**Principles of Epidemiology and Epidemiologic Methods**

"I keep six honest serving men; they taught me all I know. Their names are what, why, when, how, where and who."

**Epidemiology** is the basic science of preventive and social medicine. Although of ancient lineage, it made only slow progress up to the start of the 20th century. Epidemiology has evolved rapidly during the past few decades. Its ramifications cover not only study of disease distribution and causation (and thereby prevention), but also health and health-related events occurring in human population. Modern epidemiology has entered the most exciting phase of its evolution. By identifying risk factors of chronic disease, evaluating treatment modalities and health services, it has provided new opportunities for prevention, treatment, planning and improving the effectiveness and efficiency of health services. The current interest of medical sciences in epidemiology has given rise to newer off-shoots such as infectious disease epidemiology, chronic disease epidemiology, clinical epidemiology, serological epidemiology, cancer epidemiology, malaria epidemiology, neuro epidemiology, genetic epidemiology, occupational epidemiology, psychosocial epidemiology, and so on. This trend is bound to increase in view of the increasing importance given to the pursuit of epidemiological studies. That these studies have added substantially to the advancement of medical knowledge is indisputable. This Chapter studies the basic concepts and principles of epidemiology as an introduction to the subject.

**History**

Epidemiology began with Adam and Eve, both trying to investigate the qualities of the "forbidden fruit". Epidemiology is derived from the word epidemic (epi=among; demo=people; logos=study), which is a very old word dating back to the 3rd century B.C. The foundation of epidemiology was laid in the 19th century, when a few classic studies made a major contribution to the saving of life. Mention is made of an Epidemiological Society in London in 1850s under the presidency of the Earl of Shaftesbury (1). The Society’s main concern was the investigation of infectious diseases. The sudden growth of bacteriology had smothered the development of epidemiology in the Universities.

In the United States, Winslow and Sedgwick both lectured in epidemiology in the early 1920s, although the subject was not given departmental status. In 1927, W.H. Frost became the first professor of epidemiology in US. Later Major Greenwood became the first professor of epidemiology and medical statistics in the University of London (1). Epidemiology has grown rapidly during the past few decades. It has now become firmly established in medical education.

There appears to be almost as many definitions of epidemiology as there are authors who have written on the subject, ranging from Hippocrates to those of the present day. A short list is given below (2, 3):

1. That branch of medical science which treats epidemics (Parkin, 1873).
2. The science of the mass phenomena of infectious diseases (Frost, 1927).
3. The study of disease, any disease, as a mass phenomenon (Greenwood, 1934), and

**Definition**

Epidemiology has been defined by John M. Last in 1988 as:

"The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to the control of health problems."

The wide variety of meanings attached to epidemiology is the expression of the wide ranging subject–matter. The diseases included in the subject–matter have increased from those which occur in epidemics to include those infectious diseases which are endemic in nature, and more recently chronic diseases, accidents and mental health. Modern epidemiology has also taken within its scope the study of health-related states, events and "facts of life" occurring in human population. This includes study of the health services used by the population, and to measure their impact. Epidemiology, like public health itself, is often more concerned with the well-being of society as a whole, than with the well-being of individuals.

Although there is no single definition to which all epidemiologists subscribe, three components are common to most of them. First, studies of disease frequency; second, studies of the distribution; and third, studies of the determinants. Each of these components confers an important message.

1. **Disease frequency**

   Inherent in the definition of epidemiology is measurement of frequency of disease, disability or death, and summarizing this information in the form of rates and ratios (e.g., prevalence rate, incidence rate, death rate, etc). Thus the basic measure of disease frequency is a rate or ratio. These
rates are essential for comparing disease frequency in different populations or subgroups of the same population in relation to suspected causal factors. Such comparisons may yield important clues to disease aetiology. This is a vital step in the development of strategies for prevention or control of health problems.

Equally, epidemiology is also concerned with the measurement of health-related events and states in the community (e.g., health needs, demands, activities, tasks, health care utilization) and variables such as blood pressure, serum cholesterol, height, weight, etc. In this respect, epidemiology has the features of a quantitative science. Much of the subject matter of measurement of disease and health-related events falls in the domain of biostatistics, which is a basic tool of epidemiology.

2. Distribution of disease

It is well-known that disease, or for that matter health, is not uniformly distributed in human populations. The basic tenet of epidemiology is that the distribution of disease occurs in patterns in a community (3) and that the patterns may lead to the generation of hypotheses about causative (or risk) factors. An important function of epidemiology is to study these distribution patterns in the various subgroups of the population by time, place and person. That is, the epidemiologist examines whether there has been an increase or decrease of disease over time span; whether there is a higher concentration of disease in one geographic area than in others; whether the disease occurs more often in men or in a particular age-group, and whether most characteristics or behaviour of those affected are different from those not affected (4). Epidemiology addresses itself to a study of these variations or patterns, which may suggest or lead to measures to control or prevent the disease. An important outcome of this study is formulation of aetiological hypothesis. This aspect of epidemiology is known as "descriptive epidemiology".

3. Determinants of disease

A unique feature of epidemiology is to test aetiological hypotheses and identify the underlying causes (or risk factors) of disease. This requires the use of epidemiological principles and methods. This is the real substance of epidemiology. This aspect of epidemiology is known as "analytical epidemiology". Analytical strategies help in developing scientifically sound health programmes, interventions and policies. In recent years, analytical studies have contributed vastly to our understanding of the determinants of chronic diseases, e.g., lung cancer and cardiovascular diseases.

**Aims of epidemiology**

According to the International Epidemiological Association (IEA), epidemiology has three main aims (5):

a. to describe the distribution and magnitude of health and disease problems in human populations

b. to identify aetiological factors (risk factors) in the pathogenesis of disease; and

c. to provide the data essential to the planning, implementation and evaluation of services for the prevention, control and treatment of disease and to the setting up of priorities among those services.

In order to fulfil these aims, three rather different classes of epidemiological studies may be mentioned: descriptive studies, analytical studies, and experimental or intervention studies (6). These studies are described in the following pages.

The ultimate aim of epidemiology is to lead to effective action:

a. to eliminate or reduce the health problem or its consequences; and

b. to promote the health and well-being of society as a whole.

**Epidemiology and clinical medicine**

The basic difference between epidemiology and clinical medicine is that in epidemiology, the unit of study is a "defined population" or "population at-risk"; in clinical medicine, the unit of study is a "case" or "cases". In clinical medicine, the physician is concerned with disease in the individual patient, whereas the epidemiologist is concerned with disease patterns in the entire population. Epidemiology is thus concerned with both the sick and healthy. It has been stated that clinicians are interested in cases with the disease, the statistician with the population from which the cases are derived, and the epidemiologist is interested in the relationship between cases and the population in the form of a rate (7).

In clinical medicine, the physician seeks a diagnosis from which he derives a prognosis and prescribes specific treatment. In epidemiology, an analogous situation exists. The epidemiologist is confronted with relevant data derived from a particular epidemiological study. He seeks to identify a particular source of infection, a mode of spread or an aetiological factor in order to determine a future trend and recommend specific control measures (8). The epidemiologist also evaluates the outcome of preventive and therapeutic measures instituted which provides the necessary guidance and feedback to the health care administrator for effective management of public health programmes.

In clinical medicine, the patient comes to the doctor; in epidemiology, the investigator goes out into the community to find persons who have the disease or experience of the suspected causal factor in question. Clinical medicine is based on biomedical concepts with an ever-increasing concern for refining the technique of diagnosis and treatment at the individual level. The subject matter of clinical medicine is easily "perceived" by such techniques as clinical and laboratory examinations including postmortem reports. In contrast, the subject matter of epidemiology is "conceptual" and can only be symbolized in the form of tables and graphs (9).

Finally, it may be stated that clinical medicine and epidemiology are not antagonistic. Both are closely related, co-existent and mutually helpful. Most epidemiological enquiries could never be established without appropriate clinical consideration as to how the disease in question can be identified among individuals comprising the group under scrutiny. Likewise, a knowledge of prevalence, aetiology and prognosis derived from epidemiological research is important to the clinician for the diagnosis and management of individual patients and their families (9).

**Epidemiological approach**

The epidemiological approach to problems of health and disease is based on two major foundations:

a. Asking questions

b. Making comparisons.
a. Asking questions

Epidemiology has been defined as “a means of learning or asking questions... and getting answers that lead to further questions” (10). For example, the following questions could be asked (11):

RELATED TO HEALTH EVENTS
a. What is the event ? (the problem)
b. What is its magnitude?
c. Where did it happen?
d. When did it happen?
e. Who are affected?
f. Why did it happen?

RELATED TO HEALTH ACTION
a. What can be done to reduce this problem and its consequences ?
b. How can it be prevented in the future ?
c. What action should be taken by the community ? By the health services? By other sectors? Where and for whom these activities be carried out ?
d. What resources are required ? How are the activities to be organized ?
e. What difficulties may arise, and how might they be overcome ?

Answer to the above questions may provide clues to disease aetiology, and help the epidemiologist to guide planning and evaluation.

b. Making comparisons

The basic approach in epidemiology is to make comparisons and draw inferences. This may be comparison of two (or more groups) – one group having the disease (or exposed to risk factor) and the other group(s) not having the disease (or not exposed to risk factor), or comparison between individuals. By making comparisons, the epidemiologist tries to find out the crucial differences in the host and environmental factors between those affected and not affected. In short the epidemiologist weighs, balances and contrasts. Clues to aetiology come from such comparisons.

One of the first considerations before making comparisons is to ensure what is known as “comparability” between the study and control groups. In other words, both the groups should be similar so that “like can be compared with like”. For facts to be comparable, they must be accurate, and they must be gathered in a uniform way. For example, the study and control groups should be similar with regard to their age and sex composition, and similar other pertinent variables. The best method of ensuring comparability, in such cases, is by randomization or random allocation (see page 82). Where random allocation is not possible (as in case control and cohort studies) what is known as “matching” is done for selected characteristics that might confound the interpretation of results. Another alternative is standardization which usually has a limited application to a few characteristics such as age, sex and parity. These biostatistical concepts are elaborated in the following pages. It may be mentioned that international comparisons may be difficult because of differences in terminology. It requires standardization of definitions, classifications, criteria and nomenclature.

BASIC MEASUREMENTS IN EPIDEMIOLOGY

Epidemiology focuses, among other things, on measurement of mortality and morbidity in human populations. The first requirement is therefore definition of what is to be measured and establishment of criteria or standards by which it can be measured. This is not only a prerequisite of epidemiological studies, but also one of its goals (12). The clinician may not require a precise definition of disease (e.g., migraine) for immediate patient care, but the epidemiologist needs a definition (a) that is acceptable and applicable to its use in large populations; and (b) that is precise and valid, to enable him to identify those who have the disease from those who do not (9). Clear definitions help to minimize errors in classification of data. Standardized methods of observation and recording are therefore essential before commencing any epidemiological study.

Measurements in Epidemiology

The scope of measurements in epidemiology is very broad and unlimited and includes the following: (13)

a. Measurement of mortality
b. Measurement of morbidity
c. Measurement of disability
d. Measurement of natality
e. Measurement of the presence, absence or distribution of the characteristic or attributes of the disease
f. Measurement of medical needs, health care facilities, utilization of health services and other health-related events
g. Measurement of the presence, absence or distribution of the environmental and other factors suspected of causing the disease, and
h. Measurement of demographic variables.

Inspite of a wide range of presently available measurements, there are many areas which are not fully covered. As for example, measurement of the psycho-social aspects of health and disease. The components of well-being need to be better identified.

The basic requirements of measurements are validity, reliability, accuracy, sensitivity and specificity. These are discussed in the next chapter. Finally, measurement errors are unavoidable, no matter where and by whom measurements are taken. The purpose of quality control in measurement is, therefore, not to eliminate errors, but to reduce them as much as possible or at least to an acceptable level.

In the above connection, the following terminology needs explanation: (a) Variate: Any piece of information referring to the patient or his disease is called a variate. A variate can be discrete, that is it can be present or absent, e.g., cancer lung, broken leg, or rash in measles or it can be continuously distributed, e.g., blood pressure, serum cholesterol, height, etc. (b) Circumstance: Any factor in the environment that might be suspected of causing a disease, e.g., air pollution, polluted water, etc (9).

The frequency of a discrete variable or circumstance can be expressed as a rate in relation to population. The frequency of continuously distributed variables or circumstances is expressed in the form of a frequency distribution using the summarizing indices of mean, centiles, standard deviations, etc.
Tools of measurement

The epidemiologist usually expresses disease magnitude as a rate, ratio or proportion. A clear understanding of the term is required for proper interpretation of epidemiological data. The basic tools of measurement in epidemiology are:

1. Rates
2. Ratios, and
3. Proportions

1. RATE

When we say there were 500 deaths from motor vehicle accidents in City A during 2010, it is just nothing more than counting deaths in that city during that particular year. Such a statement might be sufficient for the municipal administrator to provide necessary health services. But it conveys no meaning to an epidemiologist who is interested in comparing the frequency of accidents in City A with that in City B. To allow such comparisons, the frequency must be expressed as a rate.

A rate measures the occurrence of some particular event (development of disease or the occurrence of death) in a population during a given time period. It is a statement of the risk of developing a condition. It indicates the change in some event that takes place in a population over a period of time. An example of a typical rate is the death rate. It is written as below:

\[
\text{Death rate} = \frac{\text{Number of deaths in one year}}{\text{Mid-year population}} \times 1000
\]

A rate comprises the following elements - numerator, denominator, time specification and multiplier. The time dimension is usually a calendar year. The rate is expressed per 1000 or some other round figure (10,000, 100,000) selected according to the convenience or convention to avoid fractions.

The various categories of rates are:

1. (1) Crude rates: These are the actual observed rates such as the birth and death rates. Crude rates are also known as unstandardized rates.
2. Specific rates: These are the actual observed rates due to specific causes (e.g., tuberculosis); or occurring in specific groups (e.g., age-sex groups) or during specific time periods (e.g., annual, monthly or weekly rates).
3. Standardized rates: These are obtained by direct or indirect method of standardization or adjustment, e.g., age and sex standardized rates (see page 58, 59).

2. RATIO

Another measure of disease frequency is a ratio. It expresses a relation in size between two random quantities. The numerator is not a component of the denominator. The numerator and denominator may involve an interval of time or may be instantaneous in time. Broadly, ratio is the result of dividing one quantity by another. It is expressed in the form of:

\[
x : y \text{ or } \frac{x}{y}
\]

Example 1:

The ratio of white blood cells relative to red cells is 1:600 or 1/600, meaning that for each white cell, there are 600 red cells.

Example 2:

The number of children with scabies at a certain time

The number of children with malnutrition at a certain time

Other examples include: sex-ratio, doctor-population ratio, child–woman ratio, etc.

3. PROPORTION

A proportion is a ratio which indicates the relation in magnitude of a part of the whole. The numerator is always included in the denominator. A proportion is usually expressed as a percentage.

\[
\frac{\text{The number of children with scabies at a certain time}}{\text{The total number of children in the village at the same time}} \times 100
\]

CONCEPT OF NUMERATOR AND DENOMINATOR

1. Numerator

Numerator refers to the number of times an event (e.g., sickness, birth, death, episodes of sickness) has occurred in a population, during a specified time–period. The numerator is a component of the denominator in calculating a rate, but not in a ratio.

2. Denominator

Numerator has little meaning unless it is related to the denominator. The epidemiologist has to choose an appropriate denominator while calculating a rate. It may be (a) related to the population, or (b) related to the total events.

a. Related to the population

The denominators related to the population comprise the following: (i) MID-YEAR POPULATION: Because the population size changes daily due to births, deaths and migration, the mid-year population is commonly chosen as a denominator. The mid-point refers to the population estimated as on the first of July of an year. (ii) POPULATION AT-RISK: This is an important concept in epidemiology because it focuses on groups at risk of disease rather than on individuals. The term is applied to all those to whom an event could have happened whether it did or not. For example, if we are determining the rate of accidents for a town, the population at risk is all the people in the town. But sometimes, it may be necessary to exclude people because they are not at risk, as for example, in food poisoning, only those who ate the food are at risk of becoming ill. Similarly in calculating “general fertility rate”, the denominator is restricted to women of child-bearing age (i.e., 15–49 years); older women and little girls are excluded because they are not “at risk” of becoming pregnant. In short, “population at risk” is restricted solely to those who are capable of having or acquiring the disease or condition in question. (iii) PERSON–TIME: In some epidemiological studies (e.g., cohort studies), persons may enter the study at different times. Consequently, they are under observation for varying time periods. In such cases, the denominator is a combination of persons and time. The most frequently used person–time is person–years. Sometimes, this may be person–months, person–weeks or man-hours. For example, if 10 persons remain in the study for 10 years, there are said to be 100 person–years of observation. The same figure would be derived if 100 persons were under observation for one year. These denominators have the advantage of summarizing the experience of persons with different
MEASUREMENT OF MORTALITY

Traditionally and universally, most epidemiological studies begin with mortality data. Mortality data are relatively easy to obtain, and, in many countries, reasonably accurate. Many countries have routine systems for collecting mortality data. Each year, information on deaths is analyzed and the resulting tabulations are made available by each government. Mortality data provide the starting point for many epidemiological studies. In fact, they are the major resource for the epidemiologist.

International Death Certificate

The basis of mortality data is the Death Certificate. So we first look at death certification for ascertaining the frequency of disease in a population. For ensuring national and international comparability, it is very necessary to have a uniform and standardized system of recording and classifying deaths. The death certificate recommended by WHO for international use is given in Fig. 1.

It will be seen from Fig. 1 that the international death certificate is in two parts. Part I deals with the immediate cause, and the underlying cause which started the whole trend of events leading to death. The underlying cause of death is recorded on line (c). In the example cited, the underlying cause of death is strangulated hernia. After operation, the patient developed bronchopneumonia as a complication which ended in death. The concept of "underlying cause" is the essence of the international death certificate. It is defined as (a) the disease or injury which initiated the train of morbid events leading directly to death or (b) the circumstances of the accident or violence which produced the fatal injury. In Part II is recorded any significant associated diseases that contributed to the death but did not directly lead to it.

Death Certificate used in India

In order to improve the quality of maternal mortality and infant mortality data and to provide alternative method of collecting data on deaths during pregnancy and infancy, a set of questions are added to the basic structure of international death certificate for use in India.

Limitations of mortality data

Mortality data are not without limitations. Problems are posed by (a) Incomplete reporting of deaths. This is not a problem in developed countries, but in India and other developing countries, this may be considerable. (b) Lack of accuracy: That is inaccuracies in the recording of age and cause of death. The practice of medical certification of death is not widespread. If it does exist, the cause of death is often inaccurate or incomplete due to such difficulties as lack of diagnostic evidence, inexperience on the part of the certifying doctor and absence of postmortem which may be important in deciding the cause of death. (c) Lack of uniformity: There is no uniform and standardized method of collection of data. This hampers national and international comparability. (d) Choosing a single cause of death: Most countries tabulate mortality data only according to the underlying cause of death. Other diseases (or risk factors) and conditions which contribute to the patient's death are not tabulated, and valuable information is thereby lost. (e) Changing: Changing coding systems and changing fashions in diagnosis may affect the validity. We also need uniform definitions and nomenclature. (f) Diseases with low fatality: Lastly, mortality statistics are virtually useless, if the disease is associated with low fatality (e.g., mental diseases, arthritis).

