1
Observational Studies

1.1 What Are Observational Studies?

William G. Cochran first presented "observational studies" as a topic defined by principles and methods of statistics. Cochran had been an author of the 1964 United States Surgeon General's Advisory Committee Report, *Smoking and Health*, which reviewed a vast literature and concluded: "Cigarette smoking is causally related to lung cancer in men; the magnitude of the effect of cigarette smoking far outweighs all other factors. The data for women, though less extensive, point in the same direction (p. 37)." Though there had been some experiments confined to laboratory animals, the direct evidence linking smoking with human health came from observational or nonexperimental studies.

In a later review, Cochran (1965) defined an observational study as an empiric investigation in which:

> ... the objective is to elucidate cause-and-effect relationships ... [in which] it is not feasible to use controlled experimentation, in the sense of being able to impose the procedures or treatments whose effects it is desired to discover, or to assign subjects at random to different procedures.

Features of this definition deserve emphasis. An observational study concerns treatments, interventions, or policies and the effects they cause, and in this respect it resembles an experiment. A study without a treatment is neither an experiment nor an observational study. Most public opinion

polls, most forecasting efforts, most studies of fairness and discrimination, and many other important empirical studies are neither experiments nor observational studies.

In an experiment, the assignment of treatments to subjects is controlled by the experimenter, who ensures that subjects receiving different treatments are comparable. In an observational study, this control is absent for one of several reasons. It may be that the treatment, perhaps cigarette smoking or radon gas, is harmful and cannot be given to human subjects for experimental purposes. Or the treatment may be controlled by a political process that, perhaps quite appropriately, will not yield control merely for an experiment, as is true of much of macroeconomic and fiscal policy. Or the treatment may be beyond the legal reach of experimental manipulation even by a government, as is true of many management decisions in a private economy. Or experimental subjects may have such strong attachments to particular treatments that they refuse to cede control to an experimenter, as is sometimes true in areas ranging from diet and exercise to bilingual education. In each case, the investigator does not control the assignment of treatments and cannot ensure that similar subjects receive different treatments.

1.2 Some Observational Studies

It is encouraging to recall cases, such as *Smoking and Health*, in which observational studies established important truths, but an understanding of the key issues in observational studies begins elsewhere. Observational data have often led competent honest scientists to false and harmful conclusions, as was the case with Vitamin C as a treatment for advanced cancer.

Vitamin C and Treatment of Advanced Cancer: An Observational Study and an Experiment Compared

In 1976, in their article in the *Proceedings of the National Academy of Sciences*, Cameron and Pauling presented observational data concerning the use of vitamin C as a treatment for advanced cancer. They gave vitamin C to 100 patients believed to be terminally ill from advanced cancer and studied subsequent survival.

For each such patient, 10 historical controls were selected of the same age and gender, the same site of primary cancer, and the same histological tumor type. This method of selecting controls is called *matched sampling*—it consists of choosing controls one at a time to be similar to individual treated subjects in terms of characteristics measured prior to treatment. Used effectively, matched sampling often creates treated and control groups that are comparable in terms of the variables used in matching, though the

groups may still differ in other ways, including ways that were not measured. Cameron and Pauling (1976, p. 3685) write: "Even though no formal process of randomization was carried out in the selection of our two groups, we believe that they come close to representing random subpopulations of the population of terminal cancer patients in the Vale of Leven Hospital." In a moment, we shall see whether this is so.

Patients receiving vitamin C were compared to controls in terms of time from "untreatability by standard therapies" to death. Cameron and Pauling found that, as a group, patients receiving vitamin C survived about four times longer than the controls. The difference was highly significant in a conventional statistical test, p-value < 0.0001, and so could not be attributed to "chance." Cameron and Pauling "conclude that there is strong evidence that treatment ... [with vitamin C] ... increases the survival time."

This study created interest in vitamin C as a treatment. In response, the Mayo Clinic (Moertel et al., 1985) conducted a careful randomized controlled experiment comparing vitamin C to placebo for patients with advanced cancer of the colon and rectum. In a *randomized experiment*, subjects are assigned to treatment or control on the basis of a chance mechanism, typically a random number generator, so it is only luck that determines who receives the treatment. They found no indication that vitamin C prolonged survival, with the placebo group surviving slightly but not significantly longer. Today, few scientists claim that vitamin C holds promise as a treatment for cancer.

