5

Establishing Causal Links

CAUSAL STUDIES

The search for causal explanations is of central importance in every area of scientific research. The first step in understanding something often involves speculating about what its cause or causes might be and then finding a way to test those speculations. In the last few years, for example, there has been a dramatic increase in the number of American children who are obese. What factors might be responsible for this increase? Too much fast food? Too little exercise? Some other factor? Some combination of factors? Causal experiments or, as they are usually called, causal studies are the main tool by which researchers confront such questions.

The design of causal studies and the assessment of their results present researchers with a series of problems over and above those we have encountered in thinking about how to design an experiment. The first stems from the fact that causation is not an all-or-nothing matter. Not every subject exposed to a cause will necessarily yield the effect. To establish a causal link, what must be shown is that significantly different levels of the effect obtained in those exposed and not exposed to the cause. And this can be a problem. Causal studies generally involve limited number of subjects. How can researchers be sure that a difference in levels of effect in their study groups is not due to the fact that relatively large differences often occur by chance in relatively small groups? How can they be sure, that is, that whatever effect a study uncovers is due to the suspected cause, and not random chance?

Second, effects are rarely associated with a single causal factor. How, then, can researchers be sure that other causal factors have not influenced the results? This problem is all the more severe because experimental and control subjects will usually be drawn from a population that may have been exposed to these other factors.

Finally, causal research cannot always be undertaken in a way that conforms to the model of a tightly controlled experiment outlined in Chapter 4. Researchers cannot, for example, investigate the possible link between diet and obesity in children by exposing young children to unhealthy food. Some other way must be found to test for this link.

Causal studies typically investigate the impact of suspected causes on the members of large populations, as in the example above. Often they will deal with issues of some considerable practical importance, like the problem of obesity in American children. For this reason, causal studies are the most widely covered of all science stories in the popular media. Near the end of the chapter we will turn out attention to the ways in which causal studies are handled in the mass media. Unfortunately, media coverage of causal research leaves a lot to be desired. Facts needed to make sense of a study will often be omitted and study outcomes oversimplified. And more often than not, these shortcomings result from a failure to appreciate the issues we will consider in this chapter.

RULING OUT CHANCE

In 2001 a study was undertaken to determine whether St. John's wort, a popular herbal remedy, can counter the effects of depression. The study involved 200 women, most in their early 40s, who had suffered from major depression for two years, though none were classified as being severely depressed. The women were assigned at random to receive either an extract of St. John's wort or a placebo. Those taking the St. John's wort began with three tablets a day, each containing a commonly recommended dosage. If they did not improve after four weeks, the dose was increased to four tablets. After eight weeks, the patients were evaluated by psychiatrists who did not know whether the patients were in the experimental or control groups. Twenty-seven percent of those who took St John's wort showed marked improvement compared with 19% in the placebo group. Is this difference large enough to show that St. John's wort is an effective treatment for depression?

Before we can answer this question, we need to know something about the thinking involved in estimating the accuracy of samples taken at random from large populations. This is because the subjects in a causal study are a sample. In the St. John's wort study, the 200 hundred subjects had been selected from the larger population composed of women suffering from depression. And the question at issue is whether, based on the study's findings about this sample, we can conclude that St. John's wort counters depression in the larger population. Or is it that the difference reported in the study is nothing more than the kind of chance variation often found in small samples?

Statisticians have been able to determine that the accuracy of a sample is, to a large extent, a function of sample size. The larger a sample, the greater the chances it will accurately mirror what is true of the population from which it was taken. This is a rough approximation of something statisticians call *the law of*

large numbers. Flip a fair coin 10 times, and it should come down "heads" half the time. But in a series of 10 flips, chances are not all that bad that "heads" will come up three or four times or perhaps six or seven times. Now flip the same coin 1,000 times. Chances are very high that the coin will come down "heads" somewhere in the neighborhood of 500 times; chances are slim that we will get a result very much higher or lower than 500. (To see why, imagine what the chances would be of flipping a coin 1,000 times and getting only one or two "heads" as opposed to flipping it 10 times with the same result.)

This fact about samples accounts for a notion that is indispensable in estimating the chances that a sample is accurate—*margin of error*. Imagine that a poll has just been taken of registered voters in your state. Five hundred voters were selected at random, telephoned, and asked if they intend to vote in the upcoming election. 52% said yes. In reporting this result, the pollsters will probably say something like "this sample has a margin of error of +/− (plus or minus) 4%." As a general rule, what this means is that if a sample of this size (500) were taken 20 times, in 19 of the 20 samples − 95% of the samples − the outcome would be within 4%, one way or the other, of the result obtained in the sample actually taken. So, in the case of our poll, there is a 95% chance that between 48% and 56% (52% +/− our 4% margin of error) of all registered voters will vote in the upcoming election.

