1

Controlled Experiments

Always do right. This will gratify some people, and astonish the rest.
—MARK TWAIN (UNITED STATES, 1835–1910)

1. THE SALK VACCINE FIELD TRIAL

A new drug is introduced. How should an experiment be designed to test its effectiveness? The basic method is *comparison*.[1] The drug is given to subjects in a *treatment group*, but other subjects are used as *controls*—they aren't treated. Then the responses of the two groups are compared. Subjects should be assigned to treatment or control *at random*, and the experiment should be run *double-blind*: neither the subjects nor the doctors who measure the responses should know who was in the treatment group and who was in the control group. These ideas will be developed in the context of an actual field trial.[2]

The first polio epidemic hit the United States in 1916, and during the next forty years polio claimed many hundreds of thousands of victims, especially children. By the 1950s, several vaccines against this disease had been discovered. The one developed by Jonas Salk seemed the most promising. In laboratory trials, it had proved safe and had caused the production of antibodies against polio. By 1954, the Public Health Service and the National Foundation for Infantile Paralysis (NFIP) were ready to try the vaccine in the real world—outside the laboratory.

Suppose the NFIP had just given the vaccine to large numbers of children. If the incidence of polio in 1954 dropped sharply from 1953, that would seem to

prove the effectiveness of the vaccine. However, polio was an epidemic disease whose incidence varied from year to year. In 1952, there were about 60,000 cases; in 1953, there were only half as many. Low incidence in 1954 could have meant that the vaccine was effective—or that 1954 was not an epidemic year.

The only way to find out whether the vaccine worked was to deliberately leave some children unvaccinated, and use them as controls. This raises a troublesome question of medical ethics, because withholding treatment seems cruel. However, even after extensive laboratory testing, it is often unclear whether the benefits of a new drug outweigh the risks.[3] Only a well-controlled experiment can settle this question.

In fact, the NFIP ran a controlled experiment to show the vaccine was effective. The subjects were children in the age groups most vulnerable to polio—grades 1, 2, and 3. The field trial was carried out in selected school districts throughout the country, where the risk of polio was high. Two million children were involved, and half a million were vaccinated. A million were deliberately left unvaccinated, as controls; half a million refused vaccination.

This illustrates the method of comparison. Only the subjects in the treatment group were vaccinated; the controls did not get the vaccine. The responses of the two groups could then be compared to see if the treatment made any difference. In the Salk vaccine field trial, the treatment and control groups were of different sizes, but that did not matter. The investigators compared the rates at which children got polio in the two groups—cases per hundred thousand. Looking at rates instead of absolute numbers adjusts for the difference in the sizes of the groups.

Children could be vaccinated only with their parents' permission. So one possible design—which also seems to solve the ethical problem—was this: the children whose parents consent would go into the treatment group and get the vaccine; the other children would be the controls. However, it was known that higher-income parents would more likely consent to treatment than lower-income parents. This design is biased against the vaccine, because children of higher-income parents are more vulnerable to polio.

That may seem paradoxical at first, because most diseases fall more heavily on the poor. But polio is a disease of hygiene. Children who live in less hygienic surroundings tend to contract mild cases of polio early in childhood while still protected by antibodies from their mother. After being infected, they generate their own antibodies which protect them against more severe infection later. Children who live in more hygienic surroundings do not develop such antibodies.

Comparing volunteers to non-volunteers biases the experiment. The statistical lesson: the treatment and control groups should be as similar as possible—except for the treatment. Then, any difference in response between the two groups is due to the treatment rather than something else. If the two groups differ with respect to some factor other than the treatment, the effect of this other factor might be *confounded* (mixed up) with the effect of treatment. Separating these effects can be difficult, and confounding is a major source of bias.

For the Salk vaccine field trial, several designs were proposed. The NFIP had originally wanted to vaccinate all grade 2 children whose parents would consent,

leaving the children in grades 1 and 3 as controls. And this design was used in many school districts. However, polio is a contagious disease, spreading through contact. So the incidence could have been higher in grade 2 than in grades 1 or 3. This would have biased the study against the vaccine. Or the incidence could have been lower in grade 2, biasing the study in favor of the vaccine. Furthermore, children in the treatment group, where parental consent was needed, were likely to have different family backgrounds from those in the control group, where parental consent was not required. With the NFIP design, the treatment group would include too many children from higher-income families. The treatment group would be more vulnerable to polio than the control group. Here was a definite bias against the vaccine.

Many public health experts saw these flaws in the NFIP design, and suggested a different design. The control group had to be chosen from the same population as the treatment group—children whose parents consented to vaccination. Otherwise, the effect of family background would be confounded with the effect of the vaccine. The next problem was assigning the children to treatment or control. Human judgment seems necessary, to make the control group like the treatment group on the relevant variables—family income as well as the children's general health, personality, and social habits.

Experience shows, however, that human judgment often results in substantial bias: it is better to rely on impersonal chance. For the Salk vaccine, the chance procedure was equivalent to tossing a coin for each child, with a 50–50 chance of assignment to the treatment group or the control group. Such a procedure is objective and impartial. The laws of chance guarantee that with enough subjects, the treatment group and the control group will resemble each other very closely with respect to all the important variables, whether or not these have been identified. When an impartial chance procedure is used to assign the subjects to treatment or control, the experiment is said to be *randomized controlled*.[4]

Another basic precaution was the use of a *placebo*: children in the control group were given an injection of salt dissolved in water. During the experiment the subjects did not know whether they were in treatment or in control, so their response was to the vaccine, not the idea of treatment. It may seem unlikely that subjects could be protected from polio just by the strength of an idea. However, hospital patients suffering from severe post-operative pain have been given a "pain killer" which was made of a completely neutral substance: about one-third of the patients experienced prompt relief.[5]

Still another precaution: diagnosticians had to decide whether the children contracted polio during the experiment. Many forms of polio are hard to diagnose, and in borderline cases the diagnosticians could have been affected by knowing whether the child was vaccinated. So the doctors were not told which group the child belonged to. This was *double blinding*: the subjects did not know whether they got the treatment or the placebo, and neither did those who evaluated the responses. This randomized controlled double-blind experiment—which is about the best design there is—was done in many school districts.

How did it all turn out? Table 1 shows the rate of polio cases (per hundred thousand subjects) in the randomized controlled experiment, for the treatment

group and the control group. The rate is much lower for the treatment group, decisive proof of the effectiveness of the Salk vaccine.

Table 1. The results of the Salk vaccine trial of 1954. Size of groups and rate of polio cases per 100,000 in each group. The numbers are rounded.

