Lecture 8: Case-control studies

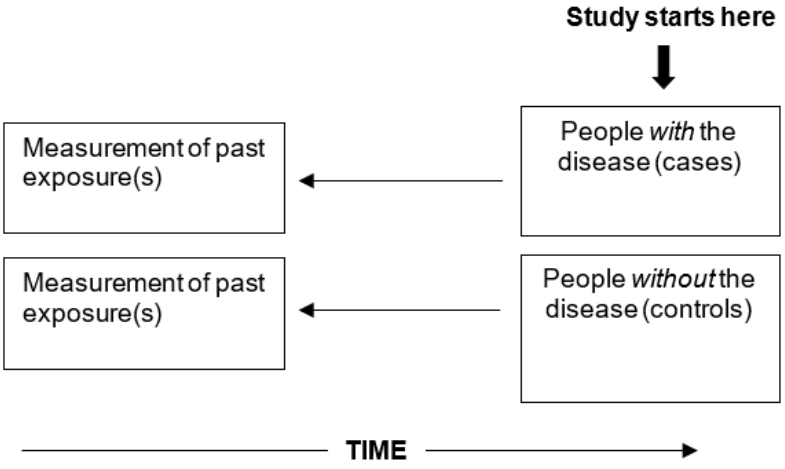
Learning objectives:

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

i. Describe the basic design features of a case-control study.
ii. Appreciate the strengths and limitations of case-control studies in comparison with other study designs.
iii. Recognise the importance of clearly defining 'cases' and 'controls'.
iv. Understand the concept of selection bias and how it might arise.
v. Appreciate the difficulties of obtaining unbiased measures of 'exposure' in this type. of study and be aware of strategies to minimise observer and responder bias.
vi. Recognise the importance of collecting information on potential confounders.
vii. Describe the advantages and disadvantages of matched and unmatched designs.
viii. Understand how data from case-control studies can be analysed.

1. The case-control study approach

A case-control study starts with the identification of a group of **cases** (individuals with a particular disease or condition) and a group of **controls** (individuals without the illness or condition). Then, the prevalence of one or more **exposures** measured and compared. If the prevalence of a given exposure is higher among the cases than among the controls, the exposure might be a **risk factor** for the disease. If the prevalence of exposure is lower among cases than among the controls, the exposure might be a **protective factor** for the disease.



2. Conducting a case-control study

2.1. Determination of the hypothesis to be tested

The hypotheses of interest must be clearly identified prior to the design of a case-control study. With the hypotheses, researchers will be able to define their exposure(s) of interest, and then design the study appropriately.

2.2. Definition and selection of cases

2.2.1 Case definition

Precise criteria for the definition of a case are essential. The criteria might be based on laboratory findings (e.g. the presence of certain histological types of a cancer), or might be clinically based (e.g. one or more hospitalizations for psychosis in the past 5 years among women aged 20-65).

2.2.2 Incident versus prevalent cases

An important issue to be considered is whether to include **prevalent** as well as **incident** cases in the study. Incident cases are new cases appearing within a fixed period of time (e.g. a year) whereas prevalent cases are all those with the disease at one point (or short period) in time. Prevalent cases will include patients who may have had the disease for some time, and may be different in terms of exposure compared to recent incident cases. Inclusion of prevalent cases may also miss people with more severe disease who may die early and whose exposure levels may differ from prevalent cases with mild disease and who are still alive. These problems are reduced by taking incident cases.

An important point to consider is that both prevalent and incident cases may have changed their habits (or "exposures") *because of* the disease – either from knowledge of the diagnosis or from symptoms/consequences of the disease. For example, lung cancer cases may have given up smoking because they had developed a cough. This change in behaviour is usually a greater concern when prevalent cases are used, than for incident cases.

Obtaining accurate recall of exposures in the past (e.g. smoking) before the onset of disease is required for both incident and prevalent cases but recall may be more difficult for prevalent cases who might have been diagnosed long ago.

2.2.3 Source of cases

The selection process for cases into a case-control study needs to be carefully considered. The study might be '**hospital-based**' and the cases taken from all patients fulfilling the case definition criteria attending a certain hospital (e.g. all cases of stillbirth delivered in the maternity department of Basingstoke District Hospital 2004-8). Alternatively, the study might be '**population-based**' and cases taken from a defined population over a fixed period of time, (e.g. all cases of salmonella food poisoning reported to North East Thames Regional Health Authority in one year). In general, population-based studies are more readily generalizable to the population but are also more logistically difficult to conduct.

It is important to consider whether the cases chosen are representative of all cases in the population. There is usually some degree of selection, and the role of **selection bias** in case-control studies may be extremely important. Issues that need to be considered are patient survival, referrals to specialist hospitals and refusals to take part in the study. For example, in a study of depression, hospital-based cases may be different from those in the community

(they may be more severe, have other co-morbidity, or social problems), cases attending general practitioners for help with depressive symptoms may be different from cases of depression who do not seek help. The selection of cases (and controls) must be made independently of their exposure status.

