Advances in Quantitative Methods in Epidemiology

ABRAHAM M. LILIENFELD, MD, MPH, DSc

Any science is as objective as its capability of measuring the events which it purports to be observing and relating. Epidemiology has not been exempt from the usual evolutionary development of this necessary aspect of its methodology.

Although initially borrowing the methods of study and the techniques of measurement (and of analyses) from other kindred sciences, epidemiology has, in the past 50 years, developed the study as art and science by means of innovative approaches to methodology and the elaboration of earlier techniques. Although methods utilized in studies of the types performed today are essentially similar to the methods of 50 years ago, the administrative and analytic sophistication developed over this period have ameliorated the biases inherent in all human population studies.

To outline for us the parallels with the past and the progress made in quantitative methods in this period is Dr. Abraham Lilienfeld, distinguished professor of epidemiology at Johns Hopkins School of Hygiene and Public Health. Dr. Lilienfeld has been the recipient of many honors for the contributions which he himself has made to epidemiology by his most careful methodologic approaches to disease investigation. His major contribution has been the expansion of the application of epidemiologic principles, derived from our study of infectious diseases, to the chronic diseases with insidious onset—an application extolled and urged by Dr. John Gordon earlier in this period of review. In more recent years, Dr. Lilienfeld has turned to the critical review of the history of epidemiology and its perspectives.—LEONARD M. SCHUMAN, MD

WHEN I ACCEPTED THIS REQUEST to review advances in quantitative methods during the half-century of the Epidemiology Section's existence, I thought it would be an easy task. But difficulties immediately arose when I tried to define "quantitative methods in epidemiology." Just as in any epidemiologic investigation where it is first necessary to define a "case" of disease, there is a need to delineate my area of discourse. The term "quantitative methods" could be regarded either narrowly in terms of statistical methods or more broadly in terms of the epidemiologic methods of study which, by the very nature of the epidemiologic enterprise, is a quantitative one.

Epidemiologists use vital statistics—that is, mortality and morbidity data—extensively. I would classify these as being part of demography, which is another discipline. Therefore, I will arbitrarily limit this discussion, with a few exceptions, to those methods of study that have been increasingly utilized by epidemiologists in the past 50 years and which can be collectively called "the Epidemiologic Study." Depicted in the diagram is the anatomy of the epidemiologic study (*1*). Some call this area "analytical epidemiology" in contrast to descriptive epidemiology which includes mortality and morbidity statistics. Calling the different types of study in the diagram collectively, "the

David E. Lilienfeld assisted the author in the preparation of this paper. It was supported in part by grant No. 1R01 LM03114 from the National Library of Medicine. Dr. Abraham Lilienfeld is the recipient of Research Career Award K06–GM13901 from the National Institute of General Medical Sciences, Public Health Service, Bethesda, Md.

Tearsheet requests to Abraham M. Lilienfeld, MD, Johns Hopkins School of Hygiene and Public Health, 615 North Wolfe St., Baltimore, Md. 21205.

