' amounts to selecting (or hypothesizing) only those worlds in which "all other causes of y" are indeed suppressed.

It is important to note, however, that the simple relationships among the three notions of causation (equations (9.11)–(9.12)) hold only under the assumption of exogeneity; the weaker relationship of (9.5) prevails in the general, nonexogenous case. Additionally, all these notions of causation are defined in terms of the global relationships $Y_x(u)$ and $Y_{x'}(u)$, which are too crude to fully characterize the many nuances of causation; the detailed structure of the causal model leading from X to Y is often needed to explicate more refined notions, such as "actual cause" (see Chapter 10).

Proof of Theorem 9.2.14

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

$$y_x = y_x \wedge (y_{x'} \vee y'_{x'}) = (y_x \wedge y_{x'}) \vee (y_x \wedge y'_{x'}) \tag{9.24}$$

and

$$y_{x'} = y_{x'} \wedge (y_x \vee y'_x) = (y_{x'} \wedge y_x) \vee (y_{x'} \wedge y'_x) = y_{x'} \wedge y_x, \tag{9.25}$$

since monotonicity entails $y_{x'} \wedge y'_x = \text{false}$. Substituting (9.25) into (9.24) yields

$$y_x = y_{x'} \vee (y_x \wedge y'_{x'}). \tag{9.26}$$

Taking the probability of (9.26) and using the disjointness of $y_{x'}$ and $y'_{x'}$, we obtain

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

or

$$P(y_x, y'_{x'}) = P(y_x) - P(y_{x'}). \tag{9.27}$$

Equation (9.27), together with the assumption of exogeneity (equation (9.8)) establishes (9.21). □

9.2.4 Identifiability under Monotonicity and Nonexogeneity

The relations established in Theorems 9.2.10–9.2.14 were based on the assumption of exogeneity. In this section, we relax this assumption and consider cases where the effect of X on Y is confounded, that is, when $P(y_x) \neq P(y|x)$. In such cases $P(y_x)$ may still be estimated by auxiliary means (e.g., through adjustment of certain covariates or through experimental studies), and the question is whether this added information can render the probability of causation identifiable. The answer is affirmative.

Theorem 9.2.15 *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:*

$$\text{PNS} = P(y_x, y'_{x'}) = P(y_x) - P(y_{x'}), \tag{9.28}$$

$$\text{PN} = P(y'_{x'}|x, y) = \frac{P(y) - P(y_{x'})}{P(x, y)}, \tag{9.29}$$

$$\text{PS} = P(y_x|x', y') = \frac{P(y_x) - P(y)}{P(x', y')}. \tag{9.30}$$

In order to appreciate the difference between equations (9.29) and (9.22), we can expand $P(y)$ and write

$$\begin{aligned} \text{PN} &= \frac{P(y|x)P(x) + P(y|x')P(x') - P(y_{x'})}{P(y|x)P(x)} \\ &= \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y_{x'})}{P(x, y)}. \end{aligned} \tag{9.31}$$

The first term on the r.h.s. of (9.31) is the familiar excess risk ratio (as in (9.22)) and represents the value of PN under exogeneity. The second term represents the *correction* needed to account for X's nonexogeneity, that is, $P(y_{x'}) \neq P(y|x')$.

Equations (9.28)–(9.30) thus provide more refined measures of causation, which can be used in situations where the causal effect $P(y_x)$ can

be identified through auxiliary means (see Example 4, Section 9.3.4). It can also be shown that expressions in (9.28)–(9.30) provide lower bounds for PNS, PN, and PS in the general, nonmonotonic case (J. Tian, personal communication).

