

Chapter 7

Types of Studies - III: Cohort Studies

The term cohort is used to describe a group of people who have something in common at the time of enrollment into the study, and who are then observed for a period of time to see what happens to them. At entry, subjects are classified according to those characteristics which might be related to outcome. They are then observed over time to see who develops the outcome.

Some examples of the criteria under which cohorts may be assembled and used are explained in table 7.1 below:

Table 7.1 Examples of types of cohorts

Characteristics in common	To assess effects of	Example
Age	Age	Life expectancy of people aged 70
Birth date	Calendar time	Tuberculosis rate for people born in a given year
Exposure	Risk factor	Lung cancer in smokers
Disease	Prognosis	Survival time for patients with breast cancer
Preventive intervention	Prevention	Reduction in incidence of measles after immunization
Therapeutic intervention	Treatment	Improvement in survival in patients treated with different regimens

Cohort studies are less prone to selection bias compared to case-control studies. On the other hand they have a potential for bias from another source. This is from the high likelihood of drop-outs

due to outmigration, change of address, death and similar other losses to follow-up during the prolonged periods of observation.

Cohort studies serve several purposes. They are primarily used for studying outcomes like cure, disability or death. They are more useful for testing rather than generating hypotheses. Large multipurpose studies can have a hypothesis generating role, for subsequent testing in smaller more detailed studies.

There are three ways in which cohort studies are usually organized (see figure 7.1):

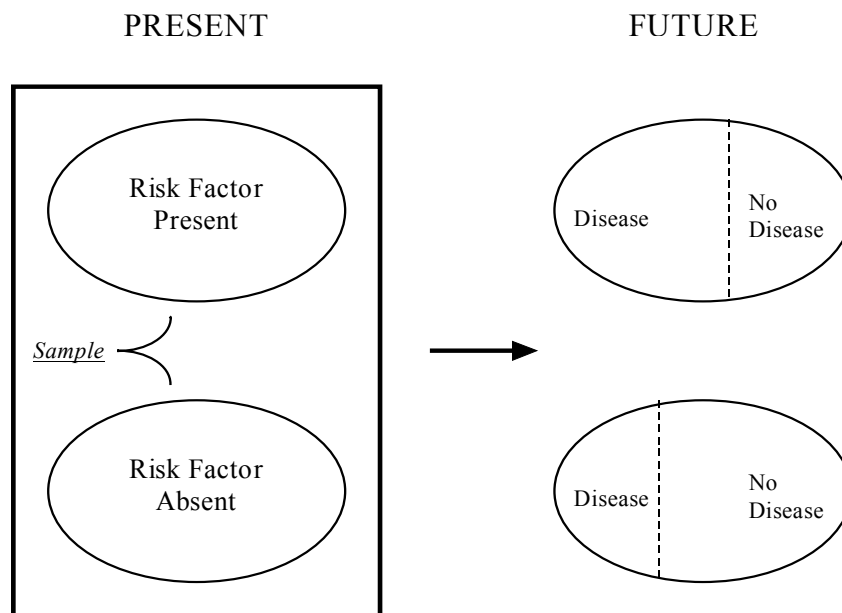


Figure 7.1: Prospective Cohort Study.

Concurrent Cohort Study. The cohort is assembled now (i.e. in the present), and followed into the future. Data collection is more satisfactory in the concurrent cohort study. Such studies are often called Prospective Cohort Studies.

Historical Cohort Study. The cohort is identified from past records and followed forward from that time to present. A good example is the Dutch famine study. During the Second World War a period of acute food shortage occurred in Holland. All the women who were pregnant at the time were identified. When their infants were born anthropometric measurements were made on the infants to see the effects

of food shortage in the mother occurring at different stages of pregnancy. The babies form a cohort which has been followed into adulthood to study the long term effects of food deprivation during foetal life. In the case of historical cohorts the data has to be sufficiently robust for obtaining meaningful answers.