What went wrong in Cameron and Pauling's observational study? Why were their findings so different from those of the randomized experiment? Could their mistake have been avoided in any way other than by conducting a true experiment?

Definite answers are not known, and in all likelihood will never be known. Evidently, the controls used in their observational study, though matched on several important variables, nonetheless differed from treated patients in some way that was important to survival.

The obvious difference between the experiment and the observational study was the random assignment of treatments. In the experiment, a single group of patients was divided into a treated and a control group using a random device. Bad luck could, in principle, make the treated and control groups differ in important ways, but it is not difficult to quantify the potential impact of bad luck and to distinguish it from an effect of the treatment. Common statistical tests and confidence intervals do precisely this. In fact, this is what it means to say that the difference could not reasonably be due to "chance." Chapter 2 discusses the link between statistical inference and random assignment of treatments.

In the observational study, subjects were not assigned to treatment or control by a random device created by an experimenter. The matched sampling ensured that the two groups were comparable in a few important ways,

but beyond this, there was little to ensure comparability. If the groups were not comparable before treatment, if they differed in important ways, then the difference in survival might be no more than a reflection of these initial differences.

It is worse than this. In the observational study, the control group was formed from records of patients already dead, while the treated patients were alive at the start of the study. The argument was that the treated patients were terminally ill, that they would all be dead shortly, so the recent records of apparently similar patients, now dead, could reasonably be used to indicate the duration of survival absent treatment with vitamin C. Nonetheless, when the results were analyzed, some patients given vitamin C were still alive; that is, their survival times were censored. This might reflect dramatic effects of vitamin C, but it might instead reflect some imprecision in judgments about who is terminally ill and how long a patient is likely to survive, that is, imprecision about the initial prognosis of patients in the treated group. In contrast, in the control group, one can say with total confidence, without reservation or caveat, that the prognosis of a patient already dead is not good. In the experiment, all patients in both treated and control groups were initially alive.

It is worse still. While death is a relatively unambiguous event, the time from "untreatability by standard therapies" to death depends also on the time of "untreatability." In the observational study, treated patients were judged, at the start of treatment with vitamin C, to be untreatable by other therapies. For controls, a date of untreatability was determined from records. It is possible that these two different processes would produce the same number, but it is by no means certain. In contrast, in the experiment, the starting date in treated and control groups was defined in the same way for both groups, simply because the starting date was determined before a subject was assigned to treatment or control.

What do we conclude from the studies of vitamin C? First, observational studies and experiments can yield very different conclusions. When this happens, the experiments tend to be believed. Chapter 2 develops some of the reasons why this tendency is reasonable. Second, matching and similar adjustments in observational studies, though often useful, do not ensure that treated and control groups are comparable in all relevant ways. More than this, the groups may not be comparable and yet the data we have may fail to reveal this. This issue is discussed extensively in later chapters. Third, while a controlled experiment uses randomization and an observational study does not, experimental control also helps in other ways. Even if we cannot randomize, we wish to exert as much experimental control as is possible, for instance, using the same eligibility criteria for treated and control groups, and the same methods for determining measurements.

Observational studies are typically conducted when experimentation is not possible. Direct comparisons of experiments and observational studies are less common, vitamin C for cancer being an exception. Another direct

comparison occurred in the Salk vaccine for polio, a story that is well told by Meier (1972). Others are discussed by Chalmers, Block, and Lee (1970), LaLonde (1986), Fraker and Maynard (1987), Zwick (1991), Friedlander and Robins (1995), and Dehejia and Wahba (1999).

Smoking and Heart Disease: An Elaborate Theory

Doll and Hill (1966) studied the mortality from heart disease of British doctors with various smoking behaviors. While dramatic associations are typically found between smoking and lung cancer, much weaker associations are found with heart disease. Still, since heart disease is a far more common cause of death, even modest increases in risk involve large numbers of deaths.

