From Association to Causation: Some Remarks on the History of Statistics

by

David Freedman, Statistics Department
University of California, Berkeley, CA 94720, USA
Technical Report No. 521 May 30, 1999
To appear in *Statistical Science*

Abstract

The "numerical method" in medicine goes back to Pierre Louis' study of pneumonia (1835), and John Snow's book on the epidemiology of cholera (1855). Snow took advantage of natural experiments and used convergent lines of evidence to demonstrate that cholera is a waterborne infectious disease. More recently, investigators in the social and life sciences have used statistical models and significance tests to deduce cause-and-effect relationships from patterns of association; an early example is Yule's study on the causes of poverty (1899). In my view, this modeling enterprise has not been successful. Investigators tend to neglect the difficulties in establishing causal relations, and the mathematical complexities obscure rather than clarify the assumptions on which the analysis is based.

Formal statistical inference is, by its nature, conditional. If maintained hypotheses A, B, C, ... hold, then H can be tested against the data. However, if A, B, C, ... remain in doubt, so must inferences about H. Careful scrutiny of maintained hypotheses should therefore be a critical part of empirical work—a principle honored more often in the breach than the observance. Snow's work on cholera will be contrasted with modern studies that depend on statistical models and tests of significance. The examples may help to clarify the limits of current statistical techniques for making causal inferences from patterns of association.

1. INTRODUCTION

In this paper, I will look at some examples from the history of statistics—examples which help to define problems of causal inference from non-experimental data. By comparing the successes with the failures, we may learn something about the causes of both; this is a primitive study design, but one that has provided useful clues to many investigators since Mill (1843). I will discuss the classical research of Pierre Louis (1835) on pneumonia, and summarize the work of John Snow (1855) on cholera. Modern epidemiology has come to rely more heavily on statistical models, which seem to have spread from the physical to the social sciences and then to epidemiology (sections 4 and 5). The modeling approach was quite successful in the physical sciences, but has been less so in the other domains, for reasons that will be suggested in sections 4–6.

Regression models are now widely used to control for the effects of confounding variables, an early paper being Yule (1899); that is the topic of section 4. Then some contemporary examples will be mentioned, including studies on asbestos in drinking water (section 5), health effects of electromagnetic fields, air pollution, the leukemia cluster at Sellafield, and cervical cancer (section 7). Section 8 discusses one of the great triumphs of

the epidemiologic method—identifying the health effects of smoking. Other points of view on modeling are briefly noted in section 9. Finally, there is a summary with conclusions.

2. LA MÉTHODE NUMÉRIQUE

In 1835, Pierre Louis published his classic study on the efficacy of the standard treatments for pneumonia: *Recherches sur les effets de la saignée dans quelques maladies inflammatoires: et sur l'action de l'émétique et des vésicatoires dans la pneumonie.* Louis was a physician in Paris. In brief, he concluded that bleeding the patient was a good treatment for pneumonia, although less effective than commonly thought:

> "Que la saignée a une heureuse influence sur la marche de la pneumonie; qu'elle en abrège la durée; que cependant cette influence est beaucoup moindre qu'on ne se l'imagine communément... [p. 62]"

His contemporaries were not all persuaded. According to one, arithmetic should not have been allowed to constrain the imagination:

> "En invoquant l'inflexibilité de l'arithmétique pour se soustraire aux empiétemens de l'imagination, on commet contre le bon sens la plus grave erreur... [p. 79]"

Pierre Louis was comparing average outcomes for patients bled early or late in the course of the disease. The critic felt that the groups were different in important respects apart from treatment. Louis replied that individual differences made it impossible to learn much from studying individual cases and necessitated the use of averages; see also Gavarret (1840). This tension has never been fully resolved, and is with us even today.

A few statistical details may be of interest. Louis reports on 78 pneumonia patients. All were bled, at different stages of the disease, and 50 survived. Among the survivors, bleeding in the first two days cut the length of the illness in half. But, Louis noted, there were differences in régime. Those treated later had not followed doctors' orders:

> "[ils] avaient commis des erreurs de régime, pris des boissons fortes, du vin chaud sucré, un ou plusieurs jours de suite, en quantité plus ou moins considérable; quelquefois même de l'eau-de-vie. [p. 13]"

From a modern perspective, there is a selection effect in Louis' analysis: those treated later in the course of an illness are likely for that reason alone to have had longer illnesses. It therefore seems better to consider outcomes for all 78 patients, including those who died, and bleeding in the first two days doubles the risk of death. Louis saw this, but dismissed it as frightening and absurd on its face:

> "Résultat effrayant, absurde en apparence. [p. 17]"

He explains that those who were bled later were older. He was also careful to point out the limitations created by a small sample.

Among other things, Louis identified two major methodological issues: (i) sampling error and (ii) confounding. These problems must be addressed in any epidemiologic study. Confounding is the more serious issue. In brief, a comparison is made between a treatment group and a control group, in order to determine the effect of treatment. If the groups differ with respect to another factor—the "confounding variable"—which influences the outcome, the estimated treatment effect will also include the effect of the confounder, leading to

a potentially serious bias. If the treatment and control groups are chosen at random, bias is minimized. Of course, in epidemiologic studies, there are many other sources of bias besides confounding. One example is "recall bias," where a respondent's answers to questions about exposure are influenced by presence or absence of disease. Another example is "selection bias," due for instance to systematic differences between subjects chosen for a study and subjects excluded from the study. Even random measurement error can create bias in estimated effects: random errors in measuring the size of a causal factor tend to create a bias toward 0, while errors in measuring a confounder create a bias in the opposite direction.

Pierre Louis' book was published in the same year as Quetelet's *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale.* Quetelet, like Louis, has had—and continues to have—an important influence over the development of our subject (sections 4 and 5).

3. SNOW ON CHOLERA

In 1855, some twenty years before Koch and Pasteur laid the foundations of modern microbiology, Snow discovered that cholera is a waterborne infectious disease. At the time, the germ theory of disease was only one of many conceptions. Imbalance in the humors of the body was an older explanation for disease. Miasma, or bad air, was often said to be the cause of epidemics. Poison in the ground was perhaps a slightly later idea.

Snow was a physician in London. By observing the course of the disease, he concluded that cholera was caused by a living organism, which entered the body with water or food, multiplied in the body, and made the body expel water containing copies of the organism. The dejecta then contaminated food or reentered the water supply, and the organism proceeded to infect other victims. The lag between infection and disease (a matter of hours or days) was explained as the time needed for the infectious agent to multiply in the body of the victim. This multiplication is characteristic of life: inanimate poisons do not reproduce themselves.

Snow developed a series of arguments in support of the germ theory. For instance, cholera spread along the tracks of human commerce. Furthermore, when a ship entered a port where cholera was prevalent, sailors contracted the disease only when they came into contact with residents of the port. These facts were easily explained if cholera was an infectious disease, but were harder to explain by the miasma theory.

There was a cholera epidemic in London in 1848. Snow identified the first or "index" case in this epidemic:

> "a seaman named John Harnold, who had newly arrived by the *Elbe* steamer from Hamburgh, where the disease was prevailing. [p. 3]"

He also identified the second case: a man named Blenkinsopp who took Harnold's room after the latter died, and presumably became infected by contact with the bedding. Next, Snow was able to find adjacent apartment buildings, one being heavily affected by cholera and one not. In each case, the affected building had a contaminated water supply; the other had relatively pure water. Again, these facts are easy to understand if cholera is an infectious disease, but hard to explain on the miasma theory.

There was an outbreak of the disease in August and September of 1854. Snow made what is now called a "spot map," showing the locations of the victims. These clustered near the Broad Street pump. (Broad Street is in Soho, London; at the time, there were public pumps used as a source of water.) However, there were a number of institutions in the area with few or no fatalities. One was a brewery. The workers seemed to have preferred ale to water: but if any wanted water, there was a private pump on the premises. Another institution free of cholera was a poor-house, which too had its own private pump. People in other areas of London contracted the disease; but in most cases, Snow was able to show they drank water from the Broad Street pump. For instance, one lady in Hampstead used to live in Soho, and so much liked the taste of the Broad Street water that she routinely sent a servant to draw water from the fatal pump.