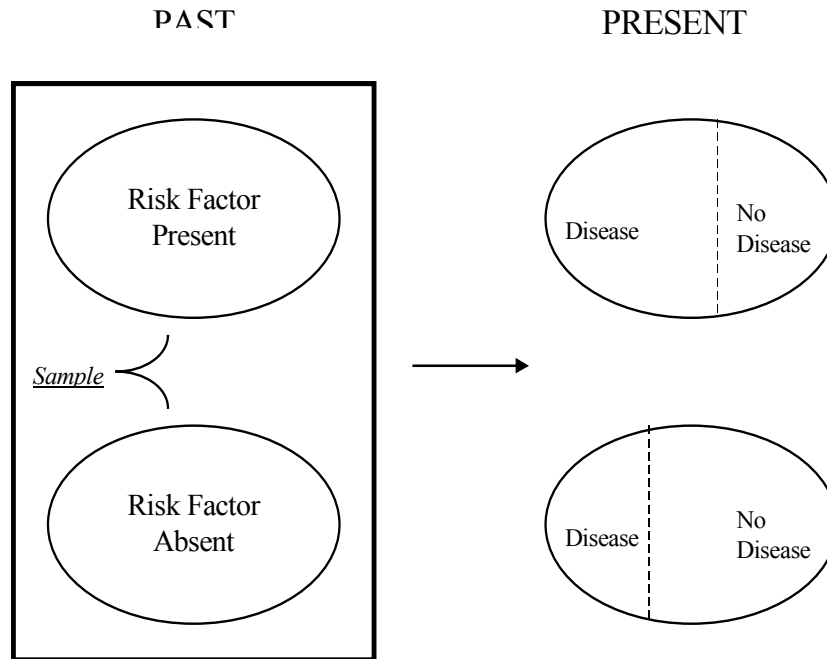


Figure 7.2: Retrospective Cohort Study.

Retrospective Cohort Study. The relevant events viz. exposure and outcome have already taken place. This type of study is in contrast to prospective cohort studies where the relevant exposure may or may not have occurred at the time of commencement, but the outcomes have certainly not yet occurred.

Cohort studies can be distinguished from case-control studies by the important characteristic that they classify subjects by presence or absence of exposure whereas in case-control studies classification is by outcome.

In a given situation the selection of case-control or cohort design will depend on the research question being asked (hypothesis being tested), the resources available and the current state of

knowledge. The objective is to choose a design that would yield the most reliable and informative answer.

In the case of cohort studies the choice of a prospective or retrospective design depends on the research question as well as logistical considerations. Retrospective designs can be conducted relatively quickly and cheaply because all relevant events have already happened. They are commonly employed for studying diseases with long latency periods for outcomes to happen. The drawback is that relevant information about exposure may not be very reliable because it was recorded in the past. Similarly, information about confounders may not be available because they were not thought of at the time of recording the information. These drawbacks are overcome in prospective studies, but the cost is also proportionately greater.

Two criteria must be fulfilled in conducting any type of a cohort study:

- Observation must take place over a meaningful period of time so that the outcome under study has a reasonable time to occur. One should also bear in mind that the longer the period of follow-up the greater is the potential for drop-outs to occur.
- All members of the cohort should be observed over the full period of follow-up. Drop-outs can distort the study, especially if the reasons for dropping out are related to the outcome.

Advantages and Disadvantages of Cohort studies

Advantages:

1. If the study groups have been properly selected they allow direct calculation of incidence or morbidity rates rather than an odds ratio. Cohort designs are a powerful strategy for measuring incidence.
2. They allow the most unambiguous determination of a temporal sequence. A potential cause gets measured before the outcome, and so there is less likelihood of bias. Moreover, all important variables can be measured carefully instead of depending on memory.
3. A study can be designed such that several different kinds of outcomes can be determined at the same time e.g. anxious attachment in infancy and educational, social and emotional outcomes in late childhood.

Cohort studies are the best available observational studies as substitutes for true experimental studies.

