The New England
Journal of Medicine

Volume 248 JUNE 11, 1953 Number 24

OBSERVATION AND EXPERIMENT*

A. Bradford Hill, C.B.E., D.Sc., Ph.D.†

LONDON, ENGLAND

TWO years ago, in his Cutter Lecture, one of my predecessors pointed out that the object of any science is "the accumulation of systematized verifiable knowledge," and that this is to be achieved through "observation, experiment and thought"—the last including both criticism and imagination. He then added, "the use of the experimental method has brilliant discoveries to its credit, whereas the method of observation has achieved little."[1] This dictum must surely prove, at least at first sight, more than a little disconcerting to the exponent of preventive medicine. In dealing with the characteristics of human populations, in sorting out the features of the environment that are detrimental from those that are beneficial, he does not often find it easy to experiment. The method of observation frequently plays a large part in the particular study of mankind that is his prerogative. Is it, then, quite so useless? Must he give it up as merely a time-wasting hobby?

Looking farther back in time I found that these questions had been considered, as indeed I had expected, by my statistical forebears and teachers in Great Britain. They did not perhaps have quite so pessimistic an outlook as the one I have quoted above, but they certainly did not underrate the difficulties of the observational approach or overlook the value of the experimental method. Thus, in 1924, Yule's[2] view was that the student of social facts could not experiment but had to deal with circumstances operating entirely beyond his control; he must accept records simply of what has happened. He wrote:

> The expert in public health, for example, must take the records of deaths as they occur, and endeavour as best he can to interpret, say, the varying incidence of death on different districts. Clearly this is a very difficult matter...The purpose of *experiment* is to replace these highly complex tangles of causation, [and] the more perfect the experiment—the more nearly the experimental ideal is attained—the less is the influence of disturbing causes, and the less necessary the use of statistical methods.

Greenwood[3] has a characteristic passage, which I quote in full since I believe that the part of it that has no close bearing on my present thesis will nevertheless more than bear repetition today;

> My conception of the statistical method in medicine has changed in the last 20 years; this is especially so with regard to the bearing of statistical method upon experiment. I used to see in the statistician the critic of the laboratory worker: it is a rôle which is gratifying to youthful vanity, for it is so easy to cheat oneself into the belief that the critic has some intellectual superiority over the criticised. I do not think even now that statistical criticism of laboratory investigations is useless, but I attach enormously more value to direct collaboration, the making of statistical experiments, and the permeation of statistical research with the experimental spirit.

The last words—written nearly thirty years ago—are, I suggest, the operative clause in the present setting—the permeation of statistical research with the experimental spirit. Although, as Yule said, facts must often, inevitably, be accepted as they occur, one does not have merely to accept facts as they are reported. One need not accept as final what some third party can give, or chooses to give—for example, a registrar-general or a census bureau. Such reported observations may, of course, prove to be a most valuable indicator of a problem; they may be, thereby, the starting point of research. But when the pattern of cause and effect is complicated they are often not likely to provide a solution. The methods of partial correlation, enthusiastically accepted a quarter of a century ago, no longer seem to have an "unlimited power to penetrate the secrets of nature."[4] One must go seek more facts, paying less attention to technics of handling the data and far more to the development and perfection of methods of obtaining them. In so doing one must have the experimental approach firmly in mind. In other words, can observations be made in such a way as to fulfil, as far as possible, experimental requirements?

Ancient Observation (the Cholera)

It was in this way, nearly a hundred years ago, that John Snow approached his problem, not only as an incomparable master of logical deduction from observations but also, it should be noted, as the

*The Cutter Lecture on Preventive Medicine, delivered at the Harvard School of Public Health, March 25, 1953.

†Professor of medical) statistics and director of the Department of Medical Statistics and Epidemiology, London School of Hygiene and Tropical Medicine; honorary director, Statistical Research Unit of the Medical Research Council, England.

constructor of observations. To recapitulate briefly, his opening arguments are based on vital statistics of the different areas of London. Using the deaths given in the first report of the Metropolitan Sanitary Commission (1847), he first shows the excessive mortality from cholera that in the epidemic of 1832 befell the districts supplied by the Southwark Water Works, a company that drew its water from the Thames at London Bridge and provided worse water, according to Snow, than any other in the metropolis. Even the order of precedence between a flea and a louse is sometimes, it appears, of importance. A death rate from cholera of 11 per 1000 inhabitants stands out starkly amidst the rates of 2, 3 and 4 for other districts of the city, but clearly that unenviable record might be explicable in terms of some quite different local characteristic. The evidence gives a lead but no more. The case is somewhat, but not at all convincingly, strengthened by the events of 1849. The highest mortality rates from cholera were again consistently to be found in the districts supplied by the Southwark Company (now combined with the South London Water Company to form the Southwark and Vauxhall) and also in those served by the Lambeth Company; both companies drew their water from the Thames in its most contaminated reaches. In 1853 there begins to appear reason to sit up and even to take notice. The Lambeth Company had removed its works from central London to Thames Ditton, where the river was wholly free from the sewage of the metropolis; the Southwark and Vauxhall Company continued to prescribe for its customers the mixture as before. In the 13 subdistricts served by the latter 192 persons died of cholera in the epidemic of 1853 — with 168,000 persons living the crude rate is thus 114 per 100,000. In 16 subdistricts served by both companies 182 perons died; among 301,000 living, that is a rate of 60 per 100,000. In three subdistricts of 15,000 persons served only by the Lambeth Company no deaths from cholera were reported.

So far do the statistical observations run; so far but not far enough. On that showing alone one might even hesitate to accept Snow's "very strong evidence" against the water supply. He himself was indeed of that mind, for "the question," he observed, "does not end *here*" (he had no intention of letting it end there). It was not said without reason that wherever cholera was visitant there was he in the midst. He noted that the Southwark and Vauxhall and Lambeth companies were competitors so that in some subdistricts the pipes of each went down all the streets and into nearly all the courts and alleys:

> Each Company supplies both rich and poor, both large houses and small...No fewer than 300,000 people of both sexes, of every age and occupation, and of every rank and station, from gentlefolk down to the very poor, were divided into two 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.

Here, then, was an unwitting experiment on the grandest scale, and Snow set himself to learn its results.

In 1854, with one medical man to assist him, up and down the streets, courts and alleys of South London he tramped in the summer's sun, learning for every cholera death the water supply of the household. Thus, by personal, persistent and accurate field work were the basic vital statistics infinitely strengthened. In 40,000 houses served by the Southwark and Vauxhall Company 286 fatal attacks were found in the first four weeks of the epidemic of 1854 — 71 deaths per 10,000 households; in 26,000 houses served by the Lambeth Company 14 fatal attacks were found — only 5 deaths per 10,000 households. In such a way was observation successfully added to observation to form a coherent and convincing whole.

It might be argued that Snow was lucky in having at hand a natural "experiment." Perhaps he was. But such "experiments" or, at the least, effective "controls" would not, I believe, really prove to be so rare if one invariably cast one's eyes round for them after vital statistics, or similar observations, had given an appropriate lead.

