Methods of Measuring Incidence and Prevalence of Disease*

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

National Institutes of Health, Bethesda, Md.

IN the following discussion I shall discuss morbidity as a group phenomenon, that is, as the study of the natural history of disease in a population. There are, of course, other uses of morbidity data, but those lie outside the present discussion. The measurement of illness as a group phenomenon requires more than a mere count of the number of persons who are sick; it requires information necessary for the calculation of morbidity rates.

A rate, as used here, has three essential elements, a population exposed to the risk of some event, an enumeration of the frequency of the event during some interval, and a unit of time in which the rate will be expressed. The definition of the population exposed to risk and the unit of time need little comment. The probability of an event happening obviously depends upon the size and composition of the population exposed to risk and the length of time this population is exposed. A morbidity rate, of course, may be expressed in a unit of time different from the interval of observation. The crux of a morbidity rate is the event to be observed.

Since we do not have a reliable index of health except absence of illness, the event to be observed is an illness. But how shall illness be defined? The distinction between the living and the dead is clear-cut and easily made; a similar distinction cannot be made between the well and the sick since these do not comprise a true dichotomy. Except for acute illness due to certain external agents the transition from health to ill health is frequently almost imperceptible.

Instead of being represented by a dichotomy, the state of health of a group of persons may more realistically be portrayed by a continuous scale ranging from the perfectly healthy to the complete absence of health, that is death. Between these two extremes, except for certain conditions, there is no sharp line of demarkation between various degrees of ill health. Unfortunately, there is no agreement concerning criteria by which the various degrees of ill health may be identified and differentiated. Nevertheless, some critical points on the scale of health which will divide the population into seemingly distinct subgroups must be selected when practical investigations of sickness are carried out, but the dependence of reported illness rates upon this selection should not be forgotten.

The following rough classification of a population with respect to health will suffice for this discussion even though it does not represent a scale:

1. Free from any defect or disease.
2. Having congenital or acquired defects, impairments or diseases which cause no current disability. Examples of these would be

* Presented at a Joint Session of the Epidemiology, Health Officers, Medical Care, and Statistics Sections of the American Public Health Association at the Seventy-eighth Annual Meeting in St. Louis, Mo., November 3, 1950.
hemophilia, visual or auditory defects, various kinds of orthopedic impairments and quiescent chronic diseases.
3. With latent or incipient disease usually unknown to the person affected. In most instances these conditions can be discovered only by special diagnostic tests. Examples of this would be early tuberculosis or diabetes.
4. Sick. This category includes the persons who recognize that they are not well. It covers a wide range on the scale of health. Moreover, an individual's belief that he is ill is affected by many factors in addition to his physiological condition. It frequently is desirable to subclassify this category according to the degree of disability caused by the illness. Many chronic diseases may become quiescent and cause no disability during a given period of observation. In the interval between acute episodes, they resemble the latent or acquired diseases mentioned above.

It is often claimed that the only illnesses which can be studied with reliability are those diagnosed by a physician. This claim must be rejected since its acceptance would exclude from study a significant proportion of cases which otherwise would fall in the last three categories of the above classification.

The three principal criteria for establishing ill health are:
1. The opinion of the individual concerned. If a person believes himself to be ill, he should be considered so, even though a physician may not be able to identify a physical cause. Of course, the patient’s statement of the cause of his illness frequently may be erroneous.
2. Clinical examination by a physician.
3. Diagnostic tests.

Some misunderstanding exists concerning the role of diagnostic tests in the study of ill health. Contrary to a widespread belief diagnostic tests, in general, do not make diagnoses. Nor do they necessarily objectively establish either the fact or the cause of illness. A diagnostic test may be qualitative, resulting in a picture, for example an x-ray film, which must be interpreted by a reader. Or a diagnostic test may yield a reading on a quantitative scale. Criteria for relating specified scale values with various degrees of ill health must be established before the results of such tests can be interpreted in terms of illness. The validity of the calibration of a test and the competence of the interpreter of the result will materially affect the number of cases of illness discovered in a population.