Disadvantages:

1. The main disadvantage is that if the outcome is infrequent, a large number of people must enter the study and be followed for a long time before results become available. This has implications for costs with regard to staff, transportation and time. Cohort studies are more efficient for outcomes which are common, and not so good for rare diseases.

Cohort studies work best when the exposure factor is relatively rare but has a high attack rate e.g. a study of the sequelae of serious head injury. A large number of controls without such a history can be assembled easily; on the other hand those with serious head injury are very likely to show some sequelae in a reasonably short time. Non-concurrent cohort studies avoid some of the constraints in tracking subjects. On the other hand there is a likelihood of bias similar to case-control studies in eliciting past exposure.

2. The second disadvantage is that the condition being studied may be present at the beginning but asymptomatic and goes clinically undiagnosed.

Steps in Planning a Cohort Study

- Assemble the initial cohort with the defined baseline characteristics.
- Devise a scheme for tracking the cohort's members, making sure that no biases are operating that could influence the determination of exposure, and that exposure did take place. A major difficulty arises from the fact that subjects have freedom of movement and of lifestyle. A great deal of time and effort gets expended in keeping track of them.
- Develop objective outcome criteria.
- Develop an unbiased method of ascertaining outcome status.
- Determine ways of measuring other factors (confounders) that may influence the outcome.

Selection of Subjects

The selection of subjects is determined by the frequency of exposure and the need to obtain complete and reliable information about both exposure and outcome on all the study subjects. For relatively common exposures like cigarette smoking, alcohol consumption, bottle feeding and so on, a sufficiently large number of subjects can be enrolled from a variety of sources. In the case of rare exposures, special groups may need to be considered, for example, exposure to lead or to radiation or industrial toxins, and so on. Such special exposure cohorts allow the evaluation of rare conditions.

Sometimes subjects are chosen because of the ease with which reliable and detailed information about exposure and outcome data can be collected. Professions like nurses, doctors, teachers, civil servants, workers in different industries, and similar well defined groups make ideal subjects for cohort study designs.

Selection of Comparison Groups

A careful choice is as important as in the case of case-control studies, the main principle being that the two groups should be as similar as possible in all respects except for the exposure factor being studied. If there is no association between exposure and the disease then the disease rates between the two populations would be similar.

If there is one single cohort initially being recruited then its members can be categorized by degree of exposure and comparison be made of disease rates between different categories of exposure.

The rates of mortality and disease occurrence in the general population may be used for comparison with the corresponding rates in a special cohort e.g. rates of a disease in the general population and amongst workers in a chemical plant. Such a comparison will underestimate the true association between exposure and disease because in such cases a part of the general population is in fact exposed. There are also other issues which muddy the results. The general population happens to contain people who are ill or disabled and therefore unable to work, and disease rates among them are generally higher than in the working population. Moreover, other factors like lifestyle, smoking, diet, housing, and so on may be different between the groups being compared. To overcome these difficulties a comparison cohort is selected from the general population which is closely matched with the study cohort except for the exposure factor. Sometimes several comparison cohorts are enrolled. If study results show similar levels of association between exposure and disease for several comparison groups the case for validity is that much stronger.

In all cohort studies followup of cohort members is the main challenge. The larger the number and the longer the period of follow-up the greater are the chances of drop-outs.

Information about Exposure

The use of pre-existing records saves time and cost. But the amount of detail may not be sufficient for the study. Nor are the details of possible confounding factors likely to be adequately recorded. In most situations details obtained from pre-existing records would need to be supplemented by interviews, questionnaires, physical examination and laboratory tests. The information gathered at the time of enrollment about changes in employment, work patterns, life-style, diet and so on, needs to be updated at each follow-up.

Information about Outcome

In all cohort studies the goal is to obtain complete and unbiased information about the outcome in every subject. Hospital records, physical examination, and laboratory tests at end-point, questionnaires and interviews are the common methods of recording outcomes. The important issue is

that whatever the procedures used for recording outcome they should be applied equally to all the groups, exposed and unexposed.

