# **Lecture 4: Chance and bias**

# **Learning Objectives**

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

1. Recognise a range of possible explanations for an observed association between an exposure and an outcome.
2. Explain the concept of random error as applied to measures of disease occurrence and measures of effect, including the idea of sampling distributions.
3. Describe how systematic error (bias) can distort estimates of an association and what can be done to prevent it.

In **analytic** epidemiology, studies are carried out to identify exposures that may increase or decrease the risk of developing a certain disease or other outcome. When interpreting findings from an epidemiologic study, it is essential to consider how much the observed association between an **exposure** and an **outcome** may have been affected by errors in the design, conduct, and analysis. The following questions should be addressed before concluding that the observed association between exposure and outcome is a true cause-effect relationship:

1. Could the observed association be due to random error (**chance**)?
2. Could the observed association be due to systematic errors (**bias**) in the way participants were selected for the study, or in the way information was obtained from them?
3. Could the observed association be due to differences between exposure groups in the distribution of another variable (**confounder**) that was not measured or was not considered in the analyses?

In this lecture we will be dealing with the first two questions.

# **1. Could the observed association be due to random error (chance)?**

Epidemiologic studies usually involve some form of sampling. We use these samples to make inferences about the population from which the sample is taken. But, because of chance, different random samples from the same population will give different results and this needs to be taken into account. The phenomenon of sampling variation caused by random error, or chance, forms the basis of the tests you have been learning about in your statistics modules.

We can illustrate sampling variation by using an example. Suppose we conduct a study in a large population in which 50% of people smoke tobacco. From the population, if we take a random sample of 100 persons and record the number who smoke, it is unlikely that *exactly* half (i.e. 50) of the sample will be smokers. The sample will give us one estimate of the percentage of smokers in the population, but that estimate will be subject to random or sampling error. When we repeat the sampling process, we will see that some samples give us estimates which are higher and some estimates which are lower than the true value of 50%. The distribution of these repeat estimates gives us a sense of the magnitude of random error for our sampling strategy.

In general, the magnitude of the random error will depend on the size of the sample we take – the larger the sample the closer the estimate is likely to be to the true proportion of smokers. In your statistics course, the concept of sampling error for means and proportions is introduced. The concept of random error is equally applicable to the rate ratios, risk ratios, odds ratios, rate differences, and risk differences we estimate using epidemiologic study designs. Whatever measure of effect is estimated, a statistician will use the random error to calculate the confidence interval and p-value for that estimate.

# **2. Could the observed association be due to systematic error (bias)?**

The word **bias** means “deviation from the truth”. In the epidemiologic context, bias is systematic error (that is, non-random error).

Bias leads to an incorrect estimate of the effect of an exposure on the outcome of interest. The observed estimate will be either greater or less than the true value, depending on the nature of the systematic error. Statistics cannot help us solve the problem of bias and much of the effort in designing an epidemiologic study is about identifying potential sources of systematic error and taking steps to minimize their impact.

Biases can be grouped into two major types: **selection bias** and **information bias**.

## **2.1. Selection bias**

**Selection bias** occurs when there is a systematic difference between the characteristics of the people who take part in a study and the characteristics of those who were eligible but did not take part, or who took part but dropped out during the study. Selection bias occurs when this systematic difference results in an observed relation between exposure and disease that is different among those who are in the study and those who are not. For example, selection bias can occur when individuals nominate themselves to take part in a research study (called self-selection bias). People who volunteer to participate in a study are likely to be different from the rest of the population in a number of demographic and lifestyle characteristics, for example because volunteers tend to be more health conscious or better-educated. Some of these characteristics may also be risk factors for the outcome of interest.

## **2.2. Information (or measurement) bias**

**Information (or measurement) bias** occurs when measurements of disease or exposure are inaccurate. This inaccuracy is known as misclassification. Misclassification may be introduced by the observer, by the participant, or by the instruments (e.g. questionnaire or device used to make the measurements). Misclassification can be of two types: differential and non-differential.

### 2.2.1. Non-differential misclassification

Non-differential misclassification occurs when an exposure or outcome classification is incorrect for equal proportions of participants in the compared groups.

* Non-differential misclassification of outcome refers to errors in classification of outcome that are unrelated to the participants’ exposure status.
* Non-differential misclassification of exposure refers to errors in classification of exposure that are unrelated to the participants’ outcome status.

In other words, non-differential misclassification of exposure/outcome/both happens to all participants equally likely. For example, occupational studies often rely upon stored records of workers (See Lecture 6 on Cohort studies). In many instances there were no specific environmental exposure measurements that would allow accurate classification of individuals. This often meant that an individual’s ‘exposure’ group was defined by job they did. These proxy variables are just crude markers of the true exposure levels and their accuracy is limited. It is, however, unlikely that the accuracy of the job records would be different for those who developed the outcome of interest and those who did not.

Another example is a study examining the association between women’s consumption of fish oil in pregnancy and the subsequent birthweight of her child. Every month during the pregnancy, researchers asked how many times the women had eaten fish as the main meal of the day. Misclassification of fish oil consumption was likely for several reasons: a woman's recall of number of fish meals may not be accurate, and the number of meals is unlikely to be a completely accurate index of fish oil consumption, as it takes no account of the amount of fish eaten in each meal or its oil content. This misclassification would not depend upon birthweight (i.e. would be non-differential), as the fish oil exposure was determined before delivery of the baby.

In most cases, non-differential misclassification will make the exposure and the unexposed groups more alike and lead to *underestimation of the strength of the association,* when a true association exists. In other words, non-differential misclassification will bias the estimate of effect towards the null hypothesis.

The implications of non-differential misclassification depend on whether the study shows an effect or not:

* When a study has a null finding, researchers must consider whether there was actually a real effect.
* When a study has a significant finding, researchers must consider whether there was actually a stronger effect.

### 2.2.2. Differential misclassification

Differential misclassification occurs when an exposure