

The Epidemiologic Approach

by Sidney Pell*

Epidemiological studies rarely provide unequivocal data on the effects of toxic substances on human health. Unlike experimental studies, pertinent variables frequently cannot be controlled and some are often unknown. Nevertheless, these limitations can be dealt with by various facets of an epidemiological investigation. These include, the choice of a study design, selection of controls or comparison populations, evaluation of available data, statistical analysis, and the drawing of appropriate inferences from the data.

Among the special problems that might be encountered in studies of the effects of cardiotoxic substances are, difficulties in establishing diagnoses, errors in death certificates, observer errors in electrocardiographic interpretations, and taking into account the effects of various risk factors known to be implicated in the etiology of cardiovascular diseases.

In spite of various methodological problems and certain inherent limitations of epidemiological studies, they can, when properly conducted, make significant contributions to knowledge of disease etiology and provide the means of reducing the risk of a disease, even when its etiology is not completely known.

When a particular chemical compound is found to be toxic to animals, data on the nature and extent of that chemical's toxicity to humans are frequently sought in an epidemiological investigation.

At the outset, it should be understood that an epidemiological study rarely provides definitive, unequivocal answers to questions concerning the effects of toxic substances on the health of humans. Data derived from the study usually permit the investigator to draw only carefully guarded inferences concerning causal relationships between the substance and health impairments observed in an exposed population.

In contrast to animal experimentation, human studies usually cannot provide adequate control of confounding variables. Subjects cannot be assigned at random to an exposed and control group, environmental conditions cannot be controlled, diets cannot be standardized, and genetic factors cannot be controlled by the use of purebred strains.

The ideal human study would be one in which the exposed and control groups are comparable with respect to all characteristics related to the diseases under investigation except that one group is or has been exposed to the toxic substance and the other has never been exposed. This ideal can rarely, if ever, be attained.

The task of the epidemiologist is to apply various

strategies and statistical techniques in such ways that he or she will come as close to achieving that ideal as the available information will allow. That task includes the following facets of an epidemiological study: study design, selection of control groups or comparison populations, evaluation of the quality of available data, statistical analyses, and the drawing of appropriate inferences from the data. I shall discuss each of these briefly.

Study Design

Epidemiologists do not entirely agree on the terminology for the classification of study designs. The terms I shall use here are those that have been adopted by MacMahon and Pugh (*1*).

The two basic types of epidemiological studies are cohort and case-control studies. In a cohort study, a study population is followed over a specified period of time to compare disease rates among those in the population having a characteristic with persons who do not have the characteristic. Comparisons can also be made with general population. It is important that at the beginning of the study period all members of the cohort are free of the disease under investigation. To study, for example, the relation of exposure to a certain chemical compound to the development of coronary heart disease, one would identify cohorts of exposed and of nonexposed persons, all free of clinical coronary heart disease, and then identify those

* Medical Division, E. I. Du Pont de Nemours and Co., Inc., Wilmington, Delaware 19898.

members of the cohorts who subsequently develop the disease.

Cohort studies may be either retrospective or prospective. In a retrospective study, the cohort is identified as of some time in the past and followed to a later date. In a prospective study, the cohort is established in the present and followed into the future. Retrospective cohort studies have also been described as historical prospective studies (2).

In a case-control study, a group of persons who have the disease is compared with a group free of the disease (the control group) to determine whether a given characteristic is more common among persons with the disease. In the example cited above, one would select a group of persons known to have coronary heart disease and a control group of persons free of the disease. Then the two groups would be compared with respect to the proportion in each group that has been exposed to the compound.

Cohort studies—especially prospective cohort studies—are preferable to case-control studies. The data from cohort studies provide direct measures of the risk of disease in exposed and nonexposed populations, whereas case-control studies would yield only estimates of risk, sometimes crude, imprecise estimates. Cohort studies, on the other hand, require a larger number of subjects and a long follow-up period. If the disease is rare, it may not be practicable to assemble the large sample size required in a cohort study. The case-control study has among its advantages the requirement of a comparatively small number of subjects and the short period of time in which results can be obtained.