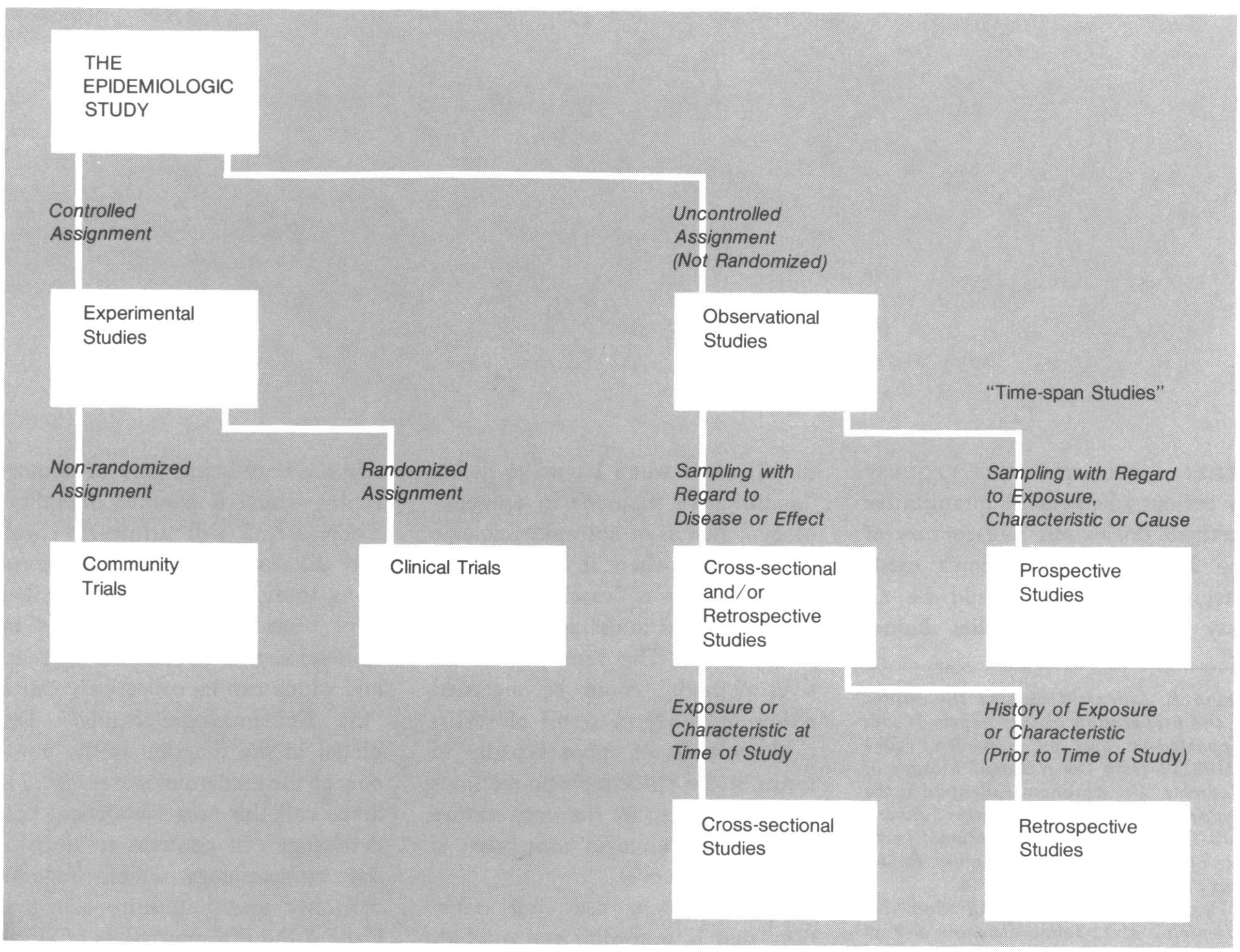
Epidemiologic Study" does not imply that those in other disciplines have not used them; actually, their origin lies in other disciplines. However, epidemiologists have been more largely concerned with these different types of study than those in other disciplines; in fact they represent the hard core of epidemiology.

An Excursion to 1929

I though it would be of interest to review those journals in which epidemiologic studies would have been published in 1929 to obtain some idea of the methods used at that time. These are the journals that I perused.

American Journal of Public Health
Public Health Reports
American Journal of Hygiene
Journal of Hygiene
Public Health
Human Biology
Biometrika
Journal of the American Statistical Association
Journal of the Royal Statistical Society
Proceedings of the National Academy of Science
Public Health Bulletin

Time does not permit an extensive review of the topics covered by these journals, which are decidedly of historical interest. I was rather surprised at the relatively modern nature of the subject matter. As one would expect, many articles were concerned with such communicable diseases as diphtheria, typhoid fever, and scarlet fever. For the most part, the papers on communicable diseases dealt with determining whether and to what extent public health measures were of value in producing observed declines in morbidity and mortality rates from these diseases. However, there were also papers entitled "The Health of Workers in Dusty Trades. II. Exposure to Silicaeous Dust" (*2*), "Effect of Repeated Daily Exposure of Several Hours to Small Amounts of Automobile Exhaust Gas" (*3*), "Physiological Responses Attending Exposure to Vapors of Methyl Bromide, Methylchloride, etc." (*4*), "Cancer as a Public Health Problem" (*5*), "Economic Status and the Incidence of Illness" (*6*), and a paper by Bradford Hill on "An Investigation of Sickness in Various Industrial Occupations" (*7*). All of these topics have a modern ring to them. I mention these



to confirm what we have shown in other historical studies, that there is no sharp dividing line between our contemporary period and the more recent past (*8*).