So far, we have persuasive anecdotal evidence that cholera is an infectious disease, spread by contact or through the water supply. Snow also made statistical studies. For instance, there were a number of water companies in the London of his time. Some took their water from heavily contaminated stretches of the Thames river; for others, the intake was relatively uncontaminated. Snow made what are now called "ecological" studies, correlating death rates from cholera in various areas of London with the quality of the water. Generally speaking, areas with contaminated water had higher death rates. One exception was the Chelsea water company. This company started with contaminated water, but had quite modern methods of purification—settling ponds, exposure to sunlight, and sand filtration. Its service area had a low death rate from cholera.

In 1852, the Lambeth water company moved its intake pipe upstream to secure relatively pure water. The Southwark and Vauxhall company left its intake pipe where it was, in a heavily contaminated stretch of the Thames. Snow made an ecological analysis comparing the areas serviced by the two companies in the epidemics of 1853–54 and in earlier years. Let him now continue in his own words.

> "Although the facts shown in the above table [the ecological analysis] afford very strong evidence of the powerful influence which the drinking of water containing the sewage of a town exerts over the spread of cholera, when that disease is present, yet the question does not end here; for the intermixing of the water supply of the Southwark and Vauxhall Company with that of the Lambeth Company, over an extensive part of London, admitted of the subject being sifted in such a way as to yield the most incontrovertible proof on one side or the other. In the subdistricts enumerated in the above table as being supplied by both Companies, the mixing of the supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition. In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies. Now it must be evident that, if the diminution of cholera, in the districts partly supplied with improved water, depended on this supply, the houses receiving it would be the houses enjoying the whole benefit of the diminution of the malady, whilst the houses supplied with the [contaminated] water from Battersea Fields would suffer the same mortality as

they would if the improved supply did not exist at all. As there is no difference whatever in the houses or the people receiving the supply of the two Water Companies, or in any of the physical conditions with which they are surrounded, it is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this, which circumstances placed ready made before the observer.

"The experiment, too, was on the grandest scale. No fewer than three hundred thousand people of both sexes, of every age and occupation, and of every rank and station, from gentlefolks down to the very poor, were divided into groups without their choice, and in most cases, without their knowledge; one group being supplied with water containing the sewage of London, and amongst it, whatever might have come from the cholera patients; the other group having water quite free from such impurity.

"To turn this grand experiment to account, all that was required was to learn the supply of water to each individual house where a fatal attack of cholera might occur. [pp. 74–75.]"

Snow's data are shown in Table 1. The denominator data—the number of houses served by each water company—were available from parliamentary records. For the numerator data, however, a house-to-house canvass was needed to determine the source of the water supply at the address of each cholera fatality. (The "bills of mortality" showed the address, but not the water source.) The death rate from the Southwark and Vauxhall water is about 9 times the death rate for the Lambeth water. This is compelling evidence.

Snow argued that the data could be analyzed as if they had resulted from an experiment of nature: there was no difference between the customers of the two water companies, except for the water. His sample was not only large but representative; therefore, it was possible to generalize to a larger population. Finally, Snow was careful to avoid the "ecological fallacy:" relationships that hold for groups may not hold for individuals (Robinson, 1950). It is the design of the study and the magnitude of the effect that compel conviction, not the elaboration of technique.

TABLE 1. Death rate from cholera by source of water. Rate per 10,000 houses. London, epidemic of 1853–54. Snow's Table IX.

	No. of Houses	Cholera Deaths	Rate per 10,000
Southwark & Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

More evidence was to come from other countries. In New York, the epidemics of 1832 and 1849 were handled according to the theories of the time. The population was exhorted to temperance and calm, since anger could increase the humor "choler" (bile), and imbalances in the humors of the body lead to disease. Pure water was brought in to wash the streets and reduce miasmas. In 1866, however, the epidemic was handled by a different method—rigorous isolation of cholera cases, with disinfection of their dejecta by lime or fire. The fatality rate was much reduced.

At the end of the 19th century, there was a burst of activity in microbiology. In 1878, Pasteur published *La théorie des germes et ses applications à la médecine et à la chirurgie.* Around that time, Pasteur and Koch isolated the anthrax bacillus and developed techniques for vaccination. The tuberculosis bacillus was next. In 1883, there was a cholera epidemic in Egypt, and Koch isolated the vibrio; he was perhaps anticipated by Filipo Pacini. There was an epidemic in Hamburg in 1892. The city fathers turned to Max von Pettenkofer, a leading figure in the German hygiene movement of the time. He did not believe Snow's theory, holding instead that cholera was caused by poison in the ground. Hamburg was a center of the slaughterhouse industry, and von Pettenkofer had the carcasses of dead animals dug up and hauled away, in order to reduce pollution of the ground. The epidemic continued its ravages, which ended only when the city lost faith in von Pettenkofer and turned in desperation to Koch.

The approach developed by Louis and Snow found many applications. For instance, Semmelweiss (1867) found the cause of puerperal fever. Around 1914, Goldberger showed that pellagra was the result of a diet deficiency. References on the history of cholera include Rosenberg (1962), Howard-Jones (1975), Evans (1987), Winkelstein (1995), Paneth et al. (1998). Terris (1964) reprints many of Goldberger's articles; also see Carpenter (1981). A useful reference on Pasteur is Dubos (1988). Today, the molecular biology of the cholera vibrio is reasonably well understood; see, for instance, Finlay, Heffron, and Fialkow (1989) or Miller, Mekalanos, and Fialkow (1989). For a synopsis, see Alberts et al. (1994, pp. 484, 738); there are recent surveys by Colwell (1996) and Raufman (1998).

4. REGRESSION MODELS IN SOCIAL SCIENCE

Legendre (1805) and Gauss (1809) developed the regression method (least absolute residuals or least squares) to fit data on the orbits of astronomical objects. In this context, the relevant variables are known and so are the functional forms of the equations connecting them. Measurement can be done to high precision, and much is known about the nature of the errors—in the measurements and the equations. Furthermore, there is ample opportunity for comparing predictions to reality.

By the turn of the century, investigators were using regression on social science data where these conditions did not hold, even to a rough approximation. One of the earliest such papers is Yule (1899), "An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades." At the time, paupers were supported either inside "poor-houses" or outside, depending on the policy of local authorities. Did the relief policy affect the number of paupers? To study this question, Yule offered a regression equation,

$$\Delta\text{Paup} = a + b \times \Delta\text{Out} + c \times \Delta\text{Old} + d \times \Delta\text{Pop} + \text{error}.$$

In this equation,

Δ is percentage change over time,
"Out" is the out-relief ratio N/D,
$N =$ number on welfare outside the poor-house,
$D =$ number inside,

"Old" is the percentage of the population over 65,

"Pop" is the population.

Data are from the English Censuses of 1871, 1881, 1891. There are two Δ's, one for 1871–81 and one for 1881–91.

Relief policy was determined separately by the local authorities in each "union," a small geographical area like a parish. At the time, there were about 600 unions, and Yule divided them into four kinds: rural, mixed, urban, metropolitan. There are $2 \times 4 = 8$ equations, one for each combination of time period and type of union. Yule assumed that the coefficients were constant for each equation, which he fitted to the data by least squares. That is, he estimated the coefficients a, b, c, and d as the values that minimized the sum of squared errors,

$$\sum \left(\Delta\text{Paup} - a - b \times \Delta\text{Out} - c \times \Delta\text{Old} - d \times \Delta\text{Pop}\right)^2.$$

The sum is taken over all unions of a given type at a given time-period.

For example, consider the metropolitan unions. Fitting the equation to the data for 1871–81 gave

$$\Delta\text{Paup} = 13.19 + 0.755\Delta\text{Out} - 0.022\Delta\text{Old} - 0.322\Delta\text{Pop} + \text{residual}.$$

For 1881–91, Yule's equation was

$$\Delta\text{Paup} = 1.36 + 0.324\Delta\text{Out} + 1.37\Delta\text{Old} - 0.369\Delta\text{Pop} + \text{residual}.$$

The framework combines the ideas of Quetelet with the mathematics of Gauss. Yule is studying the "social physics" of poverty. Nature has run an experiment, assigning different treatments to different areas. Yule is analyzing the results, using regression to isolate the effects of out-relief. His principal conclusion is that welfare outside the poor-house creates paupers—the estimated coefficient on the out-relief ratio is positive.