The randomized controlled double-blind experiment			*The NFIP study*		
	Size	*Rate*		*Size*	*Rate*
Treatment	200,000	28	Grade 2 (vaccine)	225,000	25
Control	200,000	71	Grades 1 and 3 (control)	725,000	54
No consent	350,000	46	Grade 2 (no consent)	125,000	44

Source: Thomas Francis, Jr., "An evaluation of the 1954 poliomyelitis vaccine trials—summary report," *American Journal of Public Health* vol. 45 (1955) pp. 1–63.

Table 1 also shows how the NFIP study was biased against the vaccine. In the randomized controlled experiment, the vaccine cut the polio rate from 71 to 28 per hundred thousand; the reduction in the NFIP study, from 54 to 25 per hundred thousand, is quite a bit less. The main source of the bias was confounding. The NFIP treatment group included only children whose parents consented to vaccination. However, the control group also included children whose parents would not have consented. The control group was not comparable to the treatment group.

The randomized controlled double-blind design reduces bias to a minimum—the main reason for using it whenever possible. But this design also has an important technical advantage. To see why, let us play devil's advocate and assume that the Salk vaccine had no effect. Then the difference between the polio rates for the treatment and control groups is just due to chance. How likely is that?

With the NFIP design, the results are affected by many factors that seem random: which families volunteer, which children are in grade 2, and so on. However, the investigators do not have enough information to figure the chances for the outcomes. They cannot figure the odds against a big difference in polio rates being due to accidental factors. With a randomized controlled experiment, on the other hand, chance enters in a planned and simple way—when the assignment is made to treatment or control.

The devil's-advocate hypothesis says that the vaccine has no effect. On this hypothesis, a few children are fated to contract polio; assignment to treatment or control has nothing to do with it. Each child has a 50–50 chance to be in treatment or control, just depending on the toss of a coin. Each polio case has a 50–50 chance to turn up in the treatment group or the control group.

Therefore, the number of polio cases in the two groups must be about the same. Any difference is due to the chance variability in coin tossing. Statisticians understand this kind of variability. They can figure the odds against a difference as large as the observed one. The calculation will be done in chapter 27, and the odds are astronomical—a billion to one against.

2. THE PORTACAVAL SHUNT

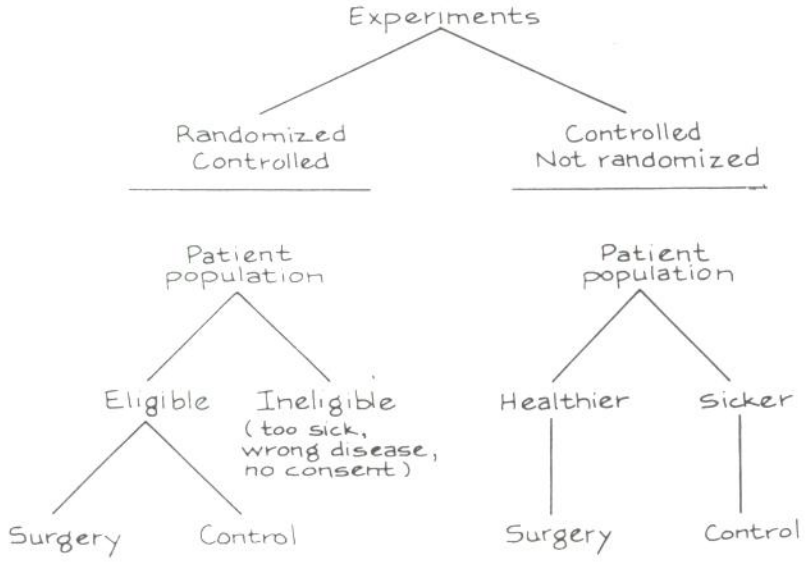
In some cases of cirrhosis of the liver, the patient may start to hemorrhage and bleed to death. One treatment involves surgery to redirect the flow of blood through a *portacaval shunt*. The operation to create the shunt is long and hazardous. Do the benefits outweigh the risks? Over 50 studies have been done to assess the effect of this surgery.[6] Results are summarized in table 2 below.

Table 2. A study of 51 studies on the portacaval shunt. The well-designed studies show the surgery to have little or no value. The poorly-designed studies exaggerate the value of the surgery.

	Degree of enthusiasm		
Design	*Marked*	*Moderate*	*None*
No controls	24	7	1
Controls, but not randomized	10	3	2
Randomized controlled	0	1	3

Source: N. D. Grace, H. Muench, and T. C. Chalmers, "The present status of shunts for portal hypertension in cirrhosis," *Gastroenterology* vol. 50 (1966) pp. 684–91.

There were 32 studies without controls (first line in the table): 24/32 of these studies, or 75%, were markedly enthusiastic about the shunt, concluding that the benefits definitely outweighed the risks. In 15 studies there were controls, but assignment to treatment or control was not randomized. Only 10/15, or 67%, were markedly enthusiastic about the shunt. But the 4 studies that were randomized controlled showed the surgery to be of little or no value. The badly designed studies exaggerated the value of this risky surgery.



A randomized controlled experiment begins with a well-defined patient population. Some are eligible for the trial. Others are ineligible: they may be too sick to undergo the treatment, or they may have the wrong kind of disease, or they may not consent to participate (see the flow chart). Eligibility is determined first;

then the eligible patients are randomized to treatment or control. That way, the comparison is made only among patients who could have received the therapy. The bottom line: the control group is like the treatment group. By contrast, with poorly-controlled studies, ineligible patients may be used as controls. Moreover, even if controls are selected among those eligible for surgery, the surgeon may choose to operate only on the healthier patients while sicker patients are put in the control group.

This sort of bias seems to have been at work in the poorly-controlled studies of the portacaval shunt. In both the well-controlled and the poorly-controlled studies, about 60% of the surgery patients were still alive 3 years after the operation (table 3). In the randomized controlled experiments, the percentage of controls who survived the experiment by 3 years was also about 60%. But only 45% of the controls in the nonrandomized experiments survived for 3 years.

In both types of studies, the surgeons seem to have used similar criteria to select patients eligible for surgery. Indeed, the survival rates for the surgery group are about the same in both kinds of studies. So, what was the crucial difference? With the randomized controlled experiments, the controls were similar in general health to the surgery patients. With the poorly controlled studies, there was a tendency to exclude sicker patients from the surgery group and use them as controls. That explains the bias in favor of surgery.

