Gordis, L., A. Lilienfeld, and R. Rodriguez. 1969. Studies in the epidemiology and preventability of rheumatic fever: I. Demographic factors and the incidence of acute attacks, and II. Socio-economic factors and the incidence of acute attacks. *J. Chron. Dis.* 21:645–666.

Gordon, T. 1957. Mortality experience among the Japanese in the United States, Hawaii and Japan. *Public Health Reports* 72:543–553.

Hill, A. B., *Principles of Medical Statistics.* (London: Oxford University Press, 1971), pp. 220–228.

Howard, J., and B. L. Holman. 1970. The effects of race and occupation on hypertension mortality. *Milbank Memorial Fund Quarterly* 48:263–296.

Lilienfeld, A. M. 1956. The relationship of cancer of the female breast to artificial menopause and marital status. *Cancer,* 9:927–934.

Limburg, C. C. 1950. Geographic distribution of multiple sclerosis and its estimated prevalence in the United States. *Proceedings of the Association for Research in Nervous and Mental Diseases,* 28:15–24, Baltimore: Williams and Wilkins.

Luby, J. P., H. J. Dewlett, and M. S. Dickerson. 1971. Measles—Dallas, Texas. *Center for Disease Control: Morbidity and Mortality Weekly Report,* 20:191–192.

Philp, J. R., T. P. Hamilton, T. J. Albert, R. S. Stone, C. F. Pait, R. R. Roberto, N. J. Fiumara, A. R. Hinman, and C. Friedmann. 1972. Hepatitis A Outbreak, Orange County, Calif. *Center for Disease Control: Hepatitis Surveillance,* Report No. 35, pp. 12–13. See also *Amer. J. Epidemiology* 97:50–54, 1973.

Sartwell, P. E., (Ed.): *Maxcy-Rosenow Preventive Medicine and Public Health, Ninth Edition.* (New York: Appleton-Century-Crofts, 1965), Chap. 4.

Segi, M., M. Kurihara, and T. Matsuyama, *Cancer Mortality for Selected Sites in 24 Countries. No. 5 (1964–1965).* Department of Public Health, Tohoku University School of Medicine. Sendai, Japan, 1969.

Silverberg, E., and A. I. Holleb. 1972. Cancer statistics, 1972. *Ca-A Cancer Journal for Clinicians,* 22:2–20.

Taylor, I., and J. Knowelden, *Principles of Epidemiology.* (Boston: Little, Brown, 1964), Chaps. 6, 7 (infectious epidemics); Chap. 12, pp. 318–321 (social class).

World Health Organization: 1967. Mortality statistics: cardiovascular diseases, annual statistics, 1955–1964, by sex and age. Epidemiological Vital Statistics Reports, 20:535–710.

Chapter 6

Prevalence Studies

In going beyond descriptive observations to delve more deeply into disease etiology, there are, as defined in Chap. 4, three basic types of observational investigations:

1 Prevalence or cross-sectional studies
2 Case-control studies
3 Incidence or cohort studies

These will be discussed in greater detail here and in the next two chapters. As will be seen, prevalence studies are, conceptually, quite straightforward, and provide a good basis for subsequent consideration and comparison of the other two study types.

How Prevalence Studies Are Carried Out

Initial Steps The question(s) for study must be clearly defined in terms of the relationship between some possible predisposing factor(s) and the disease under investigation. Then a suitable study population is identified. If this population is small enough to be studied using the human and financial resources available (e.g., students in a school, adults in a small town), the entire population can be included. If the target population is too large (e.g., children in the United States, men in a large city), then a representative sample is selected.

Sampling Methods for selecting an appropriate sample constitute an important and well-developed field of statistical study, and cannot be dealt with comprehensively in this book. The reader should be familiar with a few basic types of samples, since sampling may be necessary in any type of epidemiologic study. For a more complete discussion the reader is referred to Hansen et al. (1953) and Hill (1971).

The most elementary kind of sample is a *simple random sample* in which each person has an equal chance of being selected directly out of the entire population. One way to carry out this procedure is to assign each person a number, starting with 1, 2, 3, and so on. Then, numbers are selected at random, usually from a table of random numbers (see Arkin and Colton, 1963), until the desired sample size is attained.