Certainly, in the famous Broad Street Pump outbreak of cholera no experiment offered. Its story is too well known to need any detailed reference here, but having brought Snow into my picture, I could not bear to pass it by wholly unsung. It is not so much for persuading the local board of guardians to remove the handle of the pump that Snow here deserves credit — though for this alone it is often paid to him. In fact either through the flight of the terrified population from the stricken area (and Snow himself says that "in less than 6 days the most afflicted streets were deserted by more than three-quarters of their inhabitants") or through natural epidemiologic causes, the outbreak had been steeply declining for five or six days before the well was thus put out of action. That "experiment" provides no useful evidence.

It is again in the field work that his strength lies: the map showing the concentration of deaths around the pump with their number diminishing greatly, or ceasing altogether, at each point where it became decidedly nearer to send to another pump; the demonstration of the escape of the inmates of the workhouse, which had its own well, and, similarly, of the 70 workmen in the brewery who knew better than to drink water — or if somehow driven to do so drew from a well within the brewery. And the striking individual histories, the most conclusive of which Sherlock Holmes might well have called "the curious case of the Hampstead widow." In the weekly return of births and deaths of September 9 published by the Registrar-General of England and

Wales there appeared the following entry: "At West End [Hampstead], on 2nd September, the widow of a percussion-cap maker, aged 59 years, diarrhoea two hours, cholera epidemica sixteen hours." (The times refer to the duration of the fatal illness, then—and again now—entered by the medical practitioner upon the certificate of cause of death.) One of the factories in Broad Street made percussion caps, but on inquiry Snow found that the widow had not been in the neighborhood for many months. However, she still preferred the water from the pump to that of the more salubrious neighborhood to which she had retired, and she commissioned a carter who drove daily between the two points to bring her a large bottle. The bottle was duly delivered on August 31. She drank of it and died two days later. A niece on a visit to her likewise drank of it. She then returned to her home in Islington, where she died of cholera. There was no cholera extant in either neighborhood.

To digress for a moment, there was at least one other person who drank of that bottle. The story here is, perhaps, less well known. The first medical officer of health for Hampstead (now one of the metropolitan boroughs of London) dictated as an old man in 1889 some recollections under the title of "The Sanitary Experiences of Charles F. J. Lord, M.R.C.S." It is now held in manuscript in the Hampstead Public Library but was privately printed for circulation among the old man's friends. There is a copy in the library of the Surgeon General in Washington under the title, "Jottings: Some experience with reflections derived through life and work in Hampstead from 1827 to 1877" (Pamphlet Vol. 3807). Lord himself died before making final corrections of the proofs. On pages 36 and 37 of the printed version the following passage is included:

> A memorable case of what we may consider an imported cause of disease happened at West End; Mrs. Eley Mother of the renowned firm "Eley Brothers" had lived in Broad Street Soho, and had drunk with glorification from a deep well there situated. On leaving London, she had a big stone bottle brought daily for the use of herself at West End. Summoned hastily to see the old lady I found her in the early stage of Cholera—remedies were unavailing, though solicitously applied in every way by a daughter and one of her sons. A consultation with the highly esteemed Dr. Farre ensued, the Patient never rallied, died that night. The cause of the disease at that time was never suspected; it was proved afterwards by the untiring investigations of Dr. Snow, that the water from the Broad Street well was contaminated and produced the disease; a sort of practical joke arose among the Teetotalers of the Broad Street district; those who stuck to the Porter especially those of the Brewery were rarely victims to the disease while those who drank the water fell fast around. I myself while attending closely on the old lady, as also was her daughter, was much troubled with Diarrhoea having unsuspiciously sipped some of the imported water. This insipient [*sic*] stage of Cholera soon passed away, in the absence of full or renewed doses.

Here, then, to return to my thesis, is a masterpiece—many persons would say *the* masterpiece—of observation and logical inference, made many years before the discovery of the vibrio of cholera. It shows—as many other examples have shown—that the highest returns can be reaped by imagination in combination with a logical and critical mind, a spice of ingenuity coupled with an eye for the simple and humdrum, and a width of vision in the pursuit of facts that is allied with an attention to detail that is almost nauseating.

Modern Observation (Rubella)

A modern example of acute observation lies in the story of rubella in pregnancy unfolded, almost a hundred years later, in Australia. Again, the story is too well known to need retelling, but it has a facet perhaps less familiar and yet of great interest to the student of public health—in other words, to the observer of group phenomena. It might well be that the congenital defects observed in Australia in the years 1938 to 1941 were something new in medicine, that the rubella epidemic was of a particular virulence, or that the virus had acquired some unusual characteristic at that time. Indeed, there is so much folklore attached to events in pregnancy that if the effects of German measles were an old phenomenon one might possibly have expected to find some old-wives' tale concerning it. I know of none in Britain. That the story was not, however, new in Australia is strongly indicated by the statistical observations marshaled by Lancaster.[8] In each of the reports on the Australian censuses of 1911, 1921 and 1933 there is a section that deals with the enumerated prevalence of blindness and deaf-mutism. The incidence of the latter is revealing; it shows a maximum in each census corresponding to persons born in the years 1896-1900.

At the census of 1911 the peak lay in the age group from ten to fourteen, and the statistician, writes Lancaster, "was inclined to ascribe the maximum to the more complete enumeration of the deaf at the school ages"; most observers would, I suspect, have taken that view. When, however, in 1921, the peak shifted to the age group twenty to twenty-four the statistician considered epidemic disease as a possible cause. He suggested that the increased incidence of deafness at certain ages might synchronize with the occurrence of such illnesses as "scarlet fever, diphtheria, measles, and whooping cough." In the report on the census of 1933 infective disease was again discussed. But the lead given by the somewhat crude vital statistics was not, it appears, followed up at the time. Lancaster himself has followed it up—in 1951 and therefore, of course, after the clinical observations of 1938-41—by examining the dates of birth of children admitted to institutions for the deaf and dumb. He finds, to take a single example, that of those admitted in New South Wales 15 were born in 1898 and 16 in 1900. For the intermediate year 1899 the figure soared to 70. Furthermore, these 70 are not evenly spread throughout the year but are concentrated in the months

of April to September. On such evidence, marshaled in detail and with skill, Lancaster concludes that "deafness has appeared in epidemic form in Australia in the past, notably among children born in 1899, 1916, 1924, 1925 and in 1938-41" and that "there is some presumptive evidence that all these epidemics, with the exception of that in 1916, were caused by antecedent epidemics of rubella." It seems so easy *now*, he rightly observes, to suggest a causal relation; it is always easy to be wise after the event. Nevertheless, there was at least a legible scrawl on the wall — additional and accurate data were there for the seeking and, once sought, offered a clear case for a carefully designed field inquiry. The combined observational and statistical approach could have won the day; it could have won it quite a long time ago.