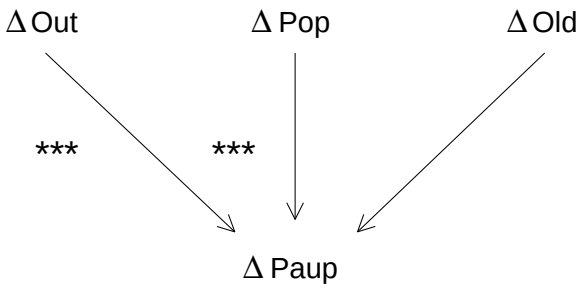
At this remove, the flaws in the argument are clear. Confounding is a salient problem. For instance, Pigou (a famous economist of the era) thought that unions with more efficient administrations were the ones building poor-houses and reducing poverty. Efficiency of administration is then a confounder, influencing both the presumed cause and its effect. Economics may be another confounder. At times, Yule seems to be using the rate of population change as a proxy for economic growth, although this is not entirely convincing. Generally, however, he pays little attention to economic activity. The explanation: "A good deal of time and labour was spent in making trial of this idea, but the results proved unsatisfactory, and finally the measure was abandoned altogether. [p. 253]"

The form of his equation is somewhat arbitrary, and the coefficients are not consistent over time and space. This is not necessarily fatal. However, if the coefficients do not exist separately from the data, how can they predict the results of interventions? There are also problems of interpretation. At best, Yule has established association. Conditional on the covariates, there is a positive association between ΔPaup and ΔOut. Is this association causal? If so, which way do the causal arrows point? These questions are not answered by the data analysis; rather, the answers are assumed *a priori*. Yule is quite concerned to parcel out changes in pauperism: so much is due to changes in the out-relief ratio, so

much to changes in other variables, and so much to random effects. However, there is one deft footnote (number 25) that withdraws all causal claims:

"Strictly speaking, for 'due to' read 'associated with.'"

FIGURE 1. Yule's Model. Metropolitan Unions, 1871–81.



Yule's approach is strikingly modern, except there is no causal diagram and no stars indicating statistical significance. Figure 1 brings him up to date. An arrow from X to Y indicates that X is included in the regression equation that explains Y. "Statistical significance" is indicated by an asterisk, and three asterisks signal a high degree of significance. The idea is that a statistically significant coefficient differs from 0, so that X has a causal influence on Y. By contrast, an insignificant coefficient is zero: then X does not exert a causal influence on Y. The reasoning is seldom made explicit, and difficulties are frequently overlooked.

Stringent assumptions are needed to determine significance from the data. Even if significance can be determined and the null hypothesis rejected or accepted, there is a much deeper problem. To make causal inferences, it must in essence be assumed that equations are invariant under proposed interventions. Verifying such assumptions—without making the interventions—is quite problematic. On the other hand, if the coefficients and error terms change when the right hand side variables are manipulated rather than being passively observed, then the equation has only a limited utility for predicting the results of interventions. These difficulties are well known in principle, but are seldom dealt with by investigators doing applied work in the social and life sciences. Despite the problems, and the disclaimer in the footnote, Yule's regression approach has become widely used in the social sciences and epidemiology.

Some formal models for causation are available, starting with Neyman (1923). See Hodges and Lehmann (1964, sec. 9.4), Rubin (1974), or Holland (1988). More recent developments will be found in Pearl (1995) or Angrist, Imbens and Rubin (1996). For critical discussion from various perspectives, see Goldthorpe (1998), Abbott (1997), Humphreys and Freedman (1996, 1999), McKim and Turner (1997), Manski (1995), Lieberson (1985), Lucas (1976), Liu (1960), or Freedman (1987, 1991, 1995). The history is discussed by Stigler (1986) and Desrosières (1993).

5. REGRESSION MODELS IN EPIDEMIOLOGY

Regression models (and variations like the Cox model) are widely used in epidemiology. The models seem to give answers, and create at least the appearance of methodological rigor. This section discusses one example, which is fairly typical of such applications and provides an interesting contrast to Snow on cholera. Snow used primitive statistical techniques, but his study designs were extraordinarily well thought out, and he made a huge effort to collect the relevant data. By contrast, many empirical papers published today, even in the leading journals, lack a sharply-focused research question; or the study design connects the hypotheses to the data collection only in a very loose way. Investigators often try to use statistical models not only to control for confounding, but also to correct basic deficiencies in the design or the data. Our example will illustrate some of these points.

Kanarek et al. (1980) asked whether asbestos fibers in the drinking water causes cancer. They studied 722 census tracts in the San Francisco Bay Area. (A census tract is a small geographical region, with several thousand inhabitants.) The investigators measured asbestos fiber concentration in the water for each tract. Perhaps surprisingly, there is enormous variation. Kanarek et al. compared the "observed" number of cancers by site with the expected number, by sex, race, and tract. The "expected" number is obtained by applying age-specific national rates to the population of the tract, age group by age group; males and females are done separately, and only whites are considered. (There are about 100 sites for which age-specific national data are available; comparison of observed to expected numbers is an example of "indirect standardization.")

Regression is used to adjust for income, education, marital status, and occupational exposure. The equation is not specified in great detail, but is of the form

$$\log \frac{\text{Obs.}}{\text{Exp.}} = A_0 + A_1 \,\text{asbestos fiber concentration} + A_2 \,\text{income} + A_3 \,\text{education} + A_4 \,\text{married} + A_5 \,\text{asbestos workers} + \text{error}.$$

Here, "income" is the median figure for persons in the tract, and "education" is the median number of years of schooling; data are available from the census. These variables adjust to some extent for socio-economic differences between tracts: usually, rates of disease go down as income and education go up. The next variable in the equation is the fraction of persons in the tract who are married; such persons are typically less subject to disease than the unmarried. Finally, there is the number of "asbestos workers" in the tract; these persons may have unusually high rates of cancer, due to exposure on the job. Thus, the variables on the right hand side of the equation are potential confounders, and the equation is an attempt to adjust for their effects. The estimate of A_1 for lung cancer in males is "highly statistically significant," with $P < .001$. A highly significant coefficient like this would nowadays be taken as strong evidence of causation, but there are serious difficulties.

Confounding. No adjustment is made for smoking habit, which was not measured in this study. Smoking is strongly but imperfectly associated with socio-economic status, and has a substantial effect on cancer rates. Thus, smoking is a confounder. The equation does not correct for the effects of smoking, and the P-value does not take this confounding into account.

FIGURE 2. Smoking as an unmeasured confounder. The non-causal association between asbestos in the water and lung cancer is explained by the associations with smoking.

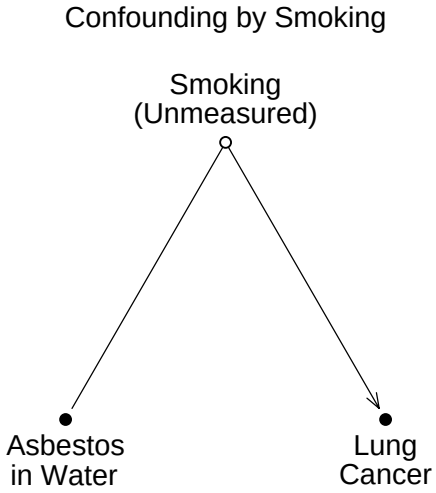


Figure 2 illustrates an alternative explanation for the data. (i) Smoking (an unmeasured confounder) is associated with the concentration of asbestos fibers in the water; indeed, both are strongly related to socio-economic variables in the tracts. The association is signaled by the straight line joining the two variables. (ii) Smoking has a strong, direct effect on lung cancer, indicated by the arrow in the figure. Associations (i) and (ii) explain the observed association between asbestos fibers in the water and lung cancer rates; this observed association is not causal. To recapitulate, a confounder is associated with the putative cause and with its effect; the confounder may explain part or all of an observed association. In epidemiology, unmeasured or poorly measured confounders are the rule rather than the exception.

Model specification. The choice of variables and functional form is somewhat arbitrary, although not completely unreasonable. The authors say that their equation is suggested by mathematical models for cancer, but the connection is rather loose; nor have the cancer models been validated (Freedman and Navidi, 1989, 1990). The models used to adjust for confounders are seldom grounded in more fundamental science.

Statistical assumptions. To compute the P-value, it is tacitly assumed that errors are statistically independent from tract to tract, and identically distributed. This assumption may be convenient, even conventional, but it lacks an empirical basis.

The search for significance. Even if we set the fundamental difficulties aside, the authors have made several hundred tests on the equations they report, without counting any preliminary data analysis that may have been done. The P-values are not adjusted for the effects of the search, which may be substantial (Dijkstra, 1988; Freedman, 1983).