Analysis of Data

Comparing Risks

Usually the researcher wishes to compare the incidence of disease in two or more cohorts which have different exposure to some possible risk factor. In order to compare risks several measures of the relationship between exposure and disease are used.

These measures of effect are as follows:

Attributable Risk. (AR). AR is the risk difference. This measure is the answer to the question "What is the additional risk (i.e. incidence of disease) following exposure over and above that experienced by people who are not exposed?"

$$\text{AR} = \text{Incidence in exposed} - \text{Incidence in not exposed.}$$

Relative Risk (RR). RR is the risk ratio This measure is the answer to the question "How many times more likely are exposed persons to get the disease relative to non-exposed persons?". It tells us nothing about the magnitude of absolute risk. In some cases even for a large RR the value of absolute risk may be very small in the case of uncommon diseases.

$$\text{RR} = \text{Incidence in exposed} \div \text{Incidence in non-exposed.}$$

In most clinical situations AR is a more meaningful measure of risk since it represents the actual additional probability of the disease in those exposed.

Population Attributable Risk. This measure is the answer to the question "How much does a risk factor contribute to the overall rates of disease in a population?". Such information helps to decide which risk factors are particularly important and which are trivial from the public health point of view.

$$\text{Population Attributable Risk} = \text{AR} \times \text{Prevalence of risk factor in the population under study.}$$

Population Attributable Risk takes into account the prevalence of a risk factor in a given community. A relatively weak risk factor which is quite prevalent in a community could account for more of the overall incidence of disease than a stronger risk factor which is rare. Conversely, a risk factor may be strongly related to a disease but may contribute less to the problem of that disease in the population if its prevalence is low. Population attributable risk is a measure of the excess incidence of disease in a community associated with the occurrence of a risk factor.

Population Attributable Fraction is the fraction of disease occurrence in a population that is associated with a particular risk factor.

Population Attributable Fraction = Population Attributable Risk ÷ Total incidence of disease in the population

Variations in Cohort. Study Designs

1). Retrospective Cohort Studies.

A cohort is assembled. Medical records are studied to obtain data about base-line measurements and personal characteristics. The cohort is followed up for observing outcome. For example, a cohort of children less than 3 years old and presenting with asthma and bronchitis is assembled. Data about feeding in infancy, family history of atopy, and any serious respiratory illness in the past are obtained. The cohort is then observed for studying the outcome (e.g. spontaneous remission).

The advantages are:

- Predictor variables preceded the outcome, and observations are made before the outcome is known. Hence there is less likelihood of bias.
- Cost is reduced since the subjects are already assembled, and base-line measurements are all made.
- All with and without the outcome are from the same population.

The disadvantages are:

- There is no quality control over measurements.
- The data about past history may be incomplete.

2). Nested Case-Control Study

This is so called because it is a case-control study 'nested' within a cohort study. Within an on-going cohort first an outcome is defined. All the members of the cohort who have developed the outcome are identified. From amongst those who have not developed the outcome a random sample is taken. Medical records of both groups before the outcome occurred are studied and predictor variables are compared between the two groups. For example, a cohort of children who were referred to hospital for febrile convulsions is assembled. In this cohort all those who go on to develop non-febrile recurrent seizures are compared with a random sample of those who do not develop this complication for defined

predictor variables like birth history, family history, characteristics of associated illness at the time of first febrile convulsion, type of convulsion, and so on.

3). Double Cohort Study.

Two distinct cohorts are assembled on the basis of exposure to a risk factor like Exposed / Not Exposed / or Low Level Exposure. For example, two cohorts of children one comprising those who received orally a traditional medicine containing lead, and another made of those who had traditional cosmetics containing lead (kohl) applied to the eyes or cheeks are compared for signs of lead toxicity with healthy children not exposed to lead.

This type of study design is not to be confused with case-control design. The samples are chosen by **exposure**. In case-control design samples are chosen by **outcome**.

