When dealing with human beings controlled experiments frequently prove to be impracticable, so for a scientific basis for our assumptions we turn to past history to reconstruct the suspected causal chain of events—and then our statistical troubles may begin, as this paper so convincingly reveals.

Philosophy of Inferences from Retrospective Studies*

HAROLD F. DORN, PH.D., F.A.P.H.A.

Chief, Biometrics Branch, National Institutes of Health, Washington, D. C.

ALL purposeful acts are based on a belief in the cause and effect relationship of events. It is difficult to conceive of existence in an environment where inferences concerning the future could not be drawn from past experience. This does not mean, however, that such inferences necessarily are correct or that the postulated cause and effect sequence of events actually exists.

The most general basis for belief in the cause and effect relationship of events is the observation that they are sequentially related in time. The first event is then thought to be the cause of the second. This belief is reinforced if the particular sequence of events is frequently observed. Until relatively recent times this method of reasoning was the principal basis of man's belief in the causal connection of events. Indeed, even today, this method is widely used. It would be exceedingly difficult to find a person who has not acted on beliefs established in this manner.

The principal error which may result from this type of reasoning is the fallacy of *post hoc, ergo propter hoc,* that is, the attribution of a causal relationship between events which are merely associated in time. The development of the experimental method provided a rational basis for detecting the existence of this fallacy and made possible the analytical study of the relationship of two or more factors under controlled conditions—much of present-day science dates from this development.

The essence of the experimental method is (a) the formulation of an hypothesis, based on existing knowledge, of the possible relationship between two or more factors; (b) the definition of the population to which the results of the experiment will be generalized; (c) testing the hypothesis under known conditions with as many as practicable of the supposititious causes of variation controlled, and (d) the formulation of conclusions based upon observation or measurement of the experimental results. Before such conclusions can be accepted as a valid verification or disproof of the original hypothesis, they must be confirmed by repetition of the original experiment. Hence, reproducibility is an integral part of the experimental method.

The controlled experiment without doubt is the most powerful method of analyzing the nature of the relationship between two or more factors which has yet been developed. It is the foundation upon which much of modern science is built. Nevertheless, there are many

* Presented before the Statistics Section of the American Public Health Association at the Eightieth Annual Meeting in Cleveland, Ohio, October 23, 1952.

natural phenomena which cannot readily be investigated in this manner. This does not mean, however, that such phenomena cannot be studied scientifically, or that there is any reason to believe that these phenomena are any less subject to the cause and effect relationship of events than are those of physics or chemistry.

There are, of course, the descriptive sciences such as paleontology, astronomy, meteorology, geology, botany, and zoology. In addition to these, the phenomena which at present seem to be least amenable to investigation by controlled experimentation are those involving human beings themselves, that is those falling in the domain of the social sciences and health and medicine.

There are many reasons why it frequently is impossible or impracticable to investigate naturally occurring phenomena by controlled experimentation. Only a few will be briefly referred to here.

In the first place, our customs and laws limit the extent to which human beings can be used in deliberate experiments which may endanger their health and well-being. Consequently, one often must await the outcome of natural uncontrolled experiments to obtain data to answer a particular question.

For example, it has been suggested that the inoculation of children against certain communicable diseases increases the likelihood that they will be attacked by paralytic poliomyelitis. It is immediately apparent that this hypothesis cannot be tested by controlled experimentation. Even if one could persuade a random group of children to be inoculated against a disease and another similar group not to be inoculated, one could not deliberately expose the two groups to poliomyelitis. As a result, the hypothesis must be tested by whatever data can be obtained from naturally occurring events.

Even in controlled experiments in biology it is often impracticable to control all possible causes of variation except the one under investigation or to make certain that a control group and a test group are identical in all characteristics which might influence the outcome of the experiment except for the factor being studied. In practice, a few of the most important factors are deliberately controlled and the possible effect of the remaining factors is equalized by assigning individuals to the test and control groups by some random method. Randomization makes possible the computation of valid estimates of sampling variation as well as being the most satisfactory method yet developed for eliminating bias in the allocation of subjects to the treated and control groups.

