Lecture 5: Confounding and Causality

Learning Objectives

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

i. Recognise a range of possible explanations for an observed association between a risk factor (an exposure) and an outcome.

ii. Identify confounding variables, and know how to control for their effects at the design stage and at the analysis stage.

iii. Explain the difference between statistical association and causality, and be aware of the basic criteria that help determine whether an association is causal.

1. Could an observed association be due to confounding?

1.1. What is confounding?

Confounding is a central concept in epidemiology, particularly in relation to observational studies. Much of epidemiology is concerned with investigating possible associations between exposures (potential risk factors) and the risk of disease. This usually involves comparing groups from two populations that are as similar as possible in all respects except the factor(s) that we are interested in studying.

In randomized controlled trials, the process of randomizing individuals (or groups) to different exposures (interventions) generally ensures that the different groups are equally balanced with respect to all relevant factors that might influence the risk of the outcome. Provided the intervention study is sufficiently large, and is conducted rigorously, if we see a difference in the incidence of the outcome between treatment groups at the end of the study, then we can conclude that this difference is caused by the treatment. For this reason, well conducted intervention (experimental) studies provide the strongest evidence of a causal association between an exposure and disease.

In observational studies (cross-sectional, case-control, cohort studies), by contrast, it is not possible for individuals to be randomly assigned to an exposure. Often, individuals who share a particular risk factor have other characteristics in common that may also influence their risk of disease. Thus, because we cannot be sure that individuals with different exposure status are similar with respect to all other relevant factors, it is more difficult to determine if the association we observe between our risk factor of interest and disease is ‘real’, or whether it is influenced by the other factors which clump, or associate with our risk factor of interest.

To illustrate this, imagine that you are interested in knowing whether smoking is associated with the incidence of coronary heart disease (CHD) among British men aged 18 to 64 years. You conduct a study by recruiting healthy men in this age group, assessing their smoking status, and following the men over time. You would expect to find a higher incidence of CHD among smokers. However, males who smoke are also likely to have a higher alcohol intake, which also increases the risk of CHD. Thus, some part of the observed effect for the smoking-CHD association is actually due to alcohol intake. Without taking account of alcohol intake between smokers and non-smokers, the observed effect for the smoking-CHD association will not reflect the true effect.

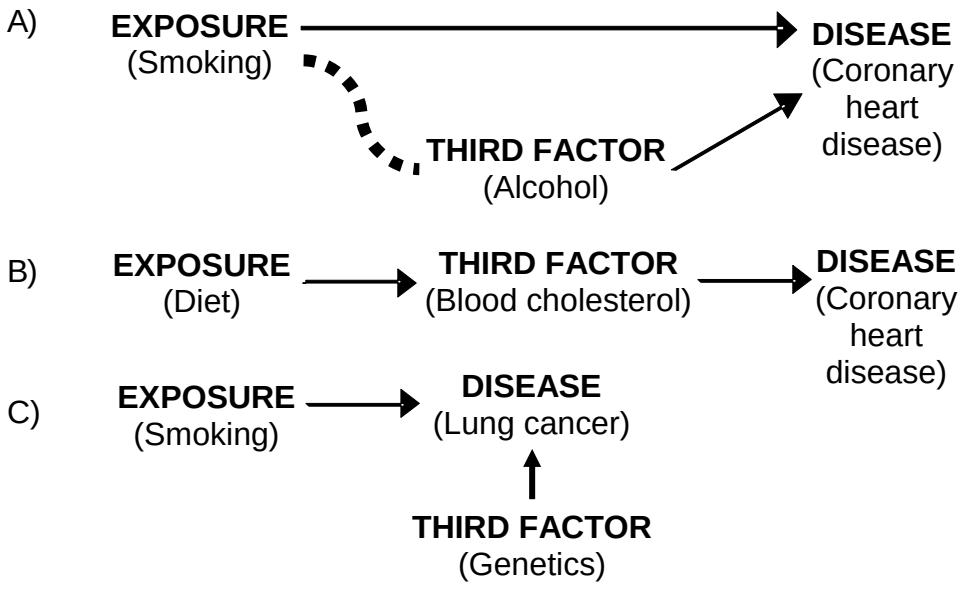
Confounding occurs when the association (or lack thereof) between an exposure and the disease is distorted by the presence of a third factor. In the example above, the association between smoking and CHD was distorted by alcohol.