Suppose instead that the sample had involved 1,500 voters with the same results. Remember, the larger a random sample, the greater the chances its results will be accurate. What this means is that as sample size increases, the margin of error decreases. For a sample of 1,500, the margin of error is about +/− 2%. Chances are 19 out of 20 that something between 50% and 54% of the registered voters will turn out in the next election.

Table 5.1 gives the margins of error for a number of common sample sizes. In each case, there is a 95% chance that the sample outcome will reflect that of the population from which it was taken. Our choice of the 95% confidence level is somewhat, though not entirely, arbitrary. For example, we could just as easily have given margins of error at a lower confidence level, say, the 80% level. As you might suspect, the margins of error for this lower confidence level would be smaller. If we are willing to tolerate a greater chance that we are wrong, we can venture a more accurate estimate. Table 5.1 tells us that in a sample of 100, for example, we can be 95% sure that the outcome will be within +/−10% of the sample outcome. Were we to restrict ourselves to a conclusion we could be 80% sure would be true, the margin of error would shrink to about 7% either way. However, most sampling is based on the 95% confidence level. Unless we have information to the contrary, it is a safe bet that a reported margin of error is at this level.

To assess the results of a causal study, we will need to make use of what we have found out about margin of error. As we noted earlier, experimental and control groups are, in a sense, samples. The 200 women in the St. John's wort study are representative of the larger population of all people who fit the profile of the study subjects: women in their 40s who suffer from moderate levels of depression. (Unless there is good reason to think this particular group is unusually susceptible to bouts of depression, the population can perhaps be

TABLE 5.1

Sample Size	Approximate Margin of Error(%)*
25	+/–22
50	+/–14
100	+/–10
250	+/–6
500	+/–4
1000	+/–3
1500	+/–2.5
2000	+/–2

*The interval surrounding the actual sample outcome containing 95% of all possible sample outcomes.

expanded to include all people who suffer moderate depression.) So the issue we must resolve is whether the difference in the outcomes for samples that make up the experimental and control group are large enough to indicate a causal link and not just the kind of random statistical variation associated with sampling.

In the St. John's wort study, 27% of 100 experimental subjects improved as did 19% of the control subjects. Is this difference enough? Look back to Table 5.1. The margin of error for samples of 100 is about +/− 10%. This tells us there is a 95% chance that in the population from which the sample was taken, somewhere between 17% and 37% could be expected to improve. In the population corresponding to the control group, somewhere between 9% and 29% would improve. Figure 5.1 shows that there is considerable overlap between these two intervals. This tells us is that chances are quite high that the difference we have discovered is due to random statistical fluctuations in the sampling process. This result does not mean that there is no link between the suspected causal agent and the effect we are testing. It is entirely possible that a causal link exists but that the level of effect is too small to measure using