Table 3. Randomized controlled experiments vs. controlled experiments that are not randomized: three-year survival rates in studies of the portacaval shunt. (Percentages are rounded.)

	Randomized	*Not randomized*
Surgery	60%	60%
Controls	60%	45%

3. HISTORICAL CONTROLS

Randomized controlled experiments are hard to do. As a result, doctors often use other designs which are not as good. For example, a new treatment can be tried out on one group of patients, who are compared to "historical controls:" patients treated the old way in the past. The problem is that the treatment group and the historical control group may differ in important ways besides the treatment. In a controlled experiment, there is a group of patients eligible for treatment at the beginning of the study. Some of these are assigned to the treatment group, the others are used as controls: assignment to treatment or control is done "contemporaneously," that is, in the same time period. Good studies use contemporaneous controls.

The poorly-controlled trials on the portacaval shunt (section 2) included some with historical controls; others had contemporaneous controls, but assignment to the control group was not randomized. As section 2 showed, design

matters. This section continues the story. Coronary bypass surgery is a widely used—and very expensive—operation for coronary artery disease. Chalmers and associates identified 29 trials of this surgery (first line of table 4). There were 8 randomized controlled trials, and 7 were quite negative about the value of the operation. By comparison, there were 21 trials with historical controls, and 16 were positive. The badly-designed studies were more enthusiastic about the value of the surgery. (The other lines in the table can be read the same way, and lead to similar conclusions about other therapies.)

Table 4. A study of studies. Four therapies evaluated by randomized controlled trials and by trials using historical controls. Conclusions of trials are summarized as positive about the value of the therapy (+), or negative (−).

Therapy	*Randomized controlled*		*Historically controlled*	
	+	−	+	−
Coronary bypass surgery	1	7	16	5
5-FU	0	5	2	0
BCG	2	2	4	0
DES	0	3	5	0

Note: 5-FU is used in chemotherapy for colon cancer; BCG is used to treat melanoma; DES, to prevent miscarriage.
Source: H. Sacks, T. C. Chalmers, and H. Smith, "Randomized versus historical controls for clinical trials," *American Journal of Medicine* vol. 72 (1982) pp. 233–40.[7]

Why are well-designed studies less enthusiastic than poorly-designed studies? In 6 of the randomized controlled experiments on coronary bypass surgery and 9 of the studies with historical controls, 3-year survival rates for the surgery group and the control group were reported (table 5). In the randomized controlled experiments, survival was quite similar in the surgery group and the control group. That is why the investigators were not enthusiastic about the operation—it did not save lives.

Table 5. Randomized controlled experiments vs. studies with historical controls: three-year survival rates for surgery patients and controls in trials of coronary bypass surgery. Randomized controlled experiments differ from trials with historical controls.

	Randomized	*Historical*
Surgery	87.6%	90.9%
Controls	83.2%	71.1%

Note: There were 6 randomized controlled experiments enrolling 9,290 patients; and 9 studies with historical controls, enrolling 18,861 patients.
Source: See table 4.

Now look at the studies with historical controls. Survival in the surgery group is about the same as before. However, the controls have much poorer

survival rates. They were not as healthy to start with as the patients chosen for surgery. Trials with historical controls are biased in favor of surgery. Randomized trials avoid that kind of bias. That explains why the design of the study matters. Tables 2 and 3 made the point for the portacaval shunt; tables 4 and 5 make the same point for other therapies.

The last line in table 4 is worth more discussion. DES (diethylstibestrol) is an artificial hormone, used to prevent spontaneous abortion. Chalmers and associates found 8 trials evaluating DES. Three were randomized controlled, and all were negative: the drug did not help. There were 5 studies with historical controls, and all were positive. These poorly-designed studies were biased in favor of the therapy.

Doctors paid little attention to the randomized controlled experiments. Even in the late 1960s, they were giving the drug to 50,000 women each year. This was a medical tragedy, as later studies showed. If administered to the mother during pregnancy, DES can have a disastrous side-effect 20 years later, causing her daughter to develop an otherwise extremely rare form of cancer (clear-cell adenocarcinoma of the vagina). DES was banned for use on pregnant women in 1971.[8]

4. SUMMARY

1. Statisticians use the *method of comparison*. They want to know the effect of a *treatment* (like the Salk vaccine) on a *response* (like getting polio). To find

out, they compare the responses of a *treatment group* with a *control group*. Usually, it is hard to judge the effect of a treatment without comparing it to something else.

2. If the control group is comparable to the treatment group, apart from the treatment, then a difference in the responses of the two groups is likely to be due to the effect of the treatment.

3. However, if the treatment group is different from the control group with respect to other factors, the effects of these other factors are likely to be *confounded* with the effect of the treatment.

4. To make sure that the treatment group is like the control group, investigators put subjects into treatment or control at random. This is done in *randomized controlled experiments*.

5. Whenever possible, the control group is given a *placebo*, which is neutral but resembles the treatment. The response should be to the treatment itself rather than to the idea of treatment.

6. In a *double-blind* experiment, the subjects do not know whether they are in treatment or in control; neither do those who evaluate the responses. This guards against bias, either in the responses or in the evaluations.

2

Observational Studies

That's not an experiment you have there, that's an experience.
—SIR R. A. FISHER (ENGLAND, 1890–1962)

1. INTRODUCTION

Controlled experiments are different from *observational studies*. In a controlled experiment, the investigators decide who will be in the treatment group and who will be in the control group. By contrast, in an observational study it is the subjects who assign themselves to the different groups: the investigators just watch what happens.

The jargon is a little confusing, because the word *control* has two senses:

- a *control* is a subject who did not get the treatment;
- a *controlled experiment* is a study where the investigators decide who will be in the treatment group and who will not.

Studies on the effects of smoking, for instance, are necessarily observational: nobody is going to smoke for ten years just to please a statistician. However, the treatment-control idea is still used. The investigators compare smokers (the treatment or "exposed" group) with non-smokers (the control group) to determine the effect of smoking.

The smokers come off badly in this comparison. Heart attacks, lung cancer, and many other diseases are more common among smokers than non-smokers. So there is a strong *association* between smoking and disease. If cigarettes

cause disease, that explains the association: death rates are higher for smokers because cigarettes kill. Thus, association is circumstantial evidence for causation. However, the proof is incomplete. There may be some hidden confounding factor which makes people smoke and also makes them get sick. If so, there is no point in quitting; that will not change the hidden factor. Association is not the same as causation.