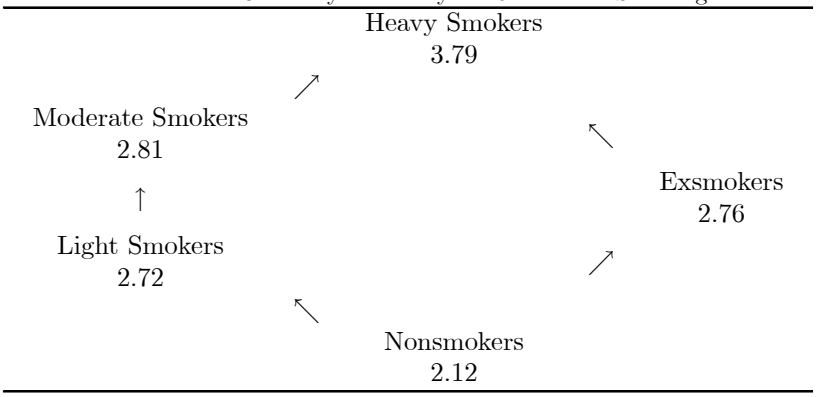
The first thing Doll and Hill did was to "adjust for age." The old are at greater risk of heart disease than the young. As a group, the smokers tended to be somewhat older than the nonsmokers, though of course there were many young smokers and many old nonsmokers. Compare smokers and nonsmokers directly, ignoring age, and you compare a somewhat older group to a somewhat younger group, so you expect a difference in coronary mortality even if smoking has no effect. In its essence, to "adjust for age" is to compare smokers and nonsmokers of the same age. Often results at different ages are combined into a single number called an age-adjusted mortality rate. Methods of adjustment and their properties are discussed in Chapters 3 and 10. For now, it suffices to say that differences in Doll and Hill's age-adjusted mortality rates cannot be attributed to differences in age, for they were formed by comparing smokers and nonsmokers of the same age. Adjustments of this sort, for age or other variables, are central to the analysis of observational data.

The second thing Doll and Hill did was to consider in detail what should be seen if, in fact, smoking causes coronary disease. Certainly, increased deaths among smokers are expected, but it is possible to be more specific. Light smokers should have mortality somewhere between that of nonsmokers and heavy smokers. People who quit smoking should also have risks between those of nonsmokers and heavy smokers, though it is not clear what to expect when comparing continuing light smokers to people who quit heavy smoking.

Why be specific? Why spell out in advance what a treatment effect should look like? The importance of highly specific theories has a long history, having been advocated in general by Sir Karl Popper (1959) and in observational studies by Sir Ronald Fisher, the inventor of randomized experiments, as quoted by Cochran (1965, §5):

> About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: 'Make your theories

TABLE 1.1. Coronary Mortality in Relation to Smoking.



> elaborate.' The reply puzzled me at first, since by Occam's razor, the advice usually given is to make theories as simple as is consistent with known data. What Sir Ronald meant, as subsequent discussion showed, was that when constructing a causal hypothesis one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold.
>
> ... this multi-phasic attack is one of the most potent weapons in observational studies.

Chapters 6 through 9 consider this advice formally and in detail.

Table 1.1 gives Doll and Hill's six age-adjusted mortality rates for death from coronary disease not associated with any other specific disease. The rates are deaths per 1000 per year, so the value 3.79 means about 4 deaths in each 1000 doctors each year. The six groups are nonsmokers, exsmokers, and light smokers of 1 to 14 cigarettes, moderate smokers of 15 to 24 cigarettes, and heavy smokers of 25 or more cigarettes per day. Doll and Hill did not separate exsmokers by the amount they had previously smoked, though this would have been interesting and would have permitted more detailed predictions. Again, differences in age do not affect these mortality rates.

Table 1.1 confirms each expectation. Mortality increases with the quantity smoked. Quitters have lower mortality than heavy smokers but higher mortality than nonsmokers. Any alternative explanation, any claim that smoking is not a cause of coronary mortality, would need to explain the entire pattern in Table 1.1. Alternative explanations are not difficult to imagine, but the pattern in Table 1.1 restricts their number.

DES and Vaginal Cancer: Sensitivity to Bias

Cancer of the vagina is a rare condition, particularly in young women. In 1971, Herbst, Ulfelder, and Poskanzer published a report describing eight cases of vaginal cancer in women aged 15 to 22. They were particularly interested in the possibility that a drug, diethylstilbestrol or DES, given to pregnant women, might be a cause of vaginal cancer in their daughters. Each of the eight cases was matched to four *referents*, that is, to four women who did not develop vaginal cancer. These four referents were born within five days of the birth of the case at the same hospital, and on the same type of service, ward or private. There were then eight cases of vaginal cancer and 32 referents, and the study compared the use of DES by their mothers.