<table>
<thead>
<tr>
<th>CAUSE OF DEATH</th>
<th>Approximate interval between onset and death</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I</strong></td>
<td></td>
</tr>
<tr>
<td>Disease or condition directly leading to death*</td>
<td>Brnochopneumonia</td>
</tr>
<tr>
<td>Antecedent causes</td>
<td>(a) due to (or as a consequence of)</td>
</tr>
<tr>
<td>Morbid conditions, if any, giving rise to the above cause, stating the underlying condition last</td>
<td>(b) due to (or as a consequence of)</td>
</tr>
<tr>
<td>(c) Strangulated Hernia</td>
<td></td>
</tr>
<tr>
<td><strong>II</strong></td>
<td></td>
</tr>
<tr>
<td>Other significant conditions contributing to the death, but not related to the disease or condition causing it</td>
<td>Diabetes</td>
</tr>
</tbody>
</table>

*This does not mean the mode of dying e.g., heart failure, ashenia, etc. It means the disease, injury, or complication which caused death

FIG. 1
International form of Death Certificate
**Uses of mortality data**

Statistics on causes of death are important and widely used for a number of purposes. They may be employed in explaining trends and differentials in overall mortality, indicating priorities for health action and the allocation of resources, in designing intervention programmes, and in the assessment and monitoring of public health problems and programmes — moreover, they give important clues for epidemiological research.

**Mortality rates and ratios**

The commonly used measures are described below:

1. **Crude death rate**

   The simplest measure of mortality is the ‘crude death rate’. It is defined as "the number of deaths (from all causes) per 1000 estimated mid-year population in one year, in a given place". It measures the rate at which deaths are occurring from various causes in a given population, during a specified period. The crude death rate is calculated from the formula:

   \[
   \text{Crude death rate} = \frac{\text{Number of deaths during the year}}{\text{Mid-year population}} \times 1000
   \]

   It is important to recognize that the crude death rate summarizes the effect of two factors:
   a. population composition
   b. age-specific death rates (which reflect the probability of dying)

   Table 1 shows the crude death rates of two populations, A and B. The crude death rate for population A is 15.2 per 1000. The crude death rate for population B is 9.9 per 1000. Apparently, population B appears healthier than population A.

   The limitation of the crude death rate is exposed, when we compare the age-specific rates between the two populations as shown in Table 1. It can be seen that population B has higher age-specific rates in all age groups. This seeming contradiction is due to differences in the age-composition of the population. The higher crude death rate in population A is due to its older population compared with population B which has a relatively younger population. Currently, this is the prevailing situation in most developing countries with low crude death rates, but high age-specific death rates.

2. **Specific death rates**

   When analysis is planned to throw light on aetiology, it is essential to use specific death rates. The specific death rates may be — (a) cause or disease specific — e.g., tuberculosis, cancer, accident; (b) related to specific groups — e.g., age-specific, sex-specific, age and sex specific, etc. Rates can also be made specific for many other variables such as income, religion, race, housing, etc. Specific death rates can help us to identify particular groups or groups “at-risk”, for preventive action. They permit comparisons between different causes within the same population. Specific death rates are obtained mainly in countries in which a satisfactory civil registration system operates and in which a high proportion of deaths is certified medically.

   Table 2 illustrates how some specific death rates in common use are computed:

   **Table 2**  
   Specific death rates

<table>
<thead>
<tr>
<th>Specific death rate due to tuberculosis</th>
<th>Number of deaths from tuberculosis during a calendar year</th>
<th>1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mid-year population</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specific death rate for males</th>
<th>Number of deaths among males during a calendar year</th>
<th>1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mid-year population of males</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Specific death rate in age group 15-20 years</th>
<th>Number of deaths of persons aged 15-20 during a calendar year</th>
<th>1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mid-year population of persons aged 15-20</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Death rate for January</th>
<th>Deaths in January × 12</th>
<th>1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Note: The deaths are multiplied by 12 in order to make the monthly death rate comparable with the annual death rate)</td>
<td>Mid-year population</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weekly death rate</th>
<th>Deaths in the week × 52</th>
<th>1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mid-year population</td>
<td></td>
</tr>
</tbody>
</table>

   In summary, the crude death rates have a major disadvantage, that is, they lack comparability for communities with populations that differ by age, sex, race, etc. However, they should always be examined first, and later the age-specific death rates which are the most useful single measures of mortality. By moving away from the crude death rate to the more detailed age-specific rates, an attractive feature of the crude death rate, that is, its ability to portray an impression in a single figure is lost.
3. Case fatality rate (Ratio)

\[
\text{Case fatality rate} = \frac{\text{Total number of deaths due to a particular disease}}{\text{Total number of cases due to the same disease}} \times 100
\]

Case fatality rate represents the killing power of a disease. It is simply the ratio of deaths to cases. The time interval is not specified. Case fatality rate is typically used in acute infectious diseases (e.g., food poisoning, cholera, measles). Its usefulness for chronic diseases is limited, because the period from onset to death is long and variable. The case fatality rate for the same disease may vary in different epidemics because of changes in the agent, host and environmental factors. Case fatality is closely related to virulence.

4. Proportional mortality rate (Ratio)

It is sometimes useful to know what proportion of total deaths are due to a particular cause (e.g., cancer) or what proportion of deaths are occurring in a particular age group (e.g., above the age of 50 years). Proportional mortality rate expresses the "number of deaths due to a particular cause (or in a specific age group) per 100 (or 1000) total deaths". Thus we have:

(a) Proportional mortality from a specific disease

\[
\text{Proportional mortality from a specific disease} = \frac{\text{Number of deaths from the specific disease in a year}}{\text{Total deaths from all causes in that year}} \times 100
\]

(b) Under-5 proportionate mortality rate

\[
\text{Under-5 proportionate mortality rate} = \frac{\text{Number of deaths under 5 years of age in the given year}}{\text{Total number of deaths during the same period}} \times 100
\]

(c) Proportional mortality rate for aged 50 years and above

\[
\text{Proportional mortality rate for aged 50 years and above} = \frac{\text{Number of deaths of persons aged 50 years and above}}{\text{Total deaths of all age groups in that year}} \times 100
\]

Proportional mortality rate is computed usually for a broad disease group (such as communicable diseases as a whole) and for a specific disease of major public health importance, such as cancer or coronary heart disease in industrialized countries (14). Proportional mortality rates are used when population data are not available. Since proportional mortality rate depends upon two variables, both of which may differ, it is of limited value in making comparison between population groups or different time periods. However, proportional rates are useful indicators within any population group of the relative importance of the specific disease or disease group, as a cause of death. Mortality from communicable diseases is especially important as it relates mostly to preventable conditions. Since the prevailing causes of death vary according to age and sex, it is desirable to compute proportionate mortality separately for each age and sex group in order to determine measures directed to particular age-sex groups for the reduction of preventable mortality (14). Proportional mortality rate does not indicate the risk of members of the population contracting or dying from the disease.

5. Survival rate

\[
\text{Survival rate} = \frac{\text{Total number of patients alive after 5 years}}{\text{Total number of patients diagnosed or treated}} \times 100
\]

6. Adjusted or standardized rates

If we want to compare the death rates of two populations with different age-composition, the crude death rate is not the right yardstick. This is because, rates are only comparable if the populations upon which they are based are comparable. And it is cumbersome to use a series of age specific death rates. The answer is "age adjustment" or "age standardization", which removes the confounding effect of different age structures and yields a single standardized or adjusted rate, by which the mortality experience can be compared directly. The adjustment can be made not only for age but also sex, race, parity, etc. Thus one can generate age-sex, and race-adjusted rates.

Standardization is carried out by one of two methods – direct or indirect standardization. Both the methods begin by choosing a "standard population", not the age-structures of the populations.

DIRECT STANDARDIZATION

Two examples of direct standardization are given. In the first, a "standard population" is selected. A standard population is defined as one for which the numbers in each age and sex group are known. A frequently used standard age-composition (14) is shown in Table 3. The standard population may also be "created" by combining 2 populations, this is shown in the second example.

The next step is to apply to the standard population, the age-specific rates of the population whose crude death rate is to be adjusted or standardized. As a result, for each age group, an "expected" number of deaths (or events) in the standard population is obtained; these are added together for all the age groups, to give the total expected deaths. The final operation is to divide the "expected" total number of deaths by the total of the standard population, which yields the standardized or age-adjusted rate.

Example 1

Example 1 shows: (a) the computation of age-specific death rates per 1000 population for city X (Table 3); and (b) application of these rates to a standard population to obtain the "expected deaths" and the standardized or age-adjusted death rate (Table 4).
TABLE 3
Calculation of age-specific death rates for City “X”

<table>
<thead>
<tr>
<th>Age</th>
<th>Mid-year population</th>
<th>Deaths in the year</th>
<th>Age-specific death rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4,000</td>
<td>60</td>
<td>15.0</td>
</tr>
<tr>
<td>1-4</td>
<td>4,500</td>
<td>20</td>
<td>4.4</td>
</tr>
<tr>
<td>5-14</td>
<td>4,000</td>
<td>12</td>
<td>3.0</td>
</tr>
<tr>
<td>15-19</td>
<td>5,000</td>
<td>15</td>
<td>3.0</td>
</tr>
<tr>
<td>20-24</td>
<td>4,000</td>
<td>16</td>
<td>4.0</td>
</tr>
<tr>
<td>25-34</td>
<td>8,000</td>
<td>25</td>
<td>3.1</td>
</tr>
<tr>
<td>35-44</td>
<td>9,000</td>
<td>48</td>
<td>5.3</td>
</tr>
<tr>
<td>45-54</td>
<td>8,000</td>
<td>100</td>
<td>12.5</td>
</tr>
<tr>
<td>55-64</td>
<td>7,000</td>
<td>150</td>
<td>21.4</td>
</tr>
<tr>
<td>Total</td>
<td>53,500</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Crude death rate per 1000 = 8.3

TABLE 4
Calculation of the standardized death rate for City “X”

<table>
<thead>
<tr>
<th>Age</th>
<th>Standard population</th>
<th>Age-specific death rates per 1000</th>
<th>Expected deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2,400</td>
<td>15.0</td>
<td>36</td>
</tr>
<tr>
<td>1-4</td>
<td>9,600</td>
<td>4.4</td>
<td>42.24</td>
</tr>
<tr>
<td>5-14</td>
<td>19,000</td>
<td>3.0</td>
<td>57</td>
</tr>
<tr>
<td>15-19</td>
<td>9,000</td>
<td>3.0</td>
<td>27</td>
</tr>
<tr>
<td>20-24</td>
<td>8,000</td>
<td>4.0</td>
<td>32</td>
</tr>
<tr>
<td>25-34</td>
<td>14,000</td>
<td>3.1</td>
<td>43.4</td>
</tr>
<tr>
<td>35-44</td>
<td>12,000</td>
<td>5.3</td>
<td>63.6</td>
</tr>
<tr>
<td>45-54</td>
<td>11,000</td>
<td>12.5</td>
<td>137.5</td>
</tr>
<tr>
<td>55-64</td>
<td>8,000</td>
<td>21.4</td>
<td>171.2</td>
</tr>
<tr>
<td>Total</td>
<td>93,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Standardized death rate per 1000 = \( \frac{609.94}{93,000} \times 1000 = 6.56 \)

It can be seen from Tables 3 and 4 that standardizing for age distribution has reduced the crude death rate from 8.3 to 6.56. The choice of the standard population is, to some extent, arbitrary. Clearly, use of a different standard population will give rise to a different value for the standardized death rate, but it must be remembered that these standardized rates have been calculated so that they can be compared between themselves – they have no intrinsic meaning other than for this purpose (15).

It is usual to use the national population as standard when inter-regional comparisons between cities within a range are made. In order that comparisons can be made over a period of years, a ‘standard population’ can be maintained for that period (15). The standard population used in Table 4 is given by WHO in its publication “Health for All” Series No. 4, on page 77 (14).

Example 2
Table 5 shows that in a study of lung cancer and smoking, 42% of cases and 18% of controls were heavy smokers.

TABLE 5
Proportion of heavy smokers in cases and controls (lung cancer)

<table>
<thead>
<tr>
<th>Age</th>
<th>CASES</th>
<th>CONTROLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>No. Heavy smokers</td>
</tr>
<tr>
<td>40-49</td>
<td>500</td>
<td>400</td>
</tr>
<tr>
<td>50-59</td>
<td>500</td>
<td>100</td>
</tr>
<tr>
<td>Total</td>
<td>1,000</td>
<td>500</td>
</tr>
</tbody>
</table>

Source : (5)

Age adjustments were carried out (a) first, by combining the number of subjects in both the age groups (500+500=1,000) to create a standard population, and (b) applying the observed age-specific proportions of heavy smokers (i.e., 50% and 10% in both cases and controls) to the same standard population. The results (or "expected" values) are shown in Table 6, which shows that the age adjusted proportions of heavy smokers are identical (30%) for cases and controls. The previously observed difference is explained entirely by the difference in age composition.

TABLE 6
Age-adjusted proportions

<table>
<thead>
<tr>
<th>Age</th>
<th>Subjects</th>
<th>Expected number of heavy smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CASES</td>
<td>CONTROLS</td>
</tr>
<tr>
<td>40-49</td>
<td>500</td>
<td>( \frac{500 \times 50}{100} = 250 )</td>
</tr>
<tr>
<td>50-59</td>
<td>500</td>
<td>( \frac{50 \times 10}{100} = 50 )</td>
</tr>
<tr>
<td>Total</td>
<td>1,000</td>
<td>300</td>
</tr>
</tbody>
</table>

Standardized rates

The direct method of standardization is feasible only if the actual specific rates in subgroups of the observed population are available, along with the number of individuals in each subgroup.

INDIRECT AGE STANDARDIZATION

1. Standardized mortality ratio (SMR)
The simplest and most useful form of indirect standardization is the Standardized Mortality Ratio (SMR). In England, it is the basis for the allocation of government money to the health regions of the country. The concept is that the regions with higher mortality also have the higher morbidity, and should therefore receive proportionately higher funding to combat ill-health (15).

Standard mortality ratio is a ratio (usually expressed as a percentage) of the total number of deaths that occur in the study group to the number of deaths that would have been expected to occur if that study group had experienced the death rates of a standard population (or other reference population). In other words, SMR compares the mortality in a study group (e.g., an occupational group) with the mortality that the occupational group would have had if they had experienced national mortality rates. In this method, the more stable rates of the larger population are
applied to the smaller study group. It gives a measure of the likely excess risk of mortality due to the occupation.

\[ *SMR = \frac{Observed\ deaths}{Expected\ deaths} \times 100 \]

If the ratio had value greater than 100, then the occupation would appear to carry a greater mortality risk than that of the whole population. If the ratio had value less than 100, then the occupation risks of mortality would seem to be proportionately less than that for the whole population.

Table 7 shows that the mortality experience of coal workers was 129 per cent, which meant that their mortality was 29 per cent more than that experienced by the national population. Values over 100 per cent represent an unfavourable mortality experience and those below 100 per cent relatively favourable mortality experience. Table 7 displays the calculations.

**TABLE 7**

Calculation of the SMR for coal workers

<table>
<thead>
<tr>
<th>Age</th>
<th>National population death rates per 1000</th>
<th>Coal workers population</th>
<th>Observed deaths</th>
<th>Expected deaths</th>
</tr>
</thead>
<tbody>
<tr>
<td>25-34</td>
<td>3.0</td>
<td>300</td>
<td>*</td>
<td>0.9</td>
</tr>
<tr>
<td>35-44</td>
<td>5.0</td>
<td>400</td>
<td>*</td>
<td>2.0</td>
</tr>
<tr>
<td>45-54</td>
<td>8.0</td>
<td>200</td>
<td>*</td>
<td>1.6</td>
</tr>
<tr>
<td>55-64</td>
<td>25.0</td>
<td>100</td>
<td>*</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1,000</td>
<td>9</td>
<td>7.0</td>
</tr>
</tbody>
</table>

\[ SMR = \frac{9}{7 \times 100} = 129 \]

* It is not necessary to know these values; only the total for the whole age-range is required

The SMR has the advantage over the direct method of age adjustment in that it permits adjustment for age and other factors where age-specific rates are not available or are unstable because of small numbers. One needs to know only the number of persons in each age group in the study population and the age-specific rates of the national population (or other reference population). It is possible to use SMR if the event of interest is occurrence of disease rather than death.

2. Other standardization techniques

(a) A more complicated method of indirect adjustment which yields absolute age adjusted rate, involves the calculation of an index death rate and a standardizing factor for each population of interest. The reader is referred to A.B. Hill's "Principles of Medical Statistics". (b) Life table is an age-adjusted summary of current all-causes mortality. (c) Regression techniques: These are an efficient means of standardization. (d) Multivariate analysis: A computer, using regression or similar methods, can standardize for many variables simultaneously (16).

**MEASUREMENT OF MORBIDITY**

Morbidity has been defined as "any departure, subjective or objective, from a state of physiological well-being" (17,18). The term is used equivalent to such terms as sickness, illness, disability etc. The WHO Expert Committee on Health Statistics noted in its 6th Report (17) that morbidity could be measured in terms of 3 units – (a) persons who were ill; (b) the illnesses (periods or spells of illness) that these persons experienced; and (c) the duration (days, weeks, etc) of these illnesses.

Three aspects of morbidity are commonly measured by morbidity rates or morbidity ratios, namely frequency, duration and severity. Disease frequency is measured by incidence and prevalence rates. The average duration per case or the disability rate, which is the average number of days of disability per person, may serve as a measure of the duration of illnesses. The case fatality rate may be used as an index of severity (19). This section focuses on incidence and prevalence rates, which are widely used to describe disease occurrence in a community.

The value of morbidity data may be summarized as follows:

a. they describe the extent and nature of the disease load in the community, and thus assist in the establishment of priorities.

b. they usually provide more comprehensive and more accurate and clinically relevant information on patient characteristics, than can be obtained from mortality data, and are therefore essential for basic research.

c. they serve as starting point for aetiological studies, and thus play a crucial role in disease prevention.

d. they are needed for monitoring and evaluation of disease control activities.

**INCIDENCE**

Incidence rate is defined as "the number of NEW cases occurring in a defined population during a specified period of time". It is given by the formula:

\[ \text{Incidence} = \frac{\text{Number of new cases of specific disease during a given time period}}{\text{Population at-risk during that period}} \times 1000 \]

For example, if there had been 500 new cases of an illness in a population of 30,000 in a year, the incidence rate would be:

\[ \frac{500}{30,000} \times 1000 = 16.7 \text{ per 1000 per year} \]

Note: Incidence rate must include the unit of time used in the final expression. If you write 16.7 per 1000, this would be inadequate. The correct expression is 16.7 per 1000 per year (20).