Cancer of the Lung

This approach seems to me to be the only one possible in another matter of community concern today — the etiology of carcinoma of the lung. The starting point is as usual the national registration system. "It is sometimes asked," says Stocks,[8] "how statistics can cure disease," and he suggests that one may counter the question by another question: "how many researches which have led to real advances in Medicine would ever have been started had there not first been some statistics to suggest that here was a problem to be investigated?" In this particular instance it is, of course, admitted that skill in, and modern adjuncts to, diagnosis make more than dubious the whole gamut of changes that the system of vital statistics reveals. But there is, in my opinion, more than enough evidence to regard some of that change as real and to justify a search for a cause of a truly rising mortality in England and Wales. Aided and abetted by the Medical Research Council, Doll and I set about that search in 1947. Our aim was to make the field observations mirror an experimental design as nearly as possible. For each patient with cancer of the lung we sought a "control" patient with some other disease — a patient of the same sex, of the same age group, in the same hospital at or about the same time, but otherwise chosen at random. In other words, we sought, as in an experiment, to limit the variables. We limited them, too, not only in this way but also by employing, in history taking, only a few skilled interviewers, each armed with a prescribed set of questions. We made, of course, no frontal attack upon smoking, which in our original questionnaire formed but one section out of nine — eleven questions out of nearly fifty.

Having admitted to a questionnaire of that magnitude I shall take this opportunity to defend myself. For I have been reported as having advocated, before a conference on the application of scientific methods to industrial and service medicine, "that nobody should be subjected to more than five questions."[9] I am, indeed, in favor of shorter and brighter forms but not always to that extent. What I said on that occasion about the problems of making observations of any value, was this: "broadly speaking, of any twenty questions asked in a field survey not more than five should be put to the surveyed, and not less than fifteen should be put to the surveyor by himself before he enters the field or, indeed, ventures to look over the gate."[9] In other words, I maintained, though doubtless somewhat clumsily, that one may ask as many questions as one believes useful — so long as the ratio one to the surveyed and three to the surveyor is maintained throughout. A basic query in the latter group will be, in every case, "is this question really necessary?" It is surprising how often that will effectively keep down the number incorporated.

On the other hand the observational approach has perhaps been somewhat discredited by a too frequent failure to keep down that number, a pathetically notable lack of the critical and imaginative thought that, as Sinclair noted, must be an integral part of the scientific method, — in other words, and more briefly, too few ideas chasing too many forms. That evil is, of course, no prerogative of the United States of America, but I cannot refrain from citing from Eric Linklater's[9] delectable book (that is, to an Englishman) *Juan in America*. Even twenty-two years ago he was moved to write that "the issuing of questionnaires had become a national habit, and work was provided for many people, who might otherwise never have found employment, in dealing with such returns: that is in docketing them, tabulating, copying, indexing, cross-indexing, rearranging them, according to ethnic, religious, social, geographic and other factors, and eventually composing a monograph on them for the Library of Congress." Perhaps Americans were quicker off the mark. I would, however, warn them that we on the other side of the Atlantic are not being backward and may even overtake them in these national vices and devices.

Returning to my theme it is, of course, possible that the relative absence of nonsmokers and the relative frequency of heavy smokers that Doll and I found in our patients with cancer of the lung (and that other workers have also noted) is really a function of some other difference between the two groups. We do not ourselves, for several reasons, believe that to be so, and it is certainly worth noting that patients with pulmonary cancer and controls are remarkably alike in other characteristics that we have recorded. Nevertheless, here lies, I admit, the weakness of the observational as compared with the experimental approach. With the former we can determine the most probable explanation of a contrast in our data; given the provision that we have taken sufficient care to remove disturbing causes, that probability can be very high. But with a well de-

signed experiment it should be possible to eliminate (or allow for) nearly all disturbing causes and thus to render the interpretation of the contrast even more certain.

Yet in this particular problem what experiment can one make? We may subject mice, or other laboratory animals, to such an atmosphere of tobacco smoke that they can — like the old man in the fairy story — neither sleep nor slumber; they can neither breed nor eat. And lung cancers may or may not develop to a significant degree. What then? We may have thus strengthened the evidence, we may even have narrowed the search, but we must, I believe, invariably return to man for the final proof or proofs.

In this instance one other method of inquiry is now being applied both in the United States and the United Kingdom: a "looking-forward" investigation. Up till now investigators have taken already marked subjects — together with a control series — and have inquired into their antecedents. That has been the method not only, of course, in this particular inquiry but in many others. It is a natural approach and one likely to yield quick returns. Adult patients with peptic ulcers are questioned concerning whether they came from broken homes; those with rheumatoid arthritis are questioned on their previous shocks and ills; and the views of the victims of neurosis upon the habits of their fathers are sought. The resulting picture, the contrast between marked and unmarked, may be clear cut, and yet it may be difficult to distinguish between effects and causes, between horse and cart. Memories may well be more profound and more retentive in the "marked," and they may indeed be more highly colored — what the adult neurotic thinks of his father may not always be the truth. Even with the method at its best one can rarely hope to make a prognosis by these means, to measure the probabilities of events. But that is what is usually needed: first to observe the broken, and unbroken, home and then to record the subsequent history of its youthful inmates. That is clearly difficult to do and calls for a considerable degree of patience, which most investigators do not possess. But if the forward approach can be employed, it is, I believe, almost always the right way to go to work; in any observational inquiry its possibility should invariably be considered.

In the particular investigation that Doll and I now have under way — broadly into the deaths in the next few years of men and women on the British medical register whose smoking habits are already characterized at a defined point of time (late 1951) —it again, of course, would not follow that any association we might find between death from carcinoma of the lung (or other causes of death) and smoking habits must be a direct association. The heavy smokers may be differentiated from the light smokers in some other way, which might have some bearing on the risks of a bronchial carcinoma. We are still faced with the most probable explanation. But we may, I submit, have further narrowed the field of possible variables, of errors of omission or commission.

The Field Experiment

There is today an increasing resort to the field experiment, a district, a town, a school or a factory being used as the laboratory. It is a striking development of the present age and, if the requirements of an efficient experiment can be met, a most valuable one. But those requirements *must* be met; a poor experiment serves no purpose. Yet it seems that the very magic in its name may serve to mislead those who worship at the experimental shrine.

As an example, in a recently reported study of vaccination against influenza, the subjects for inoculation were chosen on a voluntary basis and "without any great propaganda 32.8% of the total employees involved in the Survey voluntarily requested the inoculations." This one third, self-selected group is compared with the remaining two thirds, who, like Gallio, "cared for none of those things." Of the 1148 inoculated persons 10.80 per cent were attacked by influenza, and of the 2349 remaining population 15.02 per cent. The difference is "statistically significant" with a "P of 0.00567." And yet does this ritual and do all these decimal places mean anything at all? Admittedly, the technical test says that the two groups had experiences that differed by more than one would expect to occur by chance; equally, it tells nothing else. As it stands I do not myself believe that it gives any support whatever for the author's conclusion that here is evidence "strongly in favor of the immunization of large groups in industry." Yet I have no doubt that it will be cited in the literature under the caption "it has been shown by experiment."

In my view this is not an experiment at all. Some observations have been made of the recorded incidence of "influenza" in two groups. The investigator knew (and so incidentally did the two groups) that they differed in one respect — inoculation; they may well have differed in a score of others — even, for all one is told, in such simple respects as age and sex. None of the other possible variables of importance were controlled, and it is well known that in trials of vaccines a self-selected group is most unlikely to be a representative sample of the total. Field experiments are not, unfortunately, as easy to design and carry out as all that. In this particular field — vaccination against influenza — I speak with conviction, for the Medical Research Council has during the last winter carried out some experiments in industry, — trials of methodology, I should say, as much as of vaccines. We too, of course, have had to rely upon volunteers for our basic material. There is (fortunately) no other way of setting up a trial.