From a methodological viewpoint, there were several items of interest in the papers. One paper in the American Journal of Public Health was concerned with an explosive outbreak of typhoid fever in a rural village in Puerto Rico; the author, Morales, who was an epidemiologist in the Puerto Rican Health Department, surveyed the entire town to determine the reasons for the outbreak (*9*). What is of interest is that he made comparisons between the characteristics of the cases and what he called "controls." It was a typical modern type retrospective case-control study, in which he implicated a contaminated surface well as the source of the outbreak.

The same issue contained a paper entitled "The Factor of Chance in Diphtheria Mortality" in which the author used 95 percent confidence limits around the trends of the death rates (*10*). He stated, "The S.D. (standard deviation) is a yardstick which is invaluable to health officers as a means of measuring statistically the results of work done in a community. By means of such a measurement, one can avoid the error of attributing to health activities events that are due to pure chance and good luck."

In the American Journal of Hygiene of that year, Joseph Berkson published "A Probability Nomogram for Estimating the Significance of Rate Differences." (*11*). He illustrated its application by comparing the frequency of arthritic complications among scarlet fever patients who had or had not received serum treatment.

Several papers in 1929 also showed that correlation coefficients and regression analyses were frequently used for a variety of problems. I mention these to indicate that 50 years ago, statistical methods were being used for many purposes in epidemiologic studies, particularly with regard to infectious diseases.

Retrospective case-control studies were also being conducted in noninfectious diseases, such as the various forms of cancer. A paper published in 1929 had considerable influence on the course of epidemiologic studies during the next 15 to 20 years. This paper entitled "Cancer and Tuberculosis" by Raymond Pearl appeared in the American Journal of Hygiene (*12*). In analyzing the first 7,500 autopsies performed at the Johns Hopkins Hospital, Pearl identified 816 autopsies with mention of a malignant tumor and 816 controls who did not have a malignant tumor, matched by age at death, sex, color, and date of death. At autopsy, 16.3 percent of the controls showed active tuberculous lesions in contrast to 6.6 percent in the cancer group. The negative relationship was present regardless of the different ways in which the data were analyzed. Pearl interpreted it as indicating a negative antagonism between these two diseases. This paper aroused considerable discussion (*13,14*). It was shown later that the negative association was spurious and reflected the higher probability of autopsy for those who had had active tuberculosis. This study resulted in Berkson's algebraic analysis in 1946 of the fact that not only in autopsy series, but also in hospitalized populations, one could obtain a spurious association as a result of selection bias, generally referred to as "Berksonian bias" (*15*).

The criticisms of Pearl's work and Berkson's demonstration placed a damper on the conduct of retrospective studies of hospitalized patients, although a few such studies were carried out, particularly with regard to the relationship of cigarette smoking and lung cancer, an epidemiologic issue of extreme importance (*16,17*). But studies of hospitalized patients were conducted very infrequently and, when done, were usually criticized because of Pearl's blunder and Berkson's work.

New Statistical Tools

There was a general feeling that it would be more desirable to carry out prospective studies. This feeling persisted until Cornfield's classic paper in 1951 (*18*). He showed that under certain conditions, a study of cases and controls would provide an estimate of the relative risk, that is, the ratio of the incidence rates of an exposed group relative to that of an unexposed group. This was estimated by what is now known as the odds ratio. This demonstration—as well as the one by Woolf in England which was done independently and was concerned with the similar problem of estimating the relative risk of the relationship of blood groups and disease—provided the stimulus to the expanded use of retrospective case-control studies (*19*). This technique was further amplified by Cochran who provided a means for combining the findings of a series of 2 by 2 tables (*20*). Independently, these methods were extended by Mantel and Haenszel who used Cochran's weights for pooling relative risks over many strata, thereby obtaining a weighted relative risk; they developed significance tests and also considered using matched controls (*21*). In rereading that paper, I thought that it was interesting that Mantel and Haenszel discussed Snow's observations on cholera and Holmes' studies on puerperal fever. Relevant to this presentation is their statement in the introductory section of this paper:

Statisticians have been somewhat reluctant to discuss the analysis of data gathered by retrospective techniques, possibly because their training emphasizes the importance of defining a universe and specifying rules for counting events or drawing samples possessing certain properties. To them, proceeding from 'effect to cause', with its consequent lack of specificity of a study population at risk, seems an unnatural approach.