Weak effects. The effect being studied is weak: a 100-fold increase in asbestos fiber concentration is associated with perhaps a 5% increase in lung cancer rates. What is unusual about the present example is only the strength of the unmeasured confounder, and the weakness of the effect under investigation.

Epidemiology is best suited to the investigation of strong effects, which are hard to explain away by confounding (Cornfield et al., 1959, p. 199). As attention shifts to the weaker and less consistent effects that may be associated with low doses, difficulties will increase. Long delays between the beginning of exposure and the onset of disease are a further complication. Toxicology may be of some value but presents difficulties of its own (Freedman, Gold, and Lin, 1996; Freedman and Zeisel, 1988). The limitations of epidemiology are discussed by Taubes (1995). For detailed case studies, see Vandenbroucke and Pardoel (1989) or Taubes (1998). Other examples will be given in section 7.

6. SOME GENERAL CONSIDERATIONS

Model specification. A model is specified by choosing (i) the explanatory variables to put on the right hand side, (ii) the functional form of the equation, and (iii) the assumptions about error terms. Explanatory variables are also called "covariates," or "independent variables"; the latter term does not connote statistical independence. The functional form may be linear, log linear, and so forth. Errors may be assumed independent or autoregressive; or some other low-order covariance matrix may be assumed, with a few parameters to estimate from the data.

Epidemiologists often have binary response variables: for instance, disease is coded as "1" and health as "0." A "logit" specification is common in such circumstances. Conditional on the covariates, subjects are assumed to be independent. If Y_i is the response for subject i while X_i is a $1 \times p$ vector of covariates, the logit specification is

$$\log \frac{\text{Prob}\{Y_i = 1\}}{\text{Prob}\{Y_i = 0\}} = X_i\beta.$$

Here, β is a $p \times 1$ vector of parameters, which would be estimated from the data by maximum likelihood. For useful details on various models and estimation procedures, see Breslow and Day (1980, 1987).

Models are chosen on the basis of familiarity and convenience; there will be some effort made to avoid gross conflict with the data. Choices are generally somewhat arbitrary, although they may not be unreasonable. There will often be some preliminary data analysis: for instance, variables with insignificant coefficients are discarded, and the model refitted. Details can make a large difference in conclusions. In particular, P-values are often strongly dependent on the final specification, and the preliminary data analysis may make these P-values difficult to interpret—as discussed below.

It is sometimes argued that biases (like recall bias or selection bias) can be modeled and then corrections can be made. That might be so if the auxiliary models could themselves be validated. On the other hand, if the auxiliary models are of doubtful validity, the "corrections" they suggest may make matters worse rather than better. For more discussion, see Scharfstein, Rotnitzky and Robins (1999) or Copas and Li (1997). In the original physical-science applications, the specifications were dictated by prior theory and empirical fact (section 4). In the social sciences and epidemiology, the specifications are much more arbitrary. That is a critical distinction.

A review of P-values. It may be enough to consider one typical example. Suppose X is a random variable, distributed as $N(\mu, 1)$, so

$$\text{Prob}\{X - \mu < x\} = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}u^2}\, du.$$

The "null hypothesis" is that $\mu = 0$; the "alternative" is that $\mu \neq 0$. The "test statistic" is $|X|$. Large values of the test statistic are evidence against the null hypothesis. For instance, a value of 2.5 for $|X|$ would be quite unusual—if the null hypothesis is correct. Such large values are therefore evidence against the null.

If x is the "observed value" of X, that is, the value realized in the data, then the P-value of the test is $\Phi(-|x|) + 1 - \Phi(|x|)$. In other words, P is the chance of getting a test statistic as large as or larger than the observed one; this chance is computed on the basis of the null hypothesis. (Sometimes, P is called the "observed significance level.") If the null hypothesis is correct, then P has a uniform distribution. Otherwise, P is more concentrated near 0. Thus, small values of P argue against the null hypothesis. If $P < .05$, the result is "statistically significant"; if $P < .01$, the result is "highly significant." These distinctions are somewhat arbitrary, but have a powerful influence on the way statistical studies are received. In this example, X is an unbiased estimate of μ. If X were biased, the bias would have to be estimated from some other data, and removed from X before proceeding with the test. P is about sampling error, not bias.

The search for significance. The effect of multiple comparisons can be seen in our example. A value of 2.5 for $|X|$ is unusual. However, if 1000 independent copies of X are examined, values of 2.5 or larger are to be expected. If only the large values are noticed and the search effort is ignored when computing P, severe distortion can result. Disease clusters attributed to environmental pollution may present such analytical problems. There are many groups of people, many sources of pollution, many possible routes of exposure, and many possible health effects. Great care is needed to distinguish real effects from the effects of chance. In this context, the search effort may not be apparent, because a cluster—especially of a rare disease—can be quite salient.

The difficulty is not widely appreciated, so another example may be useful. A coin that lands heads 10 times in a row is unusual. On the other hand, if a coin is tossed 1000 times and there is at least one run of 10 heads, that is only to be expected. The latter model may be more relevant for a disease cluster, given the number of possibilities open to examination.

If adjustment for confounding is done by regression and various specifications are tried, chance capitalization again comes into play. For some empirical evidence, see Ottenbacher (1998) or Dickersin (1997). Many epidemiologists deny that problems are created by the search for significance. Some commentators are more concerned with loss of power than distortions of the P-value, because they are convinced *a priori* that the null hypothesis is untenable. Of course, it is then unclear why statistical testing and P are relevant. See, for instance, Rothman (1990) or Perneger (1998). On the other hand, Rothman's preference for estimation over testing in the epidemiologic context often seems justified, especially when there is an effect to be estimated. For more discussion, see Cox (1977, section 5); also see section 9 below.

Intermediate variables. If X and Y cause Z, but X also causes Y, the variable Y would often be treated as an "intermediate variable" along the pathway from X to Z, rather than a confounder. If the object is to estimate the total effect of X on Z, then controlling for Y is usually not advised. If the idea to estimate the direct effect of X on Z, then controlling for Y may be advised, but the matter can under some circumstances be quite delicate. See Greenland, Pearl, and Robins (1998).

7. OTHER EXAMPLES IN EPIDEMIOLOGY

This section provides more examples in epidemiology. Generally, the studies mentioned are unpersuasive, for one or more of the following reasons.

Effects are weak and inconsistent.
Endpoints are poorly defined.
There is an extensive search for statistical significance.
Important confounders are ignored.

When effects are weak or inconsistent, chance capitalization and confounding are persistent issues; poorly-defined endpoints lend themselves to a search for significance. These problems are particularly acute when studying clusters. However, the section ends on a somewhat positive note. After numerous false starts, epidemiology and molecular biology have identified the probable etiologic agent in cervical cancer.

Leukemias and sarcomas associated with exposure to electromagnetic fields. Many studies find a weak correlation between exposure to electromagnetic fields and a carcinogenic response. However, different studies find different responses in terms of tissue affected. Nor is there much consistency in measurement of dose, which would in any event be quite difficult. Some investigators try to measure dose directly, some use distance from power lines, some use "wire codes," which are summary measures of distance from transmission lines of different types. Some consider exposure to household appliances like electric blankets or microwave ovens, while some do not. The National Research Council (1997) reviewed the studies and concluded there was little evidence for a causal effect. However, those who believe in the effect continue to press their case.

Air pollution. Some investigators find an effect of air pollution on mortality rates: see Pope, Schwartz, and Ransom (1992). Styer et al. (1995) use similar data and a similar modeling strategy, but find weak or inconsistent effects; also see Gamble (1998). Estimates of risk may be determined largely by unverifiable modeling assumptions rather than data.

Sellafield. There was a leukemia cluster associated with the British nuclear facility at Sellafield. Fathers working in the facility were exposed to radiation, which was said to have damaged the sperm and caused cancer in the child after conception—the "paternal preconception irradiation" hypothesis. Two of the Sellafield leukemia victims filed suit. There was a trial with discovery and cross examination of expert witnesses, which gives a special perspective on the epidemiology. As it turned out, the leukemia cluster had been discovered by reporters. The nature and intensity of the search is unknown; P-values were not adjusted for multiple comparisons. The effects of news stories on subsequent responses to medical interviews must also be a concern. The epidemiologists who investigated the

cluster used a case-control design, but changed the definitions of cases and controls part way through the study. For such reasons among others, causation does not seem to have been demonstrated. The judge found that

> "the scales tilt decisively in favour of the defendants and the plaintiffs therefore have failed to satisfy me on the balance of probabilities that [paternal preconception irradiation] was a material contributory cause of the [Sellafield] excess. . . . [p. 209]"

The cases are Reay and Hope v. British Nuclear Fuels, 1990 R No 860, 1989 H No 3689. Sellafield is also referred to as Seascale or Windscale in the opinion, written by the Hon. Mr. Justice French of the Queen's Bench. The epidemiology is reported by Gardner et al. (1990) and Gardner (1992); also see Doll, Evans, and Darby (1994). Case-control studies will be discussed again in the next section. Chance capitalization is not a fully satisfactory explanation for the Sellafield excess. Some epidemiologists think that leukemia clusters around nuclear plants may be a real effect, caused by exposure of previously isolated populations to viruses carried by immigrants from major population centers; this hypothesis was first put forward in another context by Kinlen and John (1994).