Statisticians like Joseph Berkson and Sir R. A. Fisher did not believe the evidence against cigarettes, and suggested possible confounding variables. Epidemiologists (including Sir Richard Doll in England, and E. C. Hammond, D. Horn, H. A. Kahn in the United States) ran careful observational studies to show these alternative explanations were not plausible. Taken together, the studies make a powerful case that smoking causes heart attacks, lung cancer, and other diseases. If you give up smoking, you will live longer.[1]

Observational studies are a powerful tool, as the smoking example shows. But they can also be quite misleading. To see if confounding is a problem, it may help to find out how the controls were selected. The main issue: was the control group really similar to the treatment group—apart from the exposure of interest? If there is confounding, something has to be done about it; although perfection cannot be expected. Statisticians talk about *controlling for* confounding factors in an observational study. This is a third use of the word *control*.

One technique is to make comparisons separately for smaller and more homogeneous groups. For example, a crude comparison of death rates among smokers and non-smokers could be misleading, because smokers are disproportionately male and men are more likely than women to have heart disease anyway. The difference between smokers and non-smokers might be due to the sex difference. To eliminate that possibility, epidemiologists compare male smokers to male non-smokers, and females to females.

Age is another confounding variable. Older people have different smoking habits, and are more at risk for lung cancer. So the comparison between smokers and non-smokers is done separately by age as well as by sex. For example, male smokers age 55–59 are compared to male non-smokers age 55–59. This controls for age and sex. Good observational studies control for confounding variables. In the end, however, most observational studies are less successful than the ones on smoking. The studies may be designed by experts, but experts make mistakes too. Finding the weak points is more an art than a science, and often depends on information outside the study.

2. THE CLOFIBRATE TRIAL

The Coronary Drug Project was a randomized, controlled double-blind experiment, whose objective was to evaluate five drugs for the prevention of heart attacks. The subjects were middle-aged men with heart trouble. Of the 8,341 subjects, 5,552 were assigned at random to the drug groups and 2,789 to the control group. The drugs and the placebo (lactose) were administered in identical capsules. The patients were followed for 5 years.

One of the drugs on test was clofibrate, which reduces the levels of cholesterol in the blood. Unfortunately, this treatment did not save any lives. About 20% of the clofibrate group died over the period of followup, compared to 21% of the control group. A possible reason for this failure was suggested—many subjects in the clofibrate group did not take their medicine.

Subjects who took more than 80% of their prescribed medicine (or placebo) were called "adherers" to the protocol. For the clofibrate group, the 5-year mortality rate among the adherers was only 15%, compared to 25% among the non-adherers (table 1). This looks like strong evidence for the effectiveness of the drug. However, caution is in order. This particular comparison is observational not experimental—even though the data were collected while an experiment was going on. After all, the investigators did not decide who would adhere to protocol and who would not. The subjects decided.

Table 1. The clofibrate trial. Numbers of subjects, and percentages who died during 5 years of followup. Adherers take 80% or more of prescription.

	Clofibrate		*Placebo*	
	Number	*Deaths*	*Number*	*Deaths*
Adherers	708	15%	1,813	15%
Non-adherers	357	25%	882	28%
Total group	1,103	20%	2,789	21%

Note: Data on adherence missing for 38 subjects in the clofibrate group and 94 in the placebo group. Deaths from all causes.
Source: The Coronary Drug Project Research Group, "Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project," *New England Journal of Medicine* vol. 303 (1980) pp. 1038–41.

Maybe adherers were different from non-adherers in other ways, besides the amount of the drug they took. To find out, the investigators compared adherers and non-adherers in the control group. Remember, the experiment was double-blind: the controls did not know whether they were taking an active drug or the placebo; neither did the subjects in the clofibrate group. The psychological basis for adherence was the same in both groups.

In the control group too, the adherers did better—only 15% of them died during the 5-year period, compared to 28% among the non-adherers. The conclusions:

(i) Clofibrate does not have an effect.
(ii) Adherers are different from non-adherers.

Probably, adherers are more concerned with their health and take better care of themselves in general. That would explain why they took their capsules and why they lived longer. Observational comparisons can be quite misleading. The investigators in the clofibrate trial were unusually careful, and they found out what was wrong with comparing adherers to non-adherers.[2]

"TO ADHERE OR NOT, TO ADHERE, THAT IS THE QUESTION."

3. MORE EXAMPLES

Example 1. "*Pellagra* was first observed in Europe in the eighteenth century by a Spanish physician, Gaspar Casal, who found that it was an important cause of ill-health, disability, and premature death among the very poor inhabitants of the Asturias. In the ensuing years, numerous... authors described the same condition in northern Italian peasants, particularly those from the plain of Lombardy. By the beginning of the nineteenth century, pellagra had spread across Europe, like a belt, causing the progressive physical and mental deterioration of thousands of people in southwestern France, in Austria, in Rumania, and in the domains of the Turkish Empire. Outside Europe, pellagra was recognized in Egypt and South Africa, and by the first decade of the twentieth century it was rampant in the United States, especially in the south...."[3]

Pellagra seemed to hit some villages much more than others. Even within affected villages, many households were spared; but some had pellagra cases year after year. Sanitary conditions in diseased households were primitive; flies were everywhere. One blood-sucking fly (*Simulium*) had the same geographical range as pellagra, at least in Europe; and the fly was most active in the spring, just when most pellagra cases developed. Many epidemiologists concluded the disease was infectious, and—like malaria, yellow fever, or typhus—was transmitted from one person to another by insects. Was this conclusion justified?

Discussion. Starting around 1914, the American epidemiologist Joseph Goldberger showed by a series of observational studies and experiments that pellagra is caused by a bad diet, and is not infectious. The disease can be prevented or cured by foods rich in what Goldberger called the P-P (pellagra-preventive) factor. Since 1940, most of the flour sold in the United States is enriched with the P-P factor, among other vitamins; the P-P factor is called "niacin" on the label.

Niacin occurs naturally in meat, milk, eggs, some vegetables, and certain grains. Corn, however, contains relatively little niacin. In the pellagra areas, the poor ate corn—and not much else. Some villages and some households were poorer than others, and had even more restricted diets. That is why they were harder hit by the disease. The flies were a marker of poverty, not a cause of pellagra. Association is not the same as causation.

Example 2. Cervical cancer and circumcision. Cervical cancer was for many years one of the most common cancers among women. Many epidemiologists worked on identifying the causes of this disease. They found that in several different countries, cervical cancer was quite rare among Jews. They also found the disease to be rare among Moslems. In the 1950s, several investigators concluded that circumcision of the males was the protective factor. Were they justified?