I might add that it was the criticism of Pearl's paper mentioned earlier that produced this state of mind.

These papers made available the necessary statistical tools and provided the underlying philosophic base for an expanded use of retrospective studies and the further development of statistical methods to take into account other variables, pooling, matching, use of multiple controls, and so forth in the analysis of such studies. Unfortunately, more recently, it has resulted in a proliferation of confusing terms that is being used to describe these various methods.

Table 1. Summary of analytical methods or concepts developed or extended, 1929 to 1960 by type of epidemiologic study

Type of epidemiologic study or data	*Analytical method or concept*
Retrospective studies	Relative risk (odds ratio) Cochran's method for combining 2X2 tables Mantel-Haenszel test for combining relative risks
Prospective studies	Life-table analysis Person-years concept
Mortality data	Cohort analysis

Studying Chronic Disease

I now return to 1929 to pick up another thread of methodological development resulting from the fact that in diseases of long duration (chronic disease), the methods of analysis used in acute communicable diseases had to be extended to take into account the time factor. In 1932, Wade Hampton Frost presented this method in his studies of the "Risk of Persons in Familial Contact with Pulmonary Tuberculosis" (*22*). Frost stated this very well indeed:

For tuberculosis the requirements are essentially the same (as in acute diseases), but are more difficult to meet, chiefly because the disease is of slow evolution. . . . Observation of the exposed group must extend over a sufficient number of years to define the rates of morbidity and mortality prevailing in successive periods throughout the usual span of life. To keep a sufficiently large group of people under systematic, exact observation for such a length of time is a difficult task. However, such simple facts as lie within the knowledge and memory of the average householder may be obtained by a *retrospective* investigation, tracing familial histories backward into the past.

In a footnote, he added, "The procedure, except for the collection of data would be essentially the same in analyses of records obtained in keeping a group under planned observation." Thus, Frost presented the application of the person-years concept and the use of life tables as a method of analysis of prospective or cohort studies, whether historical or concurrent. He pointed out that he had adopted the method proposed by others—Elderton and Perry and Weinberg—who introduced it in 1910 to 1913 (*23–25*). He extended this method, and over the past 45 years it has become an important part of the armamentarium of the epidemiologist.

In 1939, after Frost's death, his paper on the cohort analysis of mortality from tuberculosis was published (*26*). Again, the method was not developed by him; actually William Farr described the concept of this method, but Frost's students and his personal influence popularized it.

At this point, it would be relevant to comment on the fact that Frost recognized the need to conduct prospective studies, again a method mentioned by William Farr, when dealing with what we term the chronic diseases. During the past 50 years we have seen the prospective method of study used extensively in studies of cancer, cardiovascular disease, and occupational diseases (*1*). It almost has become a standard method of study among epidemiologists.

The methods which I have reviewed, summarized in table 1, bring me up to 1960. These methods have become standardized and are generally used by most epidemiologists in their etiological studies, together with several statistical refinements that have been developed since they were introduced.

Logistic Analysis

The past 25 years have witnessed a marked increase in the use of prospective studies, mainly with regard to two areas—the study of the effects of cigarette smoking and the study of factors that determine the risk of developing or dying from coronary heart disease. In these latter studies, where many factors that might influence the occurrence of the disease must be taken into account, it was felt that existing methods were inadequate to analyze the data. To take care of this problem, a general method was developed, that of "logistic analysis." Time permits only a general overview of logistic analysis, which took two approaches, one known as the Truett-Cornfield and the other, the Walker-Duncan (*27,28*). The logistic function and its relevance to prospective studies are shown in table 2.