Cervical cancer. This cancer has been studied for many years. Some investigators have identified the cause as tissue irritation; others point to syphilis, or herpes, or chlamydia; still others have found circumcision of the husband to be protective. See Gagnon (1950), Røjel (1953), Aurelian et al. (1973), Hakama et al. (1993), or Wynder et al. (1954). Today, it is believed that cervical cancer is in large part a sexually transmitted disease, the agent being certain types of human papillomavirus, or HPV. There is suggestive evidence for this proposition from epidemiology and from clinical practice, as well as quite strong evidence from molecular biology. If so, the earlier investigators were misled by confounding. For example, the women with herpes were presumably more active sexually, and more likely to be exposed to HPV. The two exposures are associated, but it is HPV that is causal. For reviews, see Storey et al. (1998) or Cannistra and Niloff (1996). The history is discussed by Evans (1993, pp. 101–105); some of the papers are reprinted by Buck et al. (1989).

8. HEALTH EFFECTS OF SMOKING

In the 1920s, physicians noticed a rapid increase of death rates from lung cancer. For many years, it was debated whether the increase was real or an artifact of improvement in diagnostics. (The lungs are inaccessible, and diagnosis is not easy.) By the 1940s, there was some agreement on the reality of the increase, and the focus of the discussion shifted. What was the cause of the epidemic? Smoking was one theory. However, other experts thought that emissions from gas works were the cause. Still others believed that fumes from the tarring of roads were responsible.

Two early papers on smoking and lung cancer were Lombard and Doering (1928) and Müller (1939). Later papers attracted more attention, especially Wynder and Graham (1950) in the US and Doll and Hill (1950, 1952) in the UK. I will focus on the last, which reports on a "hospital-based case-control study." Cases were patients admitted to certain hospitals with a diagnosis of lung cancer; the controls were patients admitted for other reasons. Patients were interviewed about their exposure to cigarettes, emissions from

gas works, fumes from tarring of the roads, and various other possible etiologic agents. Interviewing was done "blind," by persons unaware of the purpose of the study. The cases and controls turned out to have rather similar exposures to suspect agents—except for smoking. Data on that exposure are shown in Table 2.

TABLE 2. Hospital-based case-control study. Smoking status for cases and controls. Doll and Hill (1952).

	Cases	Controls
Smoker	1350	1296
Nonsmoker	7	61

There were 1357 cases in the study, of whom 1350 were smokers; there were 1357 controls, of whom 1296 were smokers. In both groups, non-smokers are rare; but they are much rarer among the controls. To summarize such data, epidemiologists use the "odds ratio,"

$$\frac{1350/7}{1296/61} \approx 9.$$

Roughly speaking, lung cancer is 9 times more common among smokers than among non-smokers. (Doll and Hill matched their cases and controls, a subtlety that will be ignored here.) Interestingly enough, there some cases where the diagnosis of lung cancer turned out to be wrong; these cases smoked at the same rate as the controls—an unexpected test confirming the smoking hypothesis.

The odds ratio is a useful descriptive statistic on its own. However, there is a conventional way of doing statistical inference in this setting, which leads to confidence intervals and P-values. The basic assumption is that the cases are a random sample from the population of lung cancer cases, while the controls are a random sample (with a different sampling fraction) from the part of the population that is free of the disease. The odds ratio in the data would then estimate the odds ratio in the population.

TABLE 3. A 2×2 table for the population, classified according to presence or absence of lung cancer and smoking habit: a is the number of smokers with lung cancer, b is the number of smokers free of the disease, and so forth.

	Lung cancer	No lung cancer
Smoker	a	b
Nonsmoker	c	d

More explicitly, the population can be classified in a 2×2 table, as in Table 3, where a is the number who smoke and have lung cancer; b is the number who smoke but do not have lung cancer; similarly for c and d. Suppose the lung cancer patients in hospital are sampled at the rate ϕ from the corresponding part of the population, while the controls

are sampled at the rate ψ from the remainder of the population. With a large number of patients, the odds ratio in the study is essentially

$$\frac{(\phi a)/(\phi c)}{(\psi b)/(\psi d)} = \frac{a/c}{b/d} = \frac{a/b}{c/d}.$$

See Cornfield (1951). Since lung cancer is a rare disease even among smokers, $a/b \approx a/(a+b)$ is essentially the rate of disease among smokers, while $c/d \approx c/(c+d)$ approximates the rate among nonsmokers, and the odds ratio nearly coincides with the rate ratio. Moreover, standard errors and the like can be computed on the basis of the sampling model. For details, see Breslow and Day (1980).

The realism of the model, of course, is open to serious doubt: patients are not hospitalized at random. This limits the usefulness of confidence intervals and P-values. Scientifically, the strength of the case against smoking rests not so much on the P-values, but more on the size of the effect, its coherence, and on extensive replication both with the original research design and with many other designs. Replication guards against chance capitalization and, at least to some extent, against confounding—if there is some variation in study design (Cornfield et al., 1959; Ehrenberg and Bound, 1993).

For instance, Doll and Hill (1954) began a "cohort study," where British doctors were followed over time and mortality rates were studied in relation to smoking habit. At this point, it became clear that the smokers were dying at much faster rates than the non-smokers, not only from lung cancer but from many other diseases, notably coronary heart disease. It also became clear that the odds ratio computed from Table 2 was biased downward, because patients in a hospital are more likely to be smokers than the general population.

The results of the studies on smoking are generally coherent in the following ways. (i) There is a dose-response relationship: persons who smoke more heavily have greater risks of disease than those who smoke less. (ii) The risk from smoking increases with the duration of exposure. (iii) Among those who quit smoking, excess risk decreases with time after exposure stopped. These considerations are systematized to some degree by "Hill's postulates:" see Evans (1993, pp. 186ff). Of course, the data are not free of all difficulties. Notably, inhalation increases the risk of lung cancer only in some of the studies.

There was resistance to the idea that cigarettes could kill. The list of critics was formidable, including Berkson (1955) and Fisher (1959); for a summary of Fisher's arguments, see Cook (1980). The epidemiologists made an enormous effort to answer criticisms and to control for possible confounders that were suggested. To take only one example, Fisher advanced the "constitutional hypothesis" that there was a genetic predisposition to smoke and to have lung cancer: genotype is the confounder. If so, there is no point in giving up cigarettes, because the risk comes from the genes not the smoke. To refute Fisher, the epidemiologists studied monozygotic twins. The practical difficulties are considerable, because we need twin pairs where one smokes and the other does not; furthermore, at least one of the twins must have died from the disease of interest. Monozygotic twins are scarce, smoking-discordant twin pairs scarcer yet. And lung cancer is a very rare disease, even among heavy smokers.

Data from the Finnish twin study (Kaprio and Koskenvuo, 1989) are shown in Table 4. For example, there were 22 smoking-discordant monozygotic twin pairs where at least one

twin died. In 17 out of 22 cases, the smoker died first. Likewise, there were 9 cases where at least one twin in the pair died of coronary heart disease. In each case, the smoker won the race to death. For all-cause mortality or coronary heart disease, the constitutional hypothesis no longer seems viable. For lung cancer, the numbers are tiny. Of course, other studies could be brought into play (Carmelli and Page, 1996). The epidemiologists refuted Fisher by designing appropriate studies and collecting the relevant data, not by *a priori* arguments and modeling. For other views, see Bross (1960) or Stolley (1991).

TABLE 4. The Finnish twin study. First death by smoking status among smoking-discordant twin pairs. Kaprio and Koskenvuo (1989).

