Lecture 7: Cohort studies

Learning Objectives

By the end of this session, participants should be able to:

i. Describe the principal design features of cohort studies

ii. Explain the strengths and weaknesses of cohort studies

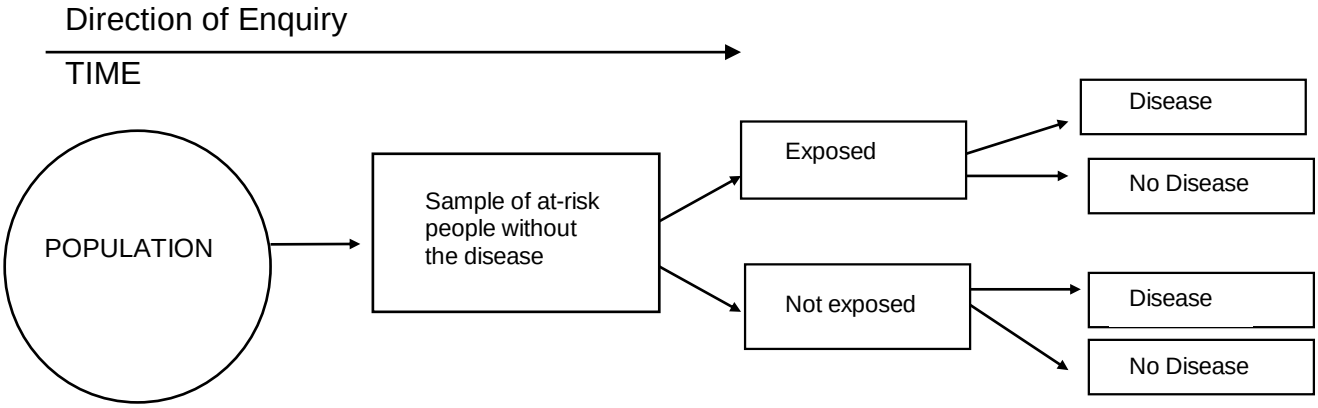
iii. Describe the basic analytical approaches commonly used for cohort studies

NOTE: Please refer to the diagram of epidemiological designs (figure 1) in Lecture 0 and in the Week 6 surgery session.

1. The cohort study approach

The essential elements of a cohort study are the:

- Identification of at-risk groups of individuals in a population
- Definition of the primary exposure and its measurement
- Definition of potential confounding variables and their measurements
- Definition of the outcome(s) and its measurement
- Collection of outcome data
- Analysis
- Interpretation



Source: Beaglehole et al., Basic Epidemiology

2. Defining exposure groups

Cohort studies take at-risk individuals and classifies them into groups according to their exposure status. Exposure in this setting means the factor that you are investigating as a possible cause of disease, e.g. smoking, an infectious agent, social media, etc.

These exposures may not be binary, and one may wish to sub-divide into degrees of exposure. Smokers can be classified by packs of cigarettes smoked per year, industrial exposures by extent of exposure (often done by place of work or job in a factory), infection by dose of agent or age at exposure, breast feeding by duration, and psychological exposures by some scale of severity. The simple situation of two groups, one exposed and one unexposed, is unusual. Many

exposures of interest are measured on a continuous scale. Using grades of exposure allows one to assess if there is a dose-response relationship to the outcome.

Cohort studies are ideal for investigation of exposures which are rarely observed in general populations, e.g. for occupational risks, new prescription drugs, experience of humanitarian disaster, etc. The public health impact of the exposure may be small, but such studies can give insight into common biological mechanisms in disease. Many of our assumptions about the impact of low-level radiation are based on the health impact on work forces with high exposure.

In cohort studies one hopes to mimic an intervention study where individuals are randomly allocated to exposure. In a well-conducted RCT the only difference between the groups is the exposure itself. Thus, in a cohort study one wants to choose exposed and unexposed groups that are as similar as possible in all respects except exposure status. In occupational cohorts this may be relatively simple. Many chemical exposures only occur amongst factory workforces. It can be assumed that the level of exposure among populations not working in this environment is virtually zero. One can therefore choose people in employment from the same geographical area who are not exposed to the chemical as a comparison group. This may be people from the same factory but working in a different job or site (an internal comparison group), or people from other factories near the study, or from census data for the same geographical area (an external comparison group).

The selection of appropriate comparison groups is one of the most difficult aspects of cohort studies (as it is in case-control studies). For example, a problem that arises is whether the comparison group are truly unexposed. People may move from one job to another within a geographical area such that workers in the comparison group may have had exposure in a previous job. Similarly, people may give up smoking but re-start later, they may have been stressed in the past but not during the period of study, or they may fail to recall exposure. Thus, the definition and measurement of exposure influences the selection of the comparison group.

