Chapter 2

Causal Variables

Causal contrasts

Imagine a four-group trial in which thousands of patients were randomized to receive a pill that contained drug A, drug B, or placebo, or to receive no pill at all. (Recipients of a pill were not told, of course, what it contained.) We'll assume that all patients adhered to their assigned treatment group. Table 2–1 shows hypothetical results.

Table 2–1. Frequency of recovery in a hypothetical trial, by treatment group

OUTCOME	TREATMENT GROUP			
	Drug A	Drug B	Placebo	No pill
Recovery within one year	20%	20%	10%	5%

What is the effect of drug A on recovery?

Anyone familiar with the rationale of trials knows that the percentage below the heading "Drug A" does not estimate the effect of that drug (whereas many people mistakenly call the percentage below the placebo heading "the placebo effect.") To estimate the effect of drug A, we have to compare the outcome of taking this drug (20% recovered) to the outcome of an alternative condition (by dividing the percentages, for example). But as Table 2–1 shows, we may choose from three alternative conditions: taking drug B, taking placebo pill, and taking no pill. The question "What is the effect of drug A on recovery?" cannot be answered until the question "As compared with what?" is answered. In other words: the effect of which causal contrast do you wish to estimate?

Although it is possible to argue that one causal contrast illuminates us more than others, we should not dismiss other contrasts too quickly: they may add to our knowledge of causal effects. Looking at Table 2–1, many of us will highlight the comparison of drug A with placebo as the effect of interest—the so-called real effect of the drug. Nonetheless, there are at least two other interesting effects that should be considered.

First, although taking drug A was better than taking placebo, it was no better than taking drug B, useful knowledge no physician would dismiss. In fact, as compared with drug B, drug A had no effect on recovery and vice versa. Second, if drug A were to be prescribed routinely to patients, its effect should also be estimated relative to not taking a pill because most physicians will either prescribe a drug or not—they will not prescribe a placebo pill instead of a drug. As a result, in the world of real medicine a patient may benefit from the sum of two separate effects: that of active ingredients and that of placebo. One effect does not prohibit the other.

The placebo effect itself can be estimated by contrasting the frequency of recovery while taking placebo (10%) with the frequency of recovery without taking a pill (5%). Now before you dismiss the placebo component of prescribing a drug as unreal, allow me to digress for a moment on a gloomy side note. Unfortunately, we know very little about the realness or magnitude or variation of the placebo effect in medicine because few trials

have estimated it; you will rarely find a third, “no pill” group in a placebo-controlled trial. And with your permission I will speculate why in one sentence. In a world of specialized research most testers of new drugs are very worried about the placebo effect but have no interest in estimating it, whereas most researchers of the placebo effect don’t test new drugs. A side effect of specialization, I think.

How many causal contrasts are there?

In some studies there is only one contrast—for example, the effect of an abnormal gene on longevity. An abnormal gene may be present or may be absent; there are no other possibilities. In a randomized trial the number of contrasts may be greater than two (as in our hypothetical trial) but it is usually small and certainly finite. In an observational study, however, that number may be large, even infinite. Consider for instance an observational study of the effect of smoking on lung cancer. We may contrast continued smoking of 20 cigarettes per day with never smoking, or with quitting smoking, or with continued smoking of 10 cigarettes per day. Each of these contrasts and many more conceivable pairs will add to our knowledge of the effect of smoking on lung cancer. Or consider the effect of systolic blood pressure on stroke. Many paired values make contrasts of interest: 180 millimeter mercury and 160; 175 and 120; 200 and 170; and so on.

Causal contrasts run through most pages of this book but they are often hidden in what is called a causal variable, our next topic.

A leap to causal variables

Speeding on the road, having an abnormal gene, systolic blood pressure of 200-millimeter mercury, and taking a drug are all called causes only because of their effects. But as we now realize their effects can be quantified only when a causal contrast is specified. The absolute effect of any of these causes, if it can be defined at all, cannot be quantified in the empirical world. So if we wish to study the effect of presumed causes, we have to study (presumed) causal contrasts.

Contributors to causal contrasts will naturally fold into a variable. Speeding and not speeding, for instance, are two categories of a variable we may label SPEEDING STATUS. By the same token an abnormal gene and a normal gene will make up the GENE variable. The four assignments in our hypothetical trial are four categories of a variable that may be called TREATMENT GROUP, whereas the values of systolic blood pressure reside in a variable we will abbreviate SBP. It becomes apparent, then, that to explore a cause we have to study a causal variable by contrasting the effects (loosely speaking) of its values or categories. And to acknowledge uncertainty of causality, we should add the adjective "presumed" in front of "causal variable." The values of a causal variable are sometimes called causal assignments, even though they are not always assigned by a human being. For example: when we study systolic blood pressure as a cause of stroke, one’s value of systolic blood pressure is one’s causal assignment.