A *stratified random sample* involves dividing the population into distinct subgroups according to some important characteristic, such as age or socioeconomic status, and selecting a random sample out of each subgroup. If the proportion of the sample drawn from each of the subgroups, or *strata*, is the same as the proportion of the total population comprised by each stratum (e.g., age group 40–59 comprises 20 percent of the population, and 20 percent of the sample comes from this age stratum), then all strata will be fairly represented with regard to numbers of persons in the sample. This proportionality is often desirable and may simplify data analysis. On the other hand, the investigator may have to take a larger proportion of his study sample out of one or a few sparsely populated strata, in order to make available for study adequate numbers of subjects with certain important characteristics.

A *cluster sample* involves (1) dividing the population into subgroups, or *clusters*, that are not necessarily (and preferably not) homogeneous, as are strata, (2) drawing a random sample of the clusters, and (3) selecting all or a random sample of the persons in each cluster. When each cluster comprises persons in a localized geographic area, such as a county, cluster sampling is especially useful for national surveys. It is obvious that many more persons can be studied for the same cost if they live in a few U.S. counties, than if they are scattered all over the country.

Finally, *systematic samples* involve first deciding what fraction of the population is to be studied—for example, one-half or one-tenth—and listing the population in order, perhaps as in a directory or on a series of index cards. Then, starting at the beginning of the list, every second or every tenth (or whatever interval is dictated by the fraction to be chosen) is selected. In order to sample in this manner, the investigator must be quite sure that the intervals do not correspond with any recurring pattern in the population. Consider what would happen if the population were made up of a series of married couples with the husband always listed first. Picking every fourth person would result in a sample of men only, if one started with the first or third subject, or of women only, if one started with the second or fourth.

Sampling can be done in multiple stages, such as sampling within strata which are, in turn, within clusters. In this manner, sampling can become quite involved and require expert assistance in planning. Experience has also revealed subtle problems and biases that might not occur to the novice. Sampling by households is a good example. If there is no one home when the interviewer arrives, he or she should come back again rather than go to the house next door, because households with a person at home in the daytime tend to differ from those without. Similarly, the first house seen as one approaches a new block should not be routinely called upon, since persons in corner houses tend to differ from those in the middle of the block.

Data Collection Once the total study population or sample is defined, the necessary data are collected. Presence of disease may be determined in a variety of ways. For example, in a small town, all or almost all the existing cases of a disease can often be found by contacting all the practicing physicians and reviewing hospital records. Or, the disease can be detected by a special examination of all the residents.

The presence of, or exposure to, the possible causative factors under investigation should also be determined by appropriate tests and measurements. For example, in considering the possible role of inhaled factors, degree of cigarette smoking can be determined by interview, and air pollution levels in various places of living or work can be determined by appropriate measuring devices.

Data Analysis The usual way to tabulate the data in a prevalence study is to subdivide the population according to the suspected predisposing factors being studied and compare the disease prevalence rates in each subgroup. If the relationship of chronic cough to number of cigarettes smoked is to be studied in a group of middle-aged men, then the group may be divided into appropriate smoking categories, such as: none, less than one-half pack per day, one-half pack or more but less than one pack, one pack or more but less than two packs, and two packs or more. The prevalence rate of chronic cough is then determined for each smoking subgroup and the rates in the subgroups are compared. Of course, before the rates are computed, strict criteria must be established for the definition of what constitutes "chronic cough."

Interpretation

In general, the prevalence study will show the presence or absence of a relationship between the study variable(s) and existing disease. *Existing* disease, as contrasted with *developing* disease found in an incidence study, implies a need for caution, since existing cases may not be representative of *all* cases of the disease.

Consider coronary heart disease, for example. One of the important manifestations of coronary heart disease is sudden unexpected death. In a prevalence survey, cases of coronary heart disease showing sudden unexpected death as their first clinical manifestation will be missed because the duration of recognizable disease is so extremely short. It would indeed be remarkable if such a case happened to occur just at the moment the individual was taking the survey examination! From this extreme example it can readily be seen that the shorter the duration of the disease, whether it kills or is cured rapidly, the less chance its victim has of being detected in a one-time prevalence survey. It follows logically, then, that cases of long duration are overrepresented in a prevalence study. The characteristics of these long-duration cases may, on the average, differ in a variety of ways from the characteristics of all cases of the disease being studied.