Choosing Cohort study Designs

Cohort studies are best suited for the following situations:

- For accurately describing the incidence or natural history of a condition.
- Often they are the only way of firmly establishing the temporal sequence.
- They are useful in studying rapidly fatal diseases.

One advantage of cohort studies is that several outcome variables may be studied simultaneously. For example, smoking causes not only heart disease but also chronic respiratory disease, lung cancer, and is associated with peptic ulcer. As different types of events accumulate it is possible to study several outcomes.

Because of their costs and the time taken to answer the research question cohort studies are undertaken when less expensive approaches fail to answer the research question satisfactorily.

Deciding about the type of cohort design will depend upon both practical and scientific considerations. If the research question can be answered by the data already existing a retrospective cohort design may suffice. On the other hand if the research question is about outcomes which are frequent in occurrence a prospective cohort design may be more fruitful.

With regard to subjects, if the research question is about describing the natural history of a condition the sample must closely resemble the target population to which the results would be applied. For example, since only the more complicated types of febrile convulsions get referred to a specialist hospital a study of the natural history of the condition is best done in the primary care setting.

If the research question is of an analytic nature, for example the cause-effect relationship, then the cohort must contain enough subjects with the major predictors and a sufficient number of outcomes to allow for meaningful analyses.

Bias in Cohort Studies

Bias can arise in cohort studies because of a number of reasons:

- The degree of accuracy with which subjects have been classified with respect to exposure and disease status.
- Information bias.
- Accuracy in diagnosing the outcome.
- Drop-outs or loss to follow-up. This may be due to causes related to exposure, outcome, or both.
- Non-participation. Those who agree to participate are likely to differ from those who refuse with regard to motivation, attitude towards health, risk factor status etc.
- Confounding bias.

In general, the likelihood of selection bias is much less in cohort compared to case-control studies. In prospective cohort studies exposure is assessed before the occurrence of the disease. Hence it is very unlikely that outcome would influence classification of exposure. This pitfall happens in retrospective studies where exposure and outcome have already taken place at the commencement of the study. If knowledge of outcome is to influence classification of exposure then selection bias may result in the same manner as it does in case-control studies.

Inaccurate classification with regard to exposure or disease status can be the major source of error in cohort studies. Careful definitions at the planning stage can help to avoid this problem.

Drop-outs are another cause of bias. Cohort studies require following individuals for a sufficient length of time to allow the outcomes to happen. By then it is inevitable that some of the subjects may have changed their address, some may have lost interest, moved away or died. If loss to follow up is in excess of 30 to 40 per cent the results would be biased. Hence choice of a stable population at the design stage of the study is important. For the drop-outs a knowledge of their backgrounds like occupation, education, income, ethnicity, life-style, and so on is helpful. If it can be shown that those who dropped out of the study are not very different from those who stayed, then the results are more valid. Another approach is to first assume that all drop-outs developed the outcome and calculate rates accordingly. Next assume that they did not, and again estimate the results. The two results provide a range between which the actual results lie.

Refusal to respond poses yet another kind of problem. If non-response is related to exposure status the results would under estimate the strength of the association.

Appendix 7.1
Sample size in cohort studies
Comparison of Exposed Group with General Population (for $\alpha = 0.05$; one-sided test)

The table shows the probability of obtaining a significant ($P < 0.05$) increased risk associated with the exposure in a cohort study in which the expected number of cases of disease in the absence of any risk is 'm' and the actual increased risk associated with the exposure is given by the risk ratio. 'C' is the minimum number of cases in the exposed group which would yield a statistically significant difference at $p < 0.05$.

For example, in a study of an exposed group in which we expect 40 cases of disease in the absence of any increased risk we would have a 52% chance of obtaining a significantly (at $p, 0.05$) increased risk if the true risk ratio associated with the exposure was 1.3, and have a 94% chance if the true risk ratio was 1.6. 'C' is the smallest number of subjects needed.