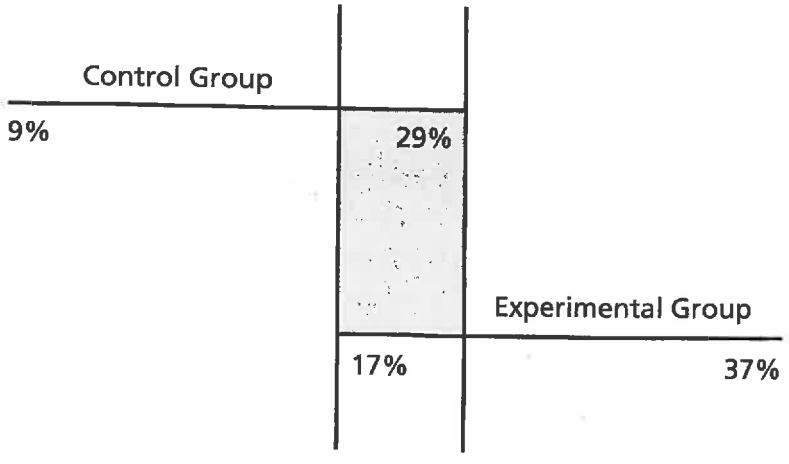


FIGURE 5.1

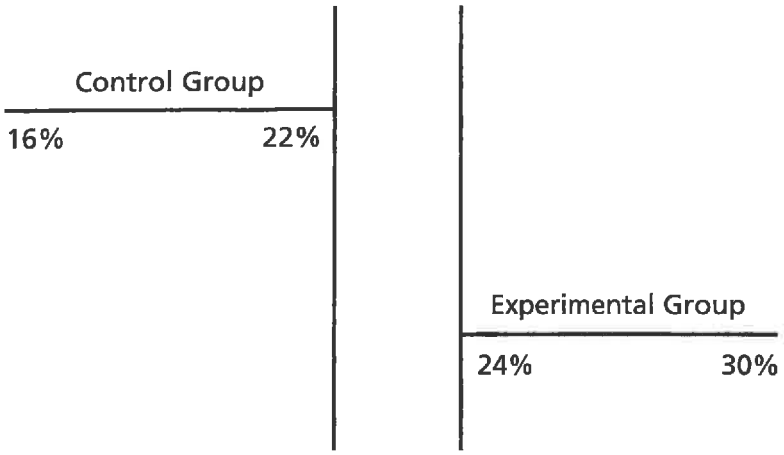


FIGURE 5.2

groups of this size. What we can conclude, however, is that this particular experiment has not conclusively established such a link. Were the difference between levels of effect in our two groups to have been 20% or more, we would have concluded that the difference is due to something other than the random statistical fluctuations associated with sampling. Quite possibly, it is due to the fact that St. John's wort can prevent depression.

Interestingly enough, the difference in levels of effect found in the St. John's wort study would have been sufficient to suggest a causal link if the experimental and control groups had been larger! Imagine if the study had been carried out on groups containing 1,000 subjects each. Table 5.1 tells us the margin of error for groups of this size is +/− 3%. The corresponding intervals are represented in Figure 5.2. Note there is a clear gap between the two intervals. It is not unusual for causal researchers to expand study sizes when their initial evidence for a causal link is tentative. A borderline result may become much less ambiguous if it can be shown to persist as larger samples are investigated. Nor is it unusual for researchers to combine the results of several small studies in an attempt to create groups of sufficient size to suggest a link that may be less clearly indicated in any of the individual studies. This approach is called *meta-analysis*. The findings of any such analysis are at best tentative, since their experimental and control groups come from many studies, and thus are likely to minimize differences in the studies over which they range.

Causal experiments do not always involve experimental and control groups of the same size. Even where the groups differ in size, we set minimal levels of difference in much the same way. Suppose, for example, that we have an experimental group of 50 subjects and a control group of 100. In constructing our intervals we need only make sure to work with the proper margins of error, which will be different in each case. Since we are working with percentages, we should encounter no difficulty in comparing the intervals.

When the results of causal experiments are reported, researchers often speak of differences that are or are not *statistically significant*. A difference in the outcome of two samples will be statistically significant when there is little or no overlap between the confidence intervals for the experimental and control

groups. Thus a difference that is statistically significant is one which is highly unlikely to be due to normal sample fluctuations; chances are slim that two groups, chosen at random, would accidentally differ by the amount we observed in our experiment. Conversely, a result that is not statistically significant suggests there is a great deal of overlap and that the observed difference in levels of effect may well be due to random sample fluctuations.

As we have noted, causal research is nearly always conducted at the 95% confidence level. A statistically significant result, then, is one that would occur by chance only one time in twenty. In working with the results of a study done at this confidence level, Table 5.1 can help us to decide whether a result is or is not statistically significant. But a note of caution is in order here. The intervals in Table 5.1 can give us a rough approximation of whether a difference in experimental and control group outcomes is significant. However, they are a bit off. The percentage difference required to achieve statistical significance is a bit less than the difference suggested by Table 5.1. For example, a difference of just over 13% will be statistically significant for groups of 100 or so. (The required differences decrease even more when levels of the effect are very near to 0% or 100%.) Table 5.1 suggests that a 20% difference would be required. The amount of overestimation in Table 5.1 decreases as the size of experimental and control groups increases. Table 5.1 suggests a 6% difference is required to achieve statistical significance for samples of about 1,000, when in fact just over a 4% difference will do the trick. We can correct for the inaccuracy in Table 5.1 if we adopt the following rules of thumb in working with reported differences between experimental and control groups:

1. If there is no overlap in the intervals for the two, the difference is statistically significant.
2. If there is some overlap in the intervals (in the intervals have less than one-third of their values in common), the difference is probably statistically significant. The greater the overlap, the smaller the chances the difference is significant.
3. If there is a good deal of overlap (more than one-third of all values), the difference is probably not statistically significant.

In the jargon of the causal researcher, failure to establish a causal link is often called a failure to reject the *null hypothesis*. The null hypothesis is simply the claim that there is no difference between levels of effect in the real populations from which the samples were taken. An experiment that succeeds in establishing a large enough difference in levels of effect between experimental and control groups will often be said to reject the null hypothesis. But in the study of St. John's wort, such a difference has not been observed. On the basis of the study, in other words, we cannot reject the null hypothesis.

Imagine now that we are about to design a causal study. Does A cause B in Cs? We select a number of Cs, assign them at random into experimental and control groups, and administer A to the experimental subjects and a placebo to the controls. How large of a difference in levels of B should we expect to find at

the end of the study if there is a causal link? The answer now should be clear. Whatever it takes to allow us to reject the null hypothesis for samples of the size with which we are working.