In many types of investigations in which human beings are involved, ethical considerations may prohibit the withholding even of a treatment the effect of which has as yet not been proved so that persons cannot be allocated at random to the control and treated groups, thus introducing biases which may becloud the results of the study. Even though a potential hazard to life or well-being may not be involved, it still may be difficult or impossible to persuade persons in administrative positions to permit randomization in the assignment of whatever is being tested. Thus, in a study of the possible effect upon health and morale of moving families from slum areas to a new housing development, administrators may feel that "practical" considerations require that the families to be moved must be selected by careful consideration of their potentiality to respond favorably to improved housing rather than by some random process.

Studies of the relationship between two characteristics of the members of a population may be complicated by a long, latent period between cause and effect. This may be illustrated by the

suggestion that smoking cigarettes may cause cancer of the lung. At first thought this appears to be a hypothesis which easily can be tested, since it is not difficult to find persons who are willing to smoke cigarettes. However, there is reason to believe that even if smoking cigarettes increases the chances of the development of lung cancer, this effect may not become manifest for as many as 20–30 years after smoking is begun. In addition, less than five per cent of persons, smokers and nonsmokers combined, who reach 20 years of age would be expected to develop cancer of the lung during their future lifetime. The inability to control exposure to other carcinogenic agents during the long interval between the time smoking is started and the time observation for cancer of the lung can be terminated; the inability to control the amount of smoking throughout this interval, in combination with the expense due to the long period of observation required, and the large number of persons which must be included due to the relatively small proportion of persons who would be expected to develop the disease, make impossible anything approaching a controlled experiment.

Faced with these and other difficulties the usual procedure is to select a group of individuals who exhibit the effect in question and then by investigation of their past history attempt to reconstruct the causal chain of events which preceded the observed effect. The individuals selected frequently are a consecutive group of patients from a single hospital. Little is known about the population of which the patients may be considered to be a sample. Thus, investigations of the relationship of smoking with lung cancer have been carried out by selecting a number of persons with cancer of the lung and interrogating each concerning his smoking habits. Another group of persons without lung cancer is selected and similarly questioned. This "control" or contrasting group usually is chosen from some convenient group of patients in the same hospital. The resulting data can be arranged as follows:

	Cancer of the Lung	*Without Cancer of the Lung*	*Total*
Smokers	A	B	A + B
Nonsmokers	C	D	C + D
Total	A + C	B + D	N

The two proportions A (A + C) and B (B + D) are compared and if A (A + C) is greater than B (B + D) it is concluded that smoking causes cancer of the lung.

The first thing to observe about this table is that it proceeds from effect to cause and not from cause to effect. The comparison is between the proportion of smokers among persons with and without lung cancer; whereas the appropriate comparison to test the hypothesis in question is the proportion of persons with cancer of the lung among smokers and nonsmokers, that is A (A + B) and C (C + D). Why is it that these latter proportions cannot be computed from these data? Primarily, due to the fact that (A + C) and (B + D) do not appear in the table in the same proportion as in the population from which the samples are selected. There are methods of correcting for this fact, but even after this is done, such data still provide an uncertain basis for causal inferences since the method of selection leaves undefined the population which the cases represent.

The second point to be noticed is that these data, as they stand, merely illustrate the relationship of two events associated in time. Suppose that the persons with cancer of the lung were largely of Scandinavian parentage, while those without cancer of the lung were largely of Italian parentage, the trait, blue eyes

and not blue eyes could be substituted for that of smoking and not smoking and a comparison of the proportions A (A + C) and B (B + D) undoubtedly would reveal that the former was considerably larger. Should we then assume that blue eyes cause lung cancer? If not, how can we differentiate between situations when it is appropriate to draw causal inferences and those in which it is not?

I know of no method of reasoning which enables one to infallibly distinguish between these two situations. Such data should be evaluated by an investigator who is thoroughly familiar with the subject, who has a strong skepticism of the possibility of interpreting events associated in time in terms of cause and effect, and who is prepared to consider the possibility that some hidden common factor may be the explanation of the observed relationship. Nevertheless, there are some precautions which may decrease the likelihood of accepting mere temporal association as evidence of a causal relationship. These may fruitfully be set forth as a series of questions to be answered before a study is initiated.