	Smokers	Non-smokers
All causes	17	5
Coronary heart disease	9	0
Lung cancer	2	0

Figure 3 shows current data from the US, with age-standardized death rates for the six most common cancers among males. Cancer is a disease of old age and the population has been getting steadily older, so standardization is essential. In brief, 1970 was chosen as a reference population. To get the standardized rates, death rates for each kind of cancer and each age group in each year are applied to the reference population.

Mathematically, the standardized death rate from cancer of type j in year t is

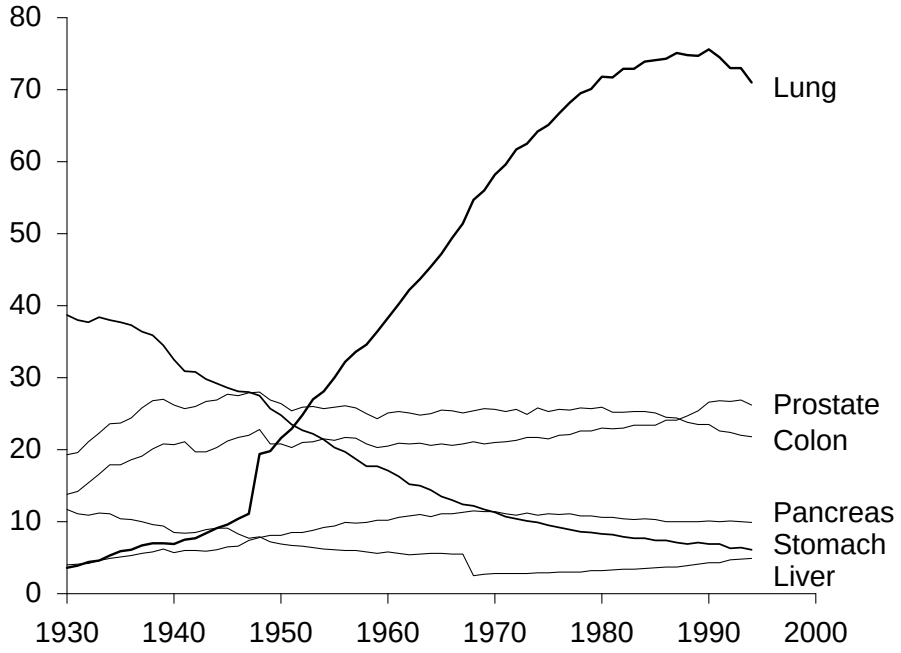
$$\sum_i n_i d_{ijt} / \sum_i n_i,$$

where n_i is the number of men in age group i in the 1970 population, and d_{ijt} is the death rate from cancer of type j among men in age group i in the population corresponding to year t. That is "direct standardization."

As will be seen, over the period 1930–1980, there is a spectacular increase in lung cancer rates. This seems to have followed by about 20 or 25 years the increase in cigarette smoking. The death rate from lung cancer starts turning down in the late 1980s, because cigarette smoking began to decrease in the late 1960s. Women started smoking later than men, and continued longer: their graph (not shown) is lower, and still rising. The data on US cigarette consumption are perhaps not quite as solid as one might like; for English data, which tell a very similar story, see Doll (1987) and Wald and Nicolaides-Bouman (1991). The initial segment of the lung cancer curve in Figure 3 was one of the first clues in the epidemiology of smoking. The downturn in the 1980s is one of the final arguments on the smoking hypothesis.

The strength of the case rests on the size and coherence of the effects, the design of the underlying epidemiologic studies, and on replication in many contexts. Great care was taken to exclude alternative explanations for the findings. Even so, the argument depends on a complex interplay among many lines of evidence. Regression models are peripheral to the enterprise. Cornfield et al. (1959) provides an interesting review of the evidence in its early stages. A summary of more recent evidence will be found in IARC (1986). Gail (1996) discusses the history.

FIGURE 3. Age-Standardized Cancer Death Rates for Males, 1930–94. Per 100,000. US Vital Statistics.



Notes: Reprinted by the permission of the American Cancer Society, Inc. Figure is redrawn from American Cancer Society (1997), using data kindly provided by the ACS. According to the ACS, "Due to changes in ICD coding, numerator information has changed over time. Rates for cancers of the liver, lung, and colon and rectum are affected by these coding changes. Denominator information for the years 1930–1959 and 1991–1993 is based on intercensal population estimates, while denominator information for the years 1960–1989 is based on postcensal recalculation of estimates. Rate estimates for 1968–1989 are most likely of better quality."

9. OTHER VIEWS

According to my near-namesake Friedman (1953, p. 15), "the relevant question to ask about the 'assumptions' of a theory is not whether they are descriptively 'realistic,' for they never are, but .... whether the theory works, which means whether it yields sufficiently accurate predictions." This argument is often used by proponents of modeling. However, the central question has been begged: how do we know whether the model is making good predictions? Fitting an equation to an existing data set is one activity; predicting the results of an intervention is quite another, and the crucial issue is getting from here to there. If regression models were generally successful in making causal inferences from associational data, that would be compelling evidence. In my experience, however, those who repeat Friedman's argument are seldom willing to engage in detailed discussions of the track record. Their reluctance is understandable.

According to Bross (1960, p. 394), "a critic who objects to a bias in the design [of a study] or a failure to control some established factor is, in fact, raising a counterhypothesis

. . . . [and] has the responsibility for showing that his counterhypothesis is tenable. In doing so, he operates under the same ground rules as [the] proponent." Also see Blau and Duncan (1967, p. 175). There is some merit to this point. Critics, like others, have an obligation to be reasonable. However, the argument is often used to shift the burden of proof from the proponent of a theory to the critic. That is perverse. Snow and his peers sought to carry the burden of proof, not to shift it. That is why their discoveries have stood the test of time.

According to some observers, regression models can be misleading in attempts to identify causes; once causation has been established, the models can be used to quantify the magnitudes of the effects. I agree, although quantification is by no means straightforward. It is not only causation that must be established, but also the specification of the model, including the identification of the principal confounders, and the form of the equation connecting the relevant factors to the outcomes of interest (section 6). The number of successes under this heading is not large.

Ken Rothman and others have expressed a preference for confidence intervals over hypothesis testing. There have been objections, on the grounds that the two forms of inference are isomorphic. These objections miss the point. The isomorphism can tell us how translate one set of mathematical theorems into another, but can scarcely dictate the form of an empirical research question. An investigator may be interested in a point estimate for some parameter, and may also want a measure of the uncertainty due to random error. For such an investigator, testing a sharp null hypothesis may be irrelevant. That would lead to confidence intervals, not P-values. Such an investigator, of course, would not care whether the confidence interval just misses—or just covers—some critical value, like 1.0 for an odds ratio. To justify his position, Rothman makes two arguments: (i) fixed-level significance testing often creates artificial dichotomies; (ii) practitioners find it easier to misinterpret P-values than point estimates. For more discussion, see Rothman (1996), Freedman, Pisani, and Purves (1997, chapter 29), Lang, Rothman, and Cann (1998), or Rothman and Greenland (1998, pp. 183–94); the last has further references to the literature.

10. SUMMARY AND CONCLUSIONS

Statisticians generally prefer to make causal inferences from randomized controlled experiments, using the techniques developed by Fisher and Neyman. In many situations, of course, experiments are impractical or unethical. Most of what we know about causation in such contexts is derived from observational studies. Sometimes, these are analyzed by regression models; sometimes, these are treated as natural experiments, perhaps after conditioning on covariates. Delicate judgments are required in order to assess the probable impact of confounders (measured and unmeasured), other sources of bias, and the adequacy of the statistical models used to make adjustments. There is much room for error in this enterprise, and much room for legitimate disagreement.

Snow's work on cholera, among other examples, shows that sound causal inferences can be drawn from non-experimental data. On the one hand, no mechanical rules can be laid down for making such inferences. Since Hume's day, that is almost a truism. On

the other hand, an enormous investment of skill, intelligence, and hard work seems to be a requirement. Many convergent lines of evidence must be developed. Natural variation needs to be identified and exploited. Data must be collected. Confounders need to be considered. Alternative explanations have to be exhaustively tested. Above all, the right question needs to be framed.

Naturally, there is a strong desire to substitute intellectual capital for labor. That is why investigators often try to base causal inference on statistical models. With this approach, P-values play a crucial role. The technology is relatively easy to use, and promises to open a wide variety of questions to the research effort. However, the appearance of methodological rigor can be deceptive. Like confidence intervals, P-values generally deal with the problem of sampling error not the problem of bias. Even with sampling error, artifactual results are likely if there is any kind of search over possible specifications for a model, or different definitions of exposure and disease. Models may be used in efforts to adjust for confounding and other sources of bias, but many somewhat arbitrary choices are made. Which variables to enter in the equation? What functional form to use? What assumptions to make about error terms? These choices are seldom dictated either by data or prior scientific knowledge. That is why judgment is so critical, the opportunity for error so large, and the number of successful applications so limited.