You will not encounter often the words “presumed causal variable” in textbooks or scientific articles. Many scientists shy away from the word cause and from any of its derivatives, assuming perhaps that semantic disguises could hide the pitfalls of causal inquiry. But there may be two other explanations: First, long terms are cumbersome to say and to write. Second, as we’ll see shortly, the variable we measure is often not the true

causal variable. Whatever the reason may be, epidemiologists have invented substitute terms for the causal variable of interest, which should be familiar to anyone who attended a basic course in epidemiology: risk factor, exposure variable, exposure status, or simply exposure. In this book, all these terms and "causal variable" are used interchangeably.

Words often invoke a mental image and the noun "exposure" might invoke the image of external harm that was delivered accidentally—like “radiation exposure”. If that’s what runs through your mind, as ran through mine as a student, you should suppress the image. Exposure is a causal variable (presumably) whose values need not be external nor should they be harmful. One’s genotype is an endogenous exposure and exposure to radiation therapy may benefit patients.

Theoretical causal variables

It is common knowledge that certain molecules in the blood cause diseases, for example LDL-cholesterol (the so-called bad type) causes coronary atherosclerosis. Now look at the previous sentence again with a critical eye, keeping in mind what a wise man has once said about the meaning of a scientific hypothesis:

“For me, a hypothesis is a statement whose truth is temporarily assumed, but whose *meaning* must be beyond all doubt.”

—Albert Einstein

The statement “LDL-cholesterol causes coronary atherosclerosis” is vague. It is vague not because the word cause is used, but because no exposure variable is specified and we are left wondering which causal contrasts should be studied to corroborate or refute that statement. Surely it’s not the presence of LDL-cholesterol molecules in the blood as compared with their absence, which is biological nonsense. Probed to clarify, we may try to improve by restating that the *amount* of LDL-cholesterol in the blood causes coronary atherosclerosis—a better formulation but still too vague for a scientific statement. What is the content of an exposure variable called “the amount of LDL-cholesterol”? The total number of circulating molecules? The number of molecules that travel through the left coronary artery in one second? The concentration in Aortic blood? Besides this ambiguity, each of these numbers change by the day, by the hour, or by the second, so a reference to time is missing.

Although this isn’t a textbook about the causes of coronary heart disease, there is something to be learned from diving deeper into that example. When I wrote the book, two exposures have attracted more attention than others: the number of LDL-cholesterol molecules that interact with—and thereby could injure—the lining of a coronary artery, and the number of molecules that get trapped in the arterial wall. Both theories, however, still require us to clarify which numbers will make up the exposure variables: cumulative number of molecules over one’s lifetime, or perhaps a complex function that integrates values over volume and space and flow and time, a function that no one has written yet.

What is the moral of the example?

When we try to refine the definition of an exposure, we are pursuing science rigorously. Thinking intensively about the true cause may help us discard overly simplified theories and might get us closer to the theoretical exposure—the center of causal inquiry. Along the way we often learn to appreciate the complexity of Nature and the imperfect nature of our causal theories. A lesson in humbleness has never hurt a scientist.

Surrogates of causal variables

Regardless of how we might specify a theoretical exposure for the effect of LDL-cholesterol on coronary atherosclerosis, we are years away from measuring any theoretical exposure of interest. And this statement holds truth for many other possible causes because technology often lags behind scientific thought and scientific imagination. Still, how would we study an exposure that cannot be measured and sometimes can be stated only vaguely?

The typical solution is to study a surrogate variable, a well-defined variable that can be measured and has a strong causal link to the theoretical exposure we have in mind. The causal link between the two may take three forms: the surrogate variable may be a strong cause of the theoretical exposure or vice versa, or both could be the effect of another cause (Figure 2–1). A combination of a direct arrow between the two and a shared cause is also possible.