One other study design that should be mentioned is the cross-sectional study. The data obtained in this type of study are gathered as they exist at a specific point in time. The study population is surveyed to identify those who have a particular disease or diseases. Then, the prevalence of disease is analyzed in relation to pertinent characteristics of that population.

It is tempting to use this study design because such a study makes use of existing data, or of data that can be developed in a short period of time, and therefore results can be obtained much sooner than in a cohort study. Cross-sectional studies, however, do have important drawbacks, particularly in studies of the relation of cardiovascular disease to exposure to toxic chemicals.

One important problem arises out of the high case fatality rate and the reduced survival time found in cardiovascular disease. Since a cross-sectional study would enumerate only living cases, failure to include those who died before the study was undertaken may seriously bias the results. If, for

example, a particular chemical causes a sudden cardiac death, or if survival time following the onset of the episode is short, most, if not all those affected by the chemical would not be enumerated in a cross-sectional study, and, as a result, the study would fail to reveal a cardiotoxic effect.

Another problem in a cross-sectional study is the difficulty in determining the temporal relation between first exposure to the substance and the onset of the disease. The former must precede the latter if there is a causal relation. Furthermore, the time interval between exposure and disease onset must be known to determine if there is a latent effect and the length of the latent period. These kinds of data are usually not available in cross-sectional studies.

Each study design has its merits and limitations. The choice depends upon an assessment of the available data, which design is most appropriate for the data, and a careful weighing of the advantages and disadvantages of each type of study.

Selection of Controls

The choice of the proper control group is a crucial element in an epidemiological study. It is by this choice that the investigator can take into account factors other than exposure to the chemical that are related to the disease. In a case-control study, for example, one control can be selected for each case so that the case and control are matched by such factors as age, sex, and race. In cohort studies, where matching is usually not feasible, the control group may consist of workers at the plant who have no history of exposure to the chemical. In this case, confounding variables can be taken into account by statistical analysis.

In cohort mortality studies, investigators frequently compare mortality in the exposed group with death rates in the U. S. population or in subdivisions of the U. S., such as the state or county in which the plant is located. It has been well established, however, that mortality in working populations is generally lower than in the general population because of selection factors that tend to exclude from the work force persons who are chronically ill (3-5). Mortality in the U. S. has been used for comparative purposes because data for employed persons have not been readily available. If one must use the U. S. as a comparison population, an assessment of differences between the U. S. and the exposed cohort must take into account the lower death rates normally found in working populations, especially among younger persons. In a comparison of death rates between the Du Pont Company and the U. S. we found that the ratio of death rates (Du Pont/U. S.), adjusted for

age, among males was 0.68. For coronary heart disease it was 0.82, and for cerebrovascular disease, 0.99 (6). Morbidity from acute myocardial infarction in the Du Pont population, however, differed little from that found in other surveyed populations (7).

Evaluation of Available Data

If the raw data available for a study are deficient with respect to their reliability, validity, and completeness, the results will be of little value, no matter how well the study is designed and no matter how sophisticated the statistical analysis may be.

There are certain problems in the diagnoses of cardiovascular diseases that have to be dealt with in epidemiological studies. One is the occurrence of asymptomatic disease, such as a "silent" myocardial infarction, i.e., where the individual was not aware of, or did not report, symptoms, but evidence of myocardial damage appears on the electrocardiogram.

Another problem is how to deal with symptoms of cardiovascular diseases in the absence of objective evidence of disease. One example is chest pain on exertion, suggesting angina pectoris, but where the electrocardiogram is normal. Another example is the occurrence of a syndrome, such as, dizziness, weakness, and loss of muscular coordination, suggesting cerebrovascular disease, but with no other information to establish a definitive diagnosis.