There has been some discussion concerning the advantages and disadvantages of these two methods and which one is preferred. Each method can be used to make estimates of relative risk and to calculate tests of significance. However, certain questions do remain with regard to both of these approaches, which are currently being investigated. Despite this, both methods are now widely used. A major reason for their employment is that the past 15 to 20 years have witnessed the entry of the computer into data analysis. The computer has had its major impact in epidemiologic studies in which large numbers of variables and large numbers of persons have been studied. Without the "number-crunching" capabilities of the computer, methods such as logistic analysis could not be used.

Before discussing a few more quantitative methods, I would like to comment on the use of computers. Admitting their many advantages, I daresay that the use of computers has had some adverse effects. First, the epidemiologist is now far more removed from the actual data than he had been in the past. Usually, he has to go through data processors, computer programers, and others to get his data analyzed. This is partly the fault of the epidemiologist because he tends not to do his own computer work. Needless to say, he obviously should be or become familiar with the technical details of his study, particularly its analysis. Second, from an educational viewpoint, I have observed that many students do not know the details of the analysis, the assumptions underlying the methods used. They collect the data and then use various programs in the program library or packaged programs, *in a mechanical way,* for their analysis. There exists a need to train epidemiologists in the use of computers to a greater extent than is being done at present.

Table 2. Logistic analysis

Logistic equation: $$f(x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}$$

The probability (P) of developing disease given factor x $$= \frac{1}{1 + e^{-(\alpha + \beta x)}}$$

Of course, this unfamiliarity probably will change as the computer revolution continues, and mini- and micro-computers are developed. When these are so available that each epidemiologist has a computer in his office, with programs that will permit him to interact directly with it, he will, I hope, become more directly familiar with what is actually going on in the analysis of his data.

The more recent decade has witnessed the application of logistic analysis to data collected by means of retrospective studies. In general, most of such data are in the form of contingency tables, 2 by 2 tables or tables with a larger number of cells, say, N by M tables (table 3). Questions have been raised concerning the appropriateness of applying logistic analysis to retrospective studies. Differences of statistical opinion do exist. However, with work being done now and probably during the next few years, it is hoped that the basis for ascertaining the proper role of these methods of logistic analysis to retrospective studies will become clarified.

Loglinear Models

More recently, another general method has been developed for the analysis of the multidimensional contingency table, namely that of loglinear models. The most systematic presentation of these methods is presented in what is known as the Green Book (*29*). These methods fit models to the data and estimate the parameters in these models, providing estimates of the magnitude of effects of interest, permitting one to judge the relative importance of these different effects. Essentially, this is done in the same manner as the usual analysis of variance, or regression analysis. Table 4 shows a simplified illustration of this model applied to a 2 by 2 table; the reason for the name is apparent. Insofar as

Table 3. General form of a two-dimensional contingency table

		FACTOR A					
		1	2	3	. . . .	c	Total
	1						N_1
	2						N_2
	3						N_3
	.						.
FACTOR B	.						.
	.						.
	.						.
	.						.
	r						N_r
Total		M_1	M_2	M_3	. . . .	M_c	T

I know, these loglinear models have not been generally used in epidemiologic studies, but I feel certain that their use will increase.

It is of interest that statisticians who have developed these methods now consider that the method of standardization for variables, the time-honored method used by epidemiologists "has been made obsolete" by the ready availability of computer programs for loglinear model analysis of multidimensional contingency tables (*30*).

Where do we stand today with regard to the availability of these relatively more complex methods of analysis? I must admit that these new methods of analysis appear quite elegant. Although I have already expressed some concern regarding our dependence on the computer, I must express a graver concern. First, I have not yet seen that these new methods have added much of epidemiologic significance to the usual methods of analysis that were used before the advent of the computer in deriving biological or etiological inferences from the data. Second, with the increased emphasis on these new methods, I see less interest and concern about the quality of the data that are being analyzed. In most studies, I have personally noted that less attention is being paid to validating the quality of the data than was true 15 to 20 years ago. Of course, one cannot place this lack of concern with the quality of data at the doorstep of these new analytical tools. One of the problems is that time for training is limited. If more time is spent on learning these techniques, there is a tendency not to conduct the studies necessary to determine data quality. It need not be so. If these new tools are more powerful in terms of the epidemiologic enterprise, we should utilize them, but we should not sacrifice our concern for the quality of the data that we are analyzing.