While we are considering the duration of illness of diseased persons in a prevalence study, it would be worthwhile to digress slightly and point out that there are two basic properties of a disease that are reflected in its prevalence. One is how much disease develops per unit of time, or incidence; the other is how long it lasts, or duration. In fact, under stable conditions, where the incidence and duration of a disease have remained constant over a period of time, the relationship between prevalence, incidence, and duration can be expressed as a simple mathematical equation: Prevalence equals incidence times mean duration ($P = I\bar{d}$). Thus, if any one of the three measures is unknown, it can be computed from the other two, provided that conditions are stable, as mentioned.

Another factor leads to "prevalence cases" being an unrepresentative sample of all cases; that is, if certain types of cases leave the community. Some affected persons may be institutionalized elsewhere or move to another city where there are special facilities for treatment, thus escaping local surveillance procedures.

When interpreting the findings of a prevalence study, care must be taken to avoid assigning an unsubstantiated time sequence to an association between a trait or other factor and the disease. If it is found, for example, that cancer patients exhibit more anxiety or other emotional problems than the unaffected members of the population, it cannot be assumed that the anxiety preceded the cancer. After all, cancer patients may have good reason to be nervous or disturbed. In contrast, there would be no doubt about the cancer being preceded by such traits as eye color, blood type, or maternal exposure to radiation.

Example I: Prevalence Studies of Chronic Respiratory Disease in Berlin, New Hampshire

In 1961, Ferris and Anderson (1962) carried out a prevalence study of chronic respiratory disease in relation to cigarette smoking and air pollution in Berlin, New Hampshire. This industrial town with almost

18,000 inhabitants is located in a valley near the Canadian border and is almost completely surrounded by mountains. The major industry and chief source of air pollution is a paper and pulp mill.

In this study, the investigators planned to diagnose three disease states—chronic bronchitis, bronchial asthma, and irreversible obstructive lung disease—using simple pulmonary function tests and a standardized interview questionnaire about respiratory symptoms. These standardized methods for assessing pulmonary disease, developed and tested in Great Britain and already used in several studies, would permit a comparison of the findings in Berlin, New Hampshire to those in British and other population groups. At that time there was great interest in the apparent disparity in the relative frequency of chronic bronchitis in Great Britain and the United States, and it was believed that differences in diagnostic criteria and fashions might have been at least partially responsible.

The investigators, in cooperation with the local health department, selected the study sample in two stages. First, using the town's tax roll book which listed the adults in alphabetical order, they randomly selected 36 pages (clusters). Second, from the 36 pages they systematically selected every second name of those in the 25–54-year age stratum and all names of persons aged 55–69. Persons aged 70 and over were listed separately in the town records, and a sample of this age stratum was randomly selected.

Before any data were collected, the local physicians and the state Health Department were contacted and the study was publicized by newspaper and radio. The study subjects were invited by letter to take the study examination at a clinic in the Health Department. Failure to respond led to a telephoned invitation, and if this, in turn, failed, the subject was visited at home and, if he agreed, the interview and physiologic testing were carried out there. Through these persistent efforts, over 95 percent of the 1,261 selected subjects were examined, with the only nonparticipants being those who were away from home during the survey and a few who refused.

Respiratory symptoms were elicited by the standardized interview. Smoking habits, occupational exposures, and previous chest illnesses were also assessed in the interview. Forced expiratory volume, both total (FEV) and during the first second ($FEV_{1.0}$), and peak flow were measured with a recording vitalometer.

The presence of disease was defined by strict criteria. For example, the diagnosis of chronic bronchitis required "the report of bringing up phlegm from the chest six times a day on four days a week for three months in a year, for the past three years or more."