The electrocardiogram is a means of obtaining objective evidence of heart disease, but although the measurements produced by the electrocardiogram are objective, their interpretation is not. Studies of the reproducibility of electrocardiographic interpretations have shown a significant amount of disagreement among observers and between repeated interpretations made by the same observer (8, 9). Although the electrocardiogram is, nevertheless, an important diagnostic tool for heart diseases, problems in its interpretation must be taken into account when it is used in epidemiological studies.

The use of death certificate information in studies of cardiovascular diseases presents us with another major problem. Errors in statements of cause of death are common. One source of error arises when there is a sudden, unattended death, and no autopsy is performed. A common practice is to attribute the death to a heart attack. The magnitude of death certificate errors has been studied where autopsy data were available. One such study by Beadenkopf et al. found that "in only 50 percent of the individuals with infarction or coronary thrombosis at autopsy was arteriosclerotic heart disease coded on the death certificate" (10). They also found, however,

that "82 percent of the death certificates reporting arteriosclerotic heart disease as the cause of death were confirmed as cases by autopsy findings."

In enumerating the occurrence of a particular type of cardiovascular disease in a cohort, all the problems discussed above must be taken into account. It is important that rigid diagnostic criteria be established and that all members of the cohort who develop the disease according to these criteria be identified. Failure to meet these requirements can result in misclassification of diseased and non-diseased persons and in biases in the data that can result in misleading conclusions. Such deficiencies in the data are apt to be found in a retrospective, rather than a prospective, cohort study, where the investigator must rely on data developed in the past and generated in such a way that they do not meet the rigid requirements of an epidemiological study.

Another problem arises in the identification of exposed persons. Work histories must be adequate to identify all persons with a history of exposure in the surveyed population, and, wherever possible, to furnish data on the level and duration of exposure. Biases can result if a certain segment of exposed persons is not identified and thereby not included in the cohort.

Statistical Analysis

In its simplest form, the statistical analysis of epidemiological data consists of computing an incidence rate or death rate in an exposed and control group, computing the ratio of the two rates to obtain a measure of relative risk, and then performing a test of significance to decide whether the difference could have occurred by chance alone, or whether the difference is probably real.

Usually, however, epidemiological data require more complex statistical analyses. Time does not permit an even cursory outline of all the analytical problems that arise in epidemiological data and how those problems are dealt with by various statistical techniques, but I would like to discuss briefly one major problem in occupational health studies, and that is the need to sort out the various factors related to the disease in order to determine to what extent, if any, the substance under investigation increases the risk of the disease among exposed workers.

It is well known that all diseases, particularly cardiovascular diseases, have a multiple etiology. Whether a particular individual will develop a given disease depends upon the influence of the three major components involved in the etiology of disease: the agent, the host, and the environment.

Fortunately, we know a great deal about the

etiology of the major cardiovascular diseases, thanks to several epidemiological studies—notably the Framingham Heart Study—that have been conducted in the past 25 years. These studies have identified risk factors for coronary heart disease, stroke, and hypertension. In addition, the data generated by these studies have been utilized to measure the extent to which the various factors increase risk, either alone or in combination. For coronary heart disease, these studies have identified three major risk factors: elevated serum cholesterol, hypertension, and cigarette smoking, as well as other risk factors, such as diabetes mellitus, overweight, sedentary living, personality factors, and family history of coronary heart disease (11). If a study is undertaken to investigate the possible role of a particular chemical compound in the development of coronary heart disease, data on the known risk factors would have to be taken into account in order to isolate the effect of the chemical.

The various statistical methods that have been applied to achieve this objective come under the general category of multiple regression analysis. One especially useful technique has been the multiple logistic function (12). By this method, one can identify significant risk factors and measure the contribution to risk made by each factor, relative to the contribution of other risk factors.

Drawing Inferences from the Results

A fundamental question that arises out of epidemiological data is whether a particular statistical association indicates a causal relationship. If, for example, the death rate from a coronary heart disease is found to be significantly greater among workers exposed to a given chemical than among those in a nonexposed control group, can one conclude that the chemical was responsible for the increased death rate?