Expected Cases m	Critical number of cases c	Risk Ratio							
		1	1.5	2	3	4	5	7.5	10
1	4	1.9	7	14	35	57	74	94	99
2	6	1.66	8	21	55	81	93	100	
3	7	3.35	17	39	79	95	99		
4	9	2.14	15	41	84	98	100		
5	10	3.18	22	54	93	100			
6	11	4.26	29	65	97	100			
7	13	2.7	26	64	98	100			
8	14	3.42	32	73	99	100			
9	15	4.15	38	79	100				
10	16	4.87	43	84	100				
11	18	3.22	39	83	100				
12	19	3.74	44	87	100				
13	20	4.27	48	90	100				
14	21	4.79	53	93	100				
15	23	3.27	49	92	100				
20	29	3.43	60	97	100				

Appendix 7.1 (cont'd).

m	c	Risk Ratio							
		1	1.1	1.2	1.3	1.4	1.5	1.6	1.7
20	29	3.43	9	18	30	45	60	73	83
25	34	4.98	13	26	42	59	74	85	92
30	40	4.63	13	27	46	64	79	89	95
35	46	4.25	13	29	49	69	83	92	97
40	52	3.87	13	30	52	72	86	94	98
45	57	4.73	16	36	60	79	91	97	99
50	63	4.24	16	37	61	81	93	98	99
60	74	4.42	18	42	69	88	96	99	100
70	85	4.48	19	47	75	92	98	100	
80	96	4.46	21	51	80	94	99	100	
90	107	4.39	22	55	83	96	99	100	
100	118	4.28	23	58	86	97	100		

Appendix 7.2
Sample size in Cohort Studies.

Comparison of exposed group with General population ($\alpha = 0.05$; one-sided test)

This table shows the risk ratio in the exposed group as compared to the general population in order to obtain a statistically significant ($p < 0.05$) increase in risk in the exposed group.

For example, in an exposed group in which we expect 40 cases of disease we would have a 90% chance of detecting a significant ($P < 0.05$) test result if the true risk was 1.54. 'C' is the smallest number of subjects required. Power probabilities are those shown at the top of the table.

Probability of obtaining significant ($P < 0.05$) test
If true risk ratio is as shown below. (Power)

Expected cases (R=1) m	Critical No. Of cases c	0.5	0.8	0.9	0.95	0.99
1	4	3.67	5.52	6.68	7.75	10.05
2	6	2.84	3.95	4.64	5.26	6.55
3	7	2.22	3.03	3.51	3.95	4.86
4	9	2.17	2.84	3.25	3.61	4.35
5	10	1.93	2.5	2.84	3.14	3.76
6	11	1.78	2.28	2.57	2.83	3.36
7	13	1.81	2.27	2.54	2.78	3.26
8	14	1.71	2.13	2.37	2.58	3.02
9	15	1.63	2.01	2.24	2.43	2.83
10	16	1.57	1.92	2.13	2.31	2.67
11	18	1.61	1.95	2.15	2.32	2.66
12	19	1.56	1.88	2.06	2.22	2.55
13	20	1.51	1.82	1.99	2.14	2.45
14	21	1.48	1.77	1.93	2.08	2.36
15	23	1.51	1.79	1.95	2.09	2.37
20	29	1.43	1.67	1.8	1.92	2.15
25	34	1.35	1.55	1.67	1.77	1.96
30	40	1.32	1.51	1.61	1.7	1.87
35	46	1.3	1.47	1.57	1.65	1.81
40	52	1.29	1.45	1.54	1.61	1.76
45	57	1.26	1.41	1.49	1.55	1.69
50	63	1.25	1.39	1.47	1.53	1.66
60	74	1.23	1.35	1.42	1.48	1.59
70	85	1.21	1.32	1.39	1.44	1.54
80	96	1.2	1.3	1.36	1.41	1.5
90	107	1.19	1.28	1.34	1.38	1.47
100	118	1.18	1.27	1.32	1.36	1.45