This sort of study is called a *case-referent study* or a *case-control study* or a *retrospective study*, no one terminology being universally accepted. In an experiment and in many observational studies, treated and control groups are followed forward in time to see how outcomes develop. In the current context, this would mean comparing two groups of women, a treated group whose mothers had received DES and a control group whose mothers had not. That sort of study is not practical because the outcome, vaginal cancer, is so rare—the treated and control groups would have to be enormous and continue for many years to yield eight cases of vaginal cancer. In a case-referent study, the groups compared are not defined by whether or not they received the treatment, but rather by whether or not they exhibit the outcome. The cases are compared to the referents to see if exposure to the treatment is more common among cases.

In general, the name "case-control" study is not ideal because the word "control" does not have its usual meaning of a person who did not receive the treatment. In fact, in most case-referent studies, many referents did receive the treatment. The name "retrospective" study is not ideal because there are observational studies in which data on entire treated and control groups are collected after treatments have been given and outcomes have appeared, that is, collected retrospectively, and yet the groups being compared are still treated and untreated groups. See MacMahon and Pugh (1970, pp. 41–46) for some detailed discussion of this terminology.

So the study compared eight cases of vaginal cancer to 32 matched referents to see if treatment with diethylstilbestrol was more common among mothers of the cases, and indeed it was. Among the mothers of the eight cases, seven had received DES during pregnancy. Among mothers of the 32 referents, none had received DES. The association between vaginal cancer and DES appears to be almost as strong as a relationship can be, though of course only eight cases have been observed. If a conventional test designed for use in a randomized experiment is used to compare cases and referents in terms of the frequency of exposure to DES, the difference is highly significant. However, experience with the first example, vitamin C and cancer, suggests caution here.

What should be concluded from the strong association observed between DES and vaginal cancer in eight cases and 32 matched referents? Unlike the case of vitamin C and cancer, it would be neither practical nor ethical to follow up with a randomized experiment. Could such a hypothetical experiment produce very different findings? That possibility can never be entirely ruled out. Still, it is possible to ask: How severe would the unseen problems in this study have to be to produce such a strong relationship if DES did not cause vaginal cancer? How far would the observational study have to depart from an experiment to produce such a relationship if DES were harmless? How does the small size of the case group, eight cases, affect these questions? Chapter 4 provides answers. As it turns out, only severe unseen problems and hidden biases, only dramatic departures from an experiment, could produce such a strong association in the absence of an effect of DES, the small sample size notwithstanding. In other words, this study is highly insensitive to hidden bias; its conclusions could be altered by dramatic biases, but not by small ones. This is by no means true of all observational studies. Chapter 4 concerns general methods for quantifying the sensitivity of findings to hidden biases, and it discusses the uses and limitations of sensitivity analyses.

Academic Achievement in Public and Catholic High Schools: Specific Responses to Specific Criticisms

A current controversy in the United States concerns the effectiveness of public or state-run schools, particularly as compared to existing privately operated schools. The 1985 paper by Hoffer, Greely, and Coleman is one of a series of observational studies of this question. They used data from the High School and Beyond Study (HSB), which includes a survey of US high-school students as sophomores with follow-up in their senior year. The HSB study provided standardized achievement test scores in several areas in sophomore and senior years, and included follow-up of students who dropped out of school, so as these things go, it is a rather complete and attractive source of data. Hoffer, Greely, and Coleman (1985) begin with a list of six objections made to their earlier studies, which had compared achievement test scores in public and Catholic schools, concluding that "... Catholic high schools are more effective than public high schools." As an illustration, objection #3 states: "Catholic schools seem to have an effect because they eliminate their disciplinary problems by expelling them from the school." The idea here is that Catholic schools eliminate difficult students while the public schools do not, so the students who remain in Catholic schools would be more likely to perform well even if there were no difference in the effectiveness of the two types of schools.