Remarkably, since PS and PN must be nonnegative, (9.29)-(9.30) provide a simple necessary test for the assumption of monotonicity:

$$P(y_x) \geq P(y) \geq P(y_{x'}), \tag{9.32}$$

which strengthen the standard inequalities (from $x \wedge y \Rightarrow y_x$)

$$P(y_x) \geq P(x, y), \quad P(y_{x'}) \geq P(x', y). \tag{9.33}$$

It can be shown that these inequalities are in fact sharp: every combination of experimental and nonexperimental data that satisfies 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, then (9.32) can be used for testing the compatibility of the experimental and non-experimental data, that is, whether subjects used in clinical trials are representative of the target population as characterized by the joint distribution $P(x, y)$.

Proof of Theorem 9.2.15

Equation (9.28) was established in (9.27). To prove (9.30), we write

$$P(y_x|x', y') = \frac{P(y_x, x', y')}{P(x', y')} = \frac{P(y_x, x', y'_{x'})}{P(x', y')}, \tag{9.34}$$

because $x' \wedge y' = x' \wedge y'_{x'}$ (by consistency). To calculate the numerator of (9.34), we conjoin (9.26) with x' to obtain

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

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

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

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

Substituting into i(9.34), we finally obtain

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

which establishes (9.30). Equation (9.29) follows via identical steps. □
One common class of models that permits the identification of $P(y_x)$ under conditions of nonexogeneity was exemplified in Chapter 3. It was shown in Section 3.2 (equation (3.13)) that, for every two variables X and Y in a positive Markovian model M, the causal effect $P(y_x)$ is identifiable and is given by

$$P(y_x) = \sum_{pa_X} P(y|pa_X, x)P(pa_X), \tag{9.35}$$

where pa_X are (realizations of) the *parents* of X in the causal graph associated with M. Thus, we can combine (9.35) with Theorem 9.2.15 to obtain a concrete condition for the identification of the probability of causation.

Corollary 9.2.16 *For any positive-Markovian model M, if the function $Y_x(u)$ is monotonic then the probabilities of causation* PNS, PS, *and* PN *are identifiable and are given by* (9.28)–(9.30), *with $P(y_x)$ as given in* (9.35).

A broader identification condition can be obtained through the use of the back-door and front-door criteria (Section 3.3), which are applicable to semi-Markovian models. These were further generalized in Galles and Pearl (1995) (see also Section 4.3.1) and lead to the following corollary.

Corollary 9.2.17 *Let **GP** be the class of semi-Markovian models that satisfy the graphical criterion of Theorem 4.3.1. If $Y_x(u)$ is monotonic, then the probabilities of causation* PNS, PS, *and* PN *are identifiable in **GP** and are given by* (9.28)–(9.30), *with $P(y_x)$ determined by the topology of $G(M)$ through the algorithm of Section 4.3.3.*

9.3 Examples and Applications

9.3.1 Example 1: Betting against a Fair Coin

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

This example is isomorphic to the clinical trial discussed in Section 1.4.4 (Figure 1.6). Let x stand for "we bet on heads," y for "we win a dollar," and u for "the coin turned up heads." The functional relationship between y, x, and u is

$$y = (x \wedge u) \vee (x' \wedge u'), \tag{9.36}$$

which is not monotonic but nevertheless permits us to compute the probabilities of causation from the basic definitions of (9.1)–(9.3). To exemplify,

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

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

$$\text{PS} = P(y_x|x', y') = P(y_x|u) = 1$$

(because $x' \wedge y' \Rightarrow u$) and

$$\begin{aligned} \text{PNS} &= P(y_x, y'_{x'}) \\ &= P(y_x, y'_{x'}|u)P(u) + P(y_x, y'_{x'}|u')P(u') \\ &= 1(0.5) + 0(0.5) = 0.5. \end{aligned}$$

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

for producing a win. The empirical content of such counterfactuals is discussed in Section 7.2.2.