Assuming that we have been able to define illness and have satisfactory criteria to establish its existence there remains the further problem of classifying the types of illness which can be identified. There are several principles by which illnesses can be classified, but the only one which is pertinent here is a generic classification.

Unlike death which occurs at a point of time, illness exists during an interval of time. Consequently a statistical measure of the amount of illness in a population must make appropriate allowance for the time dimension of morbidity. If a population is observed for an arbitrary time interval, the cases of illness which are identified can be classified with respect to time as follows:
1. Existing prior to the beginning of the interval, continuing throughout the interval and still existing at the end of the interval.
2. Existing prior to the beginning of the interval and terminating during the interval.
3. Arising during the interval and still existing at the end of the interval.
4. Arising during the interval and terminating before the end of the interval.

The unifying principle underlying these four categories of illness can be discovered by describing them in terms of the theory of population. Two functions are necessary for the characterization of the development of a population, an increment rate which determines the rate at which new members enter the population and a decrement rate which describes the rate at which members leave the population. The size and composition of a population at any time is a joint function of the increment and decrement rates. If the parameters of these two functions are known, many
characteristics of a population can readily be demonstrated.

Reexamining the categories of cases of illness enumerated above, we see that the latter two comprise all cases arising during the interval; these two categories combined may be used to define the rate at which new members are added to the population of sick persons. The increment rate of a population of sick persons is called the incidence rate of illness.

A little reflection will show that the decrement rate of a population of sick persons which is a function expressing the duration of illness cannot be derived from the four categories of cases enumerated above except under very special conditions. However, these categories combined represent the total number of persons in the sick population during the time interval. This number can be used to compute a prevalence rate of illness, that is, a measure of the aggregate size of the population of sick persons.

Unlike the incidence rate, the prevalence rate is a direct function of the length of the interval of observation. This way of looking at the prevalence rate of illness is contrary to that commonly accepted. Prevalence usually is defined not as a rate but as a ratio of two quantities and the concept of time is not involved. In other words, prevalence is said to express the proportion of a population which is sick at a specific moment of time. This concept is incomplete and inconsistent, and is not in conformity with what actually occurs.

Both the incidence rate and the duration of illness are continuous functions. The size of the population of sick persons, which is a joint function of these also, can be expressed as a continuous function. Theoretically, no matter how small the interval of observation, the four types of cases enumerated above can be identified. I know of no reason why illnesses must all terminate at midnight, although due to social customs they may appear to do so.

Prevalence then is represented by the sum of the cases of illness existing at the start of an interval and the new cases developing during the interval. The length of the interval of observation must always be specified if a prevalence rate is to be correctly interpreted, for we may speak of the number of persons who are sick at any time during a given day, week, month, year, or other arbitrary interval. We can, of course, also compute an instantaneous prevalence rate, that is the average number of persons sick at any instant during an interval. This presumably is what many people have in mind when speaking of a prevalence rate.

There are many other factors to consider in the definition of a morbidity rate. These arise from the fact that a person may have more than one attack of illness during a given time interval. Furthermore, two or more diseases may exist simultaneously or sequentially during any attack. Consequently morbidity rates may refer either to persons or to illnesses. In defining incidence rates of illnesses, account must be taken of whether a new case is to be defined as the first or initial attack during the individual’s lifetime or as a separate attack during the period of observation. However they are defined, “new” cases must be differentiated from “recurrences” and “relapses.” For some diseases, such as measles, the initial attack gives immunity to subsequent attacks; for other diseases, such as bronchitis, the condition may persist in a quiescent state causing no disability after the initial attack, but acute disabling episodes will recur from time to time; while still other diseases, for example acute upper respiratory infections, may attack the same person repeatedly with apparent complete recovery between the attacks. No further elaboration of these problems can be made at this time.