2.3. Definition and selection of controls

2.3.1 Definition of controls

Controls must fulfil the criteria defining cases, apart from those criteria relating to the disease. For example, if the cases are all females aged 14-44 years living in Manchester who have been diagnosed with rheumatoid arthritis, then the controls must be selected from women aged 14-44 years living in Manchester who are confirmed to not have rheumatoid arthritis.

2.3.2 Source of controls

The key point for recruiting control participants is that controls should represent the prevalence of exposures (and confounders) in the same population from which the cases arise. Identifying the controls is usually straightforward when the case selection is population-based. Because hospital-based cases often come from a widely dispersed locality, identifying controls may be more problematic. Many studies use “neighbourhood controls” so that for each case, a control is identified from the same local population register, or same general practice register as the case. Hospital-based controls are often used in case-control studies because they are easy to identify, and these patients are often happy to participate (e.g. taken from the hospital outpatient department). The major weakness of hospital-based controls is that they may not be representative of the prevalence of exposures in the population. For example, relative to population-based controls, hospital-based controls are often affected by other health conditions, and are more likely to be smokers or drinkers or have more unhealthy lifestyles. Population-based controls are usually preferable. However, it is often challenging to recruit population-based controls. Lower participation rates are associated with increased probability of selection bias, which is problematic if participation is somehow related to the level of exposure.

Choice of a suitable control group is the most difficult part of designing a case-control study. Some studies use more than one type of control group, but this can lead to problems: the two control groups may give different results, a situation which may be difficult to interpret, although these discrepant results may also bring insights into selection biases. On the other hand, it may be possible to examine more than one hypothesis by choosing different control groups. For example, if we have a group of healthy controls and a group of cirrhotic controls in a study of liver cancer, it is possible to examine whether risk factors act by leading to cirrhosis or are independent of this pathway. However, this must be explicit in the hypotheses being tested.

Studies may recruit more than one control per case as this increases the statistical power to detect a difference in exposure levels between cases and controls. It is common to see 2 or 3 (or more!) controls recruited for every 1 case in a case-control study, which can compensate statistically for having relatively few cases in the study.

2.3.3 Matching

Matching refers to the recruitment procedure whereby controls are selected on the basis of similarity for certain characteristics to each case who was selected into the study. Common

matching characteristics are age and sex, but other characteristics might be place of residence, socio-economic status, or parity. The characteristics chosen for matching are those which are thought to be **confounders**. Confounding is the alteration of the disease/exposure relationship brought about by the association of other factors with both the disease and the exposure (see lecture on confounding). Cases may be ***individually*** matched to one (or more) controls or ***frequency*** matched to controls (i.e. during recruitment investigators ensure that there are roughly equal number of cases and controls in each level of a matching variable, such as age group).

Matching is done to increase the statistical power of the study (i.e. a matched study would have more power to detect an association than an unmatched study with the same number of cases and controls); but it is essential that matching is done on a limited number of potential confounding factors, so that it is possible to find a match. Also note that matching results in a set of controls that are different from those who would have been selected if there was no matching, and this must be taken into account in the analysis, by conducting the appropriate matched analysis.

Note: there is a relatively common misconception that matching in case-control studies is a method to deal with confounding through the study design. This is not correct, as the statistical analysis needs to take account of the matched design or matching variables. Thus, the reason for matching in case-control studies is to improve the study precision when controlling for confounding by the matching variables.

2.4. Measuring exposures

2.4.1 Source of exposure information

Data on exposure can be gathered in many ways: for example, by personal, postal or telephone interview, by examining medical, occupational or other records, or by taking biological samples. The important issue is that the information gathered is unbiased, valid and not influenced by whether the participant is a case or a control.

2.4.2 Information (or measurement) bias

It is inevitable that there will be some inaccuracies in the information reported by respondents for some of the exposures. One critical question is whether inaccuracies in exposure measurements are different between cases and control; this is also known as differential misclassification of the exposure. These inaccuracies lead to what is known as **information (or measurement) bias** (see lecture on Bias). Observer and responder bias are the two main types of information biases.

(i) **Observer bias** occurs when the process of gathering exposure data by the investigator is systematically different for cases and controls. Ideally, the data collector / interviewer should be unaware, or '**blind**'/masked to the hypothesis under study and to who is a case and who is a control. While this level of masking is very difficult to achieve perfectly, investigators must be trained in the unbiased collection of data. Information must be collected in an objective and consistent way (e.g. the same forms and questionnaires used for both cases and controls).