Data analysis revealed a greater prevalence of respiratory disease in men than in women. Furthermore, there was a clear relationship of respiratory disease to smoking. For example, in men the age-adjusted prevalence of chronic bronchitis was:

15.0% in those who had never smoked
18.9% in exsmokers
29.8% in smokers of 1–10 cigarettes per day
34.2% in smokers of 11–20 cigarettes per day
42.3% in smokers of 21–30 cigarettes per day
61.1% in smokers of 31–40 cigarettes per day
75.3% in smokers of 41 or more cigarettes per day

The town was divided into three areas with low, intermediate, and high degrees of air pollution, according to independent measurements of air quality. Residence of study subjects in these three areas showed only an equivocal relationship to respiratory disease. However, if only nonsmokers were considered, it appeared that among men, chronic bronchitis was more apt to be found in residential areas having greater air pollution.

The planned United States–British comparison was later reported by Reid et al. (1964). The findings in Berlin, New Hampshire were compared with those derived from a random sample of urban and rural dwelling adults in Britain examined in a comparable fashion. It was found that the British exceeded the Americans very little in the prevalence of simple chronic bronchitis, characterized by chronic cough and sputum production. However, bronchitis complicated with shortness of breath and repeated acute illnesses was more prevalent in Britain, particularly in urban men.

The prevalence survey in Berlin, New Hampshire was repeated in 1967 using comparable methods (Ferris et al., 1971). It was noted that the prevalence of respiratory disease symptoms was lower in 1967 and that, on the average, there was some improvement in pulmonary function. Because there had also been a fall in air pollution between 1961 and 1967, the authors concluded that this was the probable explanation for the observed improvement. In their analysis they were careful to consider other possible explanations for the change, such as observer differences and the increasing use of filter-tip cigarettes.

The second survey in 1967 illustrates the usefulness of repeated prevalence studies in assessing time trends in disease or other population characteristics, provided that comparable measurement methods are used. The effort and expense of keeping a population under continuous long-term surveillance can often be avoided by conducting careful cross-sectional studies at fairly wide intervals.

Example 2: Cardiovascular Disease in Evans County, Georgia

In Chaps. 4 and 5, emphasis has been placed on descriptive epidemiologic findings as a source of hypotheses for further analytic studies. Another very important source of ideas and hypotheses has been clinical observations by astute and concerned health-care professionals. A physician's observations and interest proved to be a major stimulus for the epidemiologic study of cardiovascular disease in Evans County, Georgia, which began in 1960 as a prevalence study (Hames, 1971, Cassel, 1971*a* and *b*, McDonough et al., 1965).

Dr. Curtis Hames, who practiced in the area, was impressed with the difference in frequency with which he found coronary heart disease occurring in whites and blacks. Although coronary heart disease was a common problem in his white patients, he rarely saw it in blacks, despite the fact that many black patients had hypertension and appeared to consume a high animal-fat diet. Furthermore, the male-female difference in susceptibility to coronary heart disease which was so obvious in whites was not apparent in blacks.

In order to confirm and explain these differences in a systematic fashion, Hames encouraged the interest and participation of epidemiologists and other investigators. Largely due to his excellent rapport with the community, there was nearly complete participation of the selected study subjects.

Evans County is located on flat or slightly rolling terrain about 60 miles inland from the coastal port of Savannah, Georgia; its greatest diameter is 19 miles. The county's economy was, in 1960, primarily agricultural, although its extensive pine forests were a source of lumber, pulpwood, and turpentine. In 1960, the total population was 6,952, of which 66.5% were white and 33.5% were black.

The study sample consisted of a 50 percent random selection of county residents, aged 15 through 39, and all residents aged 40 and over. Of the 3,377 persons chosen, 92 percent underwent the study examination, which consisted of a medical and dietary history, physical examination, urinalysis, serum-cholesterol measurement, electrocardiogram, and chest x-ray. The social class standing of each subject was determined according to the occupation, education, and source of income of the head of the household.

The diagnosis of coronary heart disease required that a subject have either a history of angina pectoris, a history of myocardial infarction, or electrocardiographic evidence of myocardial infarction. Each of these manifestations was defined as definite, probable, possible, or absent according to standard criteria. It is essential for investigators to establish, adhere to, and describe in study reports, strict criteria for the diagnosis of disease so that others may know just what kinds of cases were included or excluded. Strict criteria also permit other investigators to reproduce the study findings, or at least to understand why their own study results might differ.