Table 4. Simplified illustration of a loglinear model

		FACTOR A		
		A (Present)	$\bar{A}$ (Absent)	Total
FACTOR B	B (Present)	P_{AB}	$P_{\bar{A}B}$	P_B
	$\bar{B}$ (Absent)	$P_{A\bar{B}}$	$P_{\overline{AB}}$	$P_{\bar{B}}$
	Total	P_A	$P_{\bar{A}}$	N

No association between factors A and B

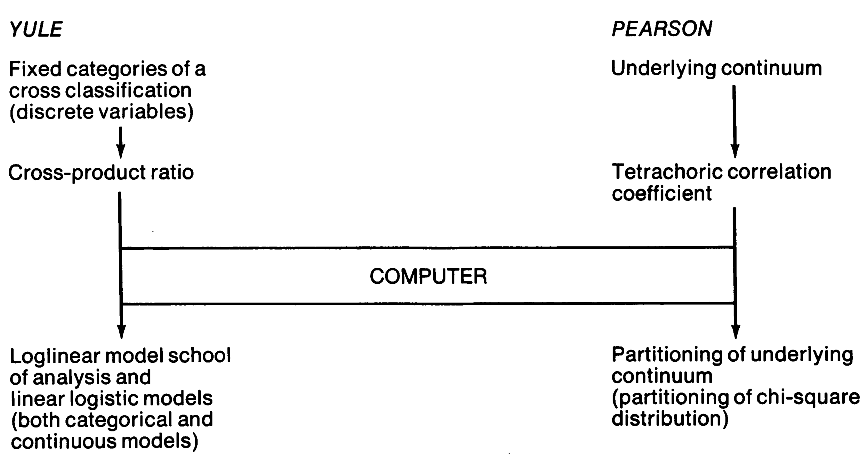
$$P_{AB} = P_A \times P_B$$

$$\ln P_{AB} = \ln P_A + \ln P_B$$

Most of this discussion has been concerned with observational studies, which has been the major area of interest for epidemiologists until the past few years. I would be remiss in this review if I did not indicate that, in the past 10 to 15 years, epidemiologists have been increasingly involved in controlled clinical trials, which have become a major activity. For the most part, these trials have been concerned with evaluating therapeutic agents, mainly drugs. But, I am impressed with the fact that such trials do contribute to our knowledge of etiological factors of disease and of the natural history and pathogenesis of the diseases being studied.

Clinical trials have brought with them a new set of quantitative methods that require solution. I arbitrarily decided to leave these developments for someone else to review. But, this omission should not be interpreted as indicating that clinical trials are not part of the epidemiologic enterprise. Far from it!

Table 5. Development of analysis of observational studies: a schematic representation of a statistician's perception



Conclusions

We have come a long way in these 50 years. The breadth and scope of the epidemiologic study has broadened considerably. Quantitative methods have become more elaborate, perhaps providing us with more analytical skills in probing the secrets of nature. But, in summarizing, I would like to emphasize that during the past 50 years many aspects of epidemiology and quantitative methods have had their origin many years before 1929. It is of interest that Fienberg has indicated that the gradual development of these methods can be traced to the beginning of this century and to certain philosophical differences concerning the underlying distributions of categorical data (*30*). I have schematically shown this in table 5. Current developments are part of a continuing historical evolution, and I am certain that today's methods will serve as the origin for further developments in the next 50 years, thereby enriching our discipline still further.