For a variable to be a confounder, it must satisfy these three criteria:

- Be associated with the risk factor (exposure) of interest, and
- Be a risk factor for disease, and
- Not be on the causal pathway between exposure and disease

Returning to our example above, the first criterion is met because smokers, on average, have a higher alcohol intake. The second criterion is met because alcohol intake is a known risk factor for CHD. Finally, alcohol intake is not on the causal pathway between smoking and CHD (smoking does cause people to drink more alcohol).

Consider the following scenarios of smoking, CHD, and a 'third factor'.



In these examples involving a third factor, only A) shows a confounding variable. In B), blood cholesterol is not a confounder but is what is known as a mediator. In C), genetics is a not a confounder but is an independent risk factor for the disease.

Potential confounding factors can include factors (e.g. age and social class) that are proxy measures of more direct causes. Drawing a diagram like those displayed above is helpful when trying to decide whether a third factor is a confounder.

1.2. How to deal with confounding

Confounding is a problem because unless it is dealt with, you get the "wrong answer" for the relationship between the exposure of interest and the disease outcome. The distorting effects of confounding can be dealt with at the design stage or the analysis stage.

1.2.1. Design stage

Two main options are available.

- Randomisation – this is a procedure whereby study participants are randomly allocated to the exposure or control groups. This is the ideal method of controlling for confounders because it ensures that the distribution of known *and of unknown* confounding variables will be similar in the groups to be compared, provided that the sample size is large. But randomization can only be used in intervention (experimental) studies and is not appropriate for all exposures of interest.

- Restriction - a procedure that limits participation in the study to people who are similar in relation to the confounder. For instance, if participation in a study is restricted to non-drinkers, any potential confounding effect of alcohol will be removed.

1.2.2. Analysis stage

It is possible to control for confounding in the analysis stage if information on confounding variables was collected during the study. You will learn how to use the following three options if you take a statistics module.

- Stratification is the method emphasised in this module:

    - First the researcher calculates an overall relative risk for the exposure-disease relationship. This is known as the crude or unadjusted relative risk (rate ratio, risk ratio, odds ratio).
    - Second, identify the third factor(s) which may be a confounding variable (e.g. sex, age, alcohol consumption etc)
    - Third, create strata of the confounding variable (e.g. male, female; 21-30 years, 31-40 years, 41-50 years; does not ever drink, drinks <u>≤14 units a week, drinks>14 units a week</u>, etc)
    - Fourth, calculate the relative risk for the exposure-disease relationship within each stratum of the confounding variable (e.g. RR among men, RR among women; OR RR among those who never drink, RR among those who drink <u>≤</u> 14 units, RR for those who drink >14 units)
    - Fifth, calculate the adjusted relative risk by taking the weighted average of the stratum-specific relative risks.

- Standardisation – is like stratification (see appendix).

- Statistical Modelling - multivariable regression models are useful when it is necessary to adjust for several confounders at once.

Accurate and precise measurement of the confounding factors will improve the extent to which the confounding can be controlled for in the analysis stage.

1.3. Stratification example

Suppose a study was set up to investigate whether an occupational exposure to Chemical X was associated with an increased risk of lung cancer. Researchers enrolled healthy participants, classified the participants by exposure status, and followed them up for five years. The results of this study are shown in Table 1.

Table 1: Incidence of lung cancer over 5 years among workers exposed or unexposed to Chemical X

		Exposure to Chemical X		Total
		Yes	No	
Lung cancer	**Yes**	480	360	840
	No	83,520	95,640	179,160
	Total	84,000	96,000	180,000

$$Crude\ incidence\ risk\ ratio = \frac{480/84000}{360/96000} = 1.52$$

The crude analysis shown in Table 1 shows that workers exposed to the Chemical X had 52% higher risk of developing lung cancer than those not exposed. However, we know from the literature that smoking is a risk factor for developing lung cancer. Before concluding that the chemical substance is associated with an increased risk of lung cancer, we need to exclude the possibility that smoking, rather than the occupational exposure, accounts for any of the observed association. In other words, we need to see whether this third factor, smoking, is a potential confounder. To do this, we need to examine the data separately for smokers and non-smokers. That is, we need to stratify the data by smoking habits.