But the volunteers were divided at random into two groups — an inoculated group given the influenza vaccine and an inoculated group given a dummy vaccine. We had their general consent to that procedure, but in the individual case it was unknown. It was also unknown to the medical practitioner diagnosing such illnesses as occurred — influenza, possibly influenza and other diseases. In such ways we have endeavored to equalize our groups *de novo* — to eliminate bias from the subsequent observations. Whether, having to cast our epidemic net wide, we have succeeded in obtaining accurate and comparable records from a score of factories and still more doctors remains to be seen. Such experiments involving human beings are, I repeat, not easy to carry out; they are, as a rule, costly. Yet in relation to the returns rendered they are relatively cheap. A well designed plan may in a few months, or years, forestall years or decades of indeterminate, unplanned observation.

Conclusion

There is one thread that runs — or it might be more accurate to say wanders — through this lecture. I have been unable — even if I would — to conceal my preference in preventive medicine for the experimental approach. At the same time that preference does not lead me to repudiate or even, I hope, to underrate the claims of accurate and designed observations. But I would place all the emphasis at my command upon those adjectives. In this field of preventive medicine I share, on the whole, the view regarding the curative aspects recently expressed by Platt,[10] professor of medicine in the University of Manchester. Records in clinical research are likely, he suggests, to be disappointing;

> Unless they have been kept with an end in view, as part of a planned experiment. . .Clinical experiment need not mean the subjection of patients to uncomfortable procedures of doubtful value or benefit. It means the planning of a line of action and the recording of observations designed to withstand critical analysis and give the answer to a clinical problem. It is an attitude of mind.

In appropriately exploiting that attitude of mind one may well need, in this age of technicalities, close and constant collaboration. Today, as Joseph Garland[11] pointed out in this city of Boston, "the mathematics of research has expressed itself in a multiplicity of graphs, charts and tables with the aid of which the average reader at a quick glance can often learn next to nothing." The biostatistician must therefore acquire a taste for lying down with the epidemiologist, and the bacteriologist with the medical officer of health (I speak in fables).

There are, of course, no grounds for antagonism between experiment and observation. The former, indeed, depends on observation but of a type that has the good fortune to be controlled at the experimenter's will. In the world of public health and preventive medicine each will — or should — constantly react beneficially upon the other. Observation in the field suggests experiment; the experiment leads back to more, and better defined, observations. However that may be, it is difficult to see how one can wholly, or ever, escape from Alexander Pope's epigram. How else but by observation upon man himself being born, living and dying, can one set about the solution of such problems as prematurity and stillbirth at one end of life and cancer and coronary thrombosis at the other? However tangled the skein of causation one must, at least at first, try to unravel it in vivo. As Pickering[12] has said: "Any work which seeks to elucidate the cause of disease, the mechanism of disease, the cure of disease, or the prevention of disease, must begin and end with observations on man, whatever the intermediate steps may be."

The observer may well have to be more patient than the experimenter — awaiting the occurrence of the natural succession of events he desires to study; he may well have to be more imaginative — sensing the correlations that lie below the surface of his observations; and he may well have to be more logical and less dogmatic — avoiding as the evil eye the fallacy of *post hoc ergo propter hoc*, the mistaking of correlation for causation.

Lastly, I quote the words of Professor William Topley,[13] a British worker for whom I had a profound admiration and from whose wisdom I endeavored to learn:

> A great part of clinical medicine, and of epidemiology, must still be observation. Nature makes the experiments, and we watch and understand them if we can. No one will deny that we should always aim at planned intervention and closer control. Here, as elsewhere, technique — the way we make our observations and check them — is half the battle; but to force experiment and observation into sharply separated categories is almost as dangerous a heresy as the science and art [of medicine] antithesis. It tends to make the clinician in the ward, the epidemiologist in the field, and the laboratory worker at his bench, think of themselves as doing different things, and bound by different rules. Actually they are all making experiments, some good, some bad. It is more difficult to make a good experiment in the ward than in the laboratory, because conditions are more difficult to control; but there is no other way of gaining knowledge... Controlled observation in the ward or in the field is an essential part of medical science, shading through almost imperceptible stages of increasing intervention into the fully developed experimental technique of the laboratory.

Mr. Winston Churchill, revisiting the Niagara Falls after more than forty years, was asked by a reporter "Do they look the same?" "Well", he is said to have replied, "the principle seems the same." General principles are obstinate things; they do tend to remain the same generation after generation. Yet one element of that sameness — their fundamental importance — perhaps justifies their being brought out into the light of day from time to time and, if one cannot weave fresh clothes, at least in a newly

Vol. 248 No. 24 1001

dyed costume. In accepting the honor of delivering this Cutter Lecture I indeed trusted that that was so. If I was wrong I must comfort myself like that charming character described by Anatole France: like Monsieur Bonnard, I have the satisfaction of believing that, in following my distinguished predecessors, I have at least "utilized to their fullest extent those mediocre faculties with which Nature endowed me."

References

1. Sinclair, H. M. Nutritional surveys of population groups. *New Eng. J. Med.* 245:39-47, 1951.
2. Yule, G. *The Function of Statistical Method in Scientific Investigation.* (Industrial Health Research Board Report.) No. 24. 14 pp. London: His Majesty's Stationery Office, 1924.
3. Greenwood, M. Is statistical method of any value in medical research? *Lancet* 2:153-158, 1924.
4. Tippett, L. H. C. *Statistics.* 136 pp. London: Oxford University Press, 1943.
5. Lancaster, H. O. Deafness as epidemic disease in Australia: note on census and institutional data. *Brit. M. J.* 2:1423-1432, 1951.
6. Stocks, P. *Modern Trends in Public Health.* Edited by A. Massey. 591 pp. London: Butterworth, 1949. Chap. XIX.
7. Himsworth, H. P. *The Application of Scientific Methods to Industrial and Service Medicine.* (Medical Research Council.) London: His Majesty's Stationery Office, 1924. P. 109.
8. Hill, A. B. Cited by Himsworth.[7] P. 7.
9. Linklater, E. *Juan in America.* 466 pp. London: Jonathan Cape, 1951.
10. Platt, R. Wisdom is not enough: reflections on art and science of medicine. *Lancet* 2:977-980, 1952.
11. Garland, J. *New England Journal of Medicine* and Massachusetts Medical Society. *New Eng. J. Med.* 246:801-806, 1952.
12. Pickering, G. W. Opportunity and universities. *Lancet* 2:895-898, 1952.
13. Topley, W. W. C. *Authority, Observation and Experiment in Medicine.* 46 pp. London: Cambridge University Press, 1940. P. 40.

Meeting January 14 1965

President's Address

The Environment and Disease: Association or Causation?

by Sir Austin Bradford Hill CBE DSC FRCP(hon) FRS
(*Professor Emeritus of Medical Statistics, University of London*)

Amongst the objects of this newly-founded Section of Occupational Medicine are firstly 'to provide a means, not readily afforded elsewhere, whereby physicians and surgeons with a special knowledge of the relationship between sickness and injury and conditions of work may discuss their problems, not only with each other, but also with colleagues in other fields, by holding joint meetings with other Sections of the Society'; and, secondly, 'to make available information about the physical, chemical and psychological hazards of occupation, and in particular about those that are rare or not easily recognized'.