1. What is the hypothesis or hypotheses to be tested? A clear formulation of the hypothesis to be tested is essential in order to decide what subjects should be included in the study and the kinds of data to be collected.

2. What specific items of information are required to verify or disprove the hypotheses? Can these be obtained with sufficient reliability to warrant starting the investigation?

3. How would the study be conducted if it were possible to do it by controlled experimentation?

The answer to this question often not only is helpful in planning a retrospective study but also may be of assistance in determining whether a retrospective study is worth doing.

4. How shall the sample of persons included in the study be selected?

The purpose of most studies of the kind under discussion here is to formulate generalizations which will be true for a larger population than the particular individuals for whom data are available. If this is to be possible, two conditions must be satisfied. First, the data collected must be of sufficient reliability and validity to permit testing the hypothesis in question. Second, the persons studied must be representative of the population to which the generalizations drawn from the study are intended to apply. All too frequently efforts are devoted almost exclusively to attempting to satisfy the first condition with only casual attention paid to the second. A balanced solution to both conditions is essential. Even if the data for every individual in the study is of the highest reliability, these individuals must be representative of some larger population if generalizations are to be drawn. Similarly, nothing is gained by placing so much emphasis on selecting a representative sample that data of sufficient reliability to answer the hypothesis in question cannot be obtained.

There are no simple methods for determining whether or not a particular sample is representative of some larger population. In general the relationship of a sample to some population may be established in two ways: (a) the population may be defined in advance and a sample chosen by a probability sampling scheme or by purposive selection, or (b) a group of persons may be selected and then the population of which it might be a representative sample is defined by a study of the characteristics of the group and of the method of selection. It is needless here to point out the dangers of purposive selection or to enumerate the advantages of selecting a probability sample other than to point out that the latter is the only method which eliminates personal bias and permits the valid estimation of sampling errors. Unfortunately, in many studies of human populations it is necessary to compromise with the principles of probability sampling. When this is true the investigator must decide

whether this compromise will decrease the generality of the inferences drawn from the sample selected to such an extent that the study would not be worth-while.

5. What biases may arise either as a result of the way the sample was selected or of a failure to obtain information from every person in the sample?

Even with a probability sample, biases may arise due to the refusal or inability of some persons to give information, the failure to find some persons included in the sample and similar reasons. If a nonprobability sample is selected the possibility of bias must be considered even more carefully.

6. How shall a control group be chosen?

Suppose that 70 per cent of a group of persons with cancer of the lung report that they have smoked 10 or more cigarettes daily for at least 15 years. In the absence of information about the smoking habits of persons without cancer of the lung no valid conclusions can be drawn concerning the association between smoking and lung cancer. In other words, a control group is required for comparison with the experimental group. The primary function of controls is to provide a basis for evaluating both the supposed explanation of the observed effect and any alternative explanations.

This type of control group differs in some important respects from the control group in a planned experiment and perhaps might preferably be called a contrasting group. In the simplest type of planned experiment one starts with a sample of individuals from some population and divides this sample into two subsamples by some random or combination of random and systematic processes so that the two subsamples will be equally affected by extraneous factors which may influence the outcome of the experiment. One group is subjected to some condition or treatment while the other, known as the control, is not, and the experimental effect is observed or measured in both groups.

By contrast, the control group in the situation when one goes from effect to cause is a sample of persons who do not exhibit the effect in question. In the example of smoking and lung cancer, the control group would be composed of persons without cancer of the lung. It contains not only persons who have not been subjected to the condition or treatment, in this case smoking, which is suspected of being the cause of the observed effect, but also those who have been subjected to the condition or treatment but who do not show the effect in question, although some may do so in the future. This distinction is important in so far as it has a bearing upon the definition of the population from which the control group is thought to be a representative sample.

The first step in selecting a control group is to clearly define the population of which it is to be a representative sample. This is not necessarily a representative sample of the general population without the effect in question. Morbidity from cancer of the lung, for example, is known to vary with age, sex, and race. Consequently, a group of control cases obtained by taking a simple, random sample of the entire population without cancer of the lung would not furnish a very precise control to say nothing of the practical difficulties of taking it.