ACKNOWLEDGMENTS

I would like to thank the editor and the anonymous referees for useful comments; thanks also to Mike Finkelstein, Mark Hansen, Ben King, Erich Lehmann, Roger Purves, Ken Rothman, and Terry Speed. This paper is based on a lecture I gave at the Académie des Sciences, Paris in 1998, and my Wald Lectures in Dallas later that year. Sections 3 and 5 are adapted from Freedman (1991); section 4, from Freedman (1997).

REFERENCES

Abbott, A. (1997). Of time and space: the contemporary relevance of the Chicago school. *Social Forces* 75 1149–82.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J. D. (1994). *Molecular Biology of the Cell,* 3rd. ed., Garland Publishing, New York.

American Cancer Society (1997). Cancer Facts & Figures—1997. Atlanta, Georgia.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91 444–72.

Aurelian, L., Schumann, B., Marcus, R. L. and Davis, H. J. (1973). Antibody to HSV-2 induced tumor specific antigens in serums from patients with with cervical carcinoma. *Science* 181 161–64.

Berkson, J. (1955). The statistical study of association between smoking and lung cancer. *Proc. Mayo Clinic* 30 319–48.

Blau, P. M. and Duncan, O. D. (1967). *The American Occupational Structure.* Wiley, New York. Chapter 5.

Breslow, N. and Day, N. E. (1980). *Statistical Methods in Cancer Research,* Vol. 1. International Agency for Research on Cancer, Lyon. Sci. Publ. No. 32. Distributed by Oxford University Press.

Breslow, N. and Day, N. E. (1987). *Statistical Methods in Cancer Research,* Vol. 2, International Agency for Research on Cancer, Lyon. Sci. Publ. No. 82. Distributed by Oxford University Press.

Bross, I. D. J. (1960). Statistical criticism. *Cancer* 13 394–400.

Buck, C., Llopis, A., Nájera, E. and Terris, M., eds. (1989). *The Challenge of Epidemiology: Issues and Selected Readings,* Sci. Publ. No. 505, World Health Organization, Geneva.

Cannistra, S. A. and Niloff, J. M. (1996). Cancer of the uterine cervix. *New Engl. J. Med.* 334 1030–38.

Carmelli, D. and Page, W. F. (1996). Twenty-four year mortality in World War II US male veteran twins discordant for cigarette smoking. *Int. J. Epidemiol.* 25 554–559.

Carpenter, K. J. (1981). *Pellagra,* Academic Press.

Colwell, R. R. (1996). Global climate and infectious disease: the cholera paradigm. *Science* 274 2025–31.

Cook, D. (1980). Smoking and lung cancer. In S. E. Fienberg and D. V. Hinkley, eds. *R. A. Fisher, An Appreciation.* Lecture notes in statistics, Vol. 1, pp. 182–91, Springer-Verlag, New York.

Copas, J. B. and Li, H. G. (1997). Inference for non-random samples. *J. Roy. Statist. Soc. Ser. B* 59 55–77.

Cornfield, J. (1951). A method for estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *J. Nat. Cancer Inst.* 11 1269–75.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* 22 173-203.

Cox, D. (1977). The role of significance tests. *Scand. J. Statist.* 4 49–70.

Desrosières, A. (1993). *La politique des grands nombres: histoire de la raison statistique.* Editions La Découverte, Paris. English translation by C. Naish (1998). *The Politics of Large Numbers: A History of Statistical Reasoning.* Harvard University Press.

Diaconis, P. (1998). A place for philosophy: the rise of modeling in statistical science. *Quarterly of Applied Mathematics* LVI 797–805.

Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *AIDS Education and Prevention* 9 Suppl. A 15–21.

Dijkstra, T. K., ed. (1988). *On Model Uncertainty and its Statistical Implications.* Lecture Notes No. 307 in Economics and Mathematical Systems, Springer.

Doll, R. (1987). Major epidemics of the 20th century: from coronary thrombosis to AIDS. *J. Roy. Statist. Soc. Ser. A* 150 373–95.

Doll, R., Evans, H. J. and Darby, S. C. (1994). Paternal exposure not to blame. *Nature* 367 678–80.

Doll, R. and Hill, A. B. (1950). Smoking and carcinoma of the lung: preliminary report. *Br. Med. J.* ii 1271–86.

Doll, R. and Hill, A. B. (1952). A study of the aetiology of carcinoma of the lung. *Br. Med. J.* ii 1271–86.

Doll, R. and Hill, A. B. (1954). The mortality of doctors in relation to their smoking habit: A preliminary report. *Br. Med. J.* i 1451–55.

Dubos, R. (1988). *Pasteur and Modern Science.* Springer.

Ehrenberg, A. S. C. and Bound, J. A. (1993). Predictability and prediction. *J. Roy. Statist. Soc. Ser. A* 156 Part 2 167–206 (with discussion).

Evans, A. S. (1993). *Causation and Disease: A Chronological Journey.* Plenum, New York.

Evans, R. J. (1987). *Death in Hamburg: Society and Politics in the Cholera Years.* Oxford University Press.

Finlay, B. B., Heffron, F. and Fialkow, S. (1989). Epithelial cell surfaces induce *Salmonella* proteins required for bacterial adherence and invasion. *Science* 243 940–42.

Fisher, R. A. (1959). *Smoking: The Cancer Controversy.* Oliver and Boyd, Edinburgh.

Freedman, D. (1983). A note on screening regression equations. *Amer. Statistician* 37 152–55.

Freedman, D. (1987). As others see us: a case study in path analysis. *J. Educational Statistics* 12 101–223.

Freedman, D. (1991). Statistical models and shoe leather. In P. Marsden, ed., *Sociol. Methodol.*.

Freedman, D. (1995). Some issues in the foundation of statistics. *Foundations of Science* 1 19–83.

Freedman, D. (1997). From association to causation via regression. *Adv. Appl. Math.* 18 59–110.

Freedman, D., Gold, L. S. and Lin, T. H. (1996). Concordance between rats and mice in bioassays for carcinogenesis. *Reg. Tox. Pharmacol.* 23 225–32.

Freedman, D. and Navidi, W. (1989). On the multistage model for carcinogenesis. *Environ. Health Perspect.* 81 169-88.

Freedman, D. and Navidi, W. (1990). Ex-smokers and the multistage model for lung cancer. *Epidemiol.* 1 21–29.

Freedman, D., Pisani, R., and Purves, R. (1997). *Statistics.* 3rd ed. Norton, New York.

Freedman, D. and Zeisel, H. (1988). From mouse to man: the quantitative assessment of cancer risks. *Statistical Science* 3 3–56 (with discussion).

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

Gagnon, F. (1950). Contribution to the study of the etiology and prevention of cancer of the cervix. *Amer. J. Obstetrics and Gynecology* 60 516–22.

Gail, M. H. (1996). Statistics in action. *J. Amer. Statist. Assoc.* 433 1–13.

Gamble, J. F. (1998). PM_{25} and mortality in long-term prospective cohort studies: cause-effect or statistical associations? *Environ. Health Perspect.* 106 535–49.

Gardner, M. J., Snee, M. P., Hall, A. J., Powell, C. A., Downes, S. and Terrell, J. D. (1990). Results of case-control study of leukaemia and lymphoma among young people near Sellafield nuclear plant in West Cumbria. *Br. Med. J.* 300 423–33. Published erratum appears in BMJ 1992 305 715, and see letter in BMJ 1991 302 907.

Gardner, M. J. (1992). Leukemia in children and paternal radiation exposure at the Sellafield nuclear site. *Mon. Nat. Cancer Inst.* 12 133–35.

Gauss, C. F. (1809). *Theoria Motus Corporum Coelestium.* Perthes et Besser, Hamburg. Reprinted in 1963 by Dover, New York.

Gavarret, J. (1840). *Principes généraux de statistique médicale, ou, Développement des règles qui doivent présider à son emploi.* Bechet jeune et Labe, Paris.

Goldthorpe, J. H. (1998). *Causation, Statistics and Sociology.* Twenthy-ninth Geary Lecture, Nuffield College, Oxford. Publ. by the Economic and Social Research Institute, Dublin, Ireland.