In studies of the natural history of disease, the fundamental morbidity rate
is incidence since this rate measures the increment to the population of sick persons and in conjunction with the decrement rate determines the size and composition of this population. From these two rates, the prevalence rate can be derived. For some simple types of populations, average prevalence equals the product of average incidence and average duration in the sick population. Prevalence, especially for chronic diseases, frequently can be determined more easily than incidence but this should not be allowed to obscure the fundamental nature of the latter.

Methods of studying the natural history of illness may be evaluated by the following principles:

1. The size and composition of the base population must be known. It is true that certain aspects of the way a disease manifests itself, such as the relative susceptibility of different organs or tissues, the length of the latent period, the stage at diagnosis and duration of symptoms, can be profitably studied from a knowledge of cases of illness alone. But incidence and prevalence and their relationship to the demographic characteristics of individuals and to environmental conditions cannot be reliably determined without knowledge of the population exposed to the risk of illness.

2. Generalization of findings beyond the specific population studied should be possible. This requires that the population studied must be a representative sample of the larger population for which the generalizations are thought to hold true. The representativeness of a sample can be insured only if the sample has been chosen in accordance with tested statistical principles.

3. The study should extend over a period of time. Except for a few conditions with abrupt termination, illness has a time dimension. The etiology, the latent period, the duration of illness, and other aspects of many diseases cannot be understood without extending observations over a time period. This may be accomplished by keeping a population under continuous observation or by repeated visits at given intervals.

4. As much as possible of the entire scale of ill health should be included. Many studies in the past have been based on the fallacious assumption that more valid information about illness would be obtained by the use of an "objective" definition such as cases discharged from hospital. This merely insures that the data will be unrepresentative of even the recognized illness in the population. A more desirable procedure is to use as inclusive a definition as possible but to differentiate between various degrees of illness.

5. It should be possible to obtain detailed verifiable information about illness. Although the affected person's statement concerning the existence and severity of illness may be accepted, more detailed information concerning etiology, diagnosis, and other aspects of morbidity should be verified by the most reliable means available. As we learn more about morbidity from a specific disease the greater becomes the demand for detailed verifiable data.

6. The volume of records should not be so large that all except the most simple analysis becomes impossible.

With these principles in mind let us briefly examine some methods which have been used to determine incidence and prevalence rates.

**UNIVERSAL REPORTING**

The most widespread method of collecting morbidity data is reports, usually compulsory from physicians. Until recently such reports have been confined almost entirely to infectious and contagious diseases and to certain occupational diseases. The increasing attention paid to chronic disease has resulted in attempts to collect morbidity for these diseases in the same way. This attempt is not likely to be generally successful.

As a source of morbidity data, universal reporting by physicians rarely has been satisfactory. Even when the information to be reported has been reduced to the barest essentials, in many instances to nothing more than the number of cases of disease, reporting has been seriously incomplete.

Chronic diseases do not occur in epidemic form. The incidence rate changes only slowly if at all. Many diseases may
persist for long periods of time and the person affected may receive medical care from several sources with the consequent necessity for the accurate identification of each case. Since so little is known about the epidemiology of many of these diseases their relationship to a wide variety of factors must be investigated.

Universal reporting, wherever it has been even partially successful, has been required for administrative or legal purposes. This is true for the registration of births, deaths, marriages, divorces, infectious diseases, and occupational diseases. The statistical data resulting from this registration arise as by-products of the system and are not the primary reason for its existence. As yet, the principal reason for the reporting of chronic diseases has been the statistical data which would be obtained. Until registration of cases of disease can be justified for reasons other than the statistics which will result, the universal reporting of chronic disease is not likely to be a useful source of morbidity data.

One justification for reporting of chronic disease is the provision of health or medical services. Unlike most contagious diseases for which the acute manifestations last only a few days or weeks, many chronic diseases persist for months or even years. The need for medical care or for periodic examination continues for corresponding periods of time.

Moreover, the onset of many chronic diseases is so slow and imperceptible that suspected cases should be kept under observation for extended periods of time. If direct services are to be given to persons with actual or suspected chronic disease such cases must be reported and records must be kept. A case register can be a very useful administrative tool in providing such services and at the same time if all cases arising in a defined population are included, can be organized so that it will yield valid morbidity data.