Discussion. There are differences between Jews or Moslems and members of other communities, besides circumcision. It turns out that cervical cancer is a sexually transmitted disease, spread by contact. Current research suggests that certain strains of HPV (human papilloma virus) are the causal agents. Some women are more active sexually than others, and have more partners; they are more likely to be exposed to the viruses causing the disease. That seems to be what makes the rate of cervical cancer higher for some groups of women. Early studies did not pay attention to this confounding variable, and reached the wrong conclusions.[4] (Cancer takes a long time to develop; sexual behavior in the 1930s or 1940s was the issue.)

Example 3. Ultrasound and low birthweight. Human babies can now be examined in the womb using ultrasound. Several experiments on lab animals have shown that ultrasound examinations can cause low birthweight. If this is true for humans, there are grounds for concern. Investigators ran an observational study to find out, at the Johns Hopkins hospital in Baltimore.

Of course, babies exposed to ultrasound differed from unexposed babies in many ways besides exposure; this was an observational study. The investigators found a number of confounding variables and adjusted for them. Even so, there was an association. Babies exposed to ultrasound in the womb had lower birthweight, on average, than babies who were not exposed. Is this evidence that ultrasound causes lower birthweight?

Discussion. Obstetricians suggest ultrasound examinations when something seems to be wrong. The investigators concluded that the ultrasound exams and low birthweights had a common cause—problem pregnancies. Later,

a randomized controlled experiment was done to get more definite evidence. If anything, ultrasound was protective.[5]

Example 4. The Samaritans and suicide. Over the period 1964–70, the suicide rate in England fell by about one-third. During this period, a volunteer welfare organization called "The Samaritans" was expanding rapidly. One investigator thought that the Samaritans were responsible for the decline in suicides. He did an observational study to prove it. This study was based on 15 pairs of towns. To control for confounding, the towns in a pair were matched on the variables regarded as important. One town in each pair had a branch of the Samaritans; the other did not. On the whole, the towns with the Samaritans had lower suicide rates. So the Samaritans prevented suicides. Or did they?

Discussion. A second investigator replicated the study, with a bigger sample and more careful matching. He found no effect. Furthermore, the suicide rate was stable in the 1970s (after the first investigator had published his paper) although the Samaritans continued to expand. The decline in suicide rates in the 1960s is better explained by a shift from coal gas to natural gas for heating and cooking. Natural gas is less toxic. In fact, about one-third of suicides in the early 1960s were by gas. At the end of the decade, there were practically no such cases, explaining the decline in suicides. The switch to natural gas was complete, so the suicide rate by gas couldn't decline much further. Finally, the suicide rate by methods other than gas was nearly constant over the 1960s—despite the Samaritans. The Samaritans were a good organization, but they do not seem to have had much effect on the suicide rate. And observational studies, no matter how carefully done, are not experiments.[6]

4. SEX BIAS IN GRADUATE ADMISSIONS

To review briefly, one source of trouble in observational studies is that subjects differ among themselves in crucial ways besides the treatment. Sometimes these differences can be adjusted for, by comparing smaller and more homogeneous subgroups. Statisticians call this technique *controlling for* the confounding factor—the third sense of the word *control*.

An observational study on sex bias in admissions was done by the Graduate Division at the University of California, Berkeley.[7] During the study period, there were 8,442 men who applied for admission to graduate school and 4,321 women. About 44% of the men and 35% of the women were admitted. Taking percents adjusts for the difference in numbers of male and female applicants: 44 out of every 100 men were admitted, and 35 out of every 100 women.

Assuming that the men and women were on the whole equally well qualified (and there is no evidence to the contrary), the difference in admission rates looks like a strong piece of evidence to show that men and women are treated differently in the admissions procedure. The university seems to prefer men, 44 to 35.

Each major did its own admissions to graduate work. By looking at them separately, the university should have been able to identify the ones which discriminated against the women. At that point, a puzzle appeared. Major by major,

"YES, ON THE SURFACE IT WOULD APPEAR TO BE SEX-BIAS BUT LET US ASK THE FOLLOWING QUESTIONS..."

there did not seem to be any bias against women. Some majors favored men, but others favored women. On the whole, if there was any bias, it ran against the men. What was going on?

Over a hundred majors were involved. However, the six largest majors together accounted for over one-third of the total number of applicants to the campus. And the pattern for these majors was typical of the whole campus. Table 2 shows the number of male and female applicants, and the percentage admitted, for each of these majors.

Table 2. Admissions data for the graduate programs in the six largest majors at University of California, Berkeley.

	Men		*Women*	
Major	*Number of applicants*	*Percent admitted*	*Number of applicants*	*Percent admitted*
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Note: University policy does not allow these majors to be identified by name.
Source: The Graduate Division, University of California, Berkeley.

In each major, the percentage of female applicants who were admitted is roughly equal to the percentage for male applicants. The only exception is major A, which appears to discriminate against men; it admitted 82% of the women but only 62% of the men. The department that looks most biased against women is E. It admitted 28% of the men and 24% of the women. This difference only amounts to 4 percentage points. However, when all six majors are taken together, they admitted 44% of the male applicants, and only 30% of the females—the difference is 14 percentage points.

This seems paradoxical, but here is the explanation:

- The first two majors were easy to get into. Over 50% of the men applied to these two majors.
- The other four majors were much harder to get into. Over 90% of the women applied to these four majors.

The men were applying to the easy majors, the women to the harder ones. There was an effect due to the choice of major, confounded with the effect due to sex. When the choice of major is controlled for, as in table 2, there is little difference in the admissions rates for men or women. The statistical lesson: relationships between percentages in subgroups (for instance, admissions rates for men and women in each department separately) can be reversed when the subgroups are combined. This is called *Simpson's paradox*.[8]

Technical note. Table 2 is hard to read because it compares twelve admissions rates. A statistician might summarize table 2 by computing one overall admissions rate for men and another for women, but adjusting for the sex difference in application rates. The procedure would be to take some kind of average admission rate separately for the men and women. An ordinary average ignores the differences in size among the departments. Instead, a *weighted average* of the admission rates could be used, the weights being the total number of applicants (male and female) to each department: see table 3.

Table 3. Total number of applicants, from table 2.