MULTIPLE CAUSAL FACTORS

As a veteran teacher with years of experiences observing students, I'm convinced that students who attend class regularly generally do better on tests that do those who attend sporadically. But then personal observation can be misleading. Maybe I have just remembered those good test takers who always came to class, since I would like to think my teaching makes some difference. Is there really a causal link between my teaching and the performance of my students? We can determine this by doing a test. I will teach two courses in the same subject next semester, each containing 100 students. The only difference between the two courses will be that in the first, attendance will be mandatory, while in the second it will be voluntary. All material to be tested will be covered either in the textbook or in lecture notes to be supplied to all students. Course grades will be based on a single, comprehensive final exam given to all students in both courses. Suppose now that we have performed this experiment, and at the end of the term we discover a statistically significant difference between the test scores of the two groups. The experimental group, the group required to attend, scored much higher, on average, than the control group, most of whom took advantage of the attendance policy and rarely appeared in class. To ensure accuracy we have excluded the five highest and lowest scores from each group and the average difference remains statistically significant.

Despite the care we have taken in designing our experiment, it nonetheless suffers from a number of shortcomings. Perhaps the most obvious is the fact that it involves no control of factors other than attendance which might influence test scores. One such factor, obviously, is the amount of time that each subject studies outside of class. Remember, tests were based solely on material available to all subjects. What if a much higher percentage of the subjects in the experimental group than the control group spent considerable time preparing for the final? If this is the case, we would expect the experimental group to do better on the final but for reasons having little to do with class attendance.

The way to avoid this sort of difficulty is by matching within the experimental and control groups for factors, other than the suspected cause, which may contribute to the level of the effect. Matching involves manipulating subjects in an attempt to ensure that all factors that may contribute to the effect are equally represented in the two groups. There are several ways of matching. One is simply to make sure that all other contributing factors are equally represented within both groups. This we might accomplish in our experiment by interviewing the students beforehand to determine the number of hours on average studied per week. Presuming we can find an accurate way of getting this information, we can then disqualify students from one or the other of our groups until we have equal numbers of good, average, and poor studiers in both

groups. Another way of matching is to eliminate all subjects who exhibit a causal factor other than that for which we are testing. Suppose we discover that a few students in each group are repeating the course. We might want to remove them altogether from our study.

The final way to match is to include only subjects who exhibit other possible causal factors. We might do this by restricting our study to students, all of whom study roughly the same amount each week. If all of our experimental and control subjects have additional factors which contribute to the effect in question, the factor for which we are testing should increase the level of the effect in the experimental group, provided that it is actually a causal factor. Matching in this last way can be problematic if there is any chance that the effect may be caused by a combination of factors. Thus we may end up with an experiment which suggests that A causes B in Cs when, in point of fact, it is A in combination with some other factor which causes B in Cs.

By matching within our two groups we can frequently account for causal factors other than the factor we are investigating. However, there is a way that unwanted causal factors can creep into an experiment which matching will not prevent. We must be on guard against the possibility that our subjects will themselves determine whether they are experimental or control subjects. Imagine, for example, a student who has enrolled in the course that requires attendance but then hears from a friend about the course that does not require attendance. It seems at least likely that poor students will opt for the course that requires less. Thus, we may find that poor students have a better chance of ending up in the control section rather than in the experimental section. We could, of course, control for this possibility by making sure students do not know the attendance policy prior to enrolling and by allowing no movement from course to course. Another problem we might have here is that poor students in the experimental group, upon hearing of the attendance policy, might drop out, again leaving us with an experimental group not well matched to the control group. In any event, it is worth taking whatever precautions are possible, in designing a causal experiment, to insure that subjects have no way to influence the composition of the experimental and control groups.

RANDOMIZED, PROSPECTIVE, AND RETROSPECTIVE CAUSAL STUDIES

Randomized Causal Studies. The causal studies we have discussed so far have several things in common. Each began with a set of subjects who, prior to the study, were very much alike. (In the second study—the one about student test performance—matching was used to make sure both groups were similar in makeup.) In particular, none had been exposed to the suspected causal agent. In both, subjects were randomly assigned to experimental and control groups. Only then were experimental subjects exposed to the suspected cause. *Randomized studies*, as studies of this sort are called, conform nicely to the criteria for a decisive experiment discussed in Chapter 4. Any pronounced differences in levels of effect

that emerge over the course of either study would be highly likely if there is a causal link and highly unlikely otherwise.