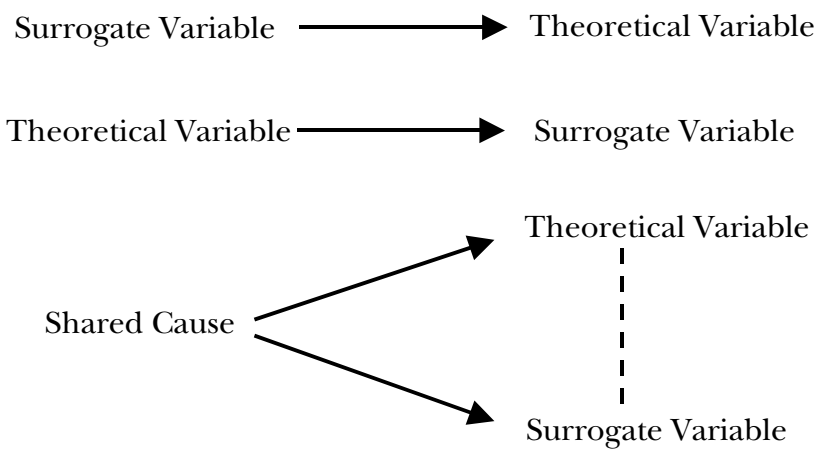

Figure 2–1. Three kinds of relations between a surrogate variable and a theoretical exposure

The logic of using a surrogate measure evolves from the idea that the actual value of the true exposure (that is, one's causal assignment) is not essential; it may be replaced by the value of another variable that preserves the true causal contrast between people. For example: if Judy's value of the theoretical exposure is 5 exposure-units and Jeff's value is 10 exposure-units on a scale that is bound at zero, the causal contrast between them will be preserved by a surrogate variable on which Judy's value is 80 mg/L and Jeff's value is

160 mg/L. On this variable, Jeff would still be ranked twice as high as Judy (or Judy half as high as Jeff.) In our example above, incidental blood concentration of LDL-cholesterol may play the role of a surrogate measure, but to claim it can do so, we need more than a simple diagram. We have to supply a chain of reasoning, one paragraphs long.

At any given moment, the concentration of LDL-cholesterol in the blood should affect the number of molecules that interact with the arterial wall. In other words, these two variables may be connected by a causal arrow. Extending that momentary relation to an infinite series of moments, we may assume that the integration of blood concentration of LDL-cholesterol over one's lifetime preserves causal contrasts on our theoretical exposure. And although the lifetime concentration of any molecule cannot be measured (yet), we may invoke, again, the assumption of preserved causal contrasts: lifetime concentration and incidental concentration should rank people reasonably similar. If Judy's lifetime concentration is "low", her incidental concentration is likely to be "low" as well. And if Jeff's is "high", his incidental concentration is likely to be "high" as well. These variables are strongly correlated, to use statistical jargon.

We have finally completed a chain of reasoning for using the incidental concentration of LDL-cholesterol in the blood to study the effect of LDL-cholesterol molecules on coronary atherosclerosis and that is exactly what epidemiologists have done. By now you might have realized, however, that the road from a theoretical exposure (interacting molecules at the arterial wall over decades) to a measured surrogate (blood concentration on February 2, 1987 at 8:53AM) is loaded with so-called reasonable assumptions about preserved causal contrasts. Only one signpost is missing—a method to distinguish between reasonable scientific assumptions and their unreasonable counterparts. If you find one (and it's not a vote), send me a note.

As I mentioned earlier, a surrogate variable need not be a cause of its theoretical exposure; a reversed causal order would work, too. Consider, for instance, the amount of fat in the abdomen—a postulated causal variable for some diseases— and one of its effects: the circumference of the waist, which is an easily measured surrogate variable. Indeed, in many epidemiological studies waist circumference (or the ratio of waist circumference to hip circumference) has substituted for the amount of abdominal fat.

Besides being the cause or effect of a theoretical exposure, a surrogate measure will be correlated with the theoretical exposure whenever the two share common causes, as illustrated in Figure 2–1 and by the following example: Contemporary technology enables us to measure the amount of abdominal fat directly by imaging methods such as computed tomography and magnetic resonance. If measured, this variable may substitute for a theoretical exposure that somehow entails lifetime variation of abdominal fat. Though a single measurement of any variable is neither the cause nor the effect of lifetime values, the two quantities share similar causes: whatever affect lifetime variation of abdominal fat (genes and diet, for example) should also affect the amount of fat on the day it was measured. And when two variables have a cause in common, they will be correlated, which means, again, that one variable will partially preserve the causal contrasts on the other. (Thinking back, we have implicitly used this reasoning earlier to substitute the incidental concentration of LDL-cholesterol for lifetime average concentration.) In chapter 6 we will discuss another important, yet unhelpful, consequence of a shared a cause—a phenomenon called confounding.