RECORDS FROM HEALTH, MEDICAL OR INSURANCE PROGRAMS

One of the first thoughts that come to the mind of many persons when the subject of morbidity statistics arises is that the records of hospitals, clinics, prepaid medical and hospitalization plans, health insurance plans, and similar organizations are the most reliable sources of data. This is the great mirage of morbidity statistics. However useful such records may be for the administration of the particular program from which they arise or as a measure of the demand for medical care, they cannot, as a rule, be expected to yield the kinds of morbidity data under discussion here. There are many reasons for this.

In some instances, in particular, hospital and clinic records, the size and composition of the population from which the sick persons come are unknown. This can be obviated by collecting records from all hospitals in an area for which the population statistics are available from the Bureau of the Census. But if this is done the volume of records may become unwieldy and unnecessarily expensive to analyze. In practically all instances, the records cover either only part of the illnesses which develop in the population or only selected groups of the population itself. Although these types of records usually are assumed to provide more precise morbidity data, particularly with reference to diagnosis, this frequently is not true since the records are kept primarily for administrative or fiscal purposes. Very little effort is devoted to maintaining accurate illness data.

POPULATION SURVEYS

Included here are all investigations wherein the selection of the individuals to be studied does not depend upon a knowledge of the presence or absence of ill health. This method of collecting morbidity data meets more of the criteria for determining the incidence and preva-
lence of disease than any of the other methods so far devised. The population exposed is defined as part of the process of obtaining information about illness. Moreover, illness can be related to a greater variety of demographic characteristics and environmental factors than if the population data must be obtained from census statistics. Perhaps the greatest advantage of the population survey is that it allows the investigator to define the kinds of illness which will be included and to design studies to answer specific questions. Its flexibility in this respect is in striking contrast to the inflexibility of obtaining morbidity data from the records of operating programs.

The population survey method, however, has disadvantages. It is expensive, relative to other methods, if a large population group is covered. This disadvantage has been partially overcome by recent advances in the knowledge and techniques of conducting sampling studies. The informant can report only on illness of which he is aware, but this usually covers a greater range of the scale of ill health than morbidity data obtained from records. The completeness of reporting decreases rapidly with an increase in the length of time for which data are requested. This source of error can be controlled by frequent visits or by having daily records kept, but these are not always practical.

The greatest drawback of the population survey is that, although it is well adapted to obtaining gross data on the total incidence and prevalence of disease and the relationship of these to a variety of demographic and environmental factors, it cannot provide reliable detailed medical information concerning the cause of illness and the way in which a disease affects the person attacked. For example, in studying morbidity from cancer we want to know the primary site, the histological type, the stage of disease at diagnosis, and the method of establishing the diagnosis. A lay informant cannot be expected to have knowledge of such facts. The same problem exists for cardiovascular diseases and a variety of other disorders.

LONGITUDINAL STUDIES
A method of studying morbidity which has been used only occasionally but which merits more serious consideration than it has received in the past is the longitudinal study of a population. Many aspects of illness, particularly illness due to chronic disease must be observed over a long interval of time. The interrelationship of various diseases and the effect of different possible agents upon the development of disease can be studied more effectively by the cohort or generation method than by the cross-section or retrospective methods. However, longitudinal studies are expensive if large numbers are included and do not yield quick results. In addition, there is some question concerning generalization of the findings to other population groups since such studies are usually confined to the population of a single community.