Table 2a: Association of Chemical X and incidence of lung cancer among smokers

		Exposure to chemical substance		Total
		Yes	No	
Lung cancer	**Yes**	80	160	240
	No	3,920	15,840	19,760
	Total	4,000	16,000	20,000

$$Incidence\ risk\ ratio\ among\ smokers = \frac{80/4000}{160/16000} = 2.0$$

Table 2b: Association of Chemical X and incidence of lung cancer among non-smokers

		Exposure to chemical substance		Total
		Yes	No	
Lung cancer	**Yes**	400	200	600
	No	79,600	79,800	159,400
	Total	80,000	80,000	160,000

$$Incidence\ risk\ ratio\ among\ nonsmokers = \frac{400/80000}{200/80000} = 2.0$$

There are statistical procedures to take the weighted average of the stratum-specific risk ratios and calculate an adjusted risk ratio, which you may learn how to do in a statistics module. In this example, both stratum-specific risk ratios are 2.0, and so the adjusted risk ratio is going to be 2.0. Therefore, the conclusion of the analysis above is that those who had exposure to Chemical X had 2.0 times the risk of lung cancer over 5 years, relative to those who were not exposed to Chemical X, after controlling for the confounding effect of smoking.

In this example the crude effect of Chemical X was to increase the incidence risk by 52% (RR=1.52) but after adjusting for smoking the risk was increased by 100% (RR=2.0). This

suggests that workers exposed to Chemical X were less likely to smoke than those who were not.

Please note that there is no formal statistical test to confirm whether confounding is present or not in any given analysis. One must establish that the third factor satisfies the three criteria for confounding, then uses stratification to calculate the stratum-specific RRs, and then observes whether the adjusted RR differs from the crude RR.

2. Could the observed effect be causal?

If chance, bias, and confounding do not seem to explain the observed association, can one conclude that the association is likely to be **causal**? In a paper published in 1965, Bradford Hill listed aspects that should be considered when assessing whether an association is likely to be causal:

2.1. Temporal relationship

This is the only *essential* criterion. For a hypothesized risk factor to be the cause of a disease, it has to precede the disease. Temporality is easier to establish from cohort studies and RCTs, as the measure of exposure happened before the outcome has occurred. Temporarily is more difficult to establish from cross-sectional or case-control studies, when measurements of the possible cause and the effect are made at the same time.

Under this heading it is useful to mention the term "**reverse causality**". Some outcomes may result in changes in exposure (rather than the other way around) and this needs to be kept in mind when interpreting associations. For example, people with colon cancer are likely to change their diet, so any association between current diet and cancer risk may be explained by reserve causality (or reverse association).

2.2. Plausibility

The association is more likely to be causal if it is consistent with other knowledge (e.g. animal experiments, biological mechanisms, etc). However, lack of plausibility may simply reflect lack of scientific knowledge.

2.3. Consistency

If similar results have been found in different populations using different study designs then the association is more likely to be causal since it is unlikely that all studies and all designs were subject to the same type of errors. However, a lack of consistency does not exclude a causal association since different exposure levels and other conditions may reduce the impact of the causal factor in certain studies.

2.4. Strength

The magnitude of the relative risk is a measure of the strength of the association, i.e the distance from 1.0. A strong effect size is stronger evidence of a causal relationship than a small effect size, as the latter is more easily accounted for by confounding or bias.

2.5. Dose-response relationship

Further evidence of a causal relationship is provided if increasing levels of exposure lead to increasing risks of disease.

2.6. Specificity

If a particular exposure increases the risk of a certain disease but not the risk of other diseases, then this is evidence in favour of a cause-effect relationship. However, one-to-one relationships between exposure and disease are rare, and lack of specificity should not be used to refute a causal relationship.

2.7. Reversibility

When the removal of a possible cause results in a reduced risk of disease, the likelihood of the association being causal is strengthened. Ideally, this should be assessed by conducting a randomised intervention trial. However, for many exposures such randomised trials are unethical, unacceptable, or impractical.

2.8. Coherence

The putative cause-effect relationship should not seriously conflict with our knowledge of the natural history and biology of the disease.

2.9. Analogy

The putative causal relationship will be further supported if there are analogies with other (well-established) cause-effect relationships.

3. Effect Modification

For a discussion of effect modification and how it is distinguished from the concept of confounding, see EPM101-FE09-Section 10 (link here).