Criticism is enormously important to observational studies. The quality of the criticism offered in a particular field is intimately connected with the

quality of the studies conducted in that field. Quality is not quantity, nor is harshness quality. What is scientifically plausible must be distinguished from what is just logically possible (Gastwirth, Krieger and Rosenbaum 1997). Cochran (1965, §5) argues that the first critic of an observational study should be its author:

> When summarizing the results of a study that shows an association consistent with the causal hypothesis, the investigator should always list and discuss all alternative explanations of his results (including different hypotheses and biases in the results) that occur to him. This advice may sound trite, but in practice is often neglected.

Criticisms of observational studies are of two kinds, the tangible and the dismissive, objection #3 being of the tangible kind. A tangible criticism is a specific and plausible alternative interpretation of the available data; indeed, a tangible criticism is itself a scientific theory, itself capable of empirical investigation. Bross (1960) writes:

> ... a critic who objects to a bias in the design or a failure to control some established factor is, in fact, raising a counterhypothesis ... [and] has the responsibility for showing [it] is tenable. In doing so, he operates under the same ground rules as the proponent ... : When a critic has shown that his counterhypothesis is tenable, his job is done, while at this point the proponent's job is just beginning. A proponent's job is not finished as long as there is a tenable hypothesis that rivals the one he asserts.

On the second page of his *The Design of Experiments*, Fisher (1935) described dismissive criticism as he argued that a theory of experimental design is needed:

> This type of criticism is usually made by what I might call a heavyweight *authority*. Prolonged experience, or at least the long possession of a scientific reputation, is almost a pre-requisite for developing successfully this line of attack. Technical details are seldom in evidence. The authoritative assertion: "His *controls* are *totally* inadequate" must have temporarily discredited many a promising line of work; and such an authoritarian method of judgement must surely continue, human nature being what it is, so long as theoretical notions of the principles of experimental design are lacking ... .

Dismissive criticism rests on the authority of the critic and is so broad and vague that its claims cannot be studied empirically. Judging the weight

of evidence is inseparable from judging the criticisms that have been or can be raised.

Concerning objection #3, Hoffer, Greely, and Coleman (1985) respond: "... the evidence from the HSB data, although indirect, does not support this objection. Among students who reported that they had been suspended during their sophomore year, those in the Catholic sector were more likely to be in the same school as seniors than those in the public sector (63 percent to 56 percent)." In other words, difficult students, or at any rate students who were suspended, remained in Catholic school more often, not less often, than in public schools. This response to objection #3, though not decisive, does gives one pause.

Successful criticism of an observational study points to ambiguity in evidence or argument, and then points to methods for removing the ambiguity. Efforts to resolve an ambiguity are sometimes undermined by efforts to win an argument. Popper (1994, p. 44) writes:

> Serious critical discussions are always difficult ... Many participants in a rational, that is, a critical, discussion find it particularly difficult to unlearn what their instincts seem to teach them (and what they are taught, incidently, by every debating society): that is, to win. For what they have to learn is that victory in debate is nothing, while even the slightest clarification of one's problem—even the smallest contribution made towards a clearer understanding of one's own position or that of one's opponent—is a great success. A discussion which you win but which fails to help you change or to clarify your mind at least a little should be regarded as a sheer loss.

1.3 Purpose of This Book

Scientific evidence is commonly and properly greeted with objections, skepticism, and doubt. Some objections come from those who simply do not like the conclusions, but setting aside such unscientific reactions, responsible scientists are responsibly skeptical. We look for failures of observation, gaps in reasoning, alternative interpretations. We compare new evidence with past evidence. This skepticism is itself scrutinized. Skepticism must be justified, defended. One needs "grounds for doubt," in Wittgenstein's (1969, §122) phrase. The grounds for doubt are themselves challenged. Objections bring forth counterobjections and more evidence. As time passes, arguments on one side or the other become strained, fewer scientists are willing to offer them, and the arguments on that side come increasingly from individuals who seem to have some stake in the outcome. In this way, questions are settled.

Scientific questions are not settled on a particular date by a single event, nor are they settled irrevocably. We speak of the weight of evidence. Eventually, the weight is such that critics can no longer lift it, or are too weary to try. Overwhelming evidence is evidence that overwhelms responsible critics.