It will be seen from the above definition that incidence rate refers

a. only to new cases

b. during a given period (usually one year)

c. in a specified population or "population at risk", unless other denominators are chosen.

d. it can also refer to new spells or episodes of disease arising in a given period of time, per 1000 population. For example, a person may suffer from common cold more than once a year. If he had suffered twice, he would contribute 2 spells of sickness in that year. The formula in this case would be:

\[ \text{Incidence rate} = \frac{\text{Number of spells of sickness starting in a defined period}}{\text{Mean number of persons exposed to risk in that period}} \times 1000 \]

Incidence measures the rate at which new cases are occurring in a population. It is not influenced by the duration of the disease. The use of incidence is generally restricted to acute conditions.
Special incidence rates

Examples include: Attack rate (case rate), Secondary attack rate, Hospital admission rate, etc.

a. Attack rate

An attack rate is an incidence rate (usually expressed as a per cent), used only when the population is exposed to risk for a limited period of time such as during an epidemic. It relates the number of cases in the population at risk and reflects the extent of the epidemic. Attack rate is given by the formula:

\[
\text{Attack rate} = \frac{\text{Number of new cases of a specified disease during a specified time interval}}{\text{Total population at risk during the same interval}} \times 100
\]

b. Secondary attack rate

It is defined as the number of exposed persons developing the disease within the range of the incubation period following exposure to a primary case. (see page 100).

USES OF INCIDENCE RATE

The incidence rate, as a health status indicator, is useful for taking action (a) to control disease, and (b) for research into aetiology and pathogenesis, distribution of diseases, and efficacy of preventive and therapeutic measures (14).

For instance, if the incidence rate is increasing, it might indicate failure or ineffectiveness of the current control programmes. Rising incidence rates might suggest the need for a new disease control or preventive programme, or that reporting practices had improved. A change or fluctuation in the incidence of disease may also mean a change in the aetiology of disease, e.g., change in the agent, host and environmental characteristics. Analysis of differences in incidence rates reported from various socio-economic groups and geographical areas may provide useful insights into the effectiveness of the health services provided (14).

PREVALENCE

The term “disease prevalence” refers specifically to all current cases (old and new) existing at a given point in time, or over a period of time in a given population. A broader definition of prevalence is as follows: “the total number of all individuals who have an attribute or disease at a particular time (or during a particular period) divided by the population at risk of having the attribute or disease at this point in time or midway through the period (2)”. Although referred to as a rate, prevalence rate is really a ratio.

Prevalence is of two types:

(a) Point prevalence

Point prevalence of a disease is defined as the number of all current cases (old and new) of a disease at one point of time, in relation to a defined population. The “point” in point prevalence, may for all practical purposes consist of a day, several days, or even a few weeks, depending upon the time it takes to examine the population sample (20).

Point prevalence is given by the formula:

\[
\text{Point prevalence} = \frac{\text{Number of all current cases (old and new)} \text{ of a specified disease existing at a given point in time}}{\text{Estimated population at the same point in time}} \times 100
\]

When the term “prevalence rate” is used, without any further qualification, it is taken to mean “point prevalence” (17).

Point prevalence can be made specific for age, sex and other relevant factors or attributes.

(b) Period prevalence

A less commonly used measure of prevalence is period prevalence. It measures the frequency of all current cases (old and new) existing during a defined period of time (e.g., annual prevalence) expressed in relation to a defined population. It includes cases arising before but extending into or through to the year as well as those cases arising during the year (Fig. 2). Period prevalence is given by the formula:

\[
\text{Period prevalence} = \frac{\text{Number of existing cases (old and new) of a specified disease during a given period of time interval}}{\text{Estimated mid-interval population at-risk}} \times 100
\]

The terms incidence and prevalence are illustrated in Fig. 2

![Diagram of prevalence and incidence](image)

Incidence would include cases - 3,4,5, and 8
Point prevalence (Jan 1) cases - 1,2, and 7
Point prevalence (Dec 31) cases - 1,3,5 and 8
Period prevalence (Jan–Dec) cases - 1,2,3,4,5,7, and 8

Relationship between prevalence and incidence

Prevalence depends upon 2 factors, the incidence and duration of illness. Given the assumption that the population is stable, and incidence and duration are unchanging, the relationship between incidence and prevalence can be expressed as:

\[
P = \text{incidence} \times \text{mean duration}
\]

Example (for a stable condition)
Incidence = 10 cases per 1000 population per year
Mean duration of disease = 5 years
Prevalence = \(10 \times 5 = 50\) per 1000 population
Conversely, it is possible to derive incidence and duration as follows:

Incidence = P/D
Duration = P/I

The above equation (P = 1 x D) shows that the longer the duration of the disease, the greater its prevalence. For example, tuberculosis has a high prevalence rate relative to incidence. This is because new cases of tuberculosis keep cropping up throughout the year, while the old ones may persist for months or years. On the other hand, if the disease is acute and of short duration either because of rapid recovery or death, the prevalence rate will be relatively low compared with the incidence rate. In some diseases (e.g., food poisoning), the disease is so short-lived, there are no "old" cases. The same is true of conditions which are rapidly fatal, such as homicides. Strictly speaking, these events have no prevalence. In other words, decrease in prevalence may take place not only from a decrease in incidence, but also from a decrease of the duration of illness through either more rapid recovery or more rapid death.

When we see a change in prevalence from one time period to another, this can result from changes in incidence, changes in duration of disease or both. For example, improvements in treatment may decrease the duration of illness and thereby decrease prevalence of a disease. But if the treatment is such that by preventing death, and at the same time not producing recovery, may give rise to the apparently paradoxical effect of an increase in prevalence. Further, if duration is decreased sufficiently, a decrease in prevalence could take place despite an increase in incidence.

Prevalence has been compared with a photograph, an instantaneous record; and incidence with a film, a continuous record. Both the terms may perhaps be better understood by taking into consideration a coffee house. After the coffee house opens in the morning, people keep entering and leaving, each one remaining inside the coffee house for a short while. At any point of time, say 10 AM, we could go into the coffee house and count people over there. This corresponds to estimating the prevalence. The rate at which people enter the coffee house, say 10 people per hour, is equivalent to the incidence. The relationship between incidence and prevalence is shown in Fig. 3 (21).

![Incidence and Prevalence Diagram](image-url)

**FIG. 3**
Relationship between incidence and prevalence

It is important to note the limitations of prevalence rate. It is not the ideal measure for studying disease aetiology or causation. We have seen that two factors determine prevalence, namely incidence and duration. Incidence is related to the occurrence of disease and duration to factors which affect the course of the disease. In other words, the element of duration reflects the prognostic factors, and incidence reflects the causal factors. Therefore, incidence rates should be optimally used in the formulation and testing of aetiologic hypotheses. When incidence rates are not available, prevalence rates (which are readily obtainable) may have to be used, but the contribution of duration element always has to be assessed.

**Uses of prevalence**

(a) Prevalence helps to estimate the magnitude of health/disease problems in the community, and identify potential high-risk populations (b) Prevalence rates are especially useful for administrative and planning purposes, e.g., hospital beds, manpower needs, rehabilitation facilities, etc.

**Epidemiologic Methods**

The primary concern of the epidemiologist is to study disease occurrence in people, who during the course of their lives are exposed to numerous factors and circumstances, some of which may have a role in disease aetiology. Unlike the clinician or the laboratory investigator, who is able to study disease conditions more precisely, the epidemiologist employs carefully designed research strategies to explore disease aetiology.

Epidemiological studies can be classified as observational studies and experimental studies with further subdivisions:

1. **Observational studies**
   a. Descriptive studies
   b. Analytical studies
      (i) Ecological or Correlational, with populations as unit of study
      (ii) Cross-sectional or Prevalence, with individuals as unit of study
      (iii) Case-control or Case-reference, with individuals as unit of study
      (iv) Cohort or Follow-up, with individuals as unit of study

2. **Experimental studies Intervention studies**
   a. Randomized controlled trials with patients as unit of study
   b. Field trials with healthy people as unit of study
   c. Community trials or Community intervention studies with communities as unit of study

These studies or methods cannot be regarded as watertight compartments; they complement one another. Observational studies allow nature to take its own course; the investigator measures but does not intervene. Descriptive study is limited to a description of the occurrence of a disease in a population. An analytical study goes further by analyzing relationship between health status and other variables. Experimental or intervention studies involve an active attempt to change a disease determinant or the progress of a disease, and are similar in design to experiments in other sciences. However, they are subject to extra constraints, since the health of the people in the study group may be at stake. The major experimental design is the randomized controlled trial using patients as subjects. Field
trials and community trials are other experimental studies in which the participants are healthy people and community respectively (22).

In all epidemiological studies, it is essential to have a clear definition of a case of the disease being investigated and of an exposed person. In absence of clear definitions of disease and exposure, great difficulties are likely to be experienced in interpreting the data.

**DESCRIPTIVE EPIDEMIOLOGY**

The best study of mankind is man. This statement emphasizes the importance of making the best use of observations on individuals or populations exposed to suspected factors of disease. Meticulous observations made in Africa by Burkitt led to the eventual incrimination of Epstein-Barr virus (EBV) as the aetiological factor (possibly conditioned by other factors such as malarial infection) of the type of cancer known as Burkitt’s lymphoma. It was the epidemiological study in New Guinea of “Kuru”, a hereditary neurological disorder, that led to the discovery of slow virus infections as the cause of chronic degenerative neurological disorders in human beings. The list is endless.

Descriptive studies are usually the first phase of an epidemiological investigation. These studies are concerned with observing the distribution of disease or health-related characteristics in human populations and identifying the characteristics with which the disease in question seems to be associated. Such studies basically ask the questions:

a. When is the disease occurring?  
   - time distribution
b. Where is it occurring?  
   - place distribution
c. Who is getting the disease?  
   - person distribution

The various procedures involved in descriptive studies may be outlined as below (Table 8).

<table>
<thead>
<tr>
<th>TABLE 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedures in descriptive studies</td>
</tr>
<tr>
<td>1. Defining the population to be studied</td>
</tr>
<tr>
<td>2. Defining the disease under study</td>
</tr>
<tr>
<td>3. Describing the disease by</td>
</tr>
<tr>
<td>a. time</td>
</tr>
<tr>
<td>b. place</td>
</tr>
<tr>
<td>c. person</td>
</tr>
<tr>
<td>4. Measurement of disease</td>
</tr>
<tr>
<td>5. Comparing with known indices</td>
</tr>
<tr>
<td>6. Formulation of an aetiological hypothesis</td>
</tr>
</tbody>
</table>

1. **Defining the population**

Descriptive studies are investigations of populations, not individuals. The first step is, therefore, to define the “population base” not only in terms of the total number, but also its composition in terms of age, sex, occupation, cultural characters and similar information needed for the study.

The “defined population” can be the whole population in a geographic area, or more often a representative sample taken from it. The defined population can also be a specially selected group such as age and sex groups, occupational groups, hospital patients, school children, small communities as well as wider groupings – in fact, wherever a group of people can be fairly accurately counted.

The defined population needs to be large enough so that age, sex and other specific rates are meaningful. The community chosen should be stable, without migration into or out of the area. It should be clear who does and who does not belong to the population, as for example, visitors and relations. Perhaps the most essential ingredient is community participation, which must be forthcoming. Furthermore, the population should not be overly different from other communities in the region. Finally, a health facility should be close enough to provide relatively easy access for patients requiring medical services. In the famous Framingham Heart Study in US, all the above criteria were taken into consideration in choosing the study population.

The concept of ‘defined population’ (or population at risk) is crucial in epidemiological studies. It provides the denominator for calculating rates which are essential to measure the frequency of disease and study its distribution and determinants. Epidemiologists therefore have been labelled as men in search of a denominator (23).

2. **Defining the disease under study**

Once the population to be studied is defined or specified, one must then define the disease or condition being investigated. Here the needs of the clinician and epidemiologist may diverge. The clinician may not need a precise definition of disease (e.g., migraine) for immediate patient care. If the diagnosis is wrong, he can revise it subsequently. But the epidemiologist, whose main concern is to obtain an accurate estimate of disease in a population, needs a definition that is both precise and valid to enable him (or observers working in field conditions) to identify those who have the disease from those who do not (9). The diagnostic methods for use in epidemiological studies must be acceptable to the population to be studied, and applicable to their use in large populations.

In other words, the epidemiologist looks out for an “operational definition”, i.e., a definition by which the disease or condition can be identified and measured in the defined population with a degree of accuracy. For example, tonsillitis might be defined clinically as an inflammation of the tonsils caused by infection, usually with streptococcus pyogenes. This definition, like many other clinical definitions (and the WHO definition of ‘health’) serves to convey particular information, but cannot be used to measure disease in the community. On the other hand, an “operational definition” spells out clearly the criteria by which the disease can be measured. Such criteria in the case of tonsillitis would include the presence of enlarged, red tonsils with white exudate, which on throat swab culture grow predominantly S. pyogenes. If the definition is not valid, it would be a powerful source of error in the presentation and comparability of measurements from different sources. With regard to certain diseases (e.g., neurological diseases) which often do not have pathognomonic signs and symptoms, disease definition is a crucial concern for the epidemiologist. In such cases, the epidemiologist frames his own definition keeping the objectives of his study in view and aiming at the same time a degree of accuracy sufficient for his purpose. Once established, the case definition must be adhered to throughout the study.

3. **Describing the disease**

The primary objective of descriptive epidemiology is to describe the occurrence and distribution of disease (or
health-related events or characteristics within populations) by time, place and person, and identifying those characteristics associated with presence or absence of disease in individuals. This involves systematic collection and analysis of data. Some of the characteristics most frequently examined by epidemiologists in descriptive studies are given in Table 9. It is only an initial separation or grouping of variables according to time, place and person and NOT a classification of causal factors.

TABLE 9
Characteristics frequently examined in descriptive studies

<table>
<thead>
<tr>
<th>Time</th>
<th>Place</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year, season</td>
<td>Climatic zones</td>
<td>Age</td>
</tr>
<tr>
<td>Month, week</td>
<td>Country, region</td>
<td>Birth order</td>
</tr>
<tr>
<td>Day, hour of onset,</td>
<td>Urban/rural, Local community</td>
<td>Marital status</td>
</tr>
<tr>
<td>Duration</td>
<td>Towns, Cities, Institutions</td>
<td>Occupation, Social status</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Blood pressure, Blood cholesterol</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Education, Personal habits</td>
</tr>
</tbody>
</table>

TIME DISTRIBUTION

The pattern of disease may be described by the time of its occurrence, i.e., by week, month, year, the day of the week, hour of onset, etc. It raises questions whether the disease is seasonal in occurrence; whether it shows periodic increase or decrease; or whether it follows a consistent time trend. Such studies may yield important clues about the source or aetiology of the disease, thereby suggesting potential preventive measures. Epidemiologists have identified three kinds of time trends or fluctuations in disease occurrence.

I. Short-term fluctuations
II. Periodic fluctuations, and
III. Long-term or secular trends

I. Short-term fluctuations

The best known short-term fluctuation in the occurrence of a disease is an epidemic. According to modern concepts an epidemic is defined as “the occurrence in a community or region of cases of an illness or other health-related events clearly in excess of normal expectancy”. The community or region, and the time period in which the cases occur, are specified precisely. Epidemicity is thus relative to usual frequency of the disease in the same area, among the specified population, at the same season of the year (2). The data in Table 10 illustrates this point.

Types of epidemics

Three major types of epidemics may be distinguished.

A. Common-source epidemics
   (a) Single exposure or “point-source” epidemics.
   (b) Continuous or multiple exposure epidemics
B. Propagated epidemics
   (a) Person-to-person
   (b) Arthropod vector
   (c) Animal reservoir
C. Slow (modern) epidemics.

A graph of the time distribution of epidemic cases is called the “epidemic curve” (Fig. 4). The epidemic curve may suggest: (1) a time relationship with exposure to a suspected source, (2) a cyclical or seasonal pattern suggestive of a particular infection, and common source or propagated spread of the disease.

A. Common-source epidemics

(a) Common-source, single exposure epidemics

These are also known as “point-source” epidemics. The exposure to the disease agent is brief and essentially simultaneous, the resultant cases all develop within one incubation period of the disease (e.g., an epidemic of food poisoning). Fig. 4 illustrates a common-source, single exposure epidemic. The curve has usually one peak. One point of interest is the “median incubation period”, it is the time required for 50 per cent of the cases to occur following exposure.

(b) Common-source, continuous or repeated exposure

Sometimes the exposure from the same source may be prolonged – continuous, repeated or intermittent – not necessarily at the same time or place. A prostitute may be a common source in a gonorrhea outbreak, but since she will infect her clients over a period of time there may be no explosive rise in the number of cases. A well of contaminated water, or a nationally distributed brand of vaccine (e.g. polio vaccine), or food, could result in similar outbreaks. In these instances, the resulting epidemics tend to be more extended or irregular. The outbreak of respiratory illness, the Legionnaire’s disease, in the summer of 1976 in Philadelphia (USA) was a common-source, continuous or repeated exposure outbreak. This outbreak, as in other outbreaks of this type, continued beyond the range of one incubation period. There was no evidence of secondary cases among persons who had contact with ill persons (24).

A variation to the above model is that an epidemic may
be initiated from a common source and then continue as a propagated epidemic. Water-borne cholera is a familiar example, the epidemic reaches a sharp peak, but tails off gradually over a longer period of time.

B. Propagated epidemics

A propagated epidemic is most often of infectious origin and results from person-to-person transmission of an infectious agent (e.g., epidemics of hepatitis A and polio). The epidemic usually shows a gradual rise and tails off over a much longer period of time. Transmission continues until the number of susceptibles is depleted or susceptible individuals are no longer exposed to infected persons or intermediary vectors. The speed of spread depends upon herd immunity, opportunities for contact and secondary attack rate. Propagated epidemics are more likely to occur where large number of susceptibles are aggregated, or where there is a regular supply of new susceptible individuals (e.g., birth, immigrants) lowering herd immunity. Fig. 5 illustrates the course of a typical propagated epidemic in which the agent is transmitted by contact between individuals.

II. Periodic fluctuations

(i) Seasonal trend: Seasonal variation is a well-known characteristic of many communicable diseases, e.g., measles, varicella, cerebro-spinal meningitis, upper respiratory infections, malaria, etc. For example, measles is usually at its height in early spring and so is varicella. Upper respiratory infections frequently show a seasonal rise during winter months. Bacterial gastrointestinal infections are prominent in summer months because of warm weather and rapid multiplication of flies. The seasonal variations of disease occurrence may be related to environmental conditions (e.g., temperature, humidity, rainfall, overcrowding, life cycle of vectors, etc.) which directly or indirectly favour disease transmission. However, in many infectious diseases (e.g., polio), the basis for seasonal variation is unknown. Non-infectious diseases and conditions may sometimes exhibit seasonal variation, e.g., sunstroke, hay fever, snakebite.

Some epidemiologists would regard seasonal trend as a form of cyclic trend. Table 10 shows a typical pattern of seasonal trend, – the outbreaks of dengue/DF starting by month of July and peaking in September, October and November, coinciding with late summer and rain.