At this first meeting of the Section and before, with however laudable intentions, we set about instructing our colleagues in other fields, it will be proper to consider a problem fundamental to our own. How in the first place do we detect these relationships between sickness, injury and conditions of work? How do we determine what are physical, chemical and psychological hazards of occupation, and in particular those that are rare and not easily recognized?

There are, of course, instances in which we can reasonably answer these questions from the general body of medical knowledge. A particular, and perhaps extreme, physical environment cannot fail to be harmful; a particular chemical is known to be toxic to man and therefore suspect on the factory floor. Sometimes, alternatively, we may be able to consider what *might* a particular environment do to man, and then see whether such consequences are indeed to be found. But more often than not we have no such guidance, no such means of proceeding; more often than not we are dependent upon our observation and enumeration of defined events for which we then seek antecedents. In other words we see that the event B is associated with the environmental feature A, that, to take a specific example, some form of respiratory illness is associated with a dust in the environment. In what circumstances can we pass from this observed *association* to a verdict of *causation*? Upon what basis should we proceed to do so?

I have no wish, nor the skill, to embark upon a philosophical discussion of the meaning of 'causation'. The 'cause' of illness may be immediate and direct, it may be remote and indirect underlying the observed association. But with the aims of occupational, and almost synonymously preventive, medicine in mind the decisive question is whether the frequency of the undesirable event B will be influenced by a change in the environmental feature A. *How* such a change exerts that influence may call for a great deal of research. However, before deducing 'causation' and taking action we shall not invariably have to sit around awaiting the results of that research. The whole chain may have to be unravelled or a few links may suffice. It will depend upon circumstances.

Disregarding then any such problem in semantics we have this situation. Our observations reveal an association between two variables, perfectly clear-cut and beyond what we would care to attribute to the play of chance. What aspects of that association should we especially consider before deciding that the most likely interpretation of it is causation?

(1) *Strength*. First upon my list I would put the strength of the association. To take a very old example, by comparing the occupations of patients with scrotal cancer with the occupations of patients presenting with other diseases, Percival Pott could reach a correct conclusion because of the *enormous* increase of scrotal cancer in the chimney sweeps. 'Even as late as the second decade of the twentieth century', writes Richard Doll (1964), 'the mortality of chimney sweeps from scrotal cancer was some 200 times that of workers who were not specially exposed to tar or mineral oils and in the eighteenth century the relative difference is likely to have been much greater.'

To take a more modern and more general example upon which I have now reflected for over fifteen years, prospective inquiries into smoking have shown that the death rate from cancer of the lung in cigarette smokers is nine to ten times the rate in non-smokers and the rate in heavy cigarette smokers is twenty to thirty times

as great. On the other hand the death rate from coronary thrombosis in smokers is no more than twice, possibly less, the death rate in non-smokers. Though there is good evidence to support causation it is surely much easier in this case to think of some features of life that may go hand-in-hand with smoking – features that might conceivably be the real underlying cause or, at the least, an important contributor, whether it be lack of exercise, nature of diet or other factors. But to explain the pronounced excess in cancer of the lung in any other environmental terms requires some feature of life so intimately linked with cigarette smoking and with the amount of smoking that such a feature should be easily detectable. If we cannot detect it or reasonably infer a specific one, then in such circumstances I think we are reasonably entitled to reject the vague contention of the armchair critic 'you can't prove it, there *may* be such a feature'.

Certainly in this situation I would reject the argument sometimes advanced that what matters is the absolute difference between the death rates of our various groups and not the ratio of one to other. That depends upon what we want to know. If we want to know how many extra deaths from cancer of the lung will take place through smoking (i.e. presuming causation), then obviously we must use the absolute differences between the death rates – 0·07 per 1,000 per year in non-smoking doctors, 0·57 in those smoking 1–14 cigarettes daily, 1·39 for 15–24 cigarettes daily and 2·27 for 25 or more daily. But it does not follow here, or in more specifically occupational problems, that this best measure of the effect upon mortality is also the best measure in relation to aetiology. In this respect the ratios of 8, 20 and 32 to 1 are far more informative. It does not, of course, follow that the differences revealed by ratios are of any practical importance. Maybe they are, maybe they are not; but that is another point altogether.

We may recall John Snow's classic analysis of the opening weeks of the cholera epidemic of 1854 (Snow 1855). The death rate that he recorded in the customers supplied with the grossly polluted water of the Southwark and Vauxhall Company was in truth quite low – 71 deaths in each 10,000 houses. What stands out vividly is the fact that the small rate is 14 times the figure of 5 deaths per 10,000 houses supplied with the sewage-free water of the rival Lambeth Company.

In thus putting emphasis upon the strength of an association we must, nevertheless, look at the obverse of the coin. We must not be too ready to dismiss a cause-and-effect hypothesis merely on the grounds that the observed association appears to be slight. There are many occasions in medicine when this is in truth so. Relatively few persons harbouring the meningococcus fall sick of meningococcal meningitis. Relatively few persons occupationally exposed to rat's urine contract Weil's disease.

(2) *Consistency:* Next on my list of features to be specially considered I would place the *consistency* of the observed association. Has it been repeatedly observed by different persons, in different places, circumstances and times?

This requirement may be of special importance for those rare hazards singled out in the Section's terms of reference. With many alert minds at work in industry today many an environmental association may be thrown up. Some of them on the customary tests of statistical significance will appear to be unlikely to be due to chance. Nevertheless whether chance is the explanation or whether a true hazard has been revealed may sometimes be answered only by a repetition of the circumstances and the observations.

Returning to my more general example, the Advisory Committee to the Surgeon-General of the United States Public Health Service found the association of smoking with cancer of the lung in 29 retrospective and 7 prospective inquiries (US Department of Health, Education & Welfare 1964). The lesson here is that broadly the same answer has been reached in quite a wide variety of situations and techniques. In other words we can justifiably infer that the association is not due to some constant error or fallacy that permeates every inquiry. And we have indeed to be on our guard against that.

Take, for instance, an example given by Heady (1958). Patients admitted to hospital for operation for peptic ulcer are questioned about recent domestic anxieties or crises that may have precipitated the acute illness. As controls, patients admitted for operation for a simple hernia are similarly quizzed. But, as Heady points out, the two groups may not be *in pari materia*. If your wife ran off with the lodger last week you still have to take your perforated ulcer to hospital without delay. But with a hernia you might prefer to stay at home for a while – to mourn (or celebrate) the event. No number of exact repetitions would remove or necessarily reveal that fallacy.

We have, therefore, the somewhat paradoxical position that the different results of a different inquiry certainly cannot be held to refute the

original evidence; yet the same results from precisely the same form of inquiry will not invariably greatly strengthen the original evidence. I would myself put a good deal of weight upon similar results reached in quite different ways, e.g. prospectively and retrospectively.