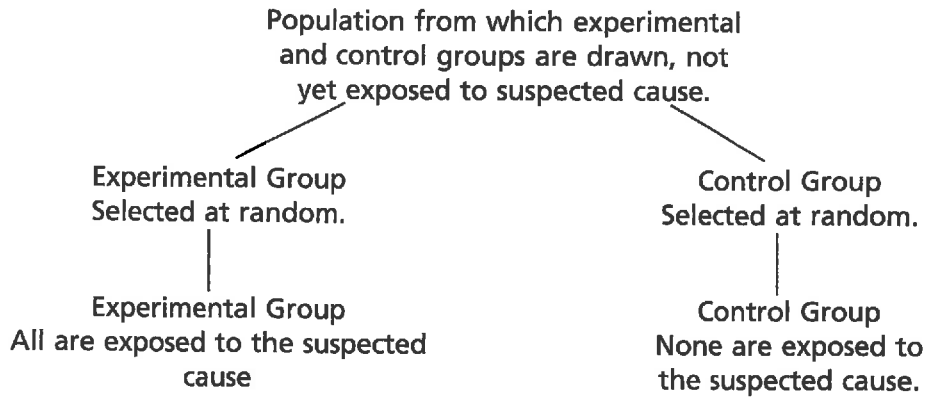
The great advantage of randomized studies is that they are capable of providing unequivocal evidence for a causal link. But they have several disadvantages as well. First, they tend to be quite expensive, particularly if it is necessary to work with large groups of subjects. Second, unless the suspected effect follows reasonably soon after exposure to the casual agent, randomized studies may take a great deal of time to carry out.

In 1981, a large-scale randomized study was begun to determine whether low doses of aspirin could decrease the risk of myocardial infarction (heart attack). The study was undertaken by a team of investigators from Harvard Medical School and a Harvard teaching hospital. Over 260,000 male U.S. physicians between the ages of 40 and 84 were invited to participate. Of the nearly 60,000 who agreed to participate, about 26,000 were excluded because they reported a history of heart problems. 33,223 willing and eligible physicians were enrolled in a "run-in" phase in which all were given low-dose aspirin. After 18 weeks, they were sent a questionnaire asking about their health status, any side effects of the aspirin, compliance with study protocols, and their willingness to continue in the trial. About 11,000 changed their mind, reported a reason for exclusion or did not reliably take their aspirin. The remaining 22,071 physicians were then randomly assigned to receive either an aspirin or an aspirin placebo. Follow-up questionnaires were sent after six and twelve months and then yearly. Though the study was scheduled to run until 1995, it was stopped after only six years because it was clear that aspirin had a significant effect on the risk of a first myocardial infarction.

The details of this story attest to the amount of time, effort, and expense that may be required to undertake an ambitious randomized causal study. The selection of candidates and subsequent efforts at matching alone took nearly two years. The study then took another six years to yield decisive results. And the entire project required millions of dollars in funding from four National Institutes of Health grants.

Time and expense are not the only impediments to randomized studies. Many suspected causal links cannot be subjected to randomized testing without putting their subject at undue risk. Do high rates of cholesterol in the blood cause heart disease? Imagine what a randomized experiment might involve. We might begin with a large number of young children. Having divided them at random into two groups, we will train one group to eat and drink lots of fatty, starchy, and generally unhealthy foods of the sort we suspect may be associated with high levels of cholesterol. I'm sure you can see the problem. Not coincidentally, much medical research is carried out on laboratory animals precisely because we tend to have much less hesitation about administering potentially hazardous substances to members of nonhuman species. But there is another approach, one that allows a more humane use of human and animal subjects: work with subjects who have been exposed to suspected causes and their effects. Such studies fall into two broad categories, called prospective and retrospective studies.

QUICK REVIEW 5.1 Randomized Studies



Prospective Causal Studies. Prospective studies begin with subjects who have already been exposed to the causal agent under investigation. Over the term of the study these experimental subjects are compared to control subjects who have not been exposed to the suspected cause. One of the most well-known prospective studies began in 1976 at Harvard Medical School. 122,000 nurses, aged 30 to 55, agreed to fill out a questionnaire every two years about diseases and heath-related topics including smoking, hormone use, and menopausal status. More than 90% of the respondents still answer a questionnaire every other year, providing information about what they eat, what medicines they take, what illness they have had, and whether they drink, smoke, exercise, or take vitamins, among other things. At various points in the study, blood samples have been obtained by sending a request and a few collection supplies to the nurses, something that could never have been done with the general public. Over the years researchers have been able to identify risk factors for diseases such as breast, colon, and lung cancer, diabetes, and heart disease. Among other things, researchers have discovered that those nurses who regularly take vitamin E have significantly lower rates of heart disease and those who have undergone hormone replacement therapy have significantly higher rates of breast cancer. As the study subjects age, researchers hope to be able to determine the extent to which diet and exercise lead to a longer, healthier life.