4. Final comments

In epidemiology, it is rare that one study alone will provide enough ‘proof’ that a certain exposure increases or decreases the risk of a particular disease. However, our degree of belief in the association will depend on the type of study design. For example, ecological studies can show associations, but because of their great potential for confounding they can never be used to establish causation. Well-conducted randomised trials are our best tools to assess causality, but their findings should always be interpreted in the context of all other available evidence, including evidence from other areas of research. For some exposure/disease relationships, randomised interventions are not possible or are unethical.

The table below gives an idea of the degree to which the results from different types of studies may be affected by bias and confounding:

Table 4. Probability of bias and confounding according to study design

Probability of:	Ecological	Cross-sectional	Case-control	Cohort	Randomised trial
Selection bias	N/A	Medium	High	Low	Low
Recall bias	N/A	High	High	Low	Low
Loss to follow-up	N/A	N/A	Low	High	Medium
Confounding	High	Medium	Medium	Low	Very Low

Modified from Beaglehole *et al* (1993).

References

Webb P and Bain C. *Essential Epidemiology: An introduction for Students and Health Professional*s. Chapter 8. Second Edition. Cambridge University Press. 2011.

Bailey L, Vardulaki K, Langham J and Chandramohan D, *Introduction to Epidemiology*. Chapter 9. Open University Press, 2005 (Understanding Public Health, Series editors: Nick Black and Rosalind Raine).

Hennekens CH & Buring JE, *Epidemiology in Medicine*, Chapter 3 and 12. Little, Brown and Company, 1987.

Beaglehole R, Bonita R & Kjellstrom T, *Basic epidemiology*, pp. 46-52, 71-81. Geneva: World Health Organisation, 1993.

Appendix to Lecture 5: Standardisation

Standardisation is a particular example of stratification. This approach is used mainly when working with routine data. For instance, when comparing mortality rates across several populations, standardisation is often used to deal with confounding variables such as age, gender, race/ethnicity, socio-economic status, area of residence, etc. Standardisation is also used in the analysis of cohort studies (often occupational cohorts) when disease rates in the cohort are compared with rates in the general population (the latter taken as the unexposed comparison group).

Direct Standardisation

Suppose we wanted to compare the mortality experience in Sweden and Panama in 1962. As can be seen in Table A1, the crude (all-ages) mortality rate for Sweden (9.8 per 1000 person years (pyrs)) was higher than that for Panama (7.7 per 1000 pyrs). However, before concluding that life in Sweden in 1962 was more risky than in Panama, we need to compare the age-specific rates (rates s*tratified* by age) for these two countries. In fact, these were greater for Panama in each age-category[1]. The reason for this discrepancy between the crude and the age-specific rates was due to the fact that these two populations had markedly different age-structures, with Sweden having a much older population than Panama (17% of the population in Sweden was 65+, whereas the corresponding figure for Panama was only 5%).

Table A1: Mortality by age-group in Sweden and Panama in 1962

	Sweden 1962			**Panama 1962**		
Age (yrs)	No of deaths	Population	Rate per 1,000 pyrs	No of deaths	Population	Rate per 1,000 pyrs
0-29	3,523	3,145,000	1.1	3,904	741,000	5.3
30-59	10,928	3,057,000	3.6	1,421	275,000	5.2
60+	59,104	1,294,000	45.7	2,956	59,000	50.1
All ages	**73,555**	**7,496,000**	**9.8**	**8,281**	**1,075,000**	**7.7**

The above example shows the dangers of drawing conclusions from comparisons of crude rates. The problem can be overcome by comparing age-specific rates as in Table A.1, but the presentation of such data is rather cumbersome, and it is often helpful to derive a single statistic that summarises the comparison, while allowing for differences in the age structure of the populations under study. Standardised or adjusted rates provide for this need. Let us assume we have a hypothetical population that we shall call 'standard' population. A frequently used hypothetical standard population is the "World Standard Population" which was developed by Segi and his collaborators (used in table A2). Suppose we wish to know **how many deaths would be expected in Sweden if it had the same age distribution as this standard population.** This is relatively easy to calculate (Table A2). All we need to do is to multiply each Swedish age-specific rate by the standard population figures in the corresponding age-group; the sum over all age categories will give the total number of expected deaths in Sweden if this country had the same age distribution as the standard. A summary mortality rate for Sweden, assuming the age-structure of the standard population, can then be obtained by dividing these total number of expected deaths (=683.1) by the total

[1] For simplicity, only three broad age-categories will be taken throughout this example.

person-years at risk in the standard population (=100,000 pyrs). This rate, of 6.8 per 1,000 pyrs, is called age-adjusted or age-standardised mortality rate.