Experiments are better than observational studies because there are fewer grounds for doubt. The ideal experiment would leave few grounds for doubt, and at times this ideal is nearly achieved, particularly in the laboratory. Experiments often settle questions faster.

Despite this, experiments are not feasible in some settings. At times, observational studies have produced overwhelming evidence, as compelling as any in science, but at other times, observational data have misled investigators to advocate harmful policies or ineffective treatments.

A statistical theory of observational studies is a framework and a set of tools that provide measures of the weight of evidence. The purpose of this book is to give an account of statistical principles and methods for the design and analysis of observational studies. An adequate account must relate observational studies to controlled experiments, showing how uncertainty about treatment effects is greater in the absence of randomization. Analytical adjustments are common in observational studies, and the account should indicate what adjustments can and cannot do. A large literature offers many devices to detect hidden biases in observational studies, for instance, the use of several control groups, and the account must show how such devices work and when they may be expected to succeed or fail. Even when it is not possible to reduce or dispel uncertainty, it is possible to be careful in discussing its magnitude. That is, even when it is not possible to remove bias through adjustment or to detect bias through careful design, it is nonetheless possible to give quantitative expression to the magnitude of uncertainties about bias, a technique called *sensitivity analysis*. The account must indicate what can and cannot be done with a sensitivity analysis.

1.4 Bibliographic Notes

Most scientific fields that study human populations conduct observational studies. Many fields have developed a literature on the design, conduct, and interpretation of observational studies, often with little reference to related work in other fields. It is not possible to do justice to these several literatures in a short bibliographic note. There follows a short and incomplete list of fine books that contain substantial general discussions of the methodology used for observational studies in epidemiology, public program evaluation, or the social sciences. A shared goal in these diverse works is evaluation of treatments, exposures, programs, or policies from nonexperimental data. The list is followed by references cited in Chapter 1.

Some Books and a Few Papers

Angrist, J. D. and Krueger, A. B. (1999) Empirical strategies in labor economics. In: *Handbook of Labor Economics*, O. Ashenfelter and D. Card, eds., Volume 3A, Chapter 23, New York: Elsevier.

Ashenfelter, O., ed. (2000) *Labor Economics*. New York: Worth.

Becker, H. S. (1997) *Tricks of the Trade*. Chicago: University of Chicago Press.

Blaug, M. (1980) *The Methodology of Economics*. New York: Cambridge University Press.

Breslow, N. and Day, N. (1980, 1987) *Statistical Methods in Cancer Research*, Volumes 1 and 2. Lyon, France: International Agency for Research on Cancer.

Campbell, D. T. (1988) *Methodology and Epistemology for Social Science: Selected Papers*. Chicago: University of Chicago Press, pp. 315—333.

Campbell, D. and Stanley, J. (1963) *Experimental and Quasi-Experimental Design for Research*. Chicago: Rand McNally.

Chamberlain, G. (1984) Panel data. In: *Handbook of Econometrics*, Chapter 22, Volume 2, Z. Griliches and M. D. Intriligator, eds., New York: Elsevier.

Cochran, W. G. (1965) The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society*, Series **A**, **128**, 134—155.

Cochran, W. (1983) *Planning and Analysis of Observational Studies*. New York: Wiley.

Cook, T. D. and Campbell, D. C. (1979) *Quasi-Experimentation*. Chicago: Rand McNally.

Cook, T. D., Campbell, D. T., and Peracchio, L. (1990) Quasi-experimentation. In: *Handbook of Industrial and Organizational Psychology*, M. Dunnette and L. Hough, eds., Palo Alto, CA: Consulting Psychologists Press, Chapter 9, pp. 491—576.

Cook, T. D. and Shadish, W. R. (1994) Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology*, **45**, 545—580.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., and Wynder, E. (1959) Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, **22**, 173—203.

Cox, D. R. (1992) Causality: Some statistical aspects. *Journal of the Royal Statistical Society*, Series **A**, **155**, 291–301.

Elwood, J. M. (1988) *Causal Relationships in Medicine*. New York: Oxford University Press.

Emerson, R. M. (1981) Observational field work. *Annual Review of Sociology*, **7**, 351—378.

Freedman, D. (1997) From association to causation via regression. *Advances in Applied Mathematics*, **18**, 59—110.