Once again looking at the obverse of the coin there will be occasions when repetition is absent or impossible and yet we should not hesitate to draw conclusions. The experience of the nickel refiners of South Wales is an outstanding example. I quote from the Alfred Watson Memorial Lecture that I gave in 1962 to the Institute of Actuaries:

'The population at risk, workers and pensioners, numbered about one thousand. During the ten years 1929 to 1938, sixteen of them had died from cancer of the lung, eleven of them had died from cancer of the nasal sinuses. At the age specific death rates of England and Wales at that time, one might have anticipated one death from cancer of the lung (to compare with the 16), and a fraction of a death from cancer of the nose (to compare with the 11). In all other bodily sites cancer had appeared on the death certificate 11 times and one would have expected it to do so 10–11 times. There had been 67 deaths from all other causes of mortality and over the ten years' period 72 would have been expected at the national death rates. Finally division of the population at risk in relation to their jobs showed that the excess of cancer of the lung and nose had fallen wholly upon the workers employed in the chemical processes.

'More recently my colleague, Dr Richard Doll, has brought this story a stage further. In the nine years 1948 to 1956 there had been, he found, 48 deaths from cancer of the lung and 13 deaths from cancer of the nose. He assessed the numbers expected at normal rates of mortality as, respectively 10 and 0·1.

'In 1923, long before any special hazard had been recognized, certain changes in the refinery took place. No case of cancer of the nose has been observed in any man who first entered the works after that year, and in these men there has been no excess of cancer of the lung. In other words, the excess in both sites is uniquely a feature in men who entered the refinery in, roughly, the first 23 years of the present century.

'No causal agent of these neoplasms has been identified. Until recently no animal experimentation had given any clue or any support to this wholly statistical evidence. Yet I wonder if any of us would hesitate to accept it as proof of a grave industrial hazard?' (Hill 1962).

In relation to my present discussion I know of no parallel investigation. We have (or certainly had) to make up our minds on a unique event; and there is no difficulty in doing so.

(3) *Specificity:* One reason, needless to say, is the specificity of the association, the third characteristic which invariably we must consider. If, as here, the association is limited to specific workers and to particular sites and types of disease and there is no association between the work and other modes of dying, then clearly that is a strong argument in favour of causation.

We must not, however, over-emphasize the importance of the characteristic. Even in my present example there is a cause and effect relationship with two different sites of cancer – the lung and the nose. Milk as a carrier of infection and, in that sense, the cause of disease can produce such a disparate galaxy as scarlet fever, diphtheria, tuberculosis, undulant fever, sore throat, dysentery and typhoid fever. Before the discovery of the underlying factor, the bacterial origin of disease, harm would have been done by pushing too firmly the need for specificity as a necessary feature before convicting the dairy.

Coming to modern times the prospective investigations of smoking and cancer of the lung have been criticized for not showing specificity – in other words the death rate of smokers is higher than the death rate of non-smokers from many causes of death (though in fact the results of Doll & Hill, 1964, do not show that). But here surely one must return to my first characteristic, the strength of the association. If other causes of death are raised 10, 20 or even 50% in smokers whereas cancer of the lung is raised 900–1,000% we have specificity – a specificity in the magnitude of the association.

We must also keep in mind that diseases may have more than one cause. It has always been possible to acquire a cancer of the scrotum without sweeping chimneys or taking to mule-spinning in Lancashire. One-to-one relationships are not frequent. Indeed I believe that multi-causation is generally more likely than single causation though possibly if we knew all the answers we might get back to a single factor.

In short, if specificity exists we may be able to draw conclusions without hesitation; if it is not apparent, we are not thereby necessarily left sitting irresolutely on the fence.

(4) *Temporality:* My fourth characteristic is the temporal relationship of the association – which is the cart and which the horse? This is a question which might be particularly relevant with diseases of slow development. Does a particular diet lead to disease or do the early stages of the disease lead to those peculiar dietetic habits? Does a

particular occupation or occupational environment promote infection by the tubercle bacillus or are the men and women who select that kind of work more liable to contract tuberculosis whatever the environment – or, indeed, have they already contracted it? This temporal problem may not arise often but it certainly needs to be remembered, particularly with selective factors at work in industry.

(5) *Biological gradient:* Fifthly, if the association is one which can reveal a biological gradient, or dose-response curve, then we should look most carefully for such evidence. For instance, the fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers. That comparison would be weakened, though not necessarily destroyed, if it depended upon, say, a much heavier death rate in light smokers and a lower rate in heavier smokers. We should then need to envisage some much more complex relationship to satisfy the cause-and-effect hypothesis. The clear dose-response curve admits of a simple explanation and obviously puts the case in a clearer light.

The same would clearly be true of an alleged dust hazard in industry. The dustier the environment the greater the incidence of disease we would expect to see. Often the difficulty is to secure some satisfactory quantitative measure of the environment which will permit us to explore this dose-response. But we should invariably seek it.

(6) *Plausibility:* It will be helpful if the causation we suspect is biologically plausible. But this is a feature I am convinced we cannot demand. What is biologically plausible depends upon the biological knowledge of the day.

To quote again from my Alfred Watson Memorial Lecture (Hill 1962), there was

'. . . no biological knowledge to support (or to refute) Pott's observation in the 18th century of the excess of cancer in chimney sweeps. It was lack of biological knowledge in the 19th that led a prize essayist writing on the value and the fallacy of statistics to conclude, amongst other "absurd" associations, that "it could be no more ridiculous for the stranger who passed the night in the steerage of an emigrant ship to ascribe the typhus, which he there contracted, to the vermin with which bodies of the sick might be infected". And coming to nearer times, in the 20th century there was no biological knowledge to support the evidence against rubella.'

In short, the association we observe may be one new to science or medicine and we must not dismiss it too light-heartedly as just too odd. As Sherlock Holmes advised Dr Watson, 'when you have eliminated the impossible, whatever remains, *however improbable*, must be the truth.'

(7) *Coherence:* On the other hand the cause-and-effect interpretation of our data should not seriously conflict with the generally known facts of the natural history and biology of the disease – in the expression of the Advisory Committee to the Surgeon-General it should have coherence.

Thus in the discussion of lung cancer the Committee finds its association with cigarette smoking coherent with the temporal rise that has taken place in the two variables over the last generation and with the sex difference in mortality – features that might well apply in an occupational problem. The known urban/rural ratio of lung cancer mortality does not detract from coherence, nor the restriction of the effect to the lung.

Personally, I regard as greatly contributing to coherence the histopathological evidence from the bronchial epithelium of smokers and the isolation from cigarette smoke of factors carcinogenic for the skin of laboratory animals. Nevertheless, while such laboratory evidence can enormously strengthen the hypothesis and, indeed, may determine the actual causative agent, the lack of such evidence cannot nullify the epidemiological observations in man. Arsenic can undoubtedly cause cancer of the skin in man but it has never been possible to demonstrate such an effect on any other animal. In a wider field John Snow's epidemiological observations on the conveyance of cholera by the water from the Broad Street pump would have been put almost beyond dispute if Robert Koch had been then around to isolate the vibrio from the baby's nappies, the well itself and the gentleman in delicate health from Brighton. Yet the fact that Koch's work was to be awaited another thirty years did not really weaken the epidemiological case though it made it more difficult to establish against the criticisms of the day – both just and unjust.