Major	*Total number of applicants*
A	933
B	585
C	918
D	792
E	584
F	714
	4,526

The weighted average admission rate for men is

$$\frac{.62\times 933 + .63\times 585 + .37\times 918 + .33\times 792 + .28\times 584 + .06\times 714}{4{,}526}$$

This works out to 39%. Similarly, the weighted average admission rate for the women is

$$\frac{.82\times933 + .68\times585 + .34\times918 + .35\times792 + .24\times584 + .07\times714}{4{,}526}$$

This works out to 43%. In these formulas, the weights are the same for the men and women; they are the totals from table 3. The admission rates are different for men and women; they are the rates from table 2. The final comparison: the weighted average admission rate for men is 39%, while the weighted average admission rate for women is 43%. The weighted averages control for the confounding factor—choice of major. These averages suggest that if anything, the admissions process is biased against the men.

5. CONFOUNDING

Hidden confounders are a major problem in observational studies. As discussed in section 1, epidemiologists found an association between exposure (smoking) and disease (lung cancer): heavy smokers get lung cancer at higher rates than light smokers; light smokers get the disease at higher rates than nonsmokers. According to the epidemiologists, the association comes about because smoking causes lung cancer. However, some statisticians—including Sir R. A. Fisher—thought the association could be explained by confounding.

Confounders have to be associated with (i) the disease and (ii) the exposure. For example, suppose there is a gene which increases the risk of lung cancer. Now, if the gene also gets people to smoke, it meets both the tests for a confounder. This gene would create an association between smoking and lung cancer. The idea is a bit subtle: a gene that causes cancer but is unrelated to smoking is not a confounder and is sideways to the argument, because it does not account for the facts—the association between smoking and cancer.[9] Fisher's "constitutional hypothesis" explained the association on the basis of genetic confounding; nowadays, there is evidence from twin studies to refute this hypothesis (review exercise 11, chapter 15).

> Confounding means a difference between the treatment and control groups—other than the treatment—which affects the responses being studied. A confounder is a third variable, associated with exposure and with disease.

Exercise Set A

1. In the U.S. in 1990, there were 2.1 million deaths from all causes, compared to 1.7 million in 1960—nearly a 25% increase.[10] True or false, and explain: the data show that the public's health got worse over the period 1960–1990.

2. Data from the Salk vaccine field trial suggest that in 1954, the school districts in the NFIP trial and in the randomized controlled experiment had similar exposures to the polio virus.
   (a) The data also show that children in the two vaccine groups (for the randomized controlled experiment and the NFIP design) came from families with similar incomes and educational backgrounds. Which two numbers in table 1 (p. 6) confirm this finding?
   (b) The data show that children in the two no-consent groups had similar family backgrounds. Which pair of numbers in the table confirm this finding?
   (c) The data show that children in the two control groups had different family backgrounds. Which pair of numbers in the table confirm this finding?
   (d) In the NFIP study, neither the control group nor the no-consent group got the vaccine. Yet the no-consent group had a lower rate of polio. Why?
   (e) To show that the vaccine works, someone wants to compare the 44/100,000 in the NFIP study with the 25/100,000 in the vaccine group. What's wrong with this idea?

3. Polio is an infectious disease; for example, it seemed to spread when children went swimming together. The NFIP study was not done blind: could that bias the results? Discuss briefly.

4. The Salk vaccine field trials were conducted only in certain experimental areas (school districts), selected by the Public Health Service in consultation with local officials.[11] In these areas, there were about 3 million children in grades 1, 2, or 3; and there were about 11 million children in those grades in the United States. In the experimental areas, the incidence of polio was about 25% higher than in the rest of the country. Did the Salk vaccine field trials cause children to get polio instead of preventing it? Answer yes or no, and explain briefly.

5. Linus Pauling thought that vitamin C prevents colds, and cures them too. Thomas Chalmers and associates did a randomized controlled double-blind experiment to find out.[12] The subjects were 311 volunteers at the National Institutes of Health. These subjects were assigned at random to 1 of 4 groups:

Group	*Prevention*	*Therapy*
1	placebo	placebo
2	vitamin C	placebo
3	placebo	vitamin C
4	vitamin C	vitamin C

All subjects were given six capsules a day for prevention, and an additional six capsules a day for therapy if they came down with a cold. However, in group 1 both sets of capsules just contained the placebo (lactose). In group 2, the prevention capsules had vitamin C while the therapy capsules were filled with the placebo. Group 3 was the reverse. And in group 4, all the capsules were filled with vitamin C.

There was quite a high dropout rate during the trial. And this rate was significantly higher in the first 3 groups than in the 4th. The investigators noticed

this, and found the reason. As it turned out, many of the subjects broke the blind. (That is quite easy to do; you just open a capsule and taste the contents; vitamin C—ascorbic acid—is sour, lactose is not.) Subjects who were getting the placebo were more likely to drop out.

The investigators analyzed the data for the subjects who remained blinded, and vitamin C had no effect. Among those who broke the blind, groups 2 and 4 had the fewest colds; groups 3 and 4 had the shortest colds. How do you interpret these results?

6. (Hypothetical.) One of the other drugs in the Coronary Drug Project (section 2) was nicotinic acid.[13] Suppose the results on nicotinic acid were as reported below. Something looks wrong. What, and why?

	Nicotinic acid		*Placebo*	
	Number	*Deaths*	*Number*	*Deaths*
Adherers	558	13%	1,813	15%
Non-adherers	487	26%	882	28%
Total group	1,045	19%	2,695	19%

7. (Hypothetical.) In a clinical trial, data collection usually starts at "baseline," when the subjects are recruited into the trial but before they are assigned to treatment or control. Data collection continues until the end of followup. Two clinical trials on prevention of heart attacks report baseline data on smoking, shown below. In one of these trials, the randomization did not work. Which one, and why?

		Number of persons	*Percent who smoked*
(i)	Teatment	1,012	49.3%
	Control	997	69.0%
(ii)	Teatment	995	59.3%
	Control	1,017	59.0%

8. Some studies find an association between liver cancer and smoking. However, alcohol consumption is a confounding variable. This means—

   (i) Alcohol causes liver cancer.
   (ii) Drinking is associated with smoking, and alcohol causes liver cancer.

   Choose one option, and explain briefly.

9. Breast cancer is one of the most common malignancies among women in the U.S. If it is detected early enough—before the cancer spreads—chances of successful treatment are much better. Do screening programs speed up detection by enough to matter?

   The first large-scale trial was run by the Health Insurance Plan of Greater New York, starting in 1963. The subjects (all members of the plan) were 62,000 women age 40 to 64. These women were divided at random into two equal groups. In the treatment group, women were encouraged to come in for annual screening, including examination by a doctor and X-rays. About 20,200 women in the treatment group did come in for the screening; but 10,800 refused. The control group was offered usual health care. All the women were followed for

many years. Results for the first 5 years are shown in the table below.[14] ("HIP" is the usual abbreviation for the Health Insurance Plan.)