Friedman, M. (1953) *Essays in Positive Economics*. Chicago: University of Chicago Press.

Gastwirth, J. (1988) *Statistical Reasoning in Law and Public Policy*. New York: Academic Press.

Gordis, L. (2000) *Epidemiology* (Second Edition) Philadelphia: Saunders.

Greenhouse, S. (1982) Jerome Cornfield's contributions to epidemiology. *Biometrics*, **28**, Supplement, 33–46.

Heckman, J. J. (2001) Micro data, heterogeneity, and the evaluation of public policy: The Nobel lecture. *Journal of Political Economy*, **109**, 673–748.

Hill, A. B. (1965) The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, **58**, 295—300.

Holland, P. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–970.

Kelsey, J., Whittemore, A., Evans, A., and Thompson. W. (1996). *Methods in Observational Epidemiology*. New York: Oxford University Press.

Khoury, M. J., Cohen, B. H., and Beaty, T. H. (1993) *Fundamentals of Genetic Epidemiology*. New York: Oxford University Press.

Kish, L. (1987) *Statistical Design for Research*. New York: Wiley.

Lilienfeld, A. and Lilienfeld, D. E. (1980) *Foundations of Epidemiology*. New York: Oxford University Press.

Lilienfeld, D. E. and Stolley, P. D. (1994) *Foundations of Epidemiology*. New York: Oxford University Press.

Lipsey, M. W. and Cordray, D. S. (2000) Evaluation methods for social intervention. *Annual Review of Psychology*, **51**, 345—375.

Little, R. J. and Rubin, D. B. (2000) Causal effects in clinical and epidemiological studies via potential outcomes. *Annual Review of Public Health*, **21**, 121—145.

Maclure, M. and Mittleman, M. A. (2000) Should we use a case-crossover design? *Annual Review of Public Health*, **21**, 193—221.

MacMahon, B. and Pugh, T. (1970) *Epidemiology*. Boston: Little, Brown.

MacMahon, B. and Trichopoulos, D. (1996) *Epidemiology*. Boston: Little, Brown.

Manski, C. (1995) *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Mantel, N. and Haenszel, W. (1959) Statistical aspects of retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719–748.

Meyer, B. D. (1995) Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics*, **13**, 151—161.

Meyer, M. and Fienberg, S., eds. (1992) *Assessing Evaluation Studies: The Case of Bilingual Education Strategies*. Washington, DC: National Academy Press.

Miettinen, O. (1985) *Theoretical Epidemiology*. New York: Wiley.

Pearl, J. (2000) *Causality: Models, Reasoning, Inference*. New York: Cambridge University Press.

Reichardt, C. S. (2000) A typology of strategies for ruling out threats to validity. In: *Research Design: Donald Campbell's Legacy*, L. Brickman, ed., Thousand Oaks, CA: Sage, Volume 2, pp., 89–115.

Reiter, J. (2000) Using statistics to determine causal relationships. *American Mathematical Monthly*, **107**, 24—32.

Robins, J. M. (1999) Association, causation, and marginal structural models. *Synthese*, **121**, 151—179.

Robins, J., Blevins, D., Ritter, G., and Wulfsohn, M. (1992) G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology,* **3**, 319—336.

Rosenthal, R. and Rosnow, R., eds. (1969) *Artifact in Behavioral Research*. New York: Academic.

Rosenzweig, M. R. and Wolpin, K. I. (2000) Natural "natural experiments" in economics. *Journal of Economic Literature*, **38**, 827—874.

Rosnow, R. L. and Rosenthal, R. (1997) *People Studying People: Artifacts and Ethics in Behavioral Research*. New York: W. H. Freeman.

Rossi, P., Freeman, H., and Lipsey, M. W. (1999) *Evaluation*. Beverly Hills, CA: Sage.

Rothman, K. and Greenland, S. (1998) *Modern Epidemiology*. Philadelphia: Lippincott-Raven.

Rubin, D. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Schlesselman, J. (1982) *Case-Control Studies*. New York: Oxford University Press.

Schulte, P. A. and Perera, F. (1993) *Molecular Epidemiology: Principles and Practices*. New York: Academic.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Shafer, G. (1996) *The Art of Causal Conjecture*. Cambridge, MA: MIT Press.