That conclusion can be drawn if all factors related to coronary heart disease were known and taken into account by the design of the study and/or the statistical analysis. It is rare that such an objective can be achieved in an epidemiological study. Although we have learned a great deal about risk factors for the major cardiovascular diseases, the etiology of these diseases is still not completely understood. Other risk factors remain to be identified. Moreover, in many epidemiological studies—especially when they are retrospective—it is not always possible to obtain all relevant information. Therefore, in almost all studies, there usually remains some doubt about the meaning of the statistical relationships.

In drawing inferences from the results of the study, one must do so by making a careful assessment of various aspects of the data to determine whether they tend to strengthen the hypothesis of a causal relationship or lend little or no support to the hypothesis. Among the factors that have to be considered are the following: the adequacy of the study design; the reliability and validity of the data; the extent of missing relevant data; the degree of the statistical association, as measured by such statistics as the relative risk, odds ratio, and correlation coefficient (the greater the degree of association, the more likely the association indicates a causal relationship); the dose-response relationship (the existence of such a relationship strengthens a causal hypothesis); whether or not the relationship between the substance and the disease can be explained by a biological mechanism; and whether or not the results of the study are consistent with the findings of other studies.

Conclusions

This discussion of the epidemiologic approach to the study of toxic substances has necessarily touched only briefly on methodology, statistical techniques, and problems in the interpretation of epidemiologic data. Because of the brevity of the discussion, portions of it may have been somewhat oversimplified.

The purpose of this paper, however, has been primarily to describe the rationale of the epidemiologic approach and to point out certain limitations of epidemiological data. One should keep in mind that, in spite of their limitations, epidemiological studies have contributed greatly to an understanding of the etiology of many diseases, and have also provided us with the capability of instituting preventive measures to reduce the risk of various diseases, even when their etiology is not completely known.

REFERENCES

1. MacMahon, B., and Pugh, T. F. *Epidemiology—Principles and Methods*, Little, Brown and Co., Boston, 1970, p. 40.
2. Mausner, J. S., and Bahn, A. K. *Epidemiology—An Introductory Text*, Saunders, Philadelphia, 1974, p. 325.
3. Enterline, P. E. The estimation of expected rates in occupational disease epidemiology. *Publ. Health Repts.* 79: 973 (1964).
4. Redmond, C. K., and Breslin, P. P. Comparison of methods for assessing occupational hazards, *J. Occup. Med.* 17: 313 (1975).
5. McMichael, A. J. Standardized mortality ratios and the "healthy worker effect:" scratching beneath the surface. *J. Occup. Med.* 18: 165 (1976).
6. Pell, S. The effect of selection factors on epidemiologic research in employed populations. In: *The Nature of Biostatistics*. University of Pittsburgh, Pittsburgh, 1969, p. 175.

7. Shapiro, S., et al. Incidence of coronary heart disease in a population insured for medical care (HIP). *Amer. J. Publ. Health* 59 (Suppl.): 12 (1969).
8. Acheson, R. M. Observer error and variation in the interpretation of electrocardiograms in an epidemiological study of coronary heart disease. *Brit. J. Prev. Soc. Med.* 14: 99 (1960).
9. Higgins, I. T. T., Kannel, W. B., and Dawber, T. R. The electrocardiogram in epidemiological studies—reproducibility, validity, and international comparison. *Brit. J. Prev. Soc. Med.* 19: 53 (1965).
10. Beadenkopf, W. G., et al. An assessment of certain medical aspects of death certificate data for epidemiologic study of arteriosclerotic heart disease. *J. Chron. Dis.* 16: 249 (1963).
11. Simborg, D. W. The status of risk factors and coronary heart disease. *J. Chron. Dis.* 22: 515 (1970).
12. Truett, J., Cornfield, J., and Kannel, W. B. A multivariate analysis of the risk of coronary heart disease in Framingham. *J. Chron. Dis.* 20: 511 (1967).