It is easy to verify that these counterfactual quantities cannot be computed from the joint probability of X and Y without knowledge of the functional relationship in (9.36), which tells us the (deterministic) policy by which a win or a loss is decided (Section 1.4.4). This can be seen, for instance, from the conditional probabilities and causal effects associated with this example,

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

because identical probabilities would be generated by a random payoff policy in which y is functionally independent of x—say, by a bookie who watches the coin and ignores our bet. In such a random policy, the probabilities of causation PN, PS, and PNS are all zero. Thus, according to our definition of identifiability (Definition 3.2.3), if two models agree on P and do not agree on a quantity Q, then Q is not identifiable. Indeed, the bounds delineated in Theorem 9.2.10 (equation (9.9)) read $0 \leq \text{PNS} \leq \frac{1}{2}$, meaning that the three probabilities of causation cannot be determined from statistical data on X and Y alone, not even in a controlled experiment; knowledge of the functional mechanism is required, as in (9.36).

It is interesting to note that whether the coin is tossed before or after the bet has no bearing on the probabilities of causation as just defined. This stands in contrast with some theories of probabilistic causality (e.g. Good 1961), which attempt to avoid deterministic mechanisms by conditioning all probabilities on "the state of the world just before" the occurrence of the cause in question (x). When applied to our betting story, the intention is to condition all probabilities on the state of the coin (u), but this is not fulfilled if the coin is tossed after the bet is placed. Attempts to enrich the conditioning set with events occurring after the cause in question have led back to deterministic relationships involving counterfactual variables (see Cartwright 1989, Eells 1991) and the discussion in Section 7.5.4).

One may argue, of course, that if the coin is tossed after the bet then it is not at all clear what our winnings would be had we bet differently; merely uttering our bet could conceivably affect the trajectory of the

coin (Dawid 1997). This objection can be diffused by placing x and u in two remote locations and tossing the coin a split second after the bet is placed but before any light ray could arrive from the betting room to the coin-tossing room. In such a hypothetical situation, the counterfactual statement "our winning would be different had we bet differently" is rather compelling, even though the conditioning event (u) occurs after the cause in question (x). We conclude that temporal descriptions such as "the state of the world just before x" cannot be used to properly identify the appropriate set of conditioning events (u) in a problem; a deterministic model of the mechanisms involved is needed for formulating the notion of "probability of causation."

9.3.2 Example 2: The Firing Squad

Consider again the firing squad of Section 7.1.2 (see Figure 9.1); A and

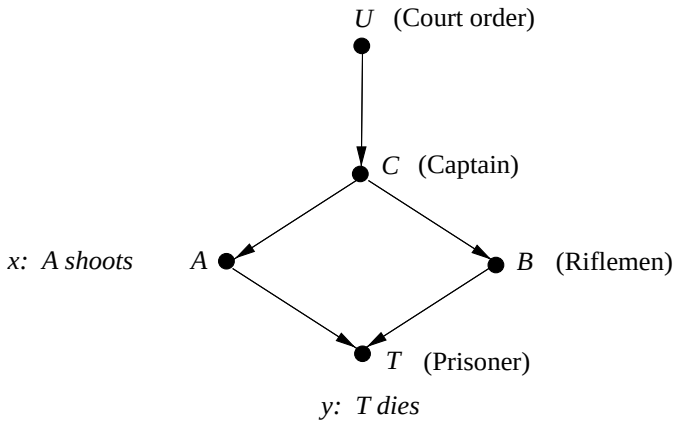

Figure 9.1: Causal relationships in the two-man firing-squad example.

B are riflemen, C is the squad's captain (who is waiting for the court order, U), and T is a condemned prisoner. Let u be the proposition that the court has ordered an execution, x the proposition stating that A pulled the trigger, and y that T is dead. We assume again that $P(u) = \frac{1}{2}$, that A and B are perfectly accurate marksmen who are alert and law-abiding, and that T is not likely to die from fright or other extraneous causes. We wish to compute the probability that x

was a necessary (or sufficient, or both) cause for y (i.e., we wish to calculate PN, PS, and PNS).