Even the best of prospective studies run up against one major stumbling block. At the onset of any prospective study, subjects in the experimental group have already been exposed to the suspected causal factor but as we have found, most things can be caused by a variety of factors. It is always possible that other causal factors are responsible for some part of the effect in both the experimental and control groups. For example, in the study above, nurses who took vitamin E had lower rates of heart disease. But it may be that those who take vitamin E tend to take good care of themselves by exercising, watching their weight, and eating a healthy diet, all factors that contribute to heart disease. And this is the problem. By concentrating on a single causal factor in the selection process, we leave open the

possibility that whatever difference in levels of effect we observe in our two groups may be due in part to other factors. This, of course, is precisely where prospective studies differ from randomized studies. By randomly dividing subjects into experimental and control groups prior to administering the suspected cause, we greatly decrease the chances that other factors will account for differences in level of effect. In prospective studies it is always possible that other factors will come into play precisely because we begin with subjects who may have been exposed to other causal factors.

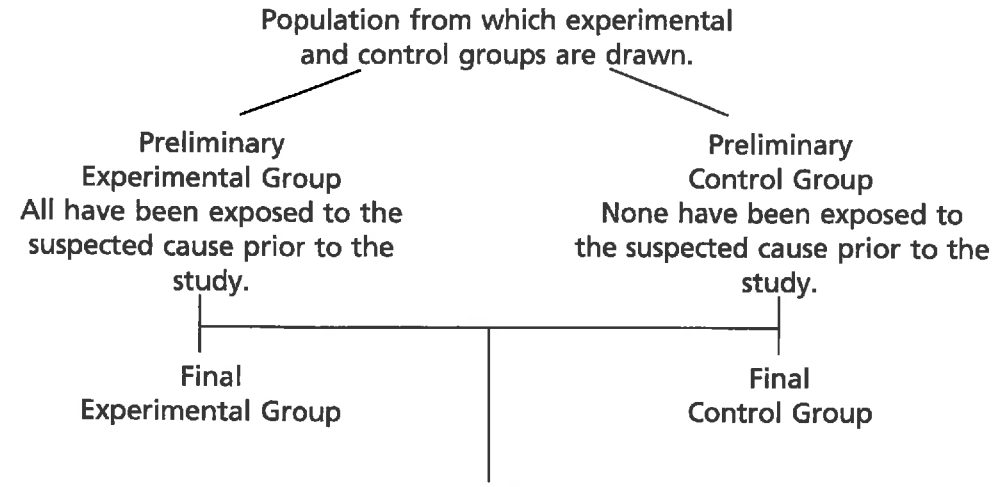
Matching can be used to control for extraneous causal factors in prospective studies. Suppose we find, in the long-term nurses study, that about 30% more experimental than control subjects exercise regularly. We can easily subtract some subjects from our experimental group or add some to the control group to achieve similar percentages of this obvious causal factor. It is not an over-simplification to say that the reliability of a prospective study is in direct proportion to the degree to which such matching is successful. Thus, in assessing the results of a prospective study, we need to know what factors have been controlled for via matching. In addition, it is always wise to be on the lookout for other factors that might influence the study's outcome yet which have not been controlled for. In general, a properly done prospective study can provide some strong indication of a causal link, though not as strong as that provided by a randomized study.

In some respects prospective studies offer advantages over randomized causal studies. For one thing, prospective studies require much less direct manipulation of experimental subjects and thus tend to be easier and less expensive to carry out. Their principle advantage, however, lies in the fact that they can involve very large groups, as in the doctors and nurses studies discussed above. Causal factors often result in differences in level of effect that are so small as to require large samples to detect reliably. Moreover, greater size alone increases the chances that the samples will be representative with respect to other causal factors. This is crucial when an effect is associated with several causal factors. If a number of factors cause B in Cs, we increase our chances of accurately representing the levels of these other factors in our two groups as we increase their size. In addition, prospective studies allow us to study potential causal links we would not want to investigate in randomized studies. For example, the nurses study discussed above uncovered a link between hormone replacement therapy and breast cancer. No researcher who suspected such a link would undertake a randomized experiment that would involve exposing women to this potential hazard. A study that involves merely tracking women who are already undergoing replacement therapy is much less objectionable.

Retrospective Causal Studies. Retrospective studies begin with two groups, our familiar experimental and control groups, but the two are composed of subjects who do and do not have the effect in question. The study then involves looking into the subjects' backgrounds in an attempt to uncover different levels of the potential causal factor in the two groups.