(8) *Experiment:* Occasionally it is possible to appeal to experimental, or semi-experimental, evidence. For example, because of an observed association some preventive action is taken. Does it in fact prevent? The dust in the workshop is reduced, lubricating oils are changed, persons stop smoking cigarettes. Is the frequency of the associated events affected? Here the strongest

QUESTIONS AND ANSWERS

Edited By ***ERNEST RUBIN***
U. S. Department of Commerce
and American University

Statistical Relationships and Proof in Medicine

Within the last two years a formidable controversy has developed as a result of investigations concerning smoking and lung cancer. Statistical data indicate that smokers have a higher incidence of lung cancer than non-smokers. The experience with the smoking—lung cancer controversy suggested the following question: Are there instances in the history of medicine in which a statistical association proved causation or has the proof of a causal relationship in medicine always depended on direct experimentation? I wish to thank Mr. Jerome Cornfield of the Office of Biometry, National Institutes of Health for preparing the following analysis of this question.

There are several preliminary issues raised by the question that need prior discussion. The first, the concept of proof, need not detain us long. Proof has a well-defined meaning in mathematics, but not in empirical science. The truth of a mathematical proposition can be demonstrated; the evidence for an empirical proposition, i.e., a statement in natural science, can be made strong or even overwhelming (despite the apparent impossibility of a satisfactory calculus of evidence). It is doubtful, however, if such propositions can ever be regarded as proved. New evidence (e.g., the discovery of black swans) may cast an entirely different light on a well-established proposition, and in an empirical science, as opposed to mathematics, there are no postulate systems which delimit the kind of new evidence that can be found. If we ask for proof in medicine, or any other empirical science, we may be asking for something that does not exist.

A second issue raised by the question is the exact nature of the distinction between a relationship based upon a statistical association and one based on direct experimentation. We all have a vague feeling that if we can make an event occur, we understand it better than if we simply observe it passively. On analysis this feeling seems to reduce to two propositions like the following: We are initially skeptical of any relationship based upon simple observation because the effects of other possibly important variables are not controlled and may account for the observed association. We are initially impressed by any relationship established by experiment because we feel that the effects of other important variables are controlled and cannot account for the association. The distinction we feel between a relationship based upon a statistical association and one based upon direct experimentation is thus a distinction between relationships that may be explained by other variables and those that cannot.

Although this statement may formalize our intuitions it is an oversimplification of the actual facts. First, there are cases in which uncontrolled observations can be so analysed as to eliminate the possibility that extraneous variables account for the observed association. The classical example of this is Snow's demonstration (1) in the middle of the nineteenth century, before the birth of bacteriology, that cholera was transmitted through polluted water. Even the most skeptical critic cannot quarrel with the conclusions drawn from his observations on the clustering of deaths about a particular source of polluted water, the famous Broad Street pump; particularly after his demonstration that mortality from cholera among subscribers of a water company that drew its supply from the Thames River was 14 times as high as that among subscribers of the competing company whose water was sewage-free. The official inquiry which followed agreed that "fecalized drinking-water . . . may breed and convey the poison [of cholera]" although with a caution that is perhaps not peculiar to the Victorian era added, "[so would] fecalized air."[1] Nor do we have to go back 100 years to find examples in which the effects of specific extraneous variables were eliminated from observational material by methods short of direct experimentation. Cross-classification of observations is an obvious, but often surprisingly powerful method of accomplishing this, for some recent examples of which references (3, 4, 5 and 6) may be instrucive.

Secondly, our intuitions may be misleading because there is no automatic guarantee in any particular instance that extraneous variables have been controlled by direct experimentation. This may seem to deny the

[1] It is not entirely irrelevant to recall at this point the experience of Max von Pettenkofer who, many years later, to prove beyond any doubt that water-borne bacteria did *not* cause cholera, drank, and induced several of his students to drink, a whole glass full of the bacilli. They not only all survived but reported nothing worse than a bellyache (2).

great virtue claimed for randomization, the automatic balancing out among treatment groups of the effects of other variables, whether or not we are aware of their existence. The denial is more apparent than real, however. Consider for example an experiment designed to study the effect of removing an organ, say the thyroid gland, on some biological response, say blood sugar level. We may randomize animals among a control and thyroidectomized group and thus eliminate in the usual probability sense the possibility that any large difference between the two treatments arose from the different characteristics of the animals treated. But we have not eliminated the possible effect of other extraneous variables in which the experimenter is equally interested such as the operation removing the thyroid, or the non-specific effect of thyroidectomy on weight loss. While it is perhaps possible to control these specific variables, for example, sham operation and under feeding the sham operated controls might be regarded as providing such a control (7), randomization by itself is insufficient. We must indicate the specific variables we wish to control and must devise specific experimental procedures to control them.

Having thus argued that there is no difference in kind between the two types of evidence it is of course necessary to add that there is a very important difference in degree. It is a good deal more difficult to control variables in observational than in experimental material, so that the experimental method has unravelled and will continue to unravel mysteries before which uncontrolled observation would be powerless. But there is no difference in principle. There are no such categories as first-class evidence and second-class evidence. There are merely associations, whether observational or experimental that, in a given state of knowledge, can be accounted for in only one way or in several different ways. If the latter, it is our obligation to state what the alternative explanations or variables might be and to see how their effects can be eliminated, while if the former it is equally our obligation to state so. To distinguish between statistical association on the one hand and relationships that are established by experimentation on the other, without any reference to alternative variables that are present in one case but not the other, seems to us to be neither good statistics, good science, nor good philosophy—though it may be good red herring.

If we consider the tobacco-lung cancer question, for example, one possible set of extraneous variables that might explain the higher incidence for smokers are those arising from self-selection. Thus, some small proportion of the population, say 5 percent, might have some special trait (or traits), say high blood levels of certain hormones, which both initiate lung cancer and make the possessors smoke. Of the remaining adult male population, 75 percent smoke for other reasons and do not develop lung cancer. It is possible to conceive but impossible to conduct an experiment that could settle this question. A large group of adolescents would be allocated at random to different smoking groups, compelled to remain on the assigned smoking schedule, followed for the 30 to 60 years required for lung cancer to develop and the lung cancer incidence computed for each group. This and, as nearly as one can see, only this, would entirely eliminate self-selection as an explanation. Short of this one must rely upon indirect evidence. If self-selection were the complete explanation of the difference, then tobacco smoke would not be a carcinogen for human lung tissue. One might consequently investigate this question by asking is it a carcinogen for any other type of tissue that one can reasonably experiment with, say human or mouse skin? If the answer had been no, this might have been regarded as some type of evidence for the self-selection hypothesis, although no one would regard the evidence as very strong. The recent induction of skin tumors in mice by tobacco tars (8) might similarly be considered evidence against the self-selection hypothesis, but again far from strong. In any event the recent announcement by the Tobacco Industry Research Committee that it would investigate psychological differences between smokers and nonsmokers suggests that we have not heard the last of the self-selection hypothesis. No matter what one's opinion on the plausibility of this as an explanation,[1] the actual investigation of whether specific differences that might arise from self-selection do in fact account for the association could be constructive, even if the results obtained were negative.