Definitions 9.2.1–9.2.3 permit us to compute these probabilities directly from the given causal model, since all functions and all probabilities are specified, with the truth value of each variable tracing that of U. Accordingly, we can write[9]

$$\begin{aligned} P(y_x) &= P(Y_x(u) = \text{true})P(u) + P(Y_x(u') = \text{true})P(u') \\ &= \frac{1}{2}(1+1) = 1. \end{aligned} \tag{9.37}$$

Similarly, we have

$$\begin{aligned} P(y_{x'}) &= P(Y_{x'}(u) = \text{true})P(u) + P(Y_{x'}(u') = \text{true})P(u') \\ &= \frac{1}{2}(1+0) = \frac{1}{2}. \end{aligned} \tag{9.38}$$

In order to compute PNS, we must evaluate the probability of the joint event $y_{x'} \wedge y_x$. Given that these two events are jointly true only when $U = \text{true}$, we have

$$\begin{aligned} \text{PNS} &= P(y_x, y_{x'}) \\ &= P(y_x, y_{x'}|u)P(u) + P(y_x, y_{x'}|u')P(u') \\ &= \frac{1}{2}(1+0) = \frac{1}{2}. \end{aligned} \tag{9.39}$$

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

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

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

[9]Recall that $P(Y_x(u') = \textit{true})$ involves the submodel M_x in which X is set to "true" independently of U. Thus, although under condition u' the captain has not given a signal, the potential outcome $Y_x(u')$ calls for hypothesizing that rifleman A pulls the trigger (x) unlawfully.

Similarly,

$$\text{PS} = P(y_x|x', y') = P(y_x|u') = 1,$$

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

Note that Theorems 9.2.10 and 9.2.11 are not applicable to this example because x is not exogenous; events x and y have a common cause (the captain's signal), which renders $P(y|x') = 0 \neq P(y_{x'}) = \frac{1}{2}$. However, the monotonicity of Y (in x) permits us to compute PNS, PS, and PN from the joint distribution $P(x, y)$ and the causal effects (using (9.28)–(9.30)), instead of consulting the functional model. Indeed, writing

$$P(x, y) = P(x', y') = \frac{1}{2} \tag{9.40}$$

and

$$P(x, y') = P(x', y) = 0, \tag{9.41}$$

we obtain

$$\text{PN} = \frac{P(y) - P(y_{x'})}{P(x, y)} = \frac{\frac{1}{2} - \frac{1}{2}}{\frac{1}{2}} = 0 \tag{9.42}$$

and

$$\text{PS} = \frac{P(y_x) - P(y)}{P(x', y')} = \frac{1 - \frac{1}{2}}{\frac{1}{2}} = 1, \tag{9.43}$$

as expected.

9.3.3 Example 3: The Effect of Radiation on Leukemia

Consider the following data (Table 9.1, adapted from[10] Finkelstein and Levin 1990) comparing leukemia deaths in children in southern Utah with high and low exposure to radiation from the fallout of nuclear tests in Nevada. Given these data, we wish to estimate the probabilities that high exposure to radiation was a necessary (or sufficient, or both) cause of death due to leukemia.

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

	Exposure	
	High (x)	Low (x')
Deaths (y)	30	16
Survivals (y')	69,130	59,010

Table 9.1:

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

$$y = f(x, u, q) = (x \wedge q) \vee u, \tag{9.44}$$

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

Since (9.44) is monotonic in x, Theorem 9.2.14 states that all three quantities would be identifiable provided X is exogenous; that is, x should be independent of q and u. Under this assumption, (9.21)–(9.23) further permit us to compute the probabilities of causation from frequency data. Taking fractions to represent probabilities, the data in Table 9.1 imply the following numerical results:

$$\text{PNS} = P(y|x) - P(y|x') = \frac{30}{30 + 69,130} - \frac{16}{16 + 59,010} = 0.0001625, \tag{9.45}$$

$$\text{PN} = \frac{\text{PNS}}{P(y|x)} = \frac{\text{PNS}}{30/(30 + 69,130)} = 0.37535, \tag{9.46}$$

$$\text{PS} = \frac{\text{PNS}}{1 - P(y|x')} = \frac{\text{PNS}}{1 - 16/(16 + 59,010)} = 0.0001625. \tag{9.47}$$

Statistically, these figures mean that:

1. There is a 1.625 in ten thousand chance that a randomly chosen child would both die of leukemia if exposed and survive if not exposed;

2. There is a 37.535% chance that an exposed child who died from leukemia would have survived had he or she not been exposed;

3. There is a 1.625 in ten thousand chance that any unexposed surviving child would have died of leukemia had he or she been exposed.