<table>
<thead>
<tr>
<th>Month</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>151</td>
<td>281</td>
<td>83</td>
</tr>
<tr>
<td>February</td>
<td>80</td>
<td>193</td>
<td>64</td>
</tr>
<tr>
<td>March</td>
<td>59</td>
<td>178</td>
<td>46</td>
</tr>
<tr>
<td>April</td>
<td>68</td>
<td>166</td>
<td>50</td>
</tr>
<tr>
<td>May</td>
<td>172</td>
<td>181</td>
<td>127</td>
</tr>
<tr>
<td>June</td>
<td>130</td>
<td>269</td>
<td>175</td>
</tr>
<tr>
<td>July</td>
<td>742</td>
<td>478</td>
<td>487</td>
</tr>
<tr>
<td>August</td>
<td>946</td>
<td>577</td>
<td>487</td>
</tr>
<tr>
<td>September</td>
<td>4,852</td>
<td>1,275</td>
<td>974</td>
</tr>
<tr>
<td>October</td>
<td>2,482</td>
<td>5,880</td>
<td>1,507</td>
</tr>
<tr>
<td>November</td>
<td>1,507</td>
<td>1,934</td>
<td>802</td>
</tr>
<tr>
<td>December</td>
<td>801</td>
<td>905</td>
<td>221</td>
</tr>
<tr>
<td>Total</td>
<td>11,990</td>
<td>12,317</td>
<td>5,023</td>
</tr>
</tbody>
</table>

Source: (25)

(ii) Cyclic trend: Some diseases occur in cycles spread over short periods of time which may be days, weeks, months or years. For example, measles in the pre-vaccination era appeared in cycles with major peaks every 2–3 years and rubella every 6–9 years. This was due to naturally occurring variations in herd immunity. A build-up of susceptibles is again required in the “herd” before there can be another attack. Influenza pandemics are known to occur at intervals of 7–10 years, due to antigenic variations. Non-infectious conditions may also show periodic fluctuations, e.g., automobile accidents in US are more frequent on week-ends, especially Saturdays. A knowledge of cyclicity of disease is useful in that it may enable communities to defend themselves.

![Course of typical propagated epidemic](Figure 5)
III. Long-term or secular trends

The term “secular trend” implies changes in the occurrence of disease (i.e., a progressive increase or decrease) over a long period of time, generally several years or decades. Although it may have short-term fluctuations imposed on it, a secular trend implies a consistent tendency to change in a particular direction or a definite movement in one direction. Examples include coronary heart disease, lung cancer and diabetes which have shown a consistent upward trend in the developed countries during the past 50 years or so, followed by a decline of such diseases as tuberculosis, typhoid fever, diphtheria and polio.

Interpretation of time-trends

By surveillance or monitoring of time-trends, the epidemiologist seeks which diseases are increasing, which decreasing, and which are the emerging health problems and of the effectiveness of measures to control old ones. He tries to formulate aetiological hypotheses, and seeks explanations whether these changes were due to changes in the aetiological agent or variations in diagnosis, reporting, case fatality or changes in age distribution, or some other determinants, specific and non-specific (e.g., changes in quality of life, socio-economic status and personal habits). For example the “time-clustering” of cases of adenocarcinoma of vagina in young women led to the incrimination of its cause, viz. in utero exposure to diethylstilbestrol (26). Even changes taking place over several years or decades can be productive of hypotheses, as in the case of lung cancer. By studying time trends, the epidemiologist seeks to provide guidelines to the health administrator in matters of prevention or control of disease.

PLACE DISTRIBUTION

(Geographical comparisons)

Studies of the geography of disease (or geographical pathology) is one of the important dimensions of descriptive epidemiology. By studying the distribution of disease in different populations we gain perspective on the fascinating differences (or variations) in disease patterns not only between countries, but also within countries. The relative importance of genes versus environment; changes with migration; and the possible roles of diet and other aetiological factors. In short geographical studies have profoundly influenced our understanding of disease, its nature, its determinants and its relation to subsequent pathology. The geographic variation in disease occurrence has been one of the stimuli to national and international studies.

The world is not a uniform unit. Cultures, standard of living and external environments vary greatly. The use of migrant studies is one way of distinguishing genetic and environmental factors. The study of the geography of diseases has developed its own special techniques, which sometimes involve complex statistical analysis. The SMR is one of them.

Geographic patterns provide an important source of clues about the causes of the disease. The range of geographic studies include those concerned with local variations. At a broader level, international comparisons may examine mortality and morbidity in relation to socio-economic factors, dietary differences and the differences in culture and behaviour. These variations may be classified as:

- a. International variations
- b. National variations
- c. Rural–urban variations
- d. Local distributions

International variations

Descriptive studies by place have shown that the pattern of disease is not the same everywhere. For example, we know that cancer exists all over the world. There is, however, a marked difference between the incidence of each cancer in different parts of the world. Thus cancer of the stomach is very common in Japan, but unusual in US. Cancers of the oral cavity and uterine cervix are exceedingly common in India as compared to industrialized countries. An international study of breast cancer showed that rates differ widely from country to country with the lowest prevalence in Japan and the highest in the western countries. Similarly, there are marked international differences in the occurrence of cardiovascular diseases. These variations have stimulated epidemiologists to search for cause–effect relationships between the environmental factors and disease. The aim is to identify factors which are crucial in the cause and prevention of disease.

National variations

It is obvious that variations in disease occurrence must also exist within countries or national boundaries. For example the distribution of endemic goitre, lymphosy, fluorosis, leprosy, nutritional deficiency diseases have all shown variations in their distribution in India, with some parts of the country more affected and others less affected or not affected at all. Such situations exist in every country. One of the functions of descriptive epidemiology is to provide data regarding the type of disease problems and their magnitude in terms of incidence, prevalence and mortality rates. Such information is needed to demarcate the affected areas and for providing appropriate health care services.

Rural–urban variations

Rural/urban variations in disease distribution are well known. Chronic bronchitis, accidents, lung cancer, cardiovascular diseases, mental illness and drug dependance are usually more frequent in urban than in rural areas. On the other hand, skin and zoonotic diseases and soil-transmitted helminths may be more frequent in rural areas than in urban areas. Death rates, especially infant and maternal mortality rates, are higher for rural than urban areas. These variations may be due to differences in population density, social class, deficiencies in medical care, levels of sanitation, education and environmental factors. The epidemiologist seeks to define groups which are at higher risk for particular diseases, and provides guidelines to the health administrator for their prevention and control.

Local distributions

Inner and outer city variations in disease frequency are well known. These variations are best studied with the aid of ‘spot maps’ or ‘shaded maps’. These maps show at a glance areas of high or low frequency, the boundaries and patterns of disease distribution. For example if the map shows “clustering” of cases, it may suggest a common source of infection or a common risk factor shared by all the cases. It was by such a study (spot map of fatal cases), John Snow of England in his classic investigation of cholera epidemic in 1854 in the Golden Square district of London was able to focus attention on the common water pump in Broad street as the source of infection (Fig. 6). Based on his descriptive
findings, Snow was able to hypothesize that cholera was a water-borne disease, long before the birth of bacteriology. It was by a spot map by "place of employment" Maxcy hypothesized a rodent reservoir for typhus fever in 1920s which led to the discovery that typhus fever was not a single disease entity, as it was earlier thought. Also, the evidence of case clustering based on sexual contact or blood product use provided the clue that AIDS (Acquired Immune Deficiency Syndrome) was an infectious disease.

In short the geographic differences in disease occurrence is an important dimension of a descriptive study. These differences are determined by the agent, host and environmental factors. The classic example of place-related diseases include yellow fever, schistosomiasis, sleeping sickness and endemic goitre. There have also been studies on asthma, cancer, cardiovascular diseases, blood groups and abnormal haemoglobins by geographic location. In short, all diseases whether acute or chronic, communicable or non-communicable, show definite patterns of geographic distribution.

The epidemiologist is interested in geographic variations in disease occurrence. Geographic distribution may provide evidence of the source of disease and its mode of spread. By relating these variations to agent, host and environmental factors, he tries to derive clues to the source of disease and its mode of spread to formulate and test aetiological hypotheses. The clinician is also benefited from knowledge that a patient comes to him from a certain geographic area which is endemic for certain infrequent diseases such as yaws or leishmaniasis, as it helps him to focus attention on these diseases to which the patient may have been exposed.

The geographic distribution of disease may change, if changes occur in the agent, host and environmental factors. The empires of malaria, plague and many other diseases have shrunk due to changes in the epidemiological triad. On the other hand, since 1961 cholera has shown an increasing geographic distribution due to changes in the disease agent. Since the mode of living and environmental factors vary from country to country, one would expect to find differences in the geographic distribution and frequency of disease.

Migration studies

Large scale migration of human populations from one country to another provides a unique opportunity to evaluate the role of the possible genetic and environmental factors in the occurrence of disease in a population. Supposing there are marked geographic differences in the occurrence of a disease in two areas, area "A" and area "B". Let us assume that the environments in these two places are very different. The question arises whether the environmental differences in the two areas account for the variations in the occurrence of the disease in question.
Ideally, samples of population in area “A” should be sent to area “B”, and vice versa to study change in incidence of disease. In human populations this is hardly possible, so we restrict our study to observation of changes in disease frequency among migrants.

Migrant studies can be carried out in two ways:

(a) comparison of disease and death rates for migrants with those of their kin who have stayed at home. This permits study of genetically similar groups but living under different environmental conditions or exposures. If the disease and death rates in migrants are similar to those of the host country, the likely explanation would be change in the environment. A special case is the use of twins who have been exposed to different environments of migration.

(b) comparison of migrants with local population of the host country provides information on genetically different groups living in a similar environment. If the migration rates of disease and death are similar to the country of origin, the likely explanation would be the genetic factors.

Migrant studies have shown that men of Japanese ancestry living in USA experience a higher rate of coronary heart disease than do the Japanese in Japan (27). Taking another example, Japan has a higher rate for stomach cancer and a lower rate for colon cancer than the United States has. However, third-generation descendants of Japanese immigrants to USA have rates of stomach and colon cancer like those of the total US population. These studies suggest that as the Japanese were probably adopting the American way of life, their susceptibility to coronary heart disease, gastric and colonic cancer was moving in the direction of that found in the Americans. Further, migrant studies may also indicate the duration of residence necessary to acquire susceptibility to the disease in question by comparing groups that left home at different ages. Studies of this kind provide a basis for further studies of specific environmental factors to which the migrants may have been exposed or of changes in their habits of life that may be of aetiological importance.

Migrant studies suffer from the usual defects of observational studies, deriving from lack of random assignment of the groups under observation. Migrants may be self-selected in that fit, vigorous and perhaps the temperamentally unstable are more likely to migrate (28). The environmental factors may only act at a certain critical point or at a certain specific age. If the incubation period of the disease is very long, migrants may not show any increased incidence or mortality from the disease for many years.

PERSON DISTRIBUTION

In descriptive studies, the disease is further characterized by defining the persons who develop the disease by age, sex, occupation, marital status, habits, social class and other host factors. These factors do not necessarily represent aetiological factors, but they contribute a good deal to our understanding of the natural history of disease. Some of the host factors basic to epidemiological studies (Table 9) are discussed below.

(a) Age: Age is strongly related to disease than any other single host factor. Certain diseases are more frequent in certain age groups than in others, e.g., measles in childhood, cancer in middle age and atherosclerosis in old age. If the attack rate of a communicable disease is uniform in all the age groups, it implies that all age groups are equally susceptible, and there was no previous immunity. Many chronic and degenerative diseases (e.g., cancer) show a progressive increase in prevalence with advancing age. This may reflect a persistent and cumulative exposure to a causal agent or risk factor (12).

Bimodality: Sometimes there may be two separate peaks instead of one in the age incidence curve of a disease as in the case of Hodgkin’s disease, leukaemia, and female breast cancer. This phenomenon is known as bimodality. Fig. 7 shows the age incidence curve for Hodgkin’s disease in USA (29). The curve is bimodal with an initial peak between the ages 15 and 35 years, and a later peak starting at age 50. Bimodality is of special interest to epidemiologists. It indicates that the study material is not homogeneous, and that two distinct sets of causal factors might be operative, even though the clinical and pathological manifestations of the disease are the same at all ages.

![Figure 7: Bimodality in Hodgkin's disease](image)

However, there are two points relating to bimodality which make their interpretation difficult: (a) small numbers of observations are a frequent source of bimodality; (b) the absence of bimodality does not signify that data have come from a homogeneous source.

(b) Sex: Sex is another host characteristic which is often studied in relation to disease, using such indices as sex-ratio, sex-specific morbidity and mortality rates. It has been found that certain chronic diseases such as diabetes, hyperthyroidism and obesity are strikingly more common in women than in men, and diseases such as lung cancer and coronary heart disease are less frequent in women.

Variations in disease frequency between sexes have been ascribed to (a) basic biological differences between the sexes, including sex-linked genetic inheritance, and (b) cultural and behavioural differences between the sexes (e.g., smoking, automobile use, alcoholism) due to different roles in social setting. In fact, it is the 4:1 male to female ratio in lung cancer that has helped to identify cigarette smoking as a causal factor. Even larger differences exist in, for example, duodenal ulcer and coronary heart disease, that are as yet unexplained (30).

(c) Ethnicity: Differences in disease occurrence have been noted between population subgroups of different racial and ethnic origin. These include tuberculosis, essential hypertension, coronary heart disease, cancer, and sickle cell anaemia. These differences, whether they are related to genetic or environmental factors, have been a stimulus to further studies.

(d) Marital status: In countries where studies on mortality
in relation to marital status have been conducted, it was found that mortality rates were always lower for married males and females than for the unmarried, of the same age and sex. According to demographers and sociologists, the reason for this phenomenon may be found in the fact that marriages are selective with respect to the health status of persons, for those who are healthy are more likely to get married, with the result that the risk of dying is also less. Besides, married persons are generally more secure and protected and they usually lead a more sober life than those who are unmarried. All these factors are thought to contribute to lower mortality rates among married persons.

Marital status can be a risk factor for some diseases and conditions. The observation that cancer cervix is rare in nuns led to the hypothesis regarding marital status and cancer cervix. Further studies led to the suggestion that cancer cervix may be associated with multiple sexual contacts and promiscuity. This in turn raised the possibility of a possible infectious agent transmitted venereally. Although the viral aetiology of cancer cervix is not yet proved, this chain of thinking serves to illustrate how an observation can be a starting point of an epidemiological enquiry.

(e) Occupation: It is now well recognized that man’s occupation from which he earns his livelihood has an important bearing on his health status. Occupation may alter the habit pattern of employees e.g., sleep, alcohol, smoking, drug addiction, night shifts etc. It is obvious that persons working in particular occupations are exposed to particular types of risks. For instance, while workers in coal mines are more likely to suffer from silicosis, those in sedentary occupations face the risk of heart disease.

(f) Social class: Epidemiological studies have shown that health and diseases are not equally distributed in social classes. Individuals in the upper social classes have a longer life expectancy and better health and nutritional status than those in the lower social classes. Certain diseases (e.g., coronary heart disease, hypertension, diabetes) have shown a higher prevalence in upper classes than in the lower classes. Social class differences have also been observed in mental illness and utilization of medical and health care services.

However, there is one snag. Social classification varies from country to country. It has different meanings for different persons. Therefore associations of disease with social class vary according to one’s concept of social class. Consequently, it is difficult to compare the results of studies in which social class has been used differently by different investigators (30).

(g) Behaviour: Human behaviour is increasingly looked upon as a risk factor in modern-day diseases such as coronary heart disease, cancer, obesity and accidents. The behavioural factors which have attracted the greatest attention are cigarette smoking, sedentary life, over-eating and drug abuse. To this must be added the mass movement of people, such as occurs in pilgrimages, which lends themselves to the transmission of infectious diseases such as cholera and diarrhoeal diseases, insect-borne and sexually transmitted diseases.

(h) Stress: Stress has been shown to affect a variety of variables related to patients response, e.g., susceptibility to disease, exacerbation of symptoms, compliance with medical regimen, etc.

(i) Migration: In India diseases like leprosy, filaria and malaria are considered to be rural problems. However, because of the movement of people from rural to urban areas these diseases have created a serious problem in urban areas also.

Human movement may be classified (i) as short-term, long-term, and permanent (ii) according to age, sex, education, occupation, (iii) internal or external (iv) urban versus rural, etc. Migration has presented challenge to control/prevention of disease.

To sum up, a study of the host factors in relation to disease occurrence is an important dimension of descriptive epidemiology. Variations in the distribution of disease in age, sex, occupation and other subgroups of the population can be the starting point for an epidemiological enquiry leading to formulation of an aetiological hypothesis for further study. Knowledge of the frequency of disease in subgroups of the population has also generated the concept of “high risk groups”.

4. Measurement of disease

It is mandatory to have a clear picture of the amount of disease (“disease load”) in the population. This information should be available in terms of mortality, morbidity, disability and so on, and should preferably be available for different subgroups of the population. Measurement of mortality is straightforward. Morbidity has two aspects – incidence and prevalence (see page 60, 61). Incidence can be obtained from “longitudinal” studies, and prevalence from “cross-sectional” studies. Descriptive epidemiology may use a cross-sectional or longitudinal design to obtain estimates of magnitude of health and disease problems in human populations.

Cross-sectional studies

Cross-sectional study is the simplest form of an observational study. It is based on a single examination of a cross-section of population at one point in time – the results of which can be projected on the whole population provided the sampling has been done correctly. Cross-sectional study is also known as “prevalence study”.

Cross-sectional studies are more useful for chronic than short-lived diseases. For example, in a study of hypertension, we can also collect data during the survey about age, sex, physical exercise, body weight, salt intake and other variables of interest. Then we can determine how prevention of hypertension is related to certain variables simultaneously measured. Such a study tells us about the distribution of a disease in population rather than its aetiology.

The most common reason that epidemiologist examines the inter-relationships between a disease, or one of its precursors, and other variables is to attempt to establish a causal chain and so give lead to possible ways of preventing that disease. A point which must be stressed is that the time sequence which is essential to the concept of causativity cannot be deduced from cross-sectional data. However, frequently there is evidence that permits ranking of events to form such a sequence. That is, the distribution patterns may suggest causal hypothesis which can be tested by analytical studies. Although a cross-sectional study provides information about disease prevalence, it provides very little information about the natural history of disease or about the rate of occurrence of new cases (incidence).

Longitudinal studies

There is an increasing emphasis on the value of longitudinal studies in which observations are repeated in the same population over a prolonged period of time by means of follow-up examinations. Cross-sectional studies
have been likened to a photograph, and longitudinal studies to a cine film. Longitudinal studies are useful (i) to study the natural history of disease and its future outcome (ii) for identifying risk factors of disease, and (iii) for finding out incidence rate or rate of occurrence of new cases of disease in the community. Longitudinal studies provide valuable information which the cross-sectional studies may not provide, but longitudinal studies are difficult to organize and more time-consuming than cross-sectional studies.

Measurement can also be extended to health states and events. For example, the study of blood pressure levels in a population will reveal the normal values, rather than abnormal ones related to disease.

5. Comparing with known indices

The essence of epidemiology is to make comparisons and ask questions. By making comparisons between different populations, and subgroups of the same population, it is possible to arrive at clues to disease aetiology. We can also identify or define groups which are at increased risk for certain diseases.

6. Formulation of a hypothesis

By studying the distribution of disease, and utilizing the techniques of descriptive epidemiology, it is often possible to formulate hypotheses relating to disease aetiology. A hypothesis is a supposition, arrived at from observation or reflection. It can be accepted or rejected, using the techniques of analytical epidemiology. An epidemiological hypothesis should specify the following (12):

- the population – the characteristics of the persons to whom the hypothesis applies
- the specific cause being considered
- the expected outcome – the disease
- the dose–response relationship – the amount of the cause needed to lead to a stated incidence of the effect
- the time–response relationship – the time period that will elapse between exposure to the cause and observation of the effect.

In other words, a hypothesis should be formulated in a manner that it can be tested taking into consideration the above elements. In practice, the components of a hypothesis are often less well-defined.

For example:

“Cigarette smoking causes lung cancer” – is an incomplete hypothesis.