Table A2. Direct age-standardised mortality rate in Sweden, 1962

Age (years)	Age-specific rate (per 1,000) (1)	No. in World Standard Population (2)	Number of Expected deaths = (1) x (2)
0-29	1.1	56,000	0.0011 x 56,000 = 61.6
30-59	3.6	33,000	0.0036 x 33,000 = 118.8
60+	45.7	11,000	0.0457 x 11,000 = 502.7
Total		**100,000**	**683.1**

Age-standardised mortality rate = 683.1 per 100,000 pyrs = 6.8 per 1,000 pyrs

Similar calculations (not shown), but based on the Panamanian age-specific rates, will yield the number of deaths that would be expected in Panama if it had the same age distribution as the World standard population (n=1,019.5). The age-adjusted rate for Panama would then be 1,019.5 / 100,000 = 10.2 per 1,000 pyrs. This is higher than the age-adjusted rate calculated for Sweden, and thus gives the "correct" relationship of mortality in the two countries.

The age-standardised rate is, in fact, a *weighted average of the age-specific rates*, the weights taken from the standard population (as in the Mantel-Haenszel method, strata with bigger numbers contribute more data). Age-adjusted rates are directly comparable provided that they refer to the same standard population, i.e. the *weights* given to the age-specific rates are the *same*. In the example given here, the age-standardised death rate for Panama was higher than that for Sweden, which agrees with the picture given by the age-specific rates. We can calculate an age-standardised rate ratio by dividing the rate for Panama by that for Sweden, which yields a rate ratio of 10.2/6.8 = 1.5 (i.e. mortality was 50% higher in Panama than in Sweden in 1962 after adjustment for differences in their age structure).

Indirect Standardisation

The direct method of standardisation requires knowledge of stratum-specific rates for each one of the populations being compared. If that is not known, one will need to use the indirect method. It is also preferable to use the indirect method in small studies as this method yields more stable estimates than the direct method.

Suppose that a researcher wants to test the hypothesis that mortality from cancer in a cohort of male rubber workers is increased. She observed a total of 99 deaths from cancer in the cohort during the years 1990-4. Is this excessive? The calculations using indirect standardisation are shown in Table A3.

Table A3: Example of Indirect Standardisation

Age (years)	Person-years (pyrs) at risk in the cohort (1)	Mean annual cancer mortality rates (per 100,000 pyrs) in reference population, 1990-4 (2)	Number of Expected cases = (1) x (2)
35 – 44	15,000	61	9
45 – 54	10,000	237	24
55 – 64	9,000	456	41
All ages			**74**

Total no. expected cases (E) =74
Total no. of observed cases (O) =99
Standardised mortality ratio (SMR) = O/E=99/74=1.34

First, the researcher needs to *stratify* time at risk by age (column 1). She then needs to choose a suitable reference (standard) population in which the age-specific rates were known (column 2). In most studies, this would usually be the national or regional death rates. In this example, the researcher used mean annual cancer mortality rates for England & Wales (the country where the cohort study was based) for the years 1990-4. Cross multiplying columns 1 and 2 gives **the expected number of deaths in each age-group assuming that the cohort had the same age-specific cancer mortality rates as the general population of England & Wales for the same time period**.

Summation over all age-groups yields the total expected number of deaths (n=74). Where 74 cancer deaths were expected, the researcher has observed 99, giving an age adjusted or *standardised mortality ratio* (SMR) of 99/74=1.34. SMRs (and standardised incidence ratios, SIRs) are adjusted rate ratios that can be expressed as percentages, that is, SMR (SIR) (%)=O/Ex100. In our example, SMR (%) is equal to 134. Cancer mortality was 34% higher in the study cohort than in the general population after adjustment for age.

Please note that the direct and indirect techniques of standardisation can be applied not only to age but also to **any factor considered to be a confounder**, such as gender, race/ethnicity, socioeconomic status and area of residence.

References

Webb P and Bain C. *Essential Epidemiology: An introduction for Students and Health Professional*s. Chapter 2, pages 49-54. Second Edition. Cambridge University Press. 2011.

dos Santos Silva I. *Cancer Epidemiology: Principles and Methods*. Chapter 4 pages 70-76. IARC: Lyon (France), 1999