DIAGNOSTIC SURVEYS
The methods so far discussed have been used to study various kinds of known illness, conditions which are recognized by the person affected or by a physician. As was pointed out above, the onset of many diseases is slow and may not be recognized at all, or if recognized, thought to be of no importance. In recent years the development of efficient diagnostic tests has made practicable the screening of large groups of supposedly well persons for unsuspected illness. Used first in screening for tuberculosis and venereal disease separately this technique recently has been extended to screening for several diseases simultaneously. If properly carried out, this technique offers great promise for further extending the observable range of the scale of ill health toward the origin of illness.
Two questions arise with reference to the use of diagnostic tests to detect hitherto unsuspected disease. First, what proportion of the total number of cases which the tests are capable of detecting are found? Second, how valid are the tests, that is, how many persons are said to have a suspected disease when in fact they do not, and conversely how many persons are said to be well when in fact they have the disease in question. Unfortunately too little attention has been devoted to obtaining answers to these questions in the surveys which have been conducted so far, even though these answers are essential for the evaluation of the techniques both as a case finding method and as a source of morbidity data.

Before the results of multiphasic examinations can be used as a measure of prevalence of disease we must know much more about the validity of the tests, the reliability of the tests, and the effect of incomplete coverage of the population. The usual method of checking the validity of a test is to reexamine suspected cases by other diagnostic procedures which are accepted as the standard for diagnosing well defined cases of disease. But this is not necessarily conclusive since the same result will be obtained irrespective of whether the diagnostic test is in error or is, in fact, detecting disease at a stage in its development when the standard test is too insensitive to discover it. The only way of deciding this question is to observe the suspected cases which are classed as false-positives for a period of time and see whether the disease progresses to the extent that it can be diagnosed by the standard techniques.

One check of the reliability of the 70 mm. film for the detection of tuberculosis by having two independent readings showed that the use of a second reader increased the number of suspected cases to be called back for verification by 46 per cent. The same study also revealed that those not reporting for examination were unrepresentative of the total population with respect to age and economic status. Based upon a sample who subsequently reported for examination, this group had a somewhat higher prevalence rate of tuberculosis. However, this problem needs much further study.

The principal conclusion to be drawn from this brief sketch of methods of measuring incidence and prevalence of disease is that no single method is satisfactory by itself. A population survey of a sample of a large population is probably the best way of obtaining incidence and prevalence rates for all forms of illness combined. If this is supplemented with verification of diagnoses by means of information from physicians and hospitals, reliable prevalence rates for specific diseases also can be obtained. To be most useful the survey should not be a one-time study but should be repeated periodically.

The range of the scale of ill health included could be extended by combining the population survey with multiphasic screening. Since the latter is not practical on a sampling basis this would have to be done by an intensive survey of the population of a local area. The data from the more extensive sampling population survey could be used to determine how widely the results of the intensive study could be generalized.

The population survey, especially if combined with some form of diagnostic screening and verification of diagnoses, can be expected to yield reliable incidence rates for all forms of illness and prevalence rates for specific diseases as well as data for studying the relationship of these to different demographic and environmental factors. It is not very efficient for measuring the incidence of specific diseases, especially diseases which do not have an abrupt onset and which persist for long periods of time. Consequently, the population survey
should be supplemented with intensive studies of illness in complete population groups. We cannot be satisfied only with prevalence data or with the gross incidence data which can be obtained by a general population survey.

The intensive investigation of morbidity by a longitudinal study of complete population groups offers real promise. Only in this way can the dynamics of illness be understood.

The question is not, Is the extensive population survey or the intensive longitudinal study of a complete population the best way of studying morbidity? Each is necessary because they are complementary. There is no single method which is the answer to the prayer for data on the incidence and prevalence of illness. The only answer to this prayer is a study of morbidity, and the two methods just mentioned have the most to offer for this purpose.

REFERENCE


Statistics Section

The Statistics Section plans to give over one of its sessions at the San Francisco A.P.H.A. Annual Meeting to contributed papers. It invites the membership of the Section, particularly, and any others who may wish to contribute, to submit an outline of the material to the Section Secretary, Carl L. Erhardt, 125 Worth Street, New York City, by April 30, 1951.

The Section would particularly like to have papers concerned with the evaluation of public health programs or with the development of specific methodology in connection with a public health program. The Section hopes in this way to increase the participation of the membership in the annual meetings and to make the membership, as a whole, more aware of the relationship of statistics to public health activities in general.