An improved formulation

“The smoking of 30–40 cigarettes per day causes lung cancer in 10 per cent of smokers after 20 years of exposure”

The improved formulation suggests data needed to test the hypothesis, i.e., the number of cigarettes smoking per day, years of exposure, and so on. The success or failure of a research project frequently depends upon the soundness of the hypothesis (12).

Uses of descriptive epidemiology

Descriptive studies: (a) provide data regarding the magnitude of the disease load and types of disease problems in the community in terms of morbidity and mortality rates and ratios (b) provide clues to disease aetiology, and help in the formulation of an aetiological hypothesis. That is, the existence of a possible causal association between a factor and a disease is usually recognized in descriptive studies. Thus, if the disease is observed to be more frequent in a particular group than in others, hypotheses are formulated to explain the increased frequency (c) provide background data for planning, organizing and evaluating preventive and curative services, and (d) they contribute to research by describing variations in disease occurrence by time, place and person.

ANALYTICAL EPIDEMIOLOGY

Analytical studies are the second major type of epidemiological studies. In contrast to descriptive studies that look at entire populations, in analytical studies, the subject of interest is the individual within the population. The object is not to formulate, but to test hypotheses. Nevertheless, although individuals are evaluated in analytical studies, the inference is not to individuals, but to the population from which they are selected.

Analytical studies comprise two distinct types of observational studies:

a. case control study
b. cohort study.

From each of these study designs, one can determine:

a. whether or not a statistical association exists between a disease and a suspected factor; and
b. if one exists, the strength of the Association.

A schematic design of case control and cohort studies is shown in Fig. 8.

![Design of a Case Control Study](Image)

**FIG. 8**

Schematic diagram of the design of case control and cohort studies

Source: (30A)
CASE CONTROL STUDY

Case control studies, often called "retrospective studies" are a common first approach to test causal hypothesis. In recent years, the case control approach has emerged as a permanent method of epidemiological investigation. The case control method has three distinct features:

a. both exposure and outcome (disease) have occurred before the start of the study
b. the study proceeds backwards from effect to cause; and

c. it uses a control or comparison group to support or refute an inference.

By definition, a case control study involves two populations – cases and controls. In case control studies, the unit is the individual rather than the group. The focus is on a disease or some other health problem that has already developed.

Case control studies are basically comparison studies. Cases and controls must be comparable with respect to known "confounding factors" such as age, sex, occupation, social status, etc. The questions asked relate to personal characteristics and antecedent exposures which may be responsible for the condition studied. For example, one can use as "cases" the immunized children and use as "controls" un-immunized children, and look for factors of interest in their past histories. Case control studies have been used effectively for studies of many cancers, and other serious conditions such as cirrhosis of the liver, lupus erythematosis, and congestive heart failure.

The basic design of a case control study is shown in Table 11. It is a 2x2 table which provides a very useful framework to discuss the various elements which make up a case control study. To illustrate, if it is our intention to test the hypothesis that "cigarette smoking causes lung cancer", using the case control method, the investigation begins by assembling a group of lung cancer cases (a+c), and a group of suitably matched controls (b+d). One then explores the past history of these two groups for the presence or absence of smoking, which is suspected to be related to the occurrence of cancer lung. If the frequency of smoking, a/(a+c) is higher in cases than in controls, b/(b+d), then smoking causes lung cancer. The possible sources from which controls can be selected include hospitals, relatives, neighbours and general population. (i) HOSPITAL CONTROLS: The controls may be selected from the same hospital as the cases, but with different illnesses other than the study disease. For example, if we are going to study cancer cervix patients, the control group may comprise patients with cancer breast, cancer of the digestive tract, or patients with non-cancerous lesions and other patients. Usually it is unwise to choose a control group from a group of patients with one disease. This is because hospital controls are often a source of "selection bias". Many hospital patients may have diseases which are also influenced by the factor under study. For example, if one was studying the

### Table 11

<table>
<thead>
<tr>
<th>Suspected or risk factors</th>
<th>Cases (Disease present)</th>
<th>Control (Disease absent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Absent</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td>a+c</td>
<td>b+d</td>
<td></td>
</tr>
</tbody>
</table>

**Basic steps**

There are four basic steps in conducting a case control study:

1. **Selection of cases and controls**
2. **Matching**
3. **Measurement of exposure, and**
4. **Analysis and interpretation.**

### 1. Selection of cases and controls

The first step is to identify a suitable group of cases and a group of controls. While identification of cases is relatively easy, selection of suitable controls may present difficulties. In this connection, definite guidelines have been laid down such as the following (4,9,12).

#### (1) SELECTION OF CASES

(a) **Definition of a case**: The prior definition of what constitutes a "case" is crucial to the case control study. It involves two specifications:

1. **DIAGNOSTIC CRITERIA**: The diagnostic criteria of the disease and the stage of disease, if any (e.g., breast cancer Stage I) to be included in the study must be specified before the study is undertaken. Supposing we are investigating cases of cancer, we should be quite clear that we have, for our cases, a group historically the same. Once the diagnostic criteria are established, they should not be altered or changed till the study is over.

2. **ELIGIBILITY CRITERIA**: The second criterion is that of eligibility. A criterion customarily employed is the requirement that only newly diagnosed (incident) cases within a specified period of time are eligible than old cases or cases in advanced stages of the disease (prevalent cases).

(b) **Sources of cases**: The cases may be drawn from:

1. **HOSPITALS**: It is often convenient to select cases from hospitals. The cases may be drawn from a single hospital or a network of hospitals, admitted during a specified period of time. The entire case series or a random sample of it is selected for study.

2. **GENERAL POPULATION**: In a population-based case control study, all cases of the study disease occurring within a defined geographic area during a specified period of time are ascertained, often through a survey, a disease registry or hospital network. The entire case series or a random sample of it is selected for study. The cases should be fairly representative of all cases in the community.

#### (2) SELECTION OF CONTROLS

The controls must be free from the disease under study. They must be as similar to the cases as possible, except for the absence of the disease under study. As a rule, a comparison group is identified before a study is done, comprising of persons who have not been exposed to the disease or some other factor whose influence is being studied. Selection of controls is an important aspect of the study, if the disease under investigation occurs in subclinical forms whose diagnosis is difficult. Selection of an appropriate control group is therefore an important prerequisite, for it is against this, we make comparisons, draw inferences and make judgements about the outcome of the investigation.

**Sources of controls**: The possible sources from which controls may be selected include hospitals, relatives, neighbours and general population. (i) HOSPITAL CONTROLS: The controls may be selected from the same hospital as the cases, but with different illnesses other than the study disease. For example, if we are going to study cancer cervix patients, the control group may comprise patients with cancer breast, cancer of the digestive tract, or patients with non-cancerous lesions and other patients. Usually, it is unwise to choose a control group from a group of patients with one disease. This is because hospital controls are often a source of "selection bias". Many hospital patients may have diseases which are also influenced by the factor under study. For example, if one was studying the
relationship of smoking and myocardial infarction and chooses bladder cancer cases as controls, the relationship between smoking and myocardial infarction may not have been demonstrated. Therefore, great care must be taken when using other patients as comparison subjects, for they differ in many ways from a normal healthy population. Ideally the controls should have undergone the same diagnostic work-up as cases, but have been found to be negative. But this may not be acceptable to most controls (ii) RELATIVES: The controls may also be taken up from relatives (spouses and siblings). Sibling controls are unsuitable where genetic conditions are under study. (iii) NEIGHBOURHOOD CONTROLS: The controls may be drawn from persons living in the same locality as cases, persons working in the same factory or children attending the same school. (iv) GENERAL POPULATION: Population controls can be obtained from defined geographic areas, by taking a random sample of individuals free of the study disease. We must use great care in the selection of controls to be certain that they accurately reflect the population that is free of the disease of interest.

2. Matching

The controls may differ from the cases in a number of factors such as age, sex, occupation, social status, etc. An important consideration is to ensure comparability between cases and controls. This involves what is known as "matching". Matching is defined as the process by which we select controls in such a way that they are similar to cases with regard to certain pertinent selected variables (e.g., age) which are known to influence the outcome of disease and which, if not adequately matched for comparability, could distort or confound the results. A "confounding factor" is defined as one which is associated both with exposure and disease, and is distributed unequally in study and control groups. More specifically a "confounding factor" is one that, although associated with "exposure" under investigation, is itself, independently of any such association, a "risk factor" for the disease. Two examples are cited to explain confounding.

(a) In the study of the role of alcohol in the aetiology of oesophageal cancer, smoking is a confounding factor because (i) it is associated with the consumption of alcohol and (ii) it is an independent risk factor for oesophageal cancer. In these conditions, the effects of alcohol consumption can be determined only if the influence of smoking is neutralized by matching (31).

(b) Age could be a confounding variable. Supposing, we are investigating the relationship between steroid contraceptive and breast cancer. If the women taking these contraceptives were younger than those in the comparison group, they would necessarily be at lower risk of breast cancer since this disease becomes increasingly common with increasing age. This "confounding" effect of age can be neutralized by matching so that both the groups have an equal proportion of each age group. In other words, matching protects against an unexpected strong association between the matching factor (e.g., age) and the disease (e.g., breast cancer). In a similar fashion other confounding variables will have to be matched.

While matching it should be borne in mind that the suspected aetiological factor or the variable we wish to measure should not be matched, because by matching, its aetiological role is eliminated in that study. The cases and controls will then become automatically alike with respect to that factor. In the above example, it would be useless to match cases and controls on steroid contraceptive use; by doing so, the aetiological role of steroid contraceptive cannot be investigated.

There are several kinds of matching procedures. One is group matching. This may be done by assigning cases to sub-categories (strata) based on their characteristics (e.g., age, occupation, social class) and then establishing appropriate controls. The frequency distribution of the matched variable must be similar in study and comparison groups. Matching is also done by pairs. For example, for each case, a control is chosen which can be matched quite closely. Thus, if we have a 50 year old mason with a particular disease, we will search for 50 year old mason without the disease as a control. Thus one can obtain pairs of patients and controls of the same sex, age, duration and severity of illness, etc. But there may be great difficulties in obtaining cases and controls matched on all characteristics, and it may be necessary to wait a considerable period of time before obtaining a sufficient number of matched pairs. Therefore, some leeway is necessary in matching for variables (32, 33). It should be noted that if matching is overdone, it may be difficult to find controls. Further with excess zeal in matching, there may be a tendency to reduce the odds ratio.

3. Measurement of exposure

Definitions and criteria about exposure (or variables which may be of aetiological importance) are just as important as those used to define cases and controls. Information about exposure should be obtained in precisely the same manner both for cases and controls. This may be obtained by interviews, by questionnaires or by studying past records of cases such as hospital records, employment records, etc. It is important to recognize that when case control studies are being used to test associations, the most important factor to be considered, even more important than the P values obtained, is the question of "bias" or systematic error which must be ruled out (see page 73).

4. Analysis

The final step is analysis, to find out

(a) Exposure rates among cases and controls to suspected factor

(b) Estimation of disease risk associated with exposure (Odds ratio)
(a) EXPOSURE RATES

A case control study provides a direct estimation of the exposure rates (frequency of exposure) to a suspected factor in disease and non-disease groups. Table 12 shows how exposure rates may be calculated from a case control study.

**TABLE 12**

A case control study of smoking and lung cancer

<table>
<thead>
<tr>
<th></th>
<th>Cases (with lung cancer)</th>
<th>Controls (without lung cancer)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smokers</td>
<td>a: 33</td>
<td>b: 55</td>
<td>a+b</td>
</tr>
<tr>
<td>(less than 5 cigarettes a day)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-smokers</td>
<td>c: 2</td>
<td>d: 27</td>
<td>c+d</td>
</tr>
<tr>
<td>Total</td>
<td>a+c=35</td>
<td>b+d=82</td>
<td>a+b+c</td>
</tr>
</tbody>
</table>

Source: (34)

**Exposure rates**

a. Cases = a/(a+c) = 33/35 = 94.2 per cent
b. Controls = b/(b+d) = 55/82 = 67.0 per cent

P < 0.001

Table 12 shows that the frequency rate of lung cancer was definitely higher among smokers than among non-smokers. The next step will be to ascertain whether there is a statistical association between exposure status and occurrence of lung cancer. This question can be resolved by calculating the \( P \) value, which in this case is less than 0.001.

The particular test of significance will depend upon the variables under investigation. If we are dealing with discrete variables, as in the present case (smoking and lung cancer; exposure and disease) the results are usually presented as rates or proportions of those present or absent in the study and in the control group. The test of significance usually adopted is the standard error of difference between two proportions or the Chi-square test. On the other hand, if we are dealing with continuous variables (e.g., age, blood pressure), the data will have to be grouped and the test of significance used is likely to be the standard error of difference between two means, or test.

According to convention, if \( P \) is less than or equal to 0.05, it is regarded as "statistically significant". The smaller the \( P \) value, the greater the statistical significance or probability that the association is not due to chance alone. However, statistical association (\( P \) value) does not imply causation. Statement of \( P \) value is thus an inadequate, although common end-point of case control studies.

(b) ESTIMATION OF RISK

The second analytical step is estimation of disease risk associated with exposure. It should be noted (Table 12) that if the exposure rate was 94.2 per cent in the study group, it does not mean that 94.2 per cent of those smoked would develop lung cancer. The estimation of disease risk associated with exposure is obtained by an index known as "Relative Risk" (RR) or "risk ratio", which is defined as the ratio between the incidence of disease among exposed persons and incidence among non-exposed. It is given by the formula:

\[
\text{Relative risk} = \frac{\text{Incidence among exposed}}{\text{Incidence among non-exposed}}
\]

A typical case control study does not provide incidence rates from which relative risk can be calculated directly, because there is no appropriate denominator or population at risk, to calculate these rates. In general, the relative risk can be exactly determined only from a cohort study.

**Odds Ratio (Cross-product ratio)**

From a case control study, we can derive what is known as Odds Ratio (OR) which is a measure of the strength of the association between risk factor and outcome. Odds ratio is closely related to relative risk. The derivation of odds ratio is based on three assumptions: (a) the disease being investigated must be relatively rare; (b) the cases must be representative of those with the disease, and (c) the controls must be representative of those without the disease. The odds ratio is the cross product of the entries in Table 11 which is reproduced below:

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Not exposed</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Odds ratio = \( \frac{ad}{bc} \)

Using the data in Table 12, the odds ratio would be estimated as follows:

\[
\text{Odds ratio} = \frac{a/b}{c/d} = \frac{ad}{bc}
\]

In the above example, smokers of less than 5 cigarettes per day showed a risk of having lung cancer 8.1 times that of non-smokers. Odds ratio is a key parameter in the analysis of case control studies.

**Bias in case control studies**

Bias is any systematic error in the determination of the association between the exposure and disease. The relative risk estimate may increase or decrease as a result of the bias; it reflects some type of non-comparability between the study and control groups. The possibility of bias must be considered when evaluating a possible cause and effect relationship.

Many varieties of bias may arise in epidemiological studies. Some of these are: (a) **Bias due to confounding** : Mention has already been made about confounding as an important source of bias. This bias can be removed by matching in case control studies. (b) **Memory or recall bias** : When cases and controls are asked questions about their past history, it may be more likely for the cases to recall the existence of certain events or factors, than the controls who are healthy persons. For example, those who have had a myocardial infarction might be more likely to remember and recall certain habits or events than those who have not. Thus cases may have a different recall of past events than controls. (c) **Selection bias** : The cases and controls may not be representative of cases and controls in the general population. There may be systematic differences in characteristics between cases and controls. The selection bias can be best controlled by its prevention (d) **Berkesonian bias** : A special example of bias is the Berkesonian bias, termed
after Dr. Joseph Berkson who recognized this problem. The bias arises because of the different rates of admission to hospitals for people with different diseases (i.e., hospital cases and controls). (e) Interviewer's bias: Bias may also occur when the interviewer knows the hypothesis and also knows who the cases are. This prior information may lead him to question the cases more thoroughly than controls regarding a positive history of the suspected causal factor. A useful check on this kind of bias can be made by noting the length of time taken to interview the average case and the average control. This type of bias can be eliminated by double-blinding (see page 83).

Advantages and disadvantages

Table 13 summarizes the advantages and disadvantages of case control studies.

<table>
<thead>
<tr>
<th>TABLE 13</th>
<th>Advantages and disadvantages of case control studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ADVANTAGES</strong></td>
<td></td>
</tr>
<tr>
<td>1. Relatively easy to carry out.</td>
<td></td>
</tr>
<tr>
<td>2. Rapid and inexpensive (compared with cohort studies).</td>
<td></td>
</tr>
<tr>
<td>3. Require comparatively few subjects.</td>
<td></td>
</tr>
<tr>
<td>4. Particularly suitable to investigate rare diseases or diseases about which little is known. But a disease which is rare in the general population (e.g., leukaemia in adolescents) may not be rare in special exposure group (e.g. prenatal X-rays).</td>
<td></td>
</tr>
<tr>
<td>5. No risk to subjects.</td>
<td></td>
</tr>
<tr>
<td>6. Allows the study of several different aetiological factors (e.g., smoking, physical activity and personality characteristics in myocardial infarction).</td>
<td></td>
</tr>
<tr>
<td>7. Risk factors can be identified. Rational prevention and control programmes can be established.</td>
<td></td>
</tr>
<tr>
<td>8. No attrition problems, because case control studies do not require follow-up of individuals into the future.</td>
<td></td>
</tr>
<tr>
<td>9. Ethical problems minimal.</td>
<td></td>
</tr>
</tbody>
</table>

| **DISADVANTAGES** | |
| 1. Problems of bias relies on memory or past records, the accuracy of which may be uncertain; validation of information obtained is difficult or sometimes impossible. | |
| 2. Selection of an appropriate control group may be difficult. | |
| 3. We cannot measure incidence, and can only estimate the relative risk. | |
| 4. Do not distinguish between causes and associated factors. | |
| 5. Not suited to the evaluation of therapy or prophylaxis of disease. | |
| 6. Another major concern is the representativeness of cases and controls. | |

Examples of case control studies

Case control studies have provided much of the current base of knowledge in epidemiology. Some of the early case control studies centred round cigarette smoking and lung cancer (34,37,38). Other studies include: maternal smoking and congenital malformations (39), radiation and leukaemia (40), oral contraceptive use and hepatocellular adenoma (41), herpes simplex and Bell palsy (42), induced abortion and spontaneous abortion (43), physical activity and coronary death (44), artificial sweeteners and bladder cancer (45), etc.

A few studies are cited in detail:

Example 1: Adenocarcinoma of vagina (26).

An excellent example of a case control study is adenocarcinoma of the vagina in young women. It is not only a rare disease, but also the usual victim is over 50 years of age. There was an unusual occurrence of this tumor in 7 young women (15 to 22 years) born in one Boston hospital between 1966 and 1969. The apparent "time clustering" of cases - 7 occurring within 4 years at a single hospital - led to this enquiry. An eighth case occurred in 1969 in a 20 year old patient who was treated at another Boston hospital in USA.

The cause of this tumor was investigated by a case control study in 1971 to find out the factors that might be associated with this tumor. As this was a rare disease, for each case, four matched controls were put up. The controls were identified from the birth records of the hospital in which each case was born. Female births occurring closest in time to each patient were selected as controls. Information was collected by personal interviews regarding (a) maternal age (b) maternal smoking (c) antenatal radiology, and (d) diethyl-stilbestrol (DES) exposure in foetal life. The results of the study are shown in Table 14 which shows that cases differed significantly from the controls in their past history. Seven of the eight cases had been exposed to DES in foetal life. This drug had been given to their mothers during the first trimester of pregnancy to prevent possible miscarriage. But none of the mothers in the control group had received DES.