Deaths in the first five years of the HIP screening trial, by cause. Rates per 1,000 women.

		Cause of Death			
		Breast cancer		*All other*	
		Number	*Rate*	*Number*	*Rate*
Treatment group					
Examined	20,200	23	1.1	428	21
Refused	10,800	16	1.5	409	38
Total	31,000	39	1.3	837	27
Control group	31,000	63	2.0	879	28

Epidemiologists who worked on the study found that (i) screening had little impact on diseases other than breast cancer; (ii) poorer women were less likely to accept screening than richer ones; and (iii) most diseases fall more heavily on the poor than the rich.

(a) Does screening save lives? Which numbers in the table prove your point?
(b) Why is the death rate from all other causes in the whole treatment group ("examined" and "refused" combined) about the same as the rate in the control group?
(c) Why is the death rate from all other causes higher for the "refused" group than the "examined" group?
(d) Breast cancer (like polio, but unlike most other diseases) affects the rich more than the poor. Which numbers in the table confirm this association between breast cancer and income?
(e) The death rate (from all causes) among women who accepted screening is about half the death rate among women who refused. Did screening cut the death rate in half? If not, what explains the difference in death rates?

10. (This continues exercise 9.)
    (a) To show that screening reduces the risk from breast cancer, someone wants to compare 1.1 and 1.5. Is this a good comparison? Is it biased against screening? For screening?
    (b) Someone claims that encouraging women to come in for breast cancer screening increases their health consciousness, so these women take better care of themselves and live longer for that reason. Is the table consistent or inconsistent with the claim?
    (c) In the first year of the HIP trial, 67 breast cancers were detected in the "examined" group, 12 in the "refused" group, and 58 in the control group. True or false, and explain briefly: screening causes breast cancer.

11. Cervical cancer is more common among women who have been exposed to the herpes virus, according to many observational studies.[15] Is it fair to conclude that the virus causes cervical cancer?

12. Physical exercise is considered to increase the risk of spontaneous abortion. Furthermore, women who have had a spontaneous abortion are more likely to have another. One observational study finds that women who exercise regularly have fewer spontaneous abortions than other women.[16] Can you explain the findings of this study?

13. A hypothetical university has two departments, A and B. There are 2,000 male applicants, of whom half apply to each department. There are 1,100 female applicants: 100 apply to department A and 1,000 to department B. Department A admits 60% of the men who apply and 60% of the women. Department B admits 30% of the men who apply and 30% of the women. "For each department, the percentage of men admitted equals the percentage of women admitted; this must be so for both departments together." True or false, and explain briefly.

Exercises 14 and 15 are designed as warm-ups for the next chapter. Do not use a calculator when working them. Just remember that "%" means "per hundred." For example, 41 people out of 398 is just about 10%. The reason: 41 out of 398 is like 40 out of 400, that's 10 out of 100, and that's 10%.

14. Say whether each of the following is about 1%, 10%, 25%, or 50%—
    (a) 39 out of 398 (b) 99 out of 407
    (c) 57 out of 209 (d) 99 out of 197

15. Among beginning statistics students in one university, 46 students out of 446 reported family incomes ranging from $40,000 to $50,000 a year.
    (a) About what percentage had family incomes in the range $40,000 to $50,000 a year?
    (b) Guess the percentage that had family incomes in the range $45,000 to $46,000 a year.
    (c) Guess the percentage that had family incomes in the range $46,000 to $47,000 a year.
    (d) Guess the percentage that had family incomes in the range $47,000 to $49,000 a year.

The answers to these exercises are on pp. A43–45.

6. REVIEW EXERCISES

Review exercises may cover material from previous chapters.

1. In 1990, four passengers were killed by crashes on commuter airlines, compared to 39 killed on scheduled carriers (like United, TWA, and so forth). True or false, and explain: the data show that if you have to fly, it is safer to do so on a commuter airline.[17]

2. The National Highway and Traffic Safety Administration analyzed thefts of new cars in 1992, compared to production figures for that year.[18]
    (a) In the Chevrolet line, 134 Corvettes were stolen out of 18,938 produced; 300 Berettas were stolen out of 47,598 produced. True or false and explain: since 300 is bigger than 134, the data show that thieves prefer Berettas.
    (b) Nissan produced 7,000 Z-cars, and 133,000 Sentras. The theft rate for the Z-cars was 9 per 1,000, while the theft rate for the Sentras was 7 per 1,000. True or false and explain: the theft rate for Sentras was lower than the rate for Z-cars because so many more Sentras were produced.

3. From table 1 in chapter 1 (p. 6), those children whose parents refused to participate in the randomized controlled Salk trial got polio at the rate of 46 per 100,000. On the other hand, those children whose parents consented to participation got polio at the slightly higher rate of 49 per 100,000 in the treatment group and control group taken together. Suppose that this field trial was repeated the following year. On the basis of the figures, some parents refused to allow their children to participate in the experiment and be exposed to this higher risk of polio. Were they right? Answer yes or no, and explain briefly.

4. The Public Health Service studied the effects of smoking on health, in a large sample of representative households.[19] For men and for women in each age group, those who had never smoked were on average somewhat healthier than the current smokers, but the current smokers were on average much healthier than those who had recently stopped smoking.
   (a) Why did they study men and women and the different age groups separately?
   (b) The lesson seems to be that you shouldn't start smoking, but once you've started, don't stop. Comment briefly.

5. There is a rare neurological disease (idiopathic hypoguesia) that makes food taste bad. It is sometimes treated with zinc sulfate. One group of investigators did two randomized controlled experiments to test this treatment. In the first trial, the subjects did not know whether they were being given the zinc sulfate or a placebo. However, the doctors doing the evaluations did know. In this trial, patients on zinc sulfate improved significantly; the placebo group showed little improvement. The second trial was run double-blind: neither the subjects nor the doctors doing the evaluation were told who had been given the drug or the placebo. In the second trial, zinc sulfate had no effect.[20] Should zinc sulfate be given to treat the disease? Answer yes or no, and explain briefly.

6. (Continue the previous exercise.) The second trial used what is called a "crossover" design. The subjects were assigned at random to one of four groups:

placebo	placebo
placebo	zinc
zinc	placebo
zinc	zinc

In the first group, the subjects stayed on the placebo through the whole experiment. In the second group, subjects began with the placebo, but halfway through the experiment they were switched to zinc sulfate. Similarly, in the third group, subjects began on zinc sulfate but were switched to placebo. In the last group, they stayed on zinc sulfate. Subjects knew the design of the study, but were not told the group to which they were assigned.

