PN *are identifiable in* ***GP*** *and are given by* (9.28)–(9.30), *with* $P(y_x)$ *determined by the topology of* $\mathbf{G}(\boldsymbol{M})$ *through the algorithm of Tian and Pearl (2002a).*

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, since the model is fully specified, permits us to compute the probabilities of causation from their definitions, (9.1)–(9.3). To exemplify,

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

because $x \wedge y \implies u$ and $Y_{x'}(u) = \text{false}$. 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,

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

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

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

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,

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

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 2000). 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 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).

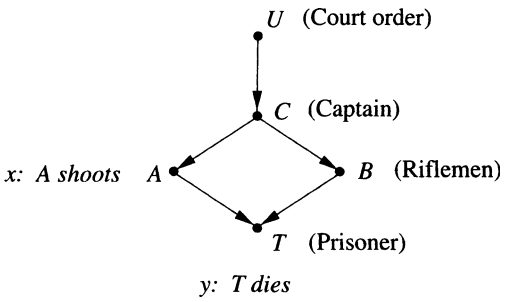

Figure 9.1 Causal relationships in the two-man firing-squad example.

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]

$$P(y_x) = P(Y_x(u) = \text{true})P(u) + P(Y_x(u') = \text{true})P(u')$$
$$= \tfrac{1}{2}(1 + 1) = 1. \tag{9.37}$$

Similarly, we have

$$P(y_{x'}) = P(Y_{x'}(u) = \text{true})P(u) + P(Y_{x'}(u') = \text{true})P(u')$$
$$= \tfrac{1}{2}(1 + 0) = \tfrac{1}{2}. \tag{9.38}$$

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

$$\text{PNS} = P(y_x, y'_{x'})$$
$$= P(y_x, y'_{x'} \mid u)P(u) + P(y_x, y'_{x'} \mid u')P(u')$$
$$= \tfrac{1}{2}(0 + 1) = \tfrac{1}{2}. \tag{9.39}$$

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

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

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.

Similarly,

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

[9] Recall that $P(Y_x(u') = \text{true})$ 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.

Table 9.1

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

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 $P(y \mid x') = 0 \neq P(y_{x'}) = \frac{1}{2}$. 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

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

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[10] from 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.

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:

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

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

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

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 37.544% 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 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 among 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.

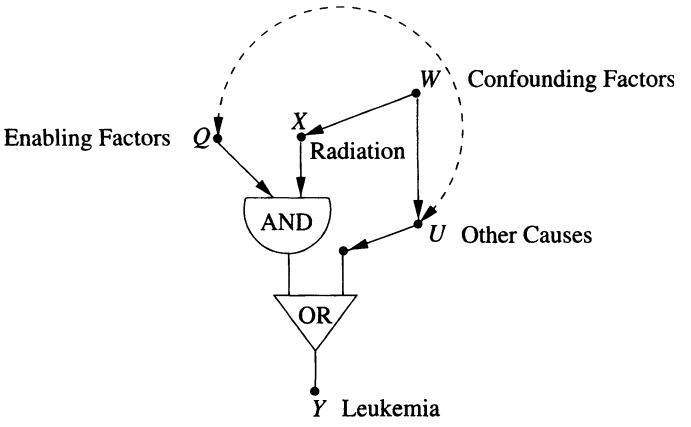

Figure 9.2 Causal relationships in the radiation–leukemia example, where W represents confounding factors.

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 $P(y_x)$ and $P(y_{x'})$. For example, assume that exposure to nuclear radiation (x) is suspected 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 $P(y_x)$ and $P(y_{x'})$ using the standard adjustment formula (equation (3.19))

$$P(y_x) = \sum_w P(y \mid x, w)P(w), \qquad P(y_{x'}) = \sum_w P(y \mid x', w)P(w) \tag{9.48}$$

(instead of $P(y \mid x)$ and $P(y \mid x')$), 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.

Table 9.2

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

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 a 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 2000); 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

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

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

Substituting these estimates in (9.29), which provides a lower bound on PN (see (11.42)), we obtain

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

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 straight-

forward 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 adequacy of the experimental procedures. For example, if the frequencies in Table 9.2 were slightly different, they could easily yield a value greater than unity for PN in (9.53) or some other violation of the fundamental inequalities of (9.33). Such violation would indicate an incompatibility of the experimental and nonexperimental groups due, perhaps, to inadequate sampling.

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 implies the inequalities of (9.33) and, if monotonicity holds, (9.32) ensues.

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 (for a nonexperimental event) 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

Table 9.3. ***PN as a Function of Assumptions and Available Data***

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

Note: ERR stands for the excess risk ratio, $1 - P(y \mid x')/P(y' \mid x')$; corrected ERR is given in (9.31).

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 right-hand side 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 $[P(y \mid x') - P(y_{x'})]/P(x, y)$, 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 $P(y \mid x') - P(y_{x'})$. However, confounding results only from association between exposure and other factors that affect the outcome; one need not be concerned with associations between such factors and susceptibility to exposure (see Figure 9.2).

The last row in Table 9.3, corresponding to no assumptions whatsoever, leads to vacuous bounds for PN, unless we have combined data. 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,

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