The following retrospective study investigated the effects of lead on children. A government health survey of 4,000 U.S. children between the ages

QUICK REVIEW 5.2 Prospective Studies

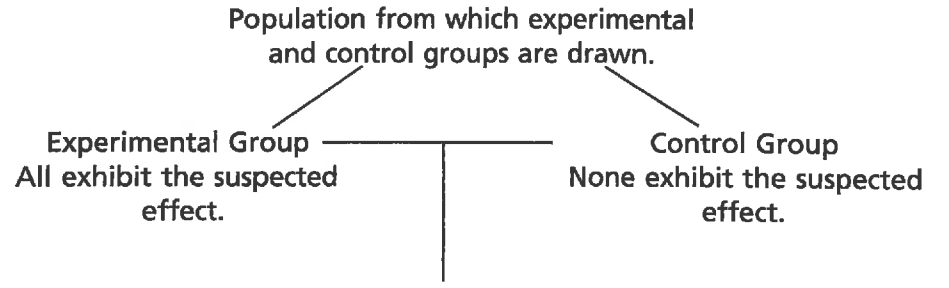


of 4 and 15, done between 1999 and 2002, included 135 children with attention deficit hyperactivity disorder (ADHA). Blood tests were done on all 4,000. Among the 135 children with ADHA, blood lead levels of more than two micrograms per deciliter occurred at a much greater rate than in the other children. The ADHA children were four times more likely to have elevated lead levels in their blood, leading researchers to conclude that exposure to lead may cause ADHA.

Even the best of retrospective studies can provide only weak evidence for a causal link. This is because, in retrospective studies, it is exceedingly difficult to control for other potential causal factors. Subjects are selected because they either do or do not have the effect in question, so other potential causal factors may automatically be built into our two groups. A kind of backwards matching is sometimes possible in retrospective studies. In the study above, if other sources of increased lead level can be identified, some attempt can be made to insure that these other factors are equally represented in the two groups. However, even if the two groups can be configured so that they exhibit similar levels of other suspected causes, we have at most very tentative evidence for the causal link in question. The process of re-configuring the two groups may leave us with a distorted picture of the extent to which the causal factor under investigation is responsible for the effect. That the two groups now appear to be alike with respect to other causal factors is, thus, largely because they are contrived to appear that way.

The particular study we are discussing suffers from an unusual weakness in the selection process for retrospective studies. Essentially, retrospective studies involve looking for something common in the background of subjects who exhibit an effect: in this case, elevated lead levels in the blood. It is at least a

QUICK REVIEW 5.3 Retrospective Studies



Sometimes modified after all data are collected to equalize the occurrence of other potential causal factors in the two groups.

possibility that cause and effect have been reversed! In the ADHA study, researchers noted that children with ADHA are more likely than others to eat old leaded paint chips or inhale leaded paint dust because of their hyperactivity. It may be that ADHA is in part responsible for the increased lead levels.

All we are in a position to conclude, as the result of a retrospective study, is that we have looked into the background of subjects who have a particular effect and we have found that a suspected cause occurs more frequently than in subjects who do not have the effect in question. Whether the effect is due to the suspected cause or whether cause and effect are reversed may be difficult to say even when pains are taken to control for other potential causal factors.

One final limitation of retrospective studies is that they provide no way of estimating the level of difference of the effect being studied. The very design of retrospective studies insures that 100% of the experimental group, but none of the control group, will have the effect. In the ADHA study, it was reported that ADHA children were "four times more likely" to have elevated lead levels in their blood. No doubt this difference is statistically significant. This tells us the difference is probably not due to chance. But it does not tell us that those children with elevated blood levels are "four times more likely" to have ADHA. Though this difference is striking, it tells us nothing about the efficacy of the suspected cause. By contrast, a prospective study—in which the experimental subjects had levels of the suspected effect four times higher than control subjects—would provide much stronger evidence of a causal link. Here, the evidence would be strengthened if other possible causes could be ruled out by matching prior to the onset of the study.

Because of these difficulties, retrospective studies are best regarded as a tool for uncovering potential causal links. The major advantage to retrospective studies, by contrast with randomized and prospective studies, is that they can be carried out quickly and inexpensively since they involve little more than careful analysis of data that is already available. And sometimes alacrity is of the essence. Imagine that there has been a rash of cases of food poisoning in your community. Before much of anything can be done, health authorities need

some sense of what might be causing the problem. Interviews with the victims reveal that most had eaten at a particular restaurant within the last few days. This bit of retrospectively gained information may provide just the clue needed to get at the cause or causes of the problem.

READING BETWEEN THE LINES

The results of causal research are reported in specialized scientific journals. Typically an article will include full information about the design of the experiment, the results, and a complete statistical analysis where appropriate. Many medical journals, like *The Journal of the American Medical Association* and *The New England Journal of Medicine*, also provide full disclosure of the funding sources of the research. Conclusions will be carefully qualified, and the article will probably contain a brief history of other relevant research. When research uncovers a result which may have an impact on the general public, it will often be reported in the popular media in newspaper, magazines, the Internet, and on television. And it is here—in the mass media—that most of us encounter the findings of causal research.