(ii) **Responder or recall bias** occurs when the way in which study participants supply information about exposure differs between cases and controls. Cases, in particular, may be influenced in their answers by their awareness of being a case. Responder/recall bias can be

minimised by keeping the study members unaware of the hypotheses under study and, where possible, ensuring that both cases and controls have similar incentives to remember past events, or using exposure information that was recorded before the disease status was known.

2.4.3 Reverse causality

This is a particular problem with case control studies in which the measurement of exposure is done after the outcome (disease) has already occurred. The exposure data may be affected by the disease itself, in which case an association between the exposure and disease is not causal, but instead the change in exposure level is caused by the disease itself.

2.4.4 Nested case control studies

A particular design of case control studies is the nested case-control design. In this approach, the case-control study is embedded ('nested') in an existing cohort study. Cases that occur as part of the follow-up of that underlying cohort study become cases for the nested case-control study, and a sample of cohort members who do not develop the outcome are selected as controls. After identifying the cases and selecting the controls, further information on exposure can be collected as required.
This is an efficient design, because it allows to spend less resources collecting additional exposure information on a subset of the cohort study, rather than the whole cohort. It also helps minimise the problems of selection bias (as we know the underlying source population for the case control study, and we can be careful to select a control group that is representative of that source population), and of reverse causality, as we know the temporal sequence between the exposure and the outcome. Note that more than one case-control study can be 'nested' within a cohort.

2.5 Analysis and interpretation of results

2.5.1 Analysis of data

In an ***unmatched study***, the numbers of cases and controls found to have been exposed and not exposed to the factor under investigation can be arranged in a 2x2 table as follows:

Table 1. 2 x 2 Table for an unmatched case-control study

	Cases	Controls	TOTAL
Exposed to factor	a	b	a+b
Not exposed to factor	c	d	c+d
TOTAL	a+c	b+d	a+b+c+d

In **cohort** and **intervention studies** it is possible to calculate all three measures of relative risk (risk ratio, rate ratio, and the odds ratio of disease) since the incidence of disease in the exposed and unexposed groups is known. **Case-control studies cannot directly estimate the incidence of disease** among the exposed and unexposed participants since the participants are selected on the basis of their disease status and not on the basis of their exposure status. It is, however, possible to calculate the **odds of exposure** among cases and the **odds of exposure** among controls to obtain an **odds ratio of exposure**. It can be shown that this 'odds ratio of exposure' among people with and without disease can also be interpreted as the 'odds ratio of disease' among those with and without exposure. This is more helpful to answer the research question of interest ("*is the frequency (odds) of disease greater*

or less among people who are exposed compared to people who are not exposed?" than the question "is the frequency (odds) of exposure greater or less among people who have the disease compared to those who do not?"

The **odds ratio of exposure** is given by the **odds of exposure** among the cases (a/c) divided by the **odds of exposure** among the controls (b/d):

$$OR = \frac{a/c}{b/d}$$

Table 2. Unmatched case-control study investigating smoking and lung cancer.

	Lung cancer (Cases)	No cancer (Controls)
Ever smoker	120	160
Never smoker	230	540
	350	700

$$OR = \frac{120/230}{160/540} = 1.76$$

Therefore, the odds of ever-smoking among cases was 1.76 times the odds of ever-smoking among controls. This can be interpreted as the odds of lung cancer among ever smoker was 1.76 times that of people who never smoked; alternatively, we can also say that people who ever smoke had 76% higher odds of lung cancer compared to those who never smoke.

Note: If controls were recruited into the study to match specific characteristics of individual cases, then a different analytic method must be used to calculate the odds ratio. For a matched case-control study, it is not possible to calculate the odds ratio using the 2x2 table method described above.

2.5.2 Interpretation of results

As with all epidemiologic studies, if an association is found between the exposure and the outcome, the investigator must consider whether the result could have arisen by chance, bias, or confounding, and consider the possibility of reverse causality.

3. Advantages of case-control studies

1. Relatively inexpensive
2. Relatively quick
3. Can test multiple hypotheses / exposures
4. Useful for rare diseases and diseases of long latency
5. Can employ expensive or time-consuming tests
6. Can test current hypotheses
7. Is usually the optimal study design to study disease outbreaks

4. Disadvantages of case-control studies

1. Prone to selection bias of cases and controls
2. Prone to information bias
3. Problems sorting out sequence of events (reverse causality)
4. Not suitable for investigating rare exposures
5. Usually cannot obtain estimates of disease incidence

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 7. Open University Press, 2005 (Understanding Public Health, Series editors: Nick Black and Rosalind Raine)

Hennekens CH & Buring JE, *Epidemiology in Medicine*, Chapter 6. Little, Brown and Company, 1987.

Dos Santos Silva, I. *Cancer Epidemiology: Principles and Methods*, Chapter 9. IARC, Lyon, France. 1999