References

1. Lilienfeld, A. M., and Lilienfeld, D. E.: Foundations of epidemiology. Ed 2. Oxford University Press, New York, 1980.
2. Russell, A. E.: Britten, R. H., Thompson, L. R., and Bloomsfield, J. J.: The health of workers in dusty trades. II. Exposure to siliceous dust (granite industry). Public Health Bulletin No. 187, Public Health Service, 1929.
3. Sayers, R. R., Yant, W. P., Levy, E., and Fulton, W. B.: Effect of repeated daily exposure of several hours to small amounts of automobile exhaust gas. Public Health Bulletin No. 186, Public Health Service, 1929.
4. Sayers, R. R., Yant, W. P., Thomas, B. G. H., and Berger, L. B.: Physiological responses attending exposure to vapors of methyl bromide, methyl chloride, ethyl bromide and ethyl chloride. Public Health Bulletin No. 185, Public Health Service, 1929.
5. Ewing, J.: Cancer as a public health problem. Public Health Rep 44: 2093–2101, Aug. 30, 1929.
6. Sydenstricker, E.: Economic status and the incidence of illness. Hagerstown morbidity studies No. X: Gross and specific illness rates by age and cause among persons classified according to family economic status. Pub Health Rep 44: 1821–1833, July 26, 1929.
7. Hill, A. B.: An investigation of sickness in various industrial occupations. J Roy Stat Soc 92: 183–238 (1929).
8. Lilienfeld, A. M., and Lilienfeld, D. E.: What else is new? An historical excursion. Am J Epidemiol 105: 169–179 (1977).
9. Morales, E. G.: Typhoid fever in a rural village of Porto Rico due to a surface well. Am J Public Health 19: 997–1004 (1929).
10. Lee, W. W.: The factor of chance in diphtheria mortality. Am J Public Health 19: 855–862 (1929).
11. Berkson, J.: A probability nomogram for estimating the significance of rate differences. Am J Hyg 9: 695–699 (1929).
12. Pearl, R.: Cancer and tuberculosis. Am J Hyg 9: 97–159 (1929).
13. Wilson, E. B.: Morbidity and the association of morbid conditions. J Prev Med 4: 27–28 (1930).
14. Wilson, E. B., and Maher, H. C.: Cancer and tuberculosis with some comments on cancer and other diseases. Am J Cancer 16: 227–250 (1932).
15. Berkson, J.: Limitations of the application of fourfold table analysis to hospital data. Biometrics 2: 47–53 (1946).
16. Schrek, R., Baker, L. A., and Dolgoff, S. Tobacco smoking as an etiologic factor in disease. I. Cancer. Cancer Res 10: 49–58 (1950).
17. Levin, M. L., Goldstein, H., and Gerhardt, P. R.: Cancer and tobacco smoking. JAMA 143: 336–338 (1950).
18. Cornfield. J.: A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. J Natl Cancer Inst 11: 1269–1275 (1951).
19. Woolf, B.: On estimating the relation between blood group and disease. Ann Hum Genetics 19: 251–253 (1955).
20. Cochran, W. G.: Some methods of strengthening the common X^2 tests. Biometrics 10: 417–451 (1954).
21. Mantel, N., and Haenszel, W.: Statistical aspects of data from retrospective studies of disease. J Natl Cancer Inst 22: 719–748 (1959).
22. Frost, W. H.: Risk of persons in familial contact with pulmonary tuberculosis. Am J Public Health 23: 426–432 (1933).
23. Elderton, W. P. and Perry, S. J.: Studies in national deterioration. VI. A third study of statistics of pulmonary tuberculosis; the mortality of the tuberculous and sanatorium treatment. Drapers' Company Research Memoirs, Cambridge University Press, London, 1910.
24. Elderton, W. P., and Perry, S. J.: Studies in national deterioration. VIII. A fourth study of the statistics of pulmonary tuberculosis; the mortality of the tuberculous, sanatorium and tuberculin treatment. Drapers' Company Research Memoirs, Cambridge University Press, London, 1913.
25. Weinberg, W.: Die Kinder der Tuberkulösen. S. Hirzel, Leipzig, 1913.
26. Frost, W. H.: The age selection of mortality from tuberculosis in successive decades. Am J Hygiene 30: 91–96 (1939).
27. Truett, J., Cornfield, J., and Kannel, W.: A multivariate analysis of the risk of coronary heart disease in Framingham. J Chronic Dis 20: 511–524 (1967).
28. Walker, S. H., and Duncan, D. B.: Estimation of the probability of an event as a function of several independent variables. Biometrika 54: 167–179 (1967).
29. Bishop, Y. M. M., et al.: Discrete multivariate analysis: theory and practice. MIT Press, Cambridge, Mass. 1974.
30. Fienberg, S. E.: Comment: the design and analysis of the observational study—a review by S. M. McKinlay. J Am Stat Assoc 70: 521–523 (1975).