Sometimes cohort studies are geographically based – as was the case with the Framingham study and the Seven Countries cohort study of diet and heart disease. Cohorts may be determined by time. A series of cohorts in the UK have been based on dates of birth, with all children in the country, or part of the country, born on specific days being eligible for enrolment in the cohort. These cohorts not only look at multiple exposures as well as multiple outcomes but are a valuable resource for investigating many different hypotheses. These hypotheses often only arise after the cohort has been established.

3. Temporal sequence

Cohort studies are sometimes referred to as *prospective* studies. Conceptually, cohort studies are prospective because they look forward in time from exposure to disease. Similarly, case-control studies can be said to be retrospective because they look back in time from disease to exposure. However, this terminology can be confusing and is best avoided.

Cohort studies are not always conducted in real time– they can be conducted using historic records, or data collected in the past, in which case they are known as *historic* cohort studies (more on these below). Recently, there has been an expansion in the use of routinely collected health data for cohort studies- these include data from electronic medical records and administrative data sources, e.g. insurance data. Equally, case-control studies can be prospective, in the sense that *cases* may be recruited over a period of time. In general, it is better to actually describe the procedure being used so that the design is clear. The key concept to remember is that in cohort studies, the exposure status is always assessed among participants *before* the outcome has occurred.

Historic and primary data collection cohort studies

Historic cohort studies rely upon previously-collected records of individuals who were at risk of the outcome. These records then allow the individual to be classified into an exposure group and

followed up for occurrence of the outcome, either in the same set of records, or in another set of records that can be linked to the exposure data. For example, in some countries it is possible to carry out historic studies using electronic health records. An advantage of such historic studies is that they are a shortcut to studying diseases with a long interval (or latency) between exposure and the disease(s) in question. Many cancers do not appear for decades after exposure to a carcinogen. In order to study these in real time, you would have to wait for many years before observing the effect of the exposure. By using a historic study, some of these questions can be investigated relatively quickly and inexpensively.

One drawback of such studies is that the investigators have to rely on historic records that may have been collected for completely different purposes, and in which information on exposure and/or outcome may be incomplete, inaccurate, or measured using outdated or sub-optimal methods. A further disadvantage is that there may be limited information about relevant confounders that need to be taken into account. One therefore has to balance the accuracy and completeness of the records against possible time and cost savings. A particularly useful historic method is where some biological sample has been stored that can be used to measure exposure in the past. For example, serum banks can be accessed to determine past infection or exposure to some chemical. This method minimises inaccuracies in exposure measurement, but the number of such biological sample banks is limited.

Collecting data on exposure *now* (using primary data collection) allows the most up-to-date methods of measurement to be used and bias in exposure classification to be minimised. However, it carries with it the problem of the time lag until outcomes occurs. If the disease is of short latency and is relatively common then the time lag may be of no importance, as is the case for many infectious diseases. An additional advantage of primary data collection or contemporary cohort studies is that "confounding" variables can also be identified and measured accurately.

In observational studies such as cohort studies the exposure has not been randomly assigned but occurs naturally. This lack of random allocation may lead to differences between the groups in terms of exposures other than the one being studied. These differences are only of importance if these other, different exposures are also risk factors for disease. Thus, if one is studying a chemical exposure and cancer, it will almost certainly be necessary to ensure that the "exposed" and "unexposed" groups have a similar smoking history, because smoking is another risk factor for cancer. If the two groups are not similar with regard to smoking history, then statistical adjustment for differences in smoking will have to be made (see Lecture 5). In order to be certain that there is no difference, or to carry out this adjustment, data on smoking for each individual are required. These data must be as accurate as possible and of similar quality to the data on the exposure of primary interest. In primary data collection cohort studies, the collection of data on potential confounding variables can be built into the design of the study, but is much more challenging for historic cohort studies. Even then, a new "cause" of the disease may be discovered during the course of the study and exposure to this new factor has not been collected. It is in these situations that the "art" of epidemiology comes into play - judgement about the size of the effect observed and that of potential confounding. Confounding will be discussed in detail in later sessions.

4. Outcome measurement

A major advantage of cohort studies is that they offer the possibility to look at a range of outcomes rather than just one. The range will be dependent on the method of measurement of outcomes. In high-income countries most studies use some form of routinely collected data (e.g. cancer registrations, death certificates or some specialised surveillance method). In this situation one is limited to the outcomes recorded and the way in which they are coded. However, in low-income countries, or when the outcome of interest is not routinely recorded, one must set up some form of surveillance of disease especially for the cohorts. For example, the follow-up may be by regular physical examination of all members of the cohort, or through self-completed questionnaires administered to cohort participants at regular intervals. Whatever the follow-up method chosen, the key point is that the method of ascertainment must be identical for those exposed and those not exposed. Ideally, the procedure used to assess the outcome will not be

influenced by the participants' exposure status in any way. An outcome assessment procedure which is influenced by the exposure status is likely to be biased (see Lecture 4).