Since this study, more cases have been reported and the association with DES has been confirmed. The case control method played a critical role in revealing exposure to DES in utero as the cause of vaginal adenocarcinoma in the exposed child 10–20 years later.

<table>
<thead>
<tr>
<th>TABLE 14</th>
<th>Association between maternal DES therapy and adenocarcinoma of vagina amongst female offspring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information acquired retrospectively</td>
<td>Cases</td>
</tr>
<tr>
<td>Maternal age</td>
<td>26.1</td>
</tr>
<tr>
<td>Maternal smoking</td>
<td>7</td>
</tr>
<tr>
<td>Antenatal radiology</td>
<td>1</td>
</tr>
<tr>
<td>Oestrogen exposure</td>
<td>7</td>
</tr>
</tbody>
</table>

Example 2: Oral contraceptives and thromboembolic disease (46,47).

By August 1965, the British Committee on Safety of Drugs had received 249 reports of adverse reactions and 16 reports of death in women taking oral contraceptives. It became apparent that epidemiological studies were needed to determine whether women who took oral contraceptives were at greater risk of developing thromboembolic disease.

In 1968 and 1969, Vassey and Doll reported the findings of their case control studies in which they interviewed women who had been admitted to hospitals with venous thrombosis or pulmonary embolism without medical cause and compared the history with that obtained from other women who had been admitted to the same hospital with other diseases and who were matched for age, marital status and parity.

It was found that out of 84, 42 (50%) of those with venous thrombosis and pulmonary embolism had been using oral contraceptives, compared with 14% of controls (Table 15). The studies confirmed that taking the pill and having pulmonary embolism co-existed more frequently than would be expected by chance. The relative risk of women who were non-users was 6.3:1. That is, the investigators found that users of oral contraceptives were about 6 times as likely as non-users to develop thromboembolic disease.
Example 3: Thalidomide tragedy (48).

Thalidomide was first marketed as a safe, non-barbiturate hypnotic in Britain in 1958. In 1961, at a congress of Gynaecologists, attention was drawn to the birth of a large number of babies with congenital abnormalities, which was previously rare. In the same year, it was suggested that thalidomide might be responsible for it.

A retrospective study of 46 mothers delivered of deformed babies showed that 41 were found to have thalidomide during their early pregnancy. This was compared with a control of 300 mothers who had delivered normal babies; none of these had taken thalidomide. Laboratory experiments confirmed that thalidomide was teratogenic in experimental studies (48).

Cohort Study

Cohort study is another type of analytical (observational) study which is usually undertaken to obtain additional evidence to refute or support the existence of an association between suspected cause and disease. Cohort study is known by a variety of names: prospective study, longitudinal study, incidence study, and forward-looking study. The most widely used term, however, is "cohort study" (4).

The distinguishing features of cohort studies are:

- a. The cohorts must be free from the disease under investigation. The basic design of a simple cohort study is shown in Table 16.
- b. Insofar as the knowledge of the disease permits, both the groups (i.e., study and control cohorts) should be equally susceptible to the disease under study, or efficiently reflect any difference in disease occurrence (for example, males over 35 years would be appropriate for studies on lung cancer).
- c. Both the groups should be comparable in respect of all the possible variables, which may influence the frequency of the disease; and
- d. The diagnostic and eligibility criteria of the disease must be defined beforehand; this will depend upon the availability of reliable methods for recognizing the disease when it develops.

The groups are then followed, under the same identical conditions, over a period of time to determine the outcome of exposure (e.g., onset of disease, disability or death) in both the groups. In chronic diseases such as cancer the time required for the follow-up may be very long.

Table 16 shows (a+b) persons were exposed to the factor under study, 'a' of which developed the disease during the follow-up period; (c+d) persons were not exposed, 'c' of which became cases (it is assumed for simplicity of presentation that there were no intermittent deaths or losses during the follow-up period). After the end of the follow-up,
the incidence rate of the disease in both the groups is determined. If it is found that the incidence of the disease in the exposed group, \( a/(a+b) \) is significantly higher than in the non-exposed group, \( c/(c+d) \), it would suggest that the disease and suspected cause are associated. Since the approach is prospective, that is, studies are planned to observe events that have not yet occurred, cohort studies are frequently referred to as "prospective" studies.

A well-designed cohort study is considered the most reliable means of showing an association between a suspected risk factor and subsequent disease because it eliminates many of the problems of the case control study and approximates the experimental model of the physical sciences.

**Types of cohort studies**

Three types of cohort studies have been distinguished on the basis of the time of occurrence of disease in relation to the time at which the investigation is initiated and continued:

1. Prospective cohort studies
2. Retrospective cohort studies
3. A combination of retrospective and prospective cohort studies.

**1. Prospective cohort studies**

A prospective cohort study (or "current" cohort study) is one in which the outcome (e.g., disease) has not yet occurred at the time the investigation begins. Most prospective studies begin in the present and continue into future. For example, the long-term effects of exposure to uranium was evaluated by identifying a group of uranium miners and a comparison group of individuals not exposed to uranium mining and by assessing subsequent development of lung cancer in both the groups. The principal finding was that the uranium miners had an excess frequency of lung cancer compared to non-miners. Since the disease had not yet occurred when the study was undertaken, this was a prospective cohort design. The US Public Health Service’s Framingham Heart Study (49), Doll and Hills (50) prospective study on smoking and lung cancer, and study of oral contraceptives and health by the Royal College of General Practitioners (51) are examples of this type of study.

**2. Retrospective cohort studies**

A retrospective cohort study (or "historical" cohort study) is one in which the outcomes have all occurred before the start of the investigation. The investigator goes back in time, sometimes 10 to 30 years, to select his study groups from existing records of past employment, medical or other records and traces them forward through time from a past date fixed on the records, usually up to the present. This type of study is known by a variety of names: retrospective cohort study, "historical" cohort study, prospective study in retrospect and non-concurrent prospective study.

The successful application of this approach is illustrated in one study undertaken in 1978 – a cohort of 17,080 babies born between January 1, 1969 and December 31, 1975 at a Boston hospital were investigated of the effects of electronic foetal monitoring during labour. The outcome measured was neonatal death. The study showed that the neonatal death rate was 1.7 times higher in unmonitored infants (52). The most notable retrospective cohort studies to date are those of occupational exposures, because the recorded information is easily available, e.g., study of the role of arsenic in human carcinogenesis, study of lung cancer in uranium miners, study of the mortality experience of groups of physicians in relation to their probable exposure to radiation (53,54,55). More recently, angiosarcoma of the liver, a very rare disease, has been reported in excess frequency in relation to poly-vinyl chloride (56). This association was picked up only because of the retrospective cohort design. Retrospective cohort studies are generally more economical and produce results more quickly than prospective cohort studies.

**3. Combination of retrospective and prospective cohort studies**

In this type of study, both the retrospective and prospective elements are combined. The cohort is identified from past records, and is assessed of date for the outcome. The same cohort is followed up prospectively into future for further assessment of outcome.

Court–Brown and Doll (1957) applied this approach to study the effects of radiation. They assembled a cohort in 1955 consisting of 13,352 patients who had received large doses of radiation therapy for ankylosing spondylitis between 1934 and 1954. The outcome evaluated was death from leukemia or aplastic anaemia between 1935 and 1954. They found that the death rate from leukemia or aplastic anaemia was substantially higher in their cohort than that of the general population. A prospective component was added to the study and the cohort was followed, as established in 1955, to identify deaths occurring in subsequent years (57).

**ELEMENTS OF A COHORT STUDY**

The elements of a cohort study are:

1. Selection of study subjects
2. Obtaining data on exposure
3. Selection of comparison groups
4. Follow-up, and
5. Analysis.

**1. Selection of study subjects**

The subjects of a cohort study are usually assembled in one of two ways – either from general population or select groups of the population that can be readily studied (e.g., persons with different degrees of exposure to the suspected causal factor).

(a) General population: When the exposure or cause of death is fairly frequent in the population, cohorts may be assembled from the general population, residing in well-defined geographical, political and administrative areas (e.g., Framingham Heart Study). If the population is very large, an appropriate sample is taken, so that the results can be generalized to the population sampled. The exposed and unexposed segments of the population to be studied should be representative of the corresponding segments of the general population.

(b) Special groups: These may be special groups or exposure groups that can readily be studied: (i) Select groups: These may be professional groups (e.g., doctors, nurses, lawyers, teachers, civil servants), insured persons, obstetric population, college alumni, government employees, volunteers, etc. These groups are usually a homogeneous population. Doll’s prospective study on smoking and lung cancer was carried out on British doctors listed in the Medical Register of the UK in 1951 (58). The study by Dorn on smoking and mortality in 293,658 veterans (i.e., former military service) in United States
having life insurance policies is another example of a study based on special groups (59). These groups are not only homogeneous, but they also offer advantages of accessibility and easy follow-up for a protracted period (ii) Exposure groups : If the exposure is rare, a more economical procedure is to select a cohort of persons known to have experienced the exposure. In other words, cohorts may be selected because of special exposure to physical, chemical and other disease agents. A readily accessible source of these groups is workers in industries and those employed in high-risk situations (e.g., radiologists exposed to X-rays).

When cohorts have been selected because of special exposure, it facilitates classification of cohort members according to the degree or duration of exposure to the suspected factor for subsequent analytical study.

2. Obtaining data on exposure

Information about exposure may be obtained directly from the (a) Cohort members : through personal interviews or mailed questionnaires. Since cohort studies involve large numbers of population, mailed questionnaires offer a simple and economic way of obtaining information. For example, Doll and Hill (60) used mailed questionnaires to collect smoking histories from British doctors. (b) Review of records : Certain kinds of information (e.g., dose of radiation, kinds of surgery, or details of medical treatment) can be obtained only from medical records. (c) Medical examination or special tests : Some types of information can be obtained only by medical examination or special tests, e.g., blood pressure, serum cholesterol, ECG. (d) Environmental surveys : This is the best source for obtaining information on exposure levels of the suspected factor in the environment where the cohort lived or worked. In fact, information may be needed from more than one or all of the above sources.

Information about exposure (or any other factor related to the development of the disease being investigated) should be collected in a manner that will allow classification of cohort members :

(a) according to whether or not they have been exposed to the suspected factor, and
(b) according to the level or degree of exposure, at least in broad classes, in the case of special exposure groups (Table 17).

In addition to the above, basic information about demographic variables which might affect the frequency of disease under investigation, should also be collected. Such information will be required for subsequent analysis.

3. Selection of comparison groups

There are many ways of assembling comparison groups :

(a) Internal comparisons

In some cohort studies, no outside comparison group is required. The comparison groups are in-built. That is, single cohort enters the study, and its members may, on the basis of information obtained, be classified into several comparison groups according to the degrees or levels of exposure to risk (e.g., smoking, blood pressure, serum cholesterol) before the development of the disease in question. The groups, so defined, are compared in terms of their subsequent morbidity and mortality rates. Table 17 illustrates this point. It shows that mortality from lung cancer increases with increasing number of cigarettes smoked reinforcing the conclusion that there is valid association between smoking and lung cancer.

<table>
<thead>
<tr>
<th>Classification of exposure (cigarettes)</th>
<th>No. of deaths</th>
<th>Death rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/2 pack</td>
<td>24</td>
<td>95.2</td>
</tr>
<tr>
<td>1/2-1 pack</td>
<td>84</td>
<td>107.8</td>
</tr>
<tr>
<td>1-2 packs</td>
<td>90</td>
<td>229.2</td>
</tr>
<tr>
<td>2 packs +</td>
<td>97</td>
<td>264.2</td>
</tr>
</tbody>
</table>

Source : (5)

(b) External comparisons

When information on degree of exposure is not available, it is necessary to put up an external control, to evaluate the experience of the exposed group, e.g., smokers and non-smokers, a cohort of radiologists compared with a cohort of ophthalmologists, etc. The study and control cohorts should be similar in demographic and possibly important variables other than those under study.

(c) Comparison with general population rates

If none is available, the mortality experience of the exposed group is compared with the mortality experience of the general population in the same geographic area as the exposed people, e.g., comparison of frequency of lung cancer among uranium mine workers with lung cancer mortality in the general population where the miners resided (54); comparison of frequency of cancer among asbestos workers with the rate in general population in the same geographic area (61).

Rates for disease occurrence in sub-groups of the control cohort by age, sex, and other variables considered important may be applied to the corresponding sub-groups of the study cohort (exposed cohort) to determine the "expected" values in the absence of exposure. The ratio of "observed" and "expected" values provides a measure of the effect of the factor under study.

The limiting factors in using general population rates for comparison are : (i) non-availability of population rates for the outcome required; and (ii) the difficulties of selecting the study and comparison groups which are representative of the exposed and non-exposed segments of the general population.

4. Follow-up

One of the problems in cohort studies is the regular follow-up of all the participants. Therefore, at the start of the study, methods should be devised depending upon the outcome to be determined (morbidity or death), to obtain data for assessing the outcome. The procedures required comprise :

(a) periodic medical examination of each member of the cohort
(b) reviewing physician and hospital records
(c) routine surveillance of death records, and
(d) mailed questionnaires, telephone calls, periodic home visits – preferably all three on an annual basis.

Of the above, periodic examination of each member of the cohort, yields greater amount of information on the individuals examined, than would the use of any other procedure.

However, inspiteof best efforts, a certain percentage of losses to follow-up are inevitable due to death, change of residence, migration or withdrawal of occupation. These losses may bias the results. It is, therefore, necessary to build into the study design a system for obtaining basic
information on outcome for those who cannot be followed up in detail for the full duration of the study (13). The safest course recommended is to achieve as close to a 95 per cent follow-up as possible (12).

5. Analysis

The data are analyzed in terms of:
(a) Incidence rates of outcome among exposed and non-exposed,
(b) Estimation of risk.

(a) Incidence rates

In a cohort study, we can determine incidence rates directly in those exposed and those not exposed. A hypothetical example is given in Table 18 showing how incidence rates may be calculated:

**TABLE 18**

<table>
<thead>
<tr>
<th>Cigarette Smoking</th>
<th>Developed Lung Cancer</th>
<th>Did Not Develop Lung Cancer</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>70</td>
<td>6930</td>
<td>7000</td>
</tr>
<tr>
<td>No</td>
<td>3</td>
<td>2997</td>
<td>3000</td>
</tr>
</tbody>
</table>

Incidence rates
(a) among smokers = 70/7000 = 10 per 1000
(b) among non-smokers = 3/3000 = 1 per 1000
Statistical significance: P < 0.001

(b) Estimation of risk

Having calculated the incidence rates, the next step is to estimate the risk of outcome (e.g., disease or death) in the exposed and non-exposed cohorts. This is done in terms of two well-known indices: (a) relative risk, (b) attributable risk.

**RELATIVE RISK**

Relative risk (RR) is the ratio of the incidence of the disease (or death) among exposed and the incidence among non-exposed. Some authors use the term “risk ratio” to refer to relative risk.

\[ RR = \frac{\text{Incidence of disease (or death) among exposed}}{\text{Incidence of disease (or death) among non-exposed}} \]

In our hypothetical example (Table 18)

RR of lung cancer = \( \frac{10}{1} = 10 \)

Estimation of relative risk (RR) is important in aetiological enquiries. It is a direct measure (or index) of the “strength” of the association between suspected cause and effect. A relative risk of one indicates no association; relative risk greater than one suggests “positive” association between exposure and the disease under study. A relative risk of 2 indicates that the incidence rate of disease is 2 times higher in the exposed group as compared with the unexposed. Equivalently, this represents a 100 per cent increase in risk. A relative risk of 0.25 indicates a 75% reduction in the incidence rate in exposed individuals as compared with the unexposed (35). It is often useful to consider the 95 per cent confidence interval of a relative risk since it provides an indication of the likely and maximum levels of risk.

In our hypothetical example (Table 18), the relative risk is 10. It implies that smokers are 10 times at greater risk of developing lung cancer than non-smokers. The larger the RR, the greater the “strength” of the association between the suspected factor and disease. It may be noted that risk does not necessarily imply causal association.

**ATTRIBUTABLE RISK**

Attributable risk (AR) is the difference in incidence rates of disease (or death) between an exposed group and non-exposed group. Some authors use the term “risk difference” to attributable risk.

Attributable risk is often expressed as a per cent. This is given by the formula:

\[ \text{Attributable risk} = \frac{\text{Incidence of disease rate among exposed} - \text{incidence of disease rate among non-exposed}}{\text{Incidence rate among exposed}} \times 100 \]

Attributable risk in our example (Table 18) would be:

\[ \frac{10 - 1}{10} \times 100 = 90 \text{ per cent} \]

Attributable risk indicates to what extent the disease under study can be attributed to the exposure. The figure in our example indicates that the association between smoking and lung cancer is causal, 90 per cent of the lung cancer among smokers was due to their smoking. This suggests the amount of disease that might be eliminated if the factor under study could be controlled or eliminated.

**POPULATION-ATTRIBUTABLE RISK**

Another concept is “population-attributable risk”. It is the incidence of the disease (or death) in the total population minus the incidence of disease (or death) among those who were not exposed to the suspected causal factor (Table 19).

**TABLE 19**

<table>
<thead>
<tr>
<th>Deaths per 100,000 person-years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heavy smokers</td>
</tr>
<tr>
<td>Non-smokers</td>
</tr>
<tr>
<td>Deaths in total population</td>
</tr>
<tr>
<td>Individual RR</td>
</tr>
<tr>
<td>Population AR</td>
</tr>
</tbody>
</table>

Source: (58)

The concept of population attributable risk is useful in that it provides an estimate of the amount by which the disease could be reduced in that population if the suspected factor was eliminated or modified. In our example (Table 19) one might expect that 86 per cent of deaths from lung cancer could be avoided if the risk factor of cigarettes were eliminated.

**Relative risk versus attributable risk**

Relative risk is important in aetiological enquiries. Its size is a better index than is attributable risk for assessing the aetiological role of a factor in disease. The larger the relative
risk, the stronger the association between cause and effect. But relative risk does not reflect the potential public health importance as does the attributable risk. That is, attributable risk gives a better idea than does relative risk of the impact of successful preventive or public health programme might have in reducing the problem.

Two examples are cited (Tables 20 and 21) to show the practical importance of distinguishing relative and absolute risk. In the first example, (Table 20) the RR of a cardiovascular complication in users of oral contraceptives is independent of age, whereas the AR is more than 5 times higher in the older age groups. This epidemiological observation has been the basis for not recommending oral contraceptive in those aged 35 years and over.

**TABLE 20**
The relative and attributable risks of cardiovascular complications in women taking oral contraceptives

<table>
<thead>
<tr>
<th>Cardiovascular risk</th>
<th>Age</th>
<th>Relative risk</th>
<th>Attributable risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>100,000 patient years</td>
<td>30–39</td>
<td>2.8</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>40–44</td>
<td>2.8</td>
<td>20.0</td>
</tr>
</tbody>
</table>

Source: (62)

The second example (Table 21) shows that smoking is attributable to 92 per cent of lung cancer, and 13.3 per cent of CHD. In CHD, both RR and AR are not very high suggesting not much of the disease could be prevented as compared to lung cancer.