Some subjects did not improve during the first half of the experiment. In each of the four groups, these subjects showed some improvement (on average) during the second half of the experiment. How can this be explained?

7. According to a study done at Kaiser Permanente in Walnut Creek, California, users of oral contraceptives have a higher rate of cervical cancer than non-users, even after adjusting for age, education, and marital status. Investigators concluded that the pill causes cervical cancer.[21]
   (a) Is this a controlled experiment or an observational study?
   (b) Why did the investigators adjust for age? education? marital status?
   (c) Women using the pill were likely to differ from non-users on another factor which affects the risk of cervical cancer. What factor is that?
   (d) Were the conclusions of the study justified by the data? Answer yes or no, and explain briefly.

8. Ads for ADT Security Systems claim[22]

   When you go on vacation, burglars go to work.... According to FBI statistics, over 25% of home burglaries occur between Memorial Day and Labor Day.

   Do the statistics prove that burglars go to work when other people go on vacation? Answer yes or no, and explain briefly.

9. People who get lots of vitamins by eating five or more servings of fresh fruit and vegetables each day (especially "cruciferous" vegetables like broccoli) have much lower death rates from colon cancer and lung cancer, according to many observational studies. These studies were so encouraging that two randomized controlled experiments were done: treatment groups were given large doses of vitamin supplements, while people in the control groups just ate their usual diet. One experiment looked at colon cancer; the other, at lung cancer.

   The first experiment found no difference in the death rate from colon cancer between the treatment group and the control group. The second experiment found that beta carotene (as a diet supplement) increased the death rate from lung cancer.[23] True or false, and explain:
   (a) The experiments confirmed the results of the observational studies.
   (b) The observational studies could easily have reached the wrong conclusions, due to confounding—people who eat lots of fruit and vegetables have lifestyles that are different in many other ways too.
   (c) The experiments could easily have reached the wrong conclusions, due to confounding—people who eat lots of fruit and vegetables have lifestyles that are different in many other ways too.

10. A study of young children found that those with more body fat tended to have more "controlling" mothers; the *San Francisco Chronicle* (November 9, 1994) concluded that "Parents of Fat Kids Should Lighten Up."[24]
    (a) Was this an observational study or a randomized controlled experiment?
    (b) Did the study find an association between mother's behavior and her child's level of body fat?
    (c) If controlling behavior by the mother causes children to eat more, would that explain an association between controlling behavior by the mother and her child's level of body fat?

(d) Suppose there is a gene which causes obesity. Would that explain the association?
(e) Can you think of another way to explain the association?
(f) Do the data support the *Chronicle*'s advice on child-rearing?

Discuss briefly.

11. California is evaluating a new program to rehabilitate prisoners before their release; the object is to reduce the recidivism rate—the percentage who will be back in prison within two years of release. The program involves several months of "boot camp"—military-style basic training with very strict discipline. Admission to the program is voluntary. According to a prison spokesman, "Those who complete boot camp are less likely to return to prison than other inmates."[25]
    (a) What is the treatment group in the prison spokesman's comparison? the control group?
    (b) Is the prison spokesman's comparison based on an observational study or a randomized controlled experiment?
    (c) True or false: the data show that boot camp worked.

    Explain your answers.

12. (Hypothetical.) A study is carried out to determine the effect of party affiliation on voting behavior in a certain city. The city is divided up into wards. In each ward, the percentage of registered Democrats who vote is higher than the percentage of registered Republicans who vote. True or false: for the city as a whole, the percentage of registered Democrats who vote must be higher than the percentage of registered Republicans who vote. If true, why? If false, give an example.

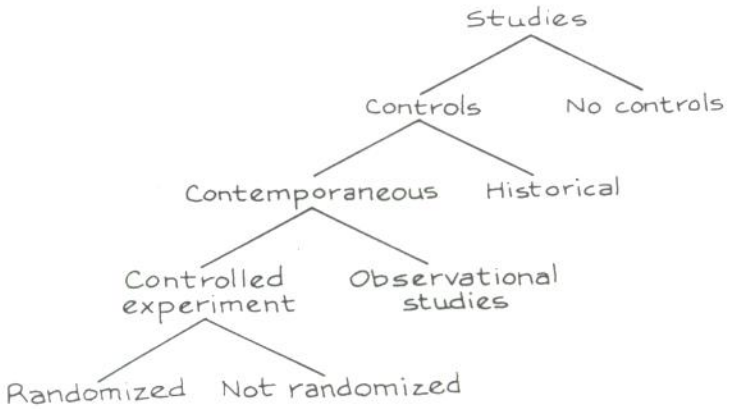
7. SUMMARY AND OVERVIEW

1. In an *observational study*, the investigators do not assign the subjects to treatment or control. Some of the subjects have the condition whose effects are being studied; this is the treatment group. The other subjects are the controls. For example, in a study on smoking, the smokers form the treatment group and the non-smokers are the controls.

2. Observational studies can establish *association*: one thing is linked to another. Association may point to causation: if exposure causes disease, then people who are exposed should be sicker than similar people who are not exposed. But association does not prove causation.

3. In an observational study, the effects of treatment may be confounded with the effects of factors that got the subjects into treatment or control in the first place. Observational studies can be quite misleading about cause-and-effect relationships, because of confounding. A *confounder* is a third variable, associated with exposure and with disease.

4. When looking at a study, ask the following questions. Was there any control group at all? Were historical controls used, or contemporaneous controls? How were subjects assigned to treatment—through a process under the control of the investigator (a controlled experiment), or a process outside the control of the investigator (an observational study)? If a controlled experiment, was the assignment made using a chance mechanism (randomized controlled), or did assignment depend on the judgment of the investigator?



5. With observational studies, and with nonrandomized controlled experiments, try to find out how the subjects came to be in treatment or in control. Are the groups comparable? different? What factors are confounded with treatment? What adjustments were made to take care of confounding? Were they sensible?

6. In an observational study, a confounding factor can sometimes be *controlled for*, by comparing smaller groups which are relatively homogeneous with respect to the factor.

7. Study design is a central issue in applied statistics. Chapter 1 introduced the idea of randomized experiments, and chapter 2 draws the contrast with observational studies. The great weakness of observational studies is confounding; randomized experiments minimize this problem. Statistical inference from randomized experiments will be discussed in chapter 27.