In most cohort studies, one cannot be certain that all individuals were disease-free at entry to the study. It is therefore usual to exclude disease events occurring in some time period following commencement of the study. For example, for cancers with long latent periods this is often the first 2-3 years of follow up.

A particularly important aspect of conducting a cohort study is the completeness of follow-up. It is essential that as high a proportion of people in the cohort as possible are followed up. Some people will migrate, some die and some change employment; but every effort must be made to include these people in the outcome measurement. All of these factors may be influenced by the exposure and so incomplete follow-up (loss-to-follow-up or LTFU) may again introduce bias (see Lecture 4).

5. Analysis

In cohort studies, it is usual to estimate the risk or rate of a disease in the different exposure groups. The risk or rate of disease in the exposed cohort is compared with that in the unexposed cohort.

The **ratio** of incidences in the exposed and unexposed groups gives rise to a risk ratio or rate ratio (see Lecture 1). If various levels of exposure are used in the cohort, then one can examine trends in incidence of disease by "dose" of exposure. This can often provide stronger evidence of an association between exposure and disease. A simplified example using incidence rates:

	Number of cases ***d***	**Person years at risk** ***Y***	**Rate per 1,000 pyrs** ***r***	**Rate Ratio** ***RR***
Unexposed	20	10,000	2	Reference group
Exposure level 1	30	10,000	3	1.5
Exposure level 2	40	10,000	4	2.0
Exposure level 3	50	10,000	5	2.5

Where information is available on confounding variables, adjustment of the rate or risk ratio will need to be made for these confounders.

Another common method of presenting the results of cohort studies is by a **Standardised Mortality (or Morbidity) Ratio**. Essentially the rates observed in the unexposed cohort (usually an external comparison group such as the total population for the region or country) are applied to the population denominators in the exposed cohort, with appropriate allowance for age, sex and other confounders to give an expected number of events (E), whether deaths, cancers or morbid events, in the exposed cohort. This is then compared to the observed (O) number of events in the exposed cohort. The ratio O/E is then calculated and multiplied by 100 to give an SMR. It is usual for the expected number not to be a whole number. Thus, in a study of liver cancer and alcohol the expected number of liver cancer cases in the exposed cohort might be 1.78, the observed number was actually 8 and the ratio 4.49, (this ratio is the rate ratio), giving an SMR of 449.

Risk or rate **differences** can also be calculated from cohort studies (see Lecture 1).

6. Interpretation

As in any observational study judgement must be used in assessing possible biases in the study, possible residual confounding and the influence that these may have on the quantitative results.

Also, it is essential that a clear hypothesis that can be accepted or refuted is formulated before the analysis has begun.

As in all observational epidemiology, one study alone cannot constitute proof. There must be consistency of findings between cohorts in different geographical locations, by different investigators and in different time periods. In addition, there should be consistent evidence from other types of studies, e.g. ecological and case-control studies. The best evidence as to validity is provided by randomised intervention studies showing a reduction in disease by removal of exposure. The issues involved in determining whether an association is causal are discussed in Lecture 5. However, for some exposure/disease relationships it is not possible or ethical to conduct an intervention study.

7. Advantages and Disadvantages

The major advantages of a cohort study are:

1. Exposure is measured before disease onset and is therefore exposure assessment is typically unbiased in terms of disease development

2. Rare exposures can be examined by appropriate selection of study cohorts

3. Multiple outcomes (diseases) can be studied for any one exposure

The main disadvantages are:

1. Since individuals must be followed up over time losing contact with them may introduce bias due to loss to follow-up (see Lecture 4).

2. For a cohort study with primary data collection: time and cost.

3. It is difficult to exclude confounding as an explanation for the result of the study. The exposed and unexposed groups differ with regard to exposure, and potentially differ by other factors which can cause the outcome. It is not possible to design or analyse a cohort study to avoid/eliminate the effect of an unknown confounder.

4. For a historic cohort study, investigators must recognize that quality and quantity of available data for exposure, outcome and confounders may not be optimal.

References

Webb P and Bain C. *Essential Epidemiology: An introduction for Students and Health Professional*s. Chapter 4. Second Edition. Cambridge University Press. 2011.

Bailey L, Vardulaki K, Langham J and Chandramohan D, *Introduction to Epidemiology*. Chapter 6. Open University Press, 2005 (Understanding Public Health, Series editors: Nick Black and Rosalind Raine)

Hennekens CH & Buring JE, *Epidemiology in Medicine*, Chapter 7. Little, Brown and Company, 1987.

Dos Santos Silva, I. *Cancer Epidemiology: Principles and Methods*, Chapter 8. IARC, Lyon, France. 1999