**TABLE 21**
Risk assessment, smokers vs non-smokers

<table>
<thead>
<tr>
<th>Cause of death</th>
<th>Death rate/1000</th>
<th>Smokers</th>
<th>Non-smokers</th>
<th>RR</th>
<th>AR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung cancer</td>
<td>0.90</td>
<td>0.07</td>
<td>12.86</td>
<td>92.2</td>
<td></td>
</tr>
<tr>
<td>CHD</td>
<td>4.87</td>
<td>4.22</td>
<td>1.15</td>
<td>13.3</td>
<td></td>
</tr>
</tbody>
</table>

Source: (63)

**Advantages and disadvantages of cohort studies**

**Advantages**

(a) Incidence can be calculated. (b) Several possible outcomes related to exposure can be studied simultaneously — that is, we can study the association of the suspected factor with many other diseases in addition to the one under study. For example, cohort studies designed to study the association between smoking and lung cancer also showed association of smoking with coronary heart disease, peptic ulcer, cancer oesophagus and several others. (c) Cohort studies provide a direct estimate of relative risk. (d) Dose–response ratios can also be calculated, and (e) Since comparison groups are formed before disease develops, certain forms of bias can be minimized like mis-classification of individuals into exposed and unexposed groups.

**Disadvantages**

Cohort studies also present a number of problems: (a) Cohort studies involve a large number of people. They are generally unsuitable for investigating uncommon diseases or diseases with low incidence in the population. (b) It takes a long time to complete the study and obtain results (20–30 years or more in cancer studies) by which time the investigators may have died or the participants may have changed their classification. Even in very common chronic diseases like coronary heart disease, cohort studies are difficult to carry out. It is difficult to keep a large number of individuals under medical surveillance indefinitely. (c) Certain administrative problems such as loss of experienced staff, loss of funding and extensive record keeping are inevitable. (d) It is not unusual to lose a substantial proportion of the original cohort — they may migrate, lose interest in the study or simply refuse to provide any required information. (e) Selection of comparison groups which are representative of the exposed and unexposed segments of the population is a limiting factor. Those who volunteer for the study may not be representative of all individuals with the characteristic of interest. (f) There may be changes in the standard methods or diagnostic criteria of the disease over prolonged follow-up. Once we have established the study protocol, it is difficult to introduce new knowledge or new tests later. (g) Cohort studies are expensive. (h) The study itself may alter people’s behaviour. If we are examining the role of smoking in lung cancer, an increased concern in the study cohort may be created. This may induce the study subjects to stop or decrease smoking. (i) With any cohort study we are faced with ethical problems of varying importance. As evidence accumulates about the implicating factor in the aetiology of disease, we are obliged to intervene and if possible reduce or eliminate this factor, and (j) Finally, in a cohort study, practical considerations dictate that we must concentrate on a limited number of factors possibly related to disease outcome.

The main differences between case control and cohort studies are summarized in Table 22.

**TABLE 22**
Main differences between case control and cohort studies

<table>
<thead>
<tr>
<th>Case control study</th>
<th>Cohort study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Proceeds from “effect to cause”.</td>
<td>Proceeds from “cause to effect”.</td>
</tr>
<tr>
<td>2. Starts with the disease.</td>
<td>Starts with people exposed to risk factor or suspected cause.</td>
</tr>
<tr>
<td>3. Tests whether the suspected cause occurs more frequently in those with the disease than among those without the disease.</td>
<td>Tests whether disease occurs more frequently in those exposed, than in those not similarly exposed.</td>
</tr>
<tr>
<td>4. Usually the first approach to testing of a hypothesis, but also useful for exploratory studies.</td>
<td>Reserved for testing of precisely formulated hypothesis.</td>
</tr>
<tr>
<td>5. Involves fewer number of subjects.</td>
<td>Involves larger number of subjects.</td>
</tr>
<tr>
<td>6. Yields relatively quick results.</td>
<td>Long follow-up period often needed, involving delayed results.</td>
</tr>
<tr>
<td>7. Suitable for the study of rare diseases.</td>
<td>Inappropriate when the disease or exposure under investigation is rare.</td>
</tr>
<tr>
<td>8. Generally yields only estimate of RR (odds ratio).</td>
<td>Yields incidence rates, RR as well as AR.</td>
</tr>
<tr>
<td>9. Cannot yield Information about diseases other than that selected for study.</td>
<td>Can yield information about more than one disease outcome.</td>
</tr>
</tbody>
</table>
Examples of cohort studies

Example 1: Smoking and lung cancer.

At least eight prospective studies on the relation of smoking to lung cancer had been done. Doll and Hill (50, 60, 64), Hammond and Horn (65,66) and Dorn (59) were the first to report their findings.

In October 1951, Doll and Hill sent a questionnaire to 59,600 British doctors listed in the Medical Register of the UK enquiring about their smoking habits. This enabled them to form two cohorts (smokers and non-smokers) who were similar in all other respects like age, education and social class. They received usable replies from 40,701 physicians - 34,494 men and 6,207 women. These were followed for 4 years and 5 months by obtaining notifications of physicians' deaths from the Registrar General, the General Medical Council and the British Medical Association. For every death certified as due to lung cancer, confirmation was obtained by writing to the physician certifying the death and also, when necessary to the hospital or consultant to whom the patient had been referred. The results of the study are shown in Table 19.

Example 2: The Framingham heart study (49).

The Framingham heart study was initiated in 1948 by the United States Public Health Service to study the relationship of a number of (risk) factors (e.g., serum cholesterol, blood pressure, weight, smoking) to the subsequent development of cardiovascular disease. The town of Framingham (Massachusetts) had a population of 28,000 in 1948. The study was planned for 20 years in view of the slow development of heart disease.

The lower and upper limits of the study population was set at 30 and 59 years. Out of 10,000 people in this age group a sample of 6,507 persons of both sexes were invited to participate in the study, out of which 5,209 participated. The initial examination revealed that 82 subjects had clinically evident CHD. These were excluded from the sample leaving a total of 5,127.

4,469 (69 per cent) of the 6,507 in the initial sample actually underwent the first examination. After the first examination, the study population was examined every 2 years for a 20 year period. Information was obtained with regard to serum cholesterol, blood pressure, weight and cigarette smoking. Although biennial examinations were the main source of follow up information, other means were also adopted to detect CHD (e.g., Death certificate records).

Among other things, the study showed increasing risk of CHD with increasing serum cholesterol levels in the 45–54 age group. The study also showed that the association between smoking and CHD varied with manifestations of the disease. Thus, smoking was more strongly associated with sudden death from CHD than with less fatal forms of the disease. Risk factors have been found to include male sex, advancing age, high serum lipoprotein concentration, high blood pressure, cigarette smoking, diabetes mellitus, obesity, low vital capacity and certain ECG abnormalities. The predictive value of serum lipids, blood pressure and cigarette smoking have been repeatedly demonstrated. The Framingham heart study became a prototype of similar studies in US and other countries.

Example 3: Oral contraceptives and health (51).

Another example is the cohort study of oral contraceptives and health conducted by the Royal College of General Practitioners in England (1974). It was initiated in 1968, after 2 years of planning. 23,000 users of the pill aged 15–49 years together with a similar number of controls using other methods or no method of contraception were brought under observation of 1400 general practitioners. During follow-up doctors recorded the diagnoses of episodes of illness, and information about pregnancies and deaths.

The study brought out the risks and benefits of oral contraceptive use. For example, the study showed that the risk of hypertension increases, and the risk of benign breast disease decreases with the dose of norethisterone acetate (progestogen) in the combined pill which is an important finding. The study found an increased mortality from diseases of cardiovascular system in pill users confirming the results of retrospective case control studies (67).

EXPERIMENTAL EPIDEMIOLOGY

In the 1920s, “experimental epidemiology” meant the study of epidemics among colonies of experimental animals such as rats and mice. In modern usage, experimental epidemiology is often equated with RANDOMIZED CONTROLLED TRIALS (2).

Experimental or intervention studies are similar in approach to cohort studies excepting that the conditions in which study is carried out are under the direct control of the investigator. Thus experimental studies involve some action, intervention or manipulation such as deliberate application or withdrawal of the suspected cause or changing one variable in the causative chain in the experimental group while making no change in the control group, and observing and comparing the outcome of the experiment in both the groups. This contrasts sharply with observational studies (e.g., descriptive, case control and cohort studies), where the epidemiologist takes no action but only observes the natural course of events or outcome.

The aims of experimental studies may be stated as follows: (a) to provide “scientific proof” of aetiological (or risk) factors which may permit the modification or control of those diseases: and (b) to provide a method of measuring the effectiveness and efficiency of health services for the prevention, control and treatment of disease and improve the health of the community.

Experimental studies have all the advantages and disadvantages of the usual prospective cohort studies plus three additional problems namely cost, ethics and feasibility. Experimental studies have become a major area of epidemiological studies. They may be conducted in animals or human beings.

Animal studies

Throughout history animals have played an important role in men's quest for knowledge about himself and his environment. Animal studies have contributed to our knowledge of anatomy, physiology, pathology, microbiology, immunology, genetics, chemotherapy and so many others. At the beginning of this century, Webster in United States and Topley, Wilson and Greenwood in England had carried out classical animal experiments. Their studies centred round inducing epidemics in animals and in studies of herd immunity under laboratory conditions.

More important application of animal experiments have been in (a) experimental reproduction of human disease in animals to confirm aetiological hypotheses and to study the
pathogenetic phenomena or mechanisms (b) testing the efficacy of preventive and therapeutic measures such as vaccines and drugs, and (c) completing the natural history of disease. For example, naturally occurring leprosy has been found in armadillos. Data obtained from studying these animals indicate that lepra bacilli might exist outside of humans.

Animal experiments have their own advantages and limitations. The advantages are that the experimental animals can be bred in laboratories and manipulated easily according to the wishes of the investigator. A more important point is that they multiply rapidly and enable the investigators to carry out certain experiments (e.g., genetic experiments) which in human population would take several years and involve many generations. The limitations of animal experiments are that not all human diseases can be reproduced in animals. Secondly, all the conclusions derived from animal experiments may not be strictly applicable to human beings. An excellent example to illustrate this point is the WHO trial of typhoid vaccine in Yugoslavia in the mid-1960s. Laboratory tests in animals showed the alcohol-killed and preserved vaccine to be more effective than the traditional heat-killed phenol-preserved vaccine. But randomized controlled trials in human beings demonstrated that, contrary to laboratory evidence, the alcohol-preserved vaccine was found to be less than half as effective in preventing typhoid fever as the traditional phenol-preserved vaccine introduced by Almorth Wright. This highlights the difficulties encountered in extrapolating findings from animal experiments in man.

Human experiments

Human experiments will always be needed to investigate disease aetiology and to evaluate the preventive and therapeutic measures. These studies are even more essential in the investigation of diseases that cannot be reproduced in animals.

Historically, in 1747, James Lind performed a human experiment (clinical trial) in which he added different substances to diet of 12 soldiers who were suffering from scurvy. He divided his patients into 6 pairs and supplemented the diets of each pair with cider, elixir vitriol, vinegar, sea water; a mixture of nutmeg, garlic, mustard and tamarind in barley water; and two oranges and one lemon daily. All the subjects were studied for 6 days. At the end of 6 days the LIMEYS recovered from scurvy and were found fit for duty. Then came Edward Jenner’s experiment with cowpox in 1796. Other classical experiments are Finlay and Reed’s experiments (1881–1900) to elucidate the mosquito–borne nature of yellow fever and Goldberger’s classical experiments in 1915 inducing pellagra by diets deficient in nicotinic acid, thereby proving pellagra to be a nutritional deficiency disease, not an infectious disease as was then supposed. Since then, human beings have participated in studies of malaria, syphilis, hepatitis, measles, polio and others. These experiments have played decisive roles in investigating disease aetiology and in testing preventive and therapeutic measures.

Although the experimental method is unquestionably the most incisive approach to scientific problem, ethical and logistic considerations often prevent its application to the study of disease in humans. Therefore, before launching human experiments, the benefits of the experiment have to be weighed against risks involved. The volunteers should be made fully aware of all possible consequences of the experiment. Thus when an illness is fatal (e.g., excessive haemorrhage) and the benefit of treatment (e.g., blood transfusion) is self-evident, it would be ethically unacceptable to prove or disprove the therapeutic value of blood transfusion. However, such instances represent only a small part of the total research effort. On the other hand, in the present era of scientific medicine, many unsound and scientifically unsound procedures are still being carried out. For instance, in the study of prescription drugs, a panel of experts in USA found that only 23 per cent of some 16,000 drugs could be classified unequivocally as "effective" (36). It is now conceded that it is equally unethical if a drug or procedure is brought into general use without establishing its effectiveness by controlled trials. The thalidomide disaster and the occurrence of carcinoma of the vagina in the offspring of pregnant women treated with diethylstilbestrol highlight the unfortunate consequence of therapy on the basis of uncontrolled observations. The WHO in 1980 has laid down a strict code of practice in connection with human trials (68).

Experimental studies are of two types:

a. Randomized controlled trials (i.e., those involving a process of random allocation); and

b. Non-randomized or "non-experimental" trials (i.e., those departing from strict randomization for practical purposes, but in such a manner that non-randomization does not seriously affect the theoretical basis of conclusions).

RANDOMIZED CONTROLLED TRIALS

Too often physicians are guided in their daily work by clinical impressions of their own or their teachers. These impressions, particularly when they are incorporated in textbooks and repeatedly quoted by reputed teachers and their students acquire authority, just as if they were proved facts. Similarly many public health measures are introduced on the basis of assumed benefits without subjecting them to rigorous testing. The history of medicine amply illustrates this. For instance, it took centuries before therapeutic blood letting and drastic purging were abandoned by the medical profession.

It is mainly in the last 35 to 40 years, determined efforts have been made to use scientific techniques to evaluate methods of treatment and prevention. An important advance in this field has been the development of an assessment method, known as Randomized Controlled Trial (RCT). It is really an epidemiologic experiment. Since its introduction, the RCT has questioned the validity of such widely used treatments as oral hypoglycaemic agents, varicose vein stripping, tonsillectomy, hospitalization of all patients with myocardial infarction, multiphasic screening, and toxicity and applicability of many preventive and therapeutic procedures.

The design of a randomized controlled trial is given in Fig. 9. For new programmes or new therapies, the RCT is the No.1 method of evaluation. The basic steps in conducting a RCT include the following:

1. Drawing up a protocol.
2. Selecting reference and experimental populations.
3. Randomization.
4. Manipulation or intervention.
5. Follow-up.
6. Assessment of outcome.
The study is conducted under a strict protocol. The protocol specifies the aims and objectives of the study, treatments to be applied when and where and how to allocate subjects into study and control groups, schedules as well as responsibilities of the parties involved in the trial, and questions to be answered, criteria for the selection of study and control groups, allocation of subjects into study and control groups, treatments to be applied — when and where and how to what kind of patients, standardization of working procedures and schedules as well as responsibilities of the parties involved in the trial, and the stage of evaluation of outcome of the study. A protocol is essential especially when a number of centres are participating in the trial.

Thus the reference population may comprise the population of a whole city, or a population of school children, industrial workers, obstetric population and so on according to the nature of the study.

(a) **Reference or target population**: The reference population is derived from the reference population. It is the actual population that participates in the experimental study. Ideally, it should be randomly chosen from the reference population, so that it has the same characteristics as the reference population. If the study population differs from the reference population, it may not be possible to generalize the findings of the study to the reference population.

When an experimental population has been defined, its members are invited to participate in the study. It is important to choose a stable population whose cooperation is assured to avoid losses to follow-up. The participants or volunteers must fulfill the following three criteria:

1. They must give “informed consent”, that is they must agree to participate in the trial after having been fully informed about the purpose, procedures and possible dangers solf the trial;
2. They should be representative of the population to which they belong (i.e., reference population); and
3. They should be qualified or eligible for the trial. That is, let us suppose, we are testing the effectiveness of a new drug for the treatment of anaemia. If the volunteers are not anaemic, we will then say, they are not eligible or qualified for the trial. Similarly, let us suppose; we are going to test the effectiveness of a new vaccine against whooping cough. If the volunteers are already immune to the disease in question, we will then say, they are not qualified for the trial. In other words, the participants must be fully susceptible to the disease under study.

It must be recognized that persons who agree to participate in a study are likely to differ from those who do not, in many ways that may affect the outcome under investigation.

**3. Randomization**

Randomization is a statistical procedure by which the participants are allocated into groups usually called “study” and “control” groups, to receive or not to receive an experimental preventive or therapeutic procedure, manoeuvre or intervention. Randomization is an attempt to eliminate “bias” and allow for comparability. Theoretically it is possible to assure comparability by matching. But when one matches, one can only match those factors which are known to be important. There may be other factors which are important but whose effect is not recognized or cannot be determined. By a process of randomization, hopefully these factors will be distributed equally between the two groups.

Randomization is the heart of a control trial. It will give the greatest confidence that the groups are comparable so that “like can be compared with like”. It ensures that the investigator has no control over allocation of participants to either study or control group, thus eliminating what is known as “selection bias”. In other words, by random allocation, every individual gets an equal chance of being allocated into either group or any of the trial groups.

It is crucial that both the groups should be alike with regard to certain variables or characteristics that might affect the outcome of the experiment (e.g., age, sex), the entire study population can be stratified into sub-groups according to the variable, and individuals within each sub-group can
then be randomly allocated into study and control groups. It is always desirable to check that the groups formed initially are basically similar in composition. Randomization is done only after the participant has entered the study, that is after having been qualified for the trial and has given his informed consent to participate in the study. Randomization is best done by using a table of random numbers (see chapter 18).

The essential difference between a randomized controlled trial and an analytical study is that in the latter, there is no randomization because a differentiation into diseased and non-diseased (exposed or non-exposed) groups has already taken place. The only option left to ensure comparability in analytical studies is by matching.

4. Manipulation

Having formed the study and control groups, the next step is to intervene or manipulate the study (experimental) group by the deliberate application or withdrawal or reduction of the suspected causal factor (e.g., this may be a drug, vaccine, dietary component, a habit, etc) as laid down in the protocol.

This manipulation creates an independent variable (e.g., drug, vaccine, a new procedure) whose effect is then determined by measurement of the final outcome, which constitutes the dependent variable (e.g., incidence of disease, survival time, recovery period).

5. Follow-up

This implies examination of the experimental and control group subjects at defined intervals of time, in a standard manner, with equal intensity, under the same given circumstances, in the same time frame till final assessment of outcome. The duration of the trial is usually based on the expectation that a significant difference (e.g., mortality) will be demonstrable at a given point in time after the start of the trial. Thus the follow-up may be short or may require many years depending upon the study undertaken.

It may be mentioned that some losses to follow-up are inevitable due to factors, such as death, migration and loss of interest. This is known as attrition. If the attrition is substantial, it may be difficult to generalise the results of the study to the reference population. Every effort, therefore, should be made to minimize the losses to follow-up.

6. Assessment

The final step is assessment of the outcome of the trial in terms of: (a) Positive results: that is, benefits of the experimental measure such as reduced incidence or severity of the disease, cost to the health service or other appropriate outcome in the study and control groups. (b) Negative results: that is, severity and frequency of side-effects and complications, if any, including death. Adverse effects may be missed if they are not sought.