Glymour (1998) analyzed this example with the aim of identifying the probability $P(q)$ (Cheng's "causal power"), which coincides with PS (see Lemma 9.2.8). Glymour concluded that $P(q)$ is identifiable and is given by (9.23), provided that x, u, and q are mutually independent. Our analysis shows that Glymour's result can be generalized in several ways. First, since Y is monotonic in X, the validity of (9.23) is assured

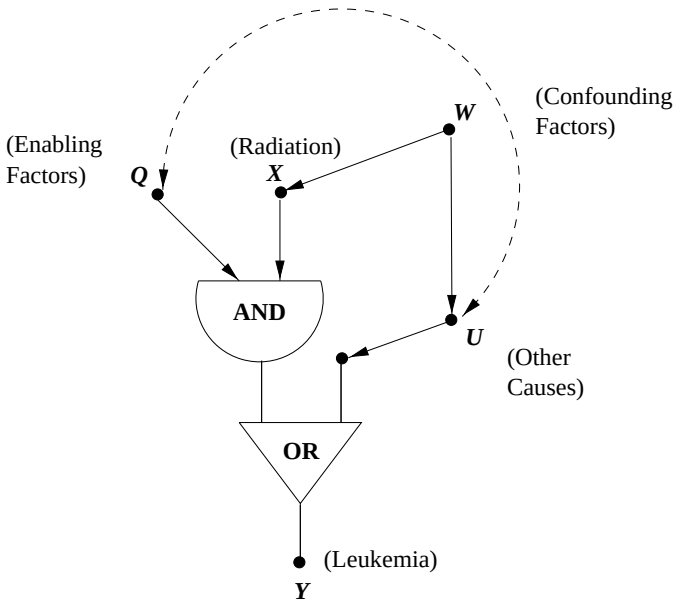

Figure 9.2: Causal relationships in the radiation-leukemia example, where W represents confounding factors.

even when q and u are dependent, because exogeneity merely requires independence between x and $\{u, q\}$ jointly. This is important in epidemiological settings, because an individual's susceptibility to nuclear radiation is likely to be associated with susceptibility to other potential causes of leukemia (e.g., natural kinds of radiation).

Second, Theorem 9.2.11 assures us that the relationships between PN, PS, and PNS (equations (9.11)–(9.12)), which Glymour derives for independent q and u, should remain valid even when u and q are dependent.

Finally, Theorem 9.2.15 assures us that PN and PS are identifiable even when x is not independent of $\{u, q\}$, provided only that the mechanism of (9.44) is embedded in a larger causal structure that permits the identification of $P(y_x)$. For example, assume that exposure to nuclear radiation (x) is suspect of being associated with terrain and altitude, which are also factors in determining exposure to cosmic radiation. A model reflecting such consideration is depicted in Figure 9.2, where W represents factors affecting both X and U. A natural way to correct for possible confounding bias in the causal effect of X on Y would be to adjust for W, that is, to calculate $P(y_x)$ using the standard adjustment formula (equation (3.19))

$$P(y_x) = \sum_w P(y|x, w)P(w) \tag{9.48}$$

(instead of $P(y|x)$), where the summation runs over levels of W. This adjustment formula, which follows from (9.35), is correct regardless of the mechanisms mediating X and Y, provided only that W represents *all* common factors affecting X and Y (see Section 3.3.1).