In general, the control group should come from a population as similar as possible to that from which the experimental group is chosen. If this is not true, differences between the two groups may, in part at least, be due to the fact that the two groups do not come from the same population. This point should be kept in mind whenever the possibility of using data from the general population to evaluate the results

obtained from an experimental group is being considered. In the smoking and lung cancer problem, the population from which the control group should be a sample is made up of persons who do not have lung cancer but who have been subjected to the same conditions relevant to the development of lung cancer, including the suspected cause, as those who have developed lung cancer. If the control group is selected entirely either from individuals who have never smoked or from individuals each of whom has smoked it is valueless as a contrasting group.

Two situations may arise when the two groups are compared. The supposed cause, for example smoking, may appear only in the group with lung cancer and be absent from the control group. Whenever this occurs, one should investigate whether or not the two groups are in fact samples from the same population before accepting the observed relationship as valid, since a clear-cut difference of this kind is infrequent. More generally, exposure to the supposed cause will be found in the histories of both groups. Evaluation of the relationship between the supposed cause, smoking, and the effect, lung cancer, is based upon the fact that a larger percentage of the lung cancer cases have ever smoked or have smoked more heavily.

This method of judging the possible association between exposure to a supposed hazard and the subsequent development of some condition breaks down when both the control and experimental group report the same degree of exposure. For example, suppose that 100 mice have been injected with a weak carcinogen which in the course of three months induces tumors in 20. A person who knows nothing about the experiment examines the history of the 20 mice with cancer and the 80 mice without cancer. Finding that each mouse in both groups has been exposed to the given agent, he has no basis for assuming that this agent is responsible for the development of cancer. The fact that the suspected agent appears with equal frequency in the history of both groups permits no conclusion concerning the existence or lack of existence of a relationship between it and the observed effect.

In investigations which proceed from effect to cause the experimental group is often chosen by some nonprobability method of sampling, such as taking the first N patients admitted to a particular hospital after a specified date. The characteristics of these persons with respect to factors which might influence the development of the effect being studied is not known until after the interviews are completed; hence, the population of which the control group should be a representative sample cannot be defined in advance. Two methods of selection of the control cases are open to the investigator: (a) matching by pairing and (b) matching without pairing. This will be discussed further in the papers by Greenberg and Cochran in this *Journal*.

7. What other hypotheses might account for the observed effect?

Before it is concluded that an observed association may represent a cause and effect relationship thorough consideration should be given to alternative explanations of the observed effect. In the smoking and lung cancer problem variation in the degree of association with duration and amount of smoking should be investigated. Persons with other forms of cancer, for example cancer of the colon, rectum, or skin could be used as a second control group. However, these should not be used as a substitute for a group of contrasting cases without cancer of any form.

Similarly, in a study of the possible association between injections and poliomyelitis the coincidence of the site of

injection and the part of the body paralyzed, as well as the relationship of the duration between injection and the onset of paralytic poliomyelitis with the known incubation period of the disease, should be analyzed. The control and the experimental group should be compared with respect to a wide variety of characteristics in order to rule out the possibility that some factor other than the one which appears most obvious may account for the apparent relationship. Finally, prior to publication, the manuscript should be given to one's severest critic.

These precautions combined with a thorough knowledge of the subject being studied, clear thinking, and strong skepticism may help to avoid some of the more obvious pitfalls in effect-to-cause investigations. In the last analysis, the validity of an experimental result can be established only by its reproducibility. Reproducibility does not necessarily establish validity, since the same mistake can be made repeatedly, but without reproducibility an experimental result becomes merely an isolated historical event and adds nothing to accumulated scientific knowledge.

Pilot Diabetes Case Finding

The West Virginia Health Department, with the help of a doctor and a nurse assigned from the U. S. Public Health Service for a year, has undertaken a pilot program in diabetes case finding. The purpose is to explore the incidence of diabetes among clinic patients of county health departments, to work out practical ways of referring previously unknown cases for proper medical care, and to find out the cost of a case-finding program.

As a preparation for this study, according to the West Virginia State Health Department's *Health Views,* the State Hygienic Laboratory has developed a practical method of preserving blood for sugar determination so that it will be satisfactory for examination even if specimens take several days to reach the laboratory.