describing variations ... 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 :

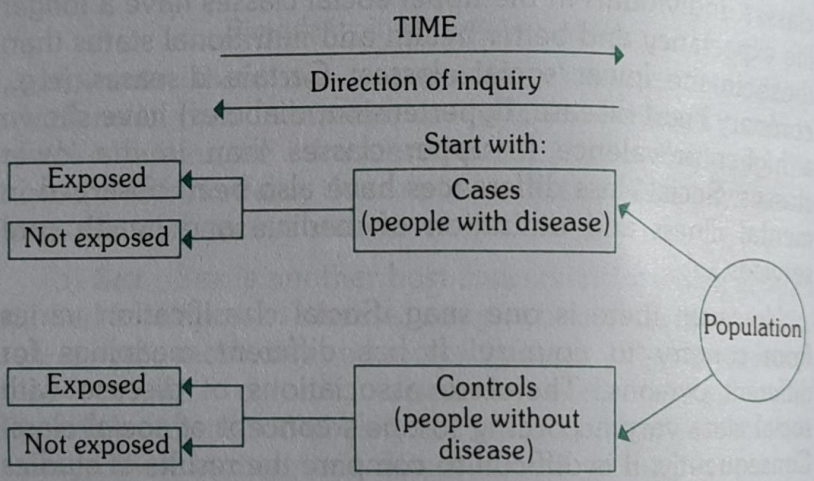
a. case control study
b. cohort study.

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**




**Design of a Cohort Study**

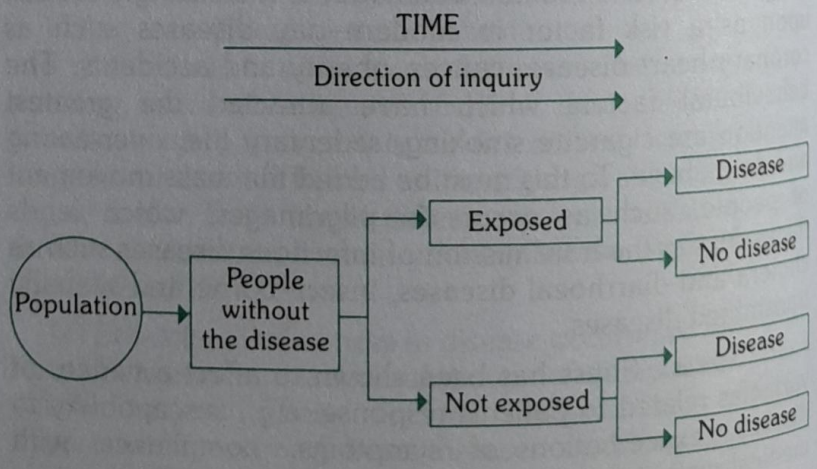



FIG. 8
Schematic diagram of the design of case control and cohort studies

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 2×2 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), an association is said to exist between smoking and lung cancer. Case control studies have their major use in the chronic disease problem when the causal pathway may span many decades.

**TABLE 11**

Framework of a case control study
(The 2 × 2 contingency table)

| Suspected or risk factors | Cases (Disease present) | Control (Disease absent) |
|---|---|---|
| Present | a | b |
| Absent | c | d |
| | a+c | b+d |

### 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 : (i) 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 histologically the same. Once the diagnostic criteria are established, they should not be altered or changed till the study is over. (ii) 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 (i) hospitals, or (ii) general population. (i) 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. (ii) 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. Difficulties may arise in the selection of controls 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 *(9)*.

*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 illneses 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.

How many controls are needed ? If many cases are available, and large study is contemplated, and if the cost to collect case and control is about equal, then one tends to use one control for each case. If the study group is small (say under 50) as many as 2,3, or even 4 controls can be selected for each study subject.

To sum up, selection of proper cases and controls is crucial to the interpretation of the results of case control studies. Some investigators select cases from one source and controls from more than one source to avoid the influence of "selection bias". Such studies are recommended by epidemiologists. It is also desired to conduct more than one case control study, preferably in different geographic areas. If the findings are consistent, it serves to increase the validity (i.e., accuracy) of the inferences. Failure to select comparable controls can introduce "bias" into results of case control studies and decrease the confidence one can place in the findings.

## 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

| | Cases (with lung cancer) | Controls (without lung cancer) | Total |
|---|---|---|---|
| Smokers (less than 5 cigarettes a day) | 33 (a) | 55 (b) | 88 (a+b) |
| Non-smokers | 2 (c) | 27 (d) | 29 (c+d) |
| Total | 35 (a+c) | 82 (b+d) | n = a+b +c+d |

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:

Incidence among exposed

$$= \frac{a}{(a+b)} + \frac{c}{(c+d)}$$

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 :

| | Diseases | |
|---|---|---|
| | *Yes* | *No* |
| Exposed | a | b |
| Not exposed | c | d |

Odds ratio = ad/bc

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

$$\text{Odds ratio} = \left(\frac{a}{b}\right)/\left(\frac{c}{d}\right) = \frac{ad}{bc}$$

$$= \frac{33 \times 27}{55 \times 2} = 8.1$$

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 Berkesonian bias, termed

after Dr. Joseph Berkeson 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 13**

