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, 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, 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 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

Contingency table applied to hypothetical cigarette smoking and lung cancer example

Cigarette smoking	Developed lung cancer	Did not develop lung cancer	Total
Yes	70 (a)	6930 (b)	7000 (a + b)
No	3 (c)	2997 (d)	3000 (c + d)

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)

$$\text{RR of lung cancer} = \frac{10}{1} = 10$$

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 :

$$= \frac{\text{Incidence of disease rate among exposed minus–incidence of disease rate among non-exposed}}{\text{Incidence rate among exposed}} \times 100$$

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

Lung cancer death rates among smokers and non-smokers : UK physicians

Deaths per 100,000 person-years		
Heavy smokers	224	Exposed to suspected factor (a)
Non-smokers	10	Non-exposed to suspected causal factor (b)
Deaths in total population	74 (c)	
Individual RR	$a/b = \frac{224}{10} = 22.40$	
Population AR	(c–b)/c = 86 per cent	

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

Cardiovascular risk 100,000 patient years	Age	
	30–39	40–44
Relative risk	2.8	2.8
Attributable risk	3.5	20.0

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

Cause of death	Death rate/1000		RR	AR (%)
	Smokers	Non-smokers		
Lung cancer	0.90	0.07	12.86	92.2
CHD	4.87	4.22	1.15	13.3

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 or factors possibly related to disease outcome.

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

TABLE 22

Main differences between case control and cohort studies

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