Sobel, M. (1995) Causal inference in the social and behavioral sciences. In: *Handbook of Statistical Modelling for the Social and Behavioral Sciences*, G. Arminger, C. Clogg, and M. Sobel, eds., New York: Plenum, 1—38.

Steenland, K., ed. (1993) *Case Studies in Occupational Epidemiology*. New York: Oxford University Press.

Strom, B. (2000) *Pharmacoepidemiology*. New York: Wiley.

Suchman, E. (1967) *Evaluation Research*. New York: Sage.

Susser, M. (1973) *Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology*. New York: Oxford University Press.

Susser, M. (1987) *Epidemiology, Health and Society: Selected Papers*. New York: Oxford University Press.

Tufte, E., ed. (1970) *The Quantitative Analysis of Social Problems*. Reading, MA: Addison-Wesley.

Weiss, C. (1997) *Evaluation*. Englewood Cliffs, NJ: Prentice-Hall.

Weiss, N. S. (1996) *Clinical Epidemiology*. New York: Oxford University Press.

Willett, W. (1998) *Nutritional Epidemiology.* New York: Oxford University Press.

Winship, C. and Morgan, S. L. (1999) The estimation of causal effects from observational data. *Annual Review of Sociology*, **25**, 659—706.

Zellner, A. (1968) *Readings in Economic Statistics and Econometrics.* Boston: Little, Brown.

1.5 References

Bross, I. D. J. (1960) Statistical criticism. *Cancer*, **13**, 394–400. Reprinted in: *The Quantitative Analysis of Social Problems*, E. Tufte, ed., Reading, MA: Addison-Wesley, pp. 97–108.

Cameron, E. and Pauling, L. (1976) Supplemental ascorbate in the supportive treatment of cancer: Prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Sciences* (USA), **73**, 3685–3689.

Chalmers, T., Block, J., and Lee, S. (1970) Controlled studies in clinical cancer research. *New England Journal of Medicine*, **287**, 75–78.

Cochran, W.G. (1965) The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society*, Series A, **128**, 134–155. Reprinted in *Readings in Economic Statistics and Econometrics*, A. Zellner, ed., 1968, Boston: Little Brown, pp. 11–36.

Dehejia, R. H. and Wahba, S. (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association,* **94**, 1053—1062.

Doll, R. and Hill, A. (1966) Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. In: *Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases*, W. Haenszel, ed., U.S. National Cancer Institute Monograph 19, Washington, DC: US Department of Health, Education, and Welfare, pp. 205–268.

Fisher, R.A. (1935, 1949) *The Design of Experiments.* Edinburgh: Oliver & Boyd.

Fraker, T. and Maynard, R. (1987) The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, **22**, 194–227.

Friedlander, D. and Robins, P. K. (1995) Evaluating program evaluations: New evidence on commonly used nonexperimental methods. *American Economic Review*, **85**, 923—937.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1997) Hypotheticals and hypotheses. *American Statistician*, **51**, 120—121.

Herbst, A., Ulfelder, H., and Poskanzer, D. (1971) Adenocarcinoma of the vagina: Association of maternal stilbestrol therapy with tumor appearance in young women. *New England Journal of Medicine*, **284**, 878–881.

Hoffer, T., Greeley, A., and Coleman, J. (1985) Achievement growth in public and Catholic schools. *Sociology of Education*, **58**, 74–97.

LaLonde, R. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, **76**, 604–620.

Meier, P. (1972) The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In: *Statistics: A Guide to the Unknown*, J. Tanur, ed., San Francisco: Holden-Day, pp. 2–13.

Moertel, C., Fleming, T., Creagan, E., Rubin, J., O'Connell, M., and Ames, M. (1985) High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: A randomized double-blind comparison. *New England Journal of Medicine*, **312**, 137–141.

Popper, K. (1959) *The Logic of Scientific Discovery*. New York: Harper & Row.

Popper, K. (1994) *The Myth of the Framework*. New York: Routledge.

United States Surgeon General's Advisory Committee Report (1964) *Smoking and Health*. Washington, DC: US Department of Health, Education and Welfare.

Wittgenstein, L. (1969) *On Certainty*. New York: Harper & Row.

Zwick, R. (1991) Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, **3**, 10–16.