This discussion of preliminary issues[2] in one sense also disposes of the main question, but the history of medicine on this point is interesting and a few words may be in order. There are numerous instances that one can cite in which the most important source of medical knowledge on a subject was supplied by statistical associations. Thus, the observation that there is a close inverse association between the amount of natural fluorides in water and the amount of dental caries among children drinking it, has induced numer-

[1] It is easy to sympathize with, even if one cannot entirely share the exasperation expressed by Greenwood and Yule on a related point (9). "[The vaccinated group] may all have been vegetarians, or nonsmokers, or red-headed, and all or any of these things may render them less likely to contract cholera; but we do not see why objections which no sensible man would allow to influence him in the ordinary affairs of life should suddenly acquire scientific importance when the question is one of interpreting statistics."

[2] There is one additional preliminary issue that deserves mention. The phrasing of the question suggests that its framer subscribes to the somewhat old-fashioned view that it is either possible or desirable to discuss knowledge without any reference to the possible actions to which it will lead. I have not challenged this view only because I share it.

ous municipalities to add fluorine to their water supply. *The resulting decline in the incidence of* dental caries in these municipalities may be considered a "direct experimental proof" of the proposition that fluorine inhibits the development of caries, but the fluorine was not added in order to study this question experimentally but rather to bring about a result indicated by the associations. It is true that the results of adding fluorine to the drinking water of experimental animals also pointed in the same direction (10), but as evidence this apparently was not given much weight. Shaw, for example, in his excellent summary of the subject (11) does not even mention the results with experimental animals.

In the study of the effects of therapy the application of modern ideas of experimental design is a very recent development, for an account of which the reader is referred to Hill's very interesting article (12). In recent years there have been several well-conceived experiments to test the efficacy of different preparations, such as gamma globulin, in protecting against the subsequent development of disease. But methods that were established in the past such as vaccination against smallpox, have never received such a carefully controlled experimental test. There are of course dozens of studies to show that individuals who had been voluntarily vaccinated developed less smallpox than others, and that when they did develop it, the outcome was less frequently fatal. But none of these studies ruled out the possibilities of self-selection any more effectively than they are now ruled out in tobacco-lung cancer studies.

In the study of infectious disease there are naturally almost no examples of direct experimental demonstration on humans. Walter Reed's experiments on yellow fever are well known, but it is difficult to find other cases. Perhaps the nearest is the ghastly episode that occurred in Lübeck in 1920, when out of 249 babies accidentally inoculated with enormous numbers of living virulent tubercle bacilli, 76 died (13). If one is willing to overlook the absence of a placebo-inoculated control group, and refrains from asking, "if the bacilli cause tuberculosis, why didn't all the inoculated children develop the disease?", this perhaps is "proof of a causal relationship." (The 173 Lübeck babies who did not die developed only minor lesions and were still free of tuberculosis when last observed 12 years later.)

In short, if we insist on direct experimental demonstration on humans there are many widely held beliefs that must be regarded as without solid foundation. If we believe that vaccination protects against the development of smallpox it is not because there has been a direct experimental demonstration but rather (a) there is a good deal of evidence that is consistent with this hypothesis, and (b) over the course of many years no evidence has been produced to support any alternative hypothesis. The truth of the matter appears to be that medical knowledge (and, one suspects, many other kinds as well) has always advanced by a combination of many different kinds of observation, some controlled, and some uncontrolled, some directly and some only tangentially relevant to the problems at hand. Although some methods of observation and analysis are clearly to be preferred to others when a choice is possible, there are no magical methods that invariably lead to the right answer. If we cannot specify exactly what has been learned in medicine from the study of statistical associations, we can at least say that we could not have accumulated the knowledge we have without them.[4]

REFERENCES

1. J. Snow, Snow on Cholera, being a reprint of two papers by John Snow, M.D., 1936, The Commonwealth Fund, New York.
2. R. H. Shryock, The Development of Modern Medicine, 1947, Alfred A. Knopf, New York, p. 282.
3. W. E. Heston, M. K. Deringer, I. R. Hughes and J. Cornfield, Interrelation of specific genes, body weight and development of tumors in mice, 1952, *J. National Cancer Institute, 12:* 1141.
4. E. L. Wynder, J. Cornfield, P. D. Schroff, K. R. Doraiswami, A Study of environmental factors in carcinoma of the cervix, 1954, *Amer. J. Obst. and Gyn., 68:* 1016.
5. W. W. Smith, R. Q. Marston, H. J. Ruth and J. Cornfield, Granulocyte count, resistance to experimental infection and spleen homogenate treatment in irradiated mice, 1954, *Amer. J. Physiol., 178:* 288.
6. W. W. Smith, L. Gonshery, I. M. Alderman and J. Cornfield, Effect of granulocyte count and litter on survival of irradiated mice, 1954, *Amer. J. Physiol., 178:* 474.
7. R. O. Scow and J. Cornfield, Effect of thyroidectomy and food intake on oral and intravenous glucose tolerance in rats, 1954, *Amer. J. Physiol., 179:* 39.
8. E. L. Wynder, E. A. Graham, A. B. Croninger, Experimental production of carcinoma with cigarette tar, 1953, *Cancer Res., 13:* 855.
9. M. Greenwood, Epidemics and Crowd Diseases, 1937, Macmillan Co., New York, p. 96.
10. I. Zipkin and F. J. McClure, Inhibitory effect of fluoride on tooth decalcification by citrate and lactate in vivo, 1949, *J. Dental Research, 28:* 151.
11. J. H. Shaw, Should fluorides be added to public water supplies?, 1954, *Scient. Monthly, 79:* 232.
12. A. B. Hill, The Clinical trial, 1952, *New England J. of Med., 247:* 113.
13. R. and J. Dubos, The White Plague, 1952, Little, Brown & Co., Boston, p. 122.
14. A. B. Hill, Observation and experiment, 1953, *New England J. Med. 248:* 995.

[4] After completing this answer my attention was called to Hill's Cutler Lecture on Preventive Medicine (14) in which much the same issues that we have covered were also considered—in, however, a more comprehensive, lucid (and reasonable) manner. The reader is enthusiastically referred to it if he is at all interested in pursuing the subject.

I–3.

Austin Bradford Hill: The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine 1965; 58:295–300.

Hill's "The Environment and Disease: Association or Causation" may be a good example of an article that has been read in quotations and paraphrases more often than in its original form. In it, Hill offered a list of nine aspects of an empirical association to consider when deciding whether an association is causal. This was not the first list of "causal criteria" to be offered, but it was perhaps the most popular. It is unfortunate that in the ensuing decades, this list or similar ones have been presented in textbooks as "criteria" for inferring causality of associations, often in such a manner as to imply that all the conditions are necessary. A careful reading of Hill shows that he did not intend to offer a list of necessary conditions; on the contrary, on page 299 he warned against laying down "hard and fast rules of evidence that *must* be obeyed before we accept cause and effect." As noted later [Rothman, 1982], Hill's only real mistake was to say that *none* of his nine aspects could be considered necessary if the association were indeed causal; in fact, temporality (No. 4) is obviously necessary, as cause must precede effect.

Perhaps the most neglected portion of the article (also on page 299) is his comment on the misuse of significance testing by both scientists and statisticians. Despite his warning against equating statistical and scientific significance, I fear that decades later the situation is little better, and so I would give this section special emphasis in any educational setting.

Reference:

Rothman KJ. Causation and causal inference. Chapter 2 in Schottenfeld D and Fraumeni JF, eds. *Cancer Epidemiology and Prevention*. Philadelphia: WB Saunders, 1982.