Unfortunately, the popular media tend to do a poor job of reporting the outcomes of causal research. Media reports will often leave out crucial information, no doubt in the name of brevity; A 20- or 30-page journal article usually will be covered in a few paragraphs. Such reports tend also to dispatch with the kind of careful qualifications that normally accompany the original write-up of the results. For these reasons, it is important to learn to read between the lines of popular reports if we are to make sense of the research on which they are based.

Here, for example, is the complete text of a newspaper story about an important piece of causal research:

> Lithium, which is widely prescribed for manic-depressive disorders, may be the first biologically effective drug treatment for alcoholism, according to studies at St. Luke's Medical Center. The new evidence indicates that the drug appears to have the unique ability to act on the brain to suppress an alcoholic's craving for alcohol. The St. Luke's study involved 84 patients, ranging from 20 to 60 years of age, who had abused alcohol for an average of 17 years. Eighty-eight percent were male. Half the patients were given lithium while the other half took a placebo, a chemically inactive substance. Seventy-five percent of the alcoholics who regularly took their daily lithium pills did not touch a drop of liquor for up to a year and a half during the follow-up phase of the experiment. This abstinence rate is at least 50% higher than that achieved by the best alcohol treatment centers one to five years after treatment. Among the alcoholics who did not take their lithium regularly, only 35% were still abstinent at the end of 18 months. Among those who stopped taking the drug altogether, all had resumed drinking

by the end of six months. (Researchers tested the level of lithium in the blood of the subjects to determine if they were taking the drug regularly.)

Just what are we to make of this story and the research it describes? Is lithium effective in the treatment of alcoholism? (Note that the story begins by claiming that lithium "may be" the first effective treatment for alcoholism.) In trying to make sense of an article like this one, it is necessary to try to answer a number of questions, all based on our findings in this chapter:

What is the causal hypothesis at issue?

What kind of causal experiment is undertaken?

What crucial facts and figures are missing from the report?

Given the information you have at your disposal, can you think of any major flaws in the design of the experiment?

Given the information available, what conclusion can be drawn about the causal hypothesis?

Let's consider again the news article about the lithium study, now in light of our five questions.

What is the causal hypothesis at issue? The hypothesis is that lithium suppresses the alcoholic's craving for alcohol.

What kind of causal experiment is undertaken? Randomized. Subjects are divided into experimental and control groups prior to the experiment and only the experimental subjects are exposed to the suspected causal agent.

What crucial facts and figures are missing from the report? The passage gives us no information about what happened to the members of the control group. Nor does it tell us the number of subjects from the experimental group who "regularly took their daily lithium pills." We know that 75% of these subjects did so, but this could be as few as three out of four. All we are told of the remaining members of the experimental group is that 35% remained abstinent and that some stopped taking the drug altogether. We are not told how many are in each of these subgroups. It is possible that the majority of experimental subjects did not remain abstinent. Given the information we have at our disposal, we just cannot say for sure, one way or the other. Though we are given no information about the control group, we are provided with some information against which to assess the results in the experimental group: we are told that the 75% abstinence rate is "at least 50% higher than that achieved by the best alcohol treatment centers one to five years after treatment." However, we are not told whether the success rate for treatment centers is a percentage of people who entered treatment or people who completed treatment. If the former is the case, there is a strong possibility treatment centers have a higher rate of success than that established in the experiment. Once again, we can draw no conclusions since we are not provided with the key comparative information.

Given the information you have at your disposal, can you think of any major flaws in the design of the experiment? One possible flaw comes to mind. It may be that the subjects who continued to take their medication (lithium or placebo) throughout the entire 18 months of the experiment were more strongly motivated to quit drinking than the other subjects. And this may have influenced the outcome of the experiment. Precautions need to be taken to ensure either that no subjects lacked this motivation or that they were equally represented in experimental and control groups. Here, information about the results of the control group would be helpful. If roughly equal numbers of people dropped out of both groups, we would have some initial reason to think that we had controlled for motivation.

Given the information available, what conclusion can be drawn about the causal hypothesis? We can conclude very little particularly because we are given no information about what happened to the control group. This is not to say that the experiment itself warrants no conclusion about the possible link between lithium and alcoholism. However the report about the study with which we have been working has presented us with so little information that we can draw no conclusion.

Media accounts of causal studies may make reference to a possible causal mechanism. The article about lithium and alcoholism does mention one: lithium, it is hypothesized, may act on the brain in some way to suppress an alcoholic's craving for alcohol. Though in this case the proposed mechanism is vague, a well-understood, well-established mechanism can provide an additional piece of information to suggest that a study is on to something.