Advantages and disadvantages of case control studies

| ADVANTAGES |
|---|
| 1. Relatively easy to carry out. |
| 2. Rapid and inexpensive (compared with cohort studies). |
| 3. Require comparatively few subjects. |
| 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). |
| 5. No risk to subjects. |
| 6. Allows the study of several different aetiological factors (e.g., smoking, physical activity and personality characteristics in myocardial infarction). |
| 7. Risk factors can be identified. Rational prevention and control programmes can be established. |
| 8. No attrition problems, because case control studies do not require follow-up of individuals into the future. |
| 9. Ethical problems minimal. |
| **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. |

Source : *(35,36)*

## 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 14**

Association between maternal DES therapy and adenocarcinoma of vagina amongst female offspring

| Information acquired retrospectively | Cases (8) | Controls (32) | Significance level |
|---|---|---|---|
| Maternal age | 26.1 | 29.3 | n.s. |
| Maternal smoking | 7 | 21 | n.s. |
| Antenatal radiology | 1 | 4 | n.s. |
| Oestrogen exposure | 7 | – | P<0.00001 |

Source : *(26)*

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 users to 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.

**TABLE 15**

Case control studies on the safety of oral contraceptives

| | No. | Per cent who used oral contraceptives |
|---|---|---|
| Cases (venous thrombosis and pulmonary embolism) | 84 | 50 |
| Controls | 168 | 14 |

Source : (46,47)

*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 are identified prior to the appearance of the disease under investigation
b. the study groups, so defined, are observed over a period of time to determine the frequency of disease among them
c. the study proceeds forward from cause to effect.

### Concept of cohort

In epidemiology, the term "cohort" is defined as a group of people who share a common characteristic or experience within a defined time period (e.g., age, occupation, exposure to a drug or vaccine, pregnancy, insured persons, etc). Thus a group of people born on the same day or in the same period of time (usually a year) form a "birth cohort". All those born in 2010 form the birth cohort of 2010. Persons exposed to a common drug, vaccine or infection within a defined period constitute an "exposure cohort". A group of males or females married on the same day or in the same period of time form a "marriage cohort". A cohort might be all those who survived a myocardial infarction in one particular year.

The comparison group may be the general population from which the cohort is drawn, or it may be another cohort of persons thought to have had little or no exposure to the substance in question, but otherwise similar.

### Indications for cohort studies

Cohort studies are indicated : (a) when there is good evidence of an association between exposure and disease, as derived from clinical observations and supported by descriptive and case control studies (b) when exposure is rare, but the incidence of disease high among exposed, e.g., special exposure groups like those in industries, exposure to X-rays, etc (c) when attrition of study population can be minimized, e.g., follow-up is easy, cohort is stable, co-operative and easily accessible, and (d) when ample funds are available.

### Framework of a cohort study

In contrast to case control studies which proceed from "effect to cause", the basic approach in cohort studies is to work from "cause to effect" (Fig. 8). That is, in a case control study, exposure and disease have already occurred when the study is initiated. In a cohort study, the exposure has occurred, but the disease has not.

The basic design of a simple cohort study is shown in Table 16. We begin with a group or cohort (a+b) exposed to a particular factor thought to be related to disease occurrence, and a group (c+d) not exposed to that particular factor. The former is known as "study cohort", and the latter "control cohort".

**TABLE 16**

Framework of a cohort study

| Cohort | Disease | | Total |
|---|---|---|---|
| | yes | no | |
| Exposed to putative aetiologic factor | a | b | a + b |
| Not exposed to putative aetiologic factor | c | d | c + d |

In assembling cohorts, the following general considerations are taken into account :

a. The cohorts must be free from the disease under study. Thus, if the disease under study is coronary heart disease, the cohort members are first examined and those who already have evidence of the disease under investigation are excluded.
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, o efficiently reflect any difference in disease occurrenc (for example, males over 35 years would b appropriate for studies on lung cancer).
c. Both the groups should be comparable in respect of a the possible variables, which may influence th frequency of the disease; and
d. The diagnostic and eligibility criteria of the disea must be defined beforehand; this will depend up the availability of reliable methods for recognizing disease when it develops.

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

Table 16 shows (a+b) persons were exposed to the fa under study, 'a' of which developed the disease durin follow-up period; (c+d) persons were not exposed, which became cases (it is assumed for simplici presentation that there were no intermittent deaths or during the follow-up period). After the end of the follo

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, and
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 leukaemia or aplastic anaemia between 1935 and 1954. They found that the death rate from leukaemia 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 17**

Age standardized death rates per 100,000 men per year by amount of current smoking

| Classification of exposure (cigarettes) | No. of deaths | Death rate |
|---|---|---|
| 1/2 pack | 24 | 95.2 |
| 1/2–1 pack | 84 | 107.8 |
| 1–2 packs | 90 | 229.2 |
| 2 packs + | 97 | 264.2 |

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 hom visits – preferably all three on an annual basis.

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

However, inspiteof best efforts, a certain percentage losses to follow-up are inevitable due to death, change residence, migration or withdrawal of occupation. The losses may bias the results. It is, therefore, necessary to bu into the study design a system for obtaining ba