Theorem 9.2.15 instructs us to evaluate PN and PS by substituting (9.48) into (9.29) and (9.30), respectively, and it assures us that the resulting expressions constitute consistent estimates of PN and PS. This consistency is guaranteed jointly by the assumption of monotonicity and by the (assumed) topology of the causal graph.

Note that monotonicity as defined in (9.20) is a global property of all pathways between x and y. The causal model may include several nonmonotonic mechanisms along these pathways without affecting the validity of (9.20). However, arguments for the validity of monotonicity must be based on substantive information, since it is not testable in general. For example, Robins and Greenland (1989) argued that exposure to nuclear radiation may conceivably be of benefit to some individuals because such radiation is routinely used clinically in treating cancer patients. The inequalities in (9.32) constitute a statistical test of monotonicity (albeit a weak one) that is based on both experimental

and observational studies.

9.3.4 Example 4: Legal Responsibility from Experimental and Nonexperimental Data

A lawsuit is filed against the manufacturer of drug x, charging that the drug is likely to have caused the death of Mr. A, who took the drug to relieve symptom S associated with disease D.

The manufacturer claims that experimental data on patients with symptom S show conclusively that drug x may cause only minor increase in death rates. However, the plaintiff argues 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 were it not for the drug. The manufacturer counterargues 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 grounds that such data may be highly confounded by extraneous factors. The court must now decide, based on both the experimental and nonexperimental studies, what the probability is that drug x was in fact the cause of Mr. A's death.

The (hypothetical) data associated with the two studies are shown in Table 9.2. The experimental data provide the estimates

$$P(y_x) = 16/1000 = 0.016, \tag{9.49}$$

$$P(y_{x'}) = 14/1000 = 0.014; \tag{9.50}$$

the nonexperimental data provide the estimates

$$P(y) = 30/2000 = 0.015, \tag{9.51}$$

$$P(y, x) = 2/2000 = 0.001. \tag{9.52}$$

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

Table 9.2:

Assuming that drug x can only cause (can never prevent) death, Theorem 9.2.15 is applicable and (9.29) yields

$$\text{PN} = \frac{P(y) - P(y_{x'})}{P(y,x)} = \frac{0.015 - 0.014}{0.001} = 1.00. \tag{9.53}$$

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

$$\frac{P(y_x) - P(y_{x'})}{P(y_x)} = \frac{0.016 - 0.014}{0.016} = 0.125. \tag{9.54}$$

Evidently, what the experimental study does not reveal is that, given a choice, terminal patients avoid 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 9.2 were obviously contrived to represent an extreme case and so facilitate a qualitative explanation of the validity of (9.29). Nevertheless, it is instructive to note that a combination of experimental and nonexperimental studies may unravel what

experimental studies alone will not reveal and, in addition, that such combination may provide a necessary test for the assumption of no-prevention, as outlined in Section 9.2.4 (equation (9.32)). For example, if the frequencies in Table 9.2 were slightly different, they could easily yield a negative value for PN in (9.53) and thus indicate violation of the fundamental inequalities of (9.32)–(9.33). Such violation might 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 wonder 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. These invariant quantities are simply the causal effects probabilities, $P(y_{x'})$ and $P(y_x)$. Although these counterfactual probabilities were not measured in the observational 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, implies the inequalities of (9.32)–(9.33).

9.3.5 Summary of results

We now summarize the results from Sections 9.2 and 9.3 that should be of value to practicing epidemiologists and policy makers. These results are shown in Table 9.3, 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 equate 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 (9.13). (The upper bound is usually unity.) The nonentries—in the

Assumptions			Data Available		
Exogeneity	Monotonicity	Additional	Experimental	Observational	Combined
+	+		ERR	ERR	ERR
+	−		bounds	bounds	bounds
−	+	covariate control	—–	corrected ERR	corrected ERR
−	+		—–	—–	corrected ERR
−	−		—–	—–	bounds

Table 9.3: *PN as Function of Assumptions and Available Data* Note: ERR stands for the excess risk ratio, $1 - P(y|x')/P(y'|x')$; corrected ERR is given in (9.31).

r.h.s. of Table 9.3 represent vacuous bounds (i.e., $0 \leq \text{PN} \leq 1$). 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 (9.31). 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 (see Figure 9.2).