Surrogate variables are often used in other aspects of causal inquiry, replacing theoretical effects and theoretical confounders that cannot be measured. I might even dare saying that surrogate variables are the norm, not the exception, in most epidemiological studies. Unfortunately, epidemiologists often neglect to state what a

surrogate variable stands for, either assuming that it's common sense knowledge or assuming that nobody dare guessing. Unlike their surrogates, theoretical variables are hard to define and commit to.

Legitimizing causal variables

One day I was walking down the hallway in my department, talking with a colleague about new results we have just received from a programmer. "Look at the sex group effect", I commented, pointing to a number on a computer printout that showed that the frequency of stroke was higher in men than in women. "You mean the *difference* between men and women", my colleague corrected me. "No", I said, "I mean the sex effect, or the gender effect if you prefer the newspeak."

My colleague is not alone in his linguistic camp. Many epidemiologists sort variables into those that are entitled to be called causes of effects and those that are not, often placing sex, race and age on the list of forbidden variables. To be called a cause, they argue, a variable must contain causal assignments that are exchangeable—it must be possible for a human being to switch from one causal assignment to another, in principle at least. One can continue to smoke or can quit smoking, take drug A or take placebo pill, have systolic blood pressure of 180 or of 120. But a man cannot be a woman, a woman cannot be a man, a white person cannot be a black person, and a black person cannot be a white person. As for age, in some minds it should join the list of non-causes for two reasons: first, one cannot exchange one's actual age with any other age. Second, aging is equivalent to the passage of time and the passage of time per se does not cause anything—so it is argued. (Go tell a homeowner in mid-January that the passage of time did not cause her 50-year old furnace to break.)

Why should we require that causal assignments be exchangeable?

I think we should not, but the disagreement may be retraceable, in part, to a choice between models of causation, and in part, to the idea that one causal assignment could be replaced by another. Contemplating a component cause of some event, the deterministic mind naturally brings up a human being and a "what if" question: "What if that component cause had been absent?" which explicitly means, "Would the event have occurred had the causal assignment been different, and all other matters unchanged?" Now, he will naturally imagine a world with a different causal assignment and judge whether it's reasonable to switch to that imaginary world. If it's not, he will be inclined to reject the candidate for the title "component cause" along with the causal variable to which it belongs.

"What if a person had not smoked?" sounds reasonable. A smoker could have switched to an alternative causal assignment, which is "not smoking". But "What if a man had been a woman?" does not sound reasonable. A man could not have been a woman. Or could he?

I dislike the deterministic trail of reasoning mainly because it forces me to discuss simple-minded questions ("Could a man have been a woman?") and invites a stingy rebuttal. Yes, a man could have been a woman in at least two ways—surgical sex change and taking female hormones. And maybe some day a man may be able to become a woman by replacing every pair of XY chromosomes in his body with XX: Who knows? We've all heard about science fiction stories of the past that have turned into respected

science of the present—gene therapy, for example. Does this mean that genotype was not a causal variable in the nineteenth century and has become one in the twenty first century? Does causal reality depend on human imagination, scientific knowledge, and today's technology?

There are other rebuttals I may offer, however. Scientists often treat variables such as sex, race and age as surrogates for theoretical exposures of interest, many of which contain exchangeable causal assignments even by contemporary thought. When a medical researcher is examining the effect of the patient's sex on the use of a diagnostic procedure, she has in mind a sexist attitude of physicians or sex-dependent symptoms, for which the patient's sex is a surrogate. When an epidemiologist is studying the effect of race on survival of stroke patients, he has in mind race-related biology, sociology, and medical care—not skin pigmentation. And age is a surrogate variable for theoretical exposures such as the error rate of biological systems, malfunction of repair mechanisms, and cumulative exposure to external hazards. All of these theoretical exposures contain exchangeable causal assignments, just in case we have to comply with that preference of the deterministic mind.

But we don't. If we hold an indeterministic model of causation, no logic requires us to create a list of illegitimate exposures. Two causal assignments may generate two different propensities to bring about an effect regardless of whether any human being may be able to switch from one assignment to the other. Nor does causal reality depend on our ability to explore its existence by a randomized trial. We may compare the causal propensity of male sex to that of female sex without phrasing the issue as "if a man were a woman" and without asking whether it is possible at this time to randomize to sex group. It is no different from comparing the gravitational force of the earth to that of the moon without asking whether the earth could have become the moon.

Finally, there is nothing to suggest empirically that some variables are entitled to be called exposures and others are not. When we inspect a computer printout, the estimated effect of sex (whatever sex represents) on the frequency of stroke is indistinguishable from that of blood pressure. If the number next to the sex variable does not estimate an effect, what other reality does it describe?