The incidence of positive/negative results is rigorously compared in both the groups, and the differences, if any, are tested for statistical significance. Techniques are available for the analysis of data as they are collected (sequential analysis), but it is more useful to analyze the results at the end of the trial.

Bias may arise from errors of assessment of the outcome due to human element. These may be from three sources: First, there may be bias on the part of the participants, who may subjectively feel better or report improvement if they knew they were receiving a new form of treatment. This is known as "subject variation". Secondly there may be observer bias, that is the investigator measuring the outcome of a therapeutic trial may be influenced if he knows beforehand the particular procedure or therapy to which the patient has been subjected. This is known as "observer bias." Thirdly, there may be bias in evaluation – that is, the investigator may subconsciously give a favourable report of the outcome of the trial. Randomization cannot guard against these sorts of bias, nor the size of the sample. In order to reduce these problems, a technique known as "blinding" is adopted, which will ensure that the outcome is assessed objectively.

Blinding: Blinding can be done in three ways –
(a) SINGLE BLIND TRIAL: The trial is so planned that the participant is not aware whether he belongs to the study group or control group. (B) DOUBLE BLIND TRIAL: The trial is so planned that neither the doctor nor the participant is aware of the group allocation and the treatment received. (C) TRIPLE BLIND TRIAL: This goes one step further. The participant, the investigator and the person analyzing the data are all "blind". Ideally, of course, triple blinding should be used; but the double blinding is the most frequently used method when a blind trial is conducted (4). When an outcome such as death is being measured, blinding is not so essential.

SOME STUDY DESIGNS

It is useful to consider here some of the study designs of controlled trials:

1. Concurrent parallel study designs

In this situation (Fig.10–a), comparisons are made between two randomly assigned groups, one group exposed to specific treatment, and the other group not exposed. Patients remain in the study group or the control group for the duration of the investigation.

2. Cross-over type of study designs

This is illustrated in Fig. 10–b. With this type of study design, each patient serves as his own control. As before, the patients are randomly assigned to a study group and control group. The study group receives the treatment under consideration. The control group receives some alternate form of active treatment or placebo. The two groups are observed over time. Then the patients in each group are taken off their medication or placebo to allow for the elimination of the medication from the body and for the possibility of any "carry over" effects, as shown in Fig. 10–b by the diagonal lines. After this period of medication (the length of this interval is determined by the pharmacologic properties of the drug being tested), the two groups are switched. Those who received the treatment under study are changed to the control group therapy or placebo, and vice versa.

Cross-over studies offer a number of advantages. With such a design, all patients can be assured that sometime during the course of investigation, they will receive the new therapy. Such studies generally economize on the total number of patients required at the expense of the time necessary to complete the study. This method of study is not suitable if the drug of interest cures the disease, if the drug is effective only during a certain stage of the disease or if the disease changes radically during the period of time required for the study.
with evaluating therapeutic agents, mainly drugs. The last decades have seen clearly the utility of clinical trials. Some treatment/supplementation before conception to prevent cardiovascular mortality and beta-carotene on cancer acute phase of myocardial infarction of the recent examples include - evaluation of beta-blockers in reducing cardiovascular mortality in patient surviving the trial of tonsillectomy and adenoidectomy without its being the value of these procedures continues to be uncertain. recurrence of neural tube defects therapy, procedure or service is introduced. Many ethical, administrative and technical problems are involved in the conduct of clinical trials. Nevertheless, they are a powerful tool and should be carried out before any new therapy, procedure or service is introduced.

1. Clinical trials

For the most part, “clinical trials" have been concerned with evaluating therapeutic agents, mainly drugs. The last decades have seen clearly the utility of clinical trials. Some of the recent examples include - evaluation of beta-blockers in reducing cardiovascular mortality in patient surviving the acute phase of myocardial infarction (69); trials of folate treatment/supplementation before conception to prevent recurrence of neural tube defects (70); trials of aspirin on cardiovascular mortality and beta-carotene on cancer incidence; efficacy of tonsillectomy for recurrent throat infection (71); randomized controlled trial of coronary bypass surgery for the prevention of myocardial infarction (72), etc. The list is endless.

Unfortunately, not all clinical trials are susceptible to being blinded. For example, there is no way to perform a clinical trial of tonsillectomy and adenoidectomy without its being obvious who received surgery and who did not, a reason why the value of these procedures continues to be uncertain. Many ethical, administrative and technical problems are involved in the conduct of clinical trials. Nevertheless, they are a powerful tool and should be carried out before any new therapy, procedure or service is introduced.

2. Preventive trials

In general usage, prevention is synonymous with primary prevention, and the term “preventive trials" implies trials of primary preventive measures. These trials are purported to prevent or eliminate disease on an experimental basis. The most frequently occurring type of preventive trials are the trials of vaccines and chemoprophylactic drugs. The basic principles of experimental design are also applicable to these trials. It may be necessary to apply the trial to groups of subjects instead of to individual subjects. For example, in 1946, the Medical Research Council of UK conducted an extensive trial (74) to test whooping cough vaccine from three manufacturers in ten separate field trials. Those children between 6–18 months who were entered into the trial were randomly allocated in study and control groups. The vaccine was given in three, monthly injections, and the children were followed up at monthly intervals to detect the occurrence of whooping cough. The study group comprised of 3,801 children who were vaccinated, and 149 developed whooping cough. The control group consisted of 3,757 unvaccinated children, and 687 of them developed the infection. This gave an attack rate of 1.45 per 1000 child months in the vaccinated group and 6.72 per 1000 child months in the control group. The difference was significant.

Analysis of a preventive trial must result in a clear statement about (a) the benefit the community will derive from the measure (b) the risks involved, and (c) the costs to the health service in terms of money, men and material resources (21). Since preventive trials involve larger number of subjects and sometimes a longer time span to obtain results, there may be greater number of practical problems in their organisation and execution.

3. Risk factor trials

A type of preventive trial is the trial of risk factors in which the investigator intervenes to interrupt the usual sequence in the development of disease for those individuals who have “risk factor” for developing the disease; often this involves risk factor modification. The concept of “risk factor" gave a new dimension to epidemiological research.

For example, the major risk factors of coronary heart disease are elevated blood cholesterol, smoking, hypertension and sedentary habits. Accordingly, the four main possibilities of intervention in coronary heart disease are : reduction of blood cholesterol, the cessation of smoking, control of hypertension and promotion of regular physical activity. Risk factor trials can be “single-factor" or “multi-factor" trials. Both the approaches are complementary, and both are needed.

The WHO (75) promoted a trial on primary prevention of coronary heart disease using clofibrate to lower serum cholesterol, which was accepted as a significant risk factor.
for CHD. This study is the largest preventive trial yet conducted, comprising more than 15,000 men of whom one-third received clofibrate and two-third received olive oil as a control treatment. The study was conducted in 3 centres in Europe (Edinburgh, Prague, and Budapest). The design was double-blind and randomization was successfully achieved. The mean observation was 9.6 years. The trial showed a significant reduction in non-fatal cardiac infarction, but unfortunately, there were 25 per cent more deaths in the clofibrate-treated group than in the control group possibly due to long-term toxic effect of the drug. The trial illustrates the kind of contribution that an epidemiological approach can make to protect the public health against possible adverse effects of long-term medication with potent drugs (75).

The other widely reported risk-factor intervention trials in coronary heart disease are: (a) The Stanford Three Community Study (b) The North Karelia Project in Finland (c) The Oslo Study, and (d) The Multiple Risk Factor Intervention Trial (MRFIT) in USA.

4. Cessation experiments

Another type of preventive trial is the cessation experiment. In this type of study, an attempt is made to evaluate the termination of a habit (or removal of suspected agent) which is considered to be causally related to a disease. If such action is followed by a significant reduction in the disease, the hypothesis of cause is greatly strengthened. The familiar example is cigarette smoking and lung cancer. If in a randomized controlled trial, one group of cigarette smokers continue to smoke and the other group has given up, the demonstration of a decrease in the incidence of lung cancer in the study group greatly strengthens the hypothesis of a causal relationship. A large randomized controlled trial has been mounted to study the role of smoking cessation in the primary prevention of coronary heart disease (76).

5. Trial of aetiological agents

One of the aims of experimental epidemiology is to confirm or refute an aetiiological hypothesis. The best known example of trial of an aetiological agent relates to retrolental fibroplasia (RLF). Retrolental fibroplasia, as a cause of blindness, was non-existent prior to 1938. It was originally observed and reported by T.L. Terry, a Boston ophthalmologist in 1942 (77), and later in many other countries outside the USA.

RLF was recognized as a leading cause of blindness by descriptive studies which showed that beginning in about 1940–1941, the incidence of the disease increased at an alarming rate (Fig. 11), and that this previously unknown disease was occurring only in premature babies. Analytical studies demonstrated its close association with administration of oxygen to premature babies. A large randomized controlled trial was mounted involving 18 hospitals in United States by Kinsey and Hemphill (78, 79) in which premature babies with birth weight of 1500 gram or less were allocated into experimental and control groups. In the experimental group, all the babies received 50 per cent oxygen therapy for 28 days, while in the control group (“curtailed oxygen group”) oxygen was used only for clinical emergency. It was later found that all of the babies in the “curtailed oxygen group” who developed RLF had received some oxygen. There were no cases among those who received none, confirming the aetiological hypothesis.

The dramatic rise and fall in frequency of RLF can be seen in Fig. 11. It will be noted that RLF reached its peak during the years 1952–53. The sharp drop in the graph after 1953 highlights the results of the decreased use of oxygen. RLF illustrates one of the problems often introduced by technological or scientific advances.

Since most diseases are fatal, disabling or unpleasant, human experiments to confirm an aetiological hypothesis are rarely possible.

6. Evaluation of health services

Randomized controlled trials have been extended to assess the effectiveness and efficiency of health services. Often, choices have to be made between alternative policies of health care delivery. The necessity of choice arises from the fact that resources are limited, and priorities must be set for the implementation of a large number of activities which could contribute to the welfare of the society. An excellent example of such an evaluation is the controlled trials in the chemotherapy of tuberculosis in India, which demonstrated that “domiciliary treatment” of pulmonary tuberculosis was as effective as the more costly “hospital or sanatorium” treatment. The results of the study have gained international acceptance and ushered in a new era – the era of domiciliary treatment, in the treatment of tuberculosis.

More recently, multiphasic screening which has achieved great popularity in some countries, was evaluated by a randomized controlled trial in South-East London. The study led to the withholding of vast outlay of resources required to mount a national programme of multiphasic screening in UK (80,81). Another example is that related to studies which have shown that many of the health care delivery tasks traditionally performed by physicians can be performed by nurses and other paramedical workers, thus saving physician time (82). These studies are also labelled as “health services research” studies.

NON-RANDOMIZED TRIALS

Although the experimental method is almost always to be preferred, it is not always possible for ethical, administrative and other reasons to resort to a randomized controlled trial in human beings. For example, smoking and lung cancer and induction of cancer by viruses have not lent themselves
to direct experimentation in human beings. Secondly, some preventive measures can be applied only to groups or on a community-wide basis (e.g., community trials of water fluoridation). Thirdly, when disease frequency is low and the natural history long (e.g., cancer cervix) randomized controlled trials require follow-up of thousands of people for a decade or more. The cost and logistics are often prohibitive. These trials are rare. In such situations, we must depend upon other study designs – these are referred to as non-randomized (or non-experimental) trials.

Where the approach is sophisticated in randomized controlled trials, it is rather crude in non-randomized trials. As there is no randomization in non-experimental trials, the degree of comparability will be low and the chances of a spurious result higher than where randomization had taken place. In other words, the validity of causal inference remains largely a matter of extra-statistical judgement. Nevertheless, vital decisions affecting public health and preventive medicine have been made by non-experimental studies. A few examples of non-randomized trials are discussed below:

1. Uncontrolled trials

There is room for uncontrolled trials (i.e., trials with no comparison group). For example, there were no randomized controlled studies of the benefits of the Pap test (cervical cancer) when it was introduced in 1920s. Today, there is indirect epidemiological evidence from well over a dozen uncontrolled studies of cervical cancer screening that the Pap test is effective in reducing mortality from this disease. Initially uncontrolled trials may be useful in evaluating whether a specific therapy appears to have any value in a particular disease, to determine an appropriate dose, to investigate adverse reactions, etc. However, even in these uncontrolled trials, one is using implied “historical controls”, i.e., the experience of earlier untreated patients affected by the same disease.

Since most therapeutic trials deal with drugs which do not produce such remarkably beneficial results, it is becoming increasingly common to employ the procedures of a double-blind controlled clinical trial in which the effects of a new drug are compared to some concurrent experience (either placebo or a currently utilized therapy).

2. Natural experiments

Where experimental studies are not possible in human populations, the epidemiologist seeks to identify “natural circumstances” that mimic an experiment. For example, in respect of cigarette smoking, people have separated themselves “naturally” into two groups, smokers and non-smokers. Epidemiologists have taken advantage of this separation and tested hypothesis regarding lung cancer and cigarette smoking. Other populations involved in natural experiments comprise the following groups: (a) migrants (b) religious or social groups (c) atomic bombing of Japan (d) famines (e) earthquakes, etc. A major earthquake in Athens in 1981 provided a “natural experiment” to epidemiologists who studied the effects of acute stress on cardiovascular mortality. They showed an excess of deaths from cardiac and external causes on the days after the major earthquake, but no excess deaths from other causes (93).

John Snow’s discovery that cholera is a water-borne disease was the outcome of a natural experiment. Snow in his “grand experiment” identified two randomly mixed populations, alike in other important respects, except the source of water supply in their households. The results of the experiment are given in Table 23.

<table>
<thead>
<tr>
<th>Sources of water supply</th>
<th>Number of houses</th>
<th>Deaths from cholera</th>
<th>Deaths in each 10,000 houses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southwark &amp; Vauxhall Co.</td>
<td>40,046</td>
<td>1263</td>
<td>315</td>
</tr>
<tr>
<td>Lambeth Co.</td>
<td>26,107</td>
<td>98</td>
<td>37</td>
</tr>
</tbody>
</table>

It will be seen from Table 23 that deaths were fewer in houses supplied by Lambeth company compared to houses supplied by Southwark and Vauxhall company. The inference was obvious – the Lambeth company water came from an intake on the River Thames well above London, whereas the Southwark and Vauxhall company water was derived from the sewage polluted water basin. The great difference in the occurrence of cholera among these two populations gave clear demonstration that cholera is a water-borne disease. This was demonstrated long before the advent of the bacteriological era; it also led to the institution of public health measures to control cholera.

3. Before and after comparison studies

These are community trials which fall into two distinct groups:

A. Before and after comparison studies without control, and
B. Before and after comparison studies with control.

A. Before and after comparison studies without control

These studies centre round comparing the incidence of disease before and after introduction of a preventive measure. The events which took place prior to the use of the new treatment or preventive procedure are used as a standard for comparison. In other words, the experiment serves as its own control; this eliminates virtually all group differences. The classic examples of “before and after comparison studies” were the prevention of scurvy among sailors by James Lind in 1750 by providing fresh fruit; studies on the transmission of cholera by John Snow in 1854; and later, prevention of polio by Salk and Sabin vaccines.

In order to establish evidence in before and after comparison studies, the following are needed: (a) data regarding the incidence of disease, before and after introduction of a preventive measure must be available (b) there should be introduction or manipulation of only one factor or change relevant to the situation, other factors remaining the same, as for example, addition of fluorine to drinking water to prevent dental caries (c) diagnostic criteria of the disease should remain the same (d) adoption of preventive measures should be over a wide area (e) reduction in the incidence must be large following the introduction of the preventive measure, because there is no control, and (f) several trials may be needed before the evaluation is considered conclusive.

Table 24 gives an example of a “before and after comparison study” in Victoria (Australia) following introduction of seat-belt legislation for prevention of deaths and injuries caused by motor vehicle accidents.
B. Before and after comparison studies with control

In the absence of a control group, comparison between observations before and after the use of a new treatment or procedure may be misleading. In such situations, the epidemiologist tries to utilize a "natural" control group i.e., the one provided by nature or natural circumstances. If preventive programme is to be applied to an entire community, he would select another community as similar as possible, particularly with respect to frequency and characteristics of the disease to be prevented. One of them is arbitrarily chosen to provide the study group and the other a control group. In the example cited (e.g., seat-belt legislation in Victoria, Australia), a natural "control" was sought by comparing the results in Victoria with other states in Australia where similar legislation was not introduced. The findings are given in Table 25.

TABLE 25
Effect of adoption of compulsory seat-belt legislation in Victoria, 1971 compared with other states where similar legislation was not introduced

<table>
<thead>
<tr>
<th></th>
<th>1970</th>
<th>1971</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths</td>
<td>564</td>
<td>464</td>
<td>-17.7</td>
</tr>
<tr>
<td>Injuries</td>
<td>14620</td>
<td>12454</td>
<td>-14.8</td>
</tr>
</tbody>
</table>

Table 24 shows a definite fall in the numbers of deaths and injuries in occupants of cars after the introduction of compulsory seat-belts in one state of Australia.

ASSOCIATION AND CAUSATION

Descriptive studies help in the identification of the disease problem in the community; and by relating disease to host, agent and environmental factors, it endeavours to suggest an aetiological hypothesis. Analytical and experimental studies test the hypotheses derived from descriptive studies and confirm or refute the observed association between suspected causes and disease. When the disease is multifactorial (e.g., coronary heart disease) numerous factors or variables become implicated in the web of causation, and the notion of "cause" becomes confused. The more associations, the more investigations to disentangle the web of causation. The epidemiologist whose primary interest is to establish a "cause and effect" relationship has to sift the husk from the grain. He proceeds from demonstration of statistical association to demonstration that the association is causal.

The terms "association" and "relationship" are often used interchangeably. Association may be defined as the concurrence of two variables more often than would be expected by chance. In other words, events are said to be associated when they occur more frequently together than one would expect by chance (2). Association does not necessarily imply a causal relationship.

It will be useful to consider here the concept of correlation. Correlation indicates the degree of association between two characteristics. The correlation coefficients range from -1.0 to +1.0. A correlation coefficient of 1.0 means that the two variables exhibit a perfect linear relationship. However, correlation cannot be used to invoke causation, because the sequence of exposure preceding disease (temporal association) cannot be assumed to have occurred. Secondly, correlation does not measure risk. It may be said that causation implies correlation, but correlation does not imply causation.

Association can be broadly grouped under three headings:

a. Spurious association
b. Indirect association
c. Direct (causal) association
   (i) one-to-one causal association
   (ii) multifactorial causation.

a. Spurious association

Sometimes an observed association between a disease and suspected factor may not be real. For example, a study in UK of 5174 births at home and 11,156 births in hospitals showed perinatal mortality rates of 5.4 per 1000 in the home births, and 27.8 per 1000 in the hospital births (84). Apparently, the perinatal mortality was higher in hospital births than in the home births. It might be concluded that homes are a safer place for delivery of births than hospitals. Such a conclusion is spurious or artifactual, because in general, hospitals attract women at high risk for delivery because of their special equipment and expertise, whereas this is not the case with home deliveries. The high perinatal mortality rate in hospitals might be due to this fact alone, and not because the quality of care was inferior. There might be other factors also such as differences in age, parity, prenatal care, home circumstances, general health and disease state between the study and control groups. This type of bias where "like" is not compared with "like" (selection bias) is very important in epidemiological studies. It may lead to a spurious association or an association when none actually existed.