Greenland, S., Pearl, J., and Robins, J. M. (1998). Causal diagrams for epidemiologic research. *Epidemiol.,* to appear.

Hakama, M., Lehtinen, M., Knekt, P., Aromaa, A., Leinikki, P., Miettinen, A., Paavonen, J., Peto, R. and Teppo, L. (1993). Serum antibodies and subsequent cervical neoplasms: A prospective study with 12 years of followup. *Amer. J. Epidemiol.* 137 166–70.

Hodges, J. L. and Lehmann, E. L. (1964). *Basic Concepts of Probability and Statistics.* Holden-Day, San Francisco.

Holland, P. (1988). Causal inference, path analysis, and recursive structural equations models. In C. Clogg, ed., *Sociol. Methodol.*.

Howard-Jones, N. (1975). *The Scientific Background of the International Sanitary Conferences 1851–1938.* World Health Organization, Geneva.

Humphreys, P. and Freedman, D. (1996). The Grand Leap. *Brit. J. Phil. Sci.* 47 113–123.

Humphreys, P. and Freedman, D. (1999). Are there algorithms that discover causal structure? Technical report no. 514, Department of Statistics, University of California, Berkeley. To appear in *Synthèse*.

IARC (1986). *Tobacco Smoking.* International Agency for Research on Cancer, Monograph 38, Lyon. Distributed by Oxford University Press.

Kanarek, M. S., Conforti, P. M., Jackson, L. A., Cooper, R. C., and Murchio, J. C. (1980). Asbestos in drinking water and cancer incidence in the San Francisco Bay Area. *Amer. J. Epidemiol.* 112 54–72.

Kaprio, J. and Koskenvuo, M. (1989). Twins, smoking and mortality: a 12-year prospective study of smoking-discordant twin pairs. *Social Science and Medicine* 29 1083–89.

Kinlen, L. J. and John, S. M. (1994). Wartime evacuation and mortality from childhood leukaemia in England and Wales in 1945–9. *Br. Med. J.* 309 1197–1201.

Lang, J. M., Rothman, K. J., Cann, C. I. (1998). That confounded P-value. *Epidemiology* 9 7–8.

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes.* Courcier, Paris. Reprinted in 1959 by Dover, New York.

Lieberson, S. (1985). *Making it Count.* University of California Press, Berkeley.

Liu, T. C. (1960). Under-identification, structural estimation, and forecasting. *Econometrica* 28 855-65.

Lombard, H. L. and Doering, C. R. (1928). Cancer studies in Massachusetts: Habits, characteristics and environment of individuals with and without lung cancer. *New Engl. J. Med.* 198 481–87.

Louis, P. (1835). *Recherches sur les effets de la saignée dans quelques maladies inflammatoires: et sur l'action de l'émétique et des vésicatoires dans la pneumonie.* J. B. Baillière, Paris. Reprinted by The Classics of Medicine Library, Birmingham, Alabama, 1986.

Lucas, R. E. Jr. (1976). Econometric policy evaluation: a critique. In K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*, vol. 1 of the Carnegie-Rochester Conferences on Public Policy, supplementary series to the Journal of Monetary Economics, North-Holland, Amsterdam, pp. 19–64. (With discussion.)

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

McKim, V. and Turner, S., eds. (1997). *Causality in Crisis? Proceedings of the Notre Dame Conference on Causality,* Notre Dame Press.

Mill, J. S. (1843) *A System of Logic, Ratiocinative and Inductive.* John W. Parker, London. 8th ed. reprinted by Longman, Green and Co., Ltd., London (1965). See especially Book III Chapter VIII. Reprinted in 1974 by the University of Toronto Press.

Miller, J. F., Mekalanos, J. J. and Fialkow, S. (1989). Coordinate regulation and sensory transduction in the control of bacterial virulence. *Science* 243 916–22.

Müller, F. H. (1939). Tabakmissbrauch und Lungcarcinom. *Zeitschrift fur Krebsforsuch* 49 57–84.

National Research Council (1997). *Possible Health Effects of Exposure to Residential Electric and Magnetic Fields.* National Academy of Science, Washington, DC.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczki* 10 1–51, in Polish. English translation by D. Dabrowska and T. Speed, 1990. *Statistical Science* 5 463–80.

Ottenbacher, K. J. (1998). Quantitative evaluation of multiplicity in epidemiology and public health research. *Amer. J. Epidemiol.* 147 615–19.

Paneth, N., Vinten-Johansen, P., Brody, H. and Rip, M. (1998). A rivalry of foulness: official and unofficial investigations of the London cholera epidemic of 1854. *Amer. J. Publ. Health* 88 1545–53.

Pasteur, L. (1878). *La théorie des germes et ses applications à la médecine et à la chirurgie,* lecture faite à l'Academie de Médecine le 30 avril 1878, par M. Pasteur en son nom et au nom de MM. Joubert et Chamberland, G. Masson, Paris.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* 82 689–709.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *Br. Med. J.* 316 1236–38.

Pope, C. A., Schwartz, J. and Ransom, M. R. (1992). Daily mortality and PM_{10} pollution in Utah Valley. *Archives of Environmental Health* 47 211-17.

Quetelet, A. (1835). *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale.* Bachelier, Paris

Raufman, J. P. (1998). Cholera. *Amer. J. Med.* 104 386–94.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *Amer. Sociol. Rev.* 15 351–7.

Rosenberg, C. E. (1962). *The Cholera Years.* Chicago University Press.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiol.* 1 43–46.

Rothman, K. J. (1996). Lessons from John Graunt. *Lancet* 347 37–39.

Rothman, K. J. and Greenland, S., eds. (1998). *Modern Epidemiology,* 2nd. ed. Lippincott-Raven.

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

Røjel, J. (1953). *The Interrelation between Uterine Cancer and Syphilis.* Copenhagen.

Semmelweiss, I. (1867). *The Etiology, Concept, and Prophylaxis of Childbed Fever.* Translated by K. C. Carter, University of Wisconsin Press, 1983.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *J. Amer. Statist. Assoc.* to appear.

Snow, J. (1855). *On the Mode of Communication of Cholera.* Churchill, London. Reprinted by Hafner, New York, 1965.

Stigler, S. M. (1986). *The History of Statistics.* Harvard University Press.

Stolley, P. (1991). When genius errs. *Amer. J. Epidemiol.* 133 416–25.

Storey, A., Thomas, M, Kalita, A., Harwood, C., Gardiol, D., Mantovani, F., Breuer, J., Leigh, I. M., Matlashewski, G. and Banks, L. (1998). Role of a p53 polymorphism in the development of human papillomavirus-associated cancer. *Nature* 393 229–34.

Styer, P., McMillan, N., Gao, F., Davis, J. and Sacks, J. (1995). Effect of outdoor airborne particulate matter on daily death counts. *Environ. Health Perspect.* 103 490–97.

Taubes, G. (1995). Epidemiology faces its limits. *Science* 269, 14 July 1995, pp. 164–9. Letters: 8 Sep 1995, pp. 1325–8.

Taubes, G. (1998). The (political) science of salt. *Science* 281, 14 August 1998, pp. 898–907.

Terris, M., ed. (1964). *Goldberger on Pellagra.* Louisiana State University Press.

Vandenbroucke, J. P. and Pardoel, V. P. (1989). An autopsy of epidemiologic methods: the case of 'poppers' in the early epidemic of the acquired immunodeficiency syndrome (AIDS). *Amer. J. Epidemiol.* 129 455–7; and see comments.

Wald, N. and Nicolaides-Bouman, A., eds. (1991). *UK Smoking Statistics.* 2nd ed., Oxford University Press.

Winkelstein, W. (1995). A new perspective on John Snow's communicable disease theory. *Amer. J. Epidemiol.* 142 (9 Suppl.) S3-9.

Wynder, E. L. and Graham, E. A. (1950). Tobacco smoking as a possible etiological factor in bronchogenic carcinoma: a study of six hundred and eight-four proved cases. *J. Amer. Med. Assoc.* 143 329-36

Wynder, E. L., Cornfield, J., Schroff, P. D. and Doraiswami, K. R. (1954). A study of environmental factors in carcinoma of the cervix. *American Journal of Obstetrics and Gynecology* 68 1016–52.

Yule, G. U. (1899). An investigation into the causes of changes in pauperism in England, chiefly during the last two intercensal decades. *J. Roy. Statist. Soc.* 62 249–95.