The major findings of the Evans County prevalence study included confirmation of the initial clinical observations. Coronary heart disease prevalence was indeed lower in blacks than in whites, the difference occurring only in men. Part of this black-white difference could be explained by social class, since white men of lower socioeconomic status had coronary heart disease prevalence rates approaching the low levels in blacks, almost all of whom were in the lower social bracket. The age-adjusted prevalence rates were:

97 per thousand in high-social-class whites
40 per thousand in low-social-class whites
21 per thousand in blacks

The investigators could not explain these racial and social class differences by taking into account differences in other risk factors, including blood pressure, serum-cholesterol levels, body weight, cigarette smoking, and dietary intake. However, it was noted that habitual physical activity, as estimated by type of occupation, was inversely related to coronary heart disease prevalence. Men in occupations involving the most physical exertion (e.g., manual labor, sharecropping) showed the lowest prevalence of coronary heart disease. Since these occupations were primarily engaged in by blacks and low-social-class whites, it appeared that physical activity might explain their relatively low prevalence of coronary heart disease.

As with the Berlin, New Hampshire study, described above, a second examination procedure was carried out several years later, beginning in 1967. However, this was *not* for the purpose of repeating the prevalence study. Rather, the second round of examinations was applied only to the initially examined cohort as part of the follow-up for an incidence study. An initial prevalence survey can be, and often is, used as the first stage of an incidence study, in that it defines and characterizes a *population at risk*—those initially free of the disease being studied. As will be described in Chap. 8, this population at risk can then be followed up for the development of the disease.

The incidence study confirmed the black-white difference in the occurrence of coronary heart disease, but the social class difference in whites was no longer evident. It appeared that this was due to a rapid catching up of the lower-class men to the upper-class men in coronary heart disease risk, during a period when Evans County was changing from an agrarian to an industrial economy. The only subgroup of white men with the same low risk as the blacks were sharecroppers, again suggesting a protective effect of high levels of physical activity.

REFERENCES

Arkin, H., and R. R. Colton, *Tables for Statisticians, 2d ed.* (New York: Barnes and Noble, 1963), pp. 26–27, 158–161.

Cassel, J. C. 1971*a*. Summary of major findings of the Evans County cardiovascular studies. *Arch. Intern. Med.*, **128**:887–889.

Cassel, J. C.: 1971*b*. Review of the 1960 through 1962 cardiovascular disease prevalence study. *Arch. Intern. Med.*, **128**:890–895.

Ferris, B. G., Jr., and D. O. Anderson. 1962. The prevalence of chronic respiratory disease in a New Hampshire town. *Am. Rev. Resp. Dis.*, **86**:165–177.

Ferris, B. G., Jr., I. T. T. Higgins, M. W. Higgins, J. M. Peters, W. F. van Ganse, and M. D. Goldman. 1971. Chronic nonspecific respiratory disease, Berlin, New Hampshire, 1961–1967: a cross-sectional study. *Am. Rev. Resp. Dis.*, **104**:232–244.

Hames, C. G. 1971. Evans County cardiovascular and cerebrovascular epidemiologic study: Introduction. *Arch. Intern. Med.*, **128**:883–886.

Hansen, M. H., W. N. Hurwitz, and W. G. Madow, *Sampling Survey Methods and Theory.* vol. I., Methods and Application. (New York:Wiley 1953), Chaps. 1–3.

Hill, A. B., *Principles of Medical Statistics.* (London: Oxford University Press, 1971), Chaps. 2, 3.

McDonough, J. R., C. G. Hames, S. C. Stulb, and G. E. Garrison. 1965. Coronary heart disease among Negroes and whites in Evans County, Georgia. *J. Chron. Dis.*, **18**:443–468.

Reid, D. D., D. O. Anderson, B. G. Ferris, Jr., and C. M. Fletcher. 1964. An Anglo-American comparison of the prevalence of bronchitis. *Brit. Med. J.*, **2**:1487–1491.