The last row in Table 9.3, corresponding to no assumptions whatsoever, leads to vacuous bounds for PN. This does not mean, however, that justifiable assumptions *other* than monotonicity and exogeneity could not be helpful in rendering PN identifiable. The use of such assumptions is explored in the next section.

9.4 Identification in Nonmonotonic Models

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

Our first step would be to study under what conditions the function $P(u)$ can be identified, thus rendering the entire model identifiable. If M is Markovian, then the problem can be analyzed by considering each parents-child family separately. Consider any arbitrary equation in M

$$\begin{aligned} y &= f(pa_Y, u_Y) \\ &= f(x_1, x_2, \ldots, x_k, u_1, \ldots, u_m), \end{aligned} \tag{9.55}$$

where $U_Y = \{U_1, \ldots, U_m\}$ is the set of background (possibly dependent) variables that appear in the equation for Y. In general, the domain of U_Y can be arbitrary, discrete, or continuous, since these variables represent unobserved factors that were omitted from the model. However, since the observed variables are binary, there is only a finite number $(2^{(2^k)})$ of functions from PA_Y to Y and, for any point $U_Y = u$, only one of those functions is realized. This defines a partition of the domain of U_Y into a set S of equivalence classes, where each equivalence class $s \in S$ induces the same function $f^{(s)}$ from PA_Y to Y (see Section 8.2.2). Thus, as u varies over its domain, a set S of such functions is realized, and we can regard S as a new background variable whose values correspond to the set $\{f^{(s)} : s \in S\}$ of functions from PA_Y to Y that are realizable in U_Y. The number of such functions will usually be smaller than $2^{(2^k)}$.[11]

For example, consider the model described in Figure 9.2. As the background variables (Q, U) vary over their respective domains, the relation between X and Y spans three distinct functions:

$$f^{(1)} : Y = \text{ true}, \quad f^{(2)} : Y = \text{ false}, \quad \text{and} \quad f^{(3)} : Y = X.$$

[11] Balke and Pearl (1994a,b) called these S variables "response variables," as in Section 8.2.2; Heckerman and Shachter (1995) called them "mapping variables."

The fourth possible function, $Y \neq X$, is never realized because $f_Y(\cdot)$ is monotonic. The cells (q, u) and (q', u) induce the same function between X and Y; hence they belong to the same equivalence class.

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

$$P(y|pa_Y) = \sum_{s:f^{(s)}(pa_Y) \,=\, \text{true}} P(s). \tag{9.56}$$

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

$$\vec{p} = \boldsymbol{R}\vec{q}, \tag{9.57}$$

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

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

$$Y = U_0 \bigvee_{i=1,\ldots,k} (X_i \wedge U_i), \tag{9.58}$$

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

$$Y = U_0 \bigvee_{i \in T} X_i \tag{9.59}$$

and, for $U_0 =$ false, this equation defines a distinct function for each T. The number of induced functions is $2^k + 1$, which (subtracting 1 for

normalization) is exactly the number of distinct realizations of PA_Y. Moreover, it is easy to show that the matrix connecting $\boldsymbol{p}$ and $\boldsymbol{q}$ is invertible. We thus conclude that the probability of every counterfactual sentence can be identified in any Markovian model composed of noisy OR mechanisms, regardless of whether the background variables in each family are mutually independent. The same holds, of course, for noisy AND mechanisms or any combination thereof (including negating mechanisms), provided that each family consists of one type of mechanism.

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

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

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

In contrast, the mechanism governed by the following function, although nonmonotonic, is invertible:

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

where XOR$(\cdot)$ stands for exclusive OR. This equation induces only two functions from PA_Y to Y:

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

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

We summarize these considerations with a theorem.

Definition 9.4.1 (Local Invertibility)
A model M is said to be locally invertible *if, for every variable $V_i \in V$, the set of $2^k + 1$ equations*

$$P(y|pa_i) = \sum_{s: f^{(s)}(pa_i) = \text{ true}} q_i(s), \tag{9.60}$$

$$\sum_s q_i(s) = 1 \tag{9.61}$$

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

Theorem 9.4.2 *Given a Markovian model $M = \langle U, V, \{f_i\}\rangle$ in which the functions $\{f_i\}$ are known and the exogenous variables U are unobserved, if M is locally invertible then the probability of every counterfactual sentence is identifiable from the joint probability $P(v)$.*

Proof
If (9.60) has a unique solution for $q_i(s)$, then we can replace U with S and obtain an equivalent model as follows:

$$M' = \langle S, V, \{f_i'\}\rangle, \quad \text{where} \quad f_i' = f_i^{(s)}(pa_i).$$

The model M', together with $q_i(s)$, completely specifies a probabilistic causal model $\langle M', P(s)\rangle$ (owing to the Markov property), from which probabilities of counterfactuals are derivable by definition. □ Theorem 9.4.2 provides a sufficient condition for identifying probabilities of causation, but of course it does not exhaust the spectrum of assumptions that are helpful in achieving identification. In many cases we might be justified in hypothesizing additional structure on the model—for example, that the U variables entering each family are themselves independent. In such cases, additional constraints are imposed on the probabilities $P(s)$, (9.60) may be solved even when the cardinality of S far exceeds the number of conditional probabilities $P(y|pa_Y)$.

9.5 Conclusions

This chapter has explicated and analyzed the interplay between the necessary and sufficient components of causation. Using counterfactual interpretations that rest on structural model semantics, we demonstrated

how simple techniques of computing probabilities of counterfactuals can be used in computing probabilities of causes, deciding questions of identification, uncovering conditions under which probabilities of causes can be estimated from statistical data, and devising tests for assumptions that are routinely made (often unwittingly) by analysts and investigators.

On the practical side, we have offered several useful tools (partly summarized in Table 9.3) for epidemiologists and health scientists. This chapter formulates and calls attention to subtle assumptions that must be ascertained before statistical measures such as excess risk ratio can be used to represent causal quantities such as attributable risk or probability of causes (Theorem 9.2.14). It shows how data from both experimental and nonexperimental studies can be combined to yield information that neither study alone can reveal (Theorem 9.2.15 and Section 9.3.4). Finally, it provides tests for the commonly made assumption of "no prevention" and for the often asked question of whether a clinical study is representative of its target population (equation (9.32)).

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

The importance of the sufficiency component can be uncovered in examples where the necessary component is either dormant or ensured. Why do we consider striking a match to be a more adequate explanation (of a fire) than the presence of oxygen? Recasting the question in the language of PN and PS, we note that, since both explanations are necessary for the fire, each will command a PN of unity. (In fact, 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.

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

However, what weight should the law assign to the necessary versus the sufficient component of causation? This question obviously lies beyond the scope of our investigation, and it is not at all clear who would be qualified to tackle the issue or whether our legal system would be prepared to implement the recommendation. I am hopeful, however, that whoever undertakes to consider such questions will find the analysis in this chapter to be of some use. The next chapter combines aspects of necessity and sufficiency in explicating a more refined notion: "actual cause."

Acknowledgments

I am indebted to Sander Greenland for many suggestions and discussions concerning the treatment of attribution in the epidemiological literature and the potential applications of our results in practical epidemiological studies. Donald Michie and Jack Good are responsible for shifting my attention from PN to PS and PNS. Clark Glymour and Patricia Cheng helped to unravel some of the mysteries of causal power theory, and Michelle Pearl provided useful pointers to the epidemiological literature. Blai Bonet corrected omissions from earlier versions of Lemmas 9.2.7 and 9.2.8, and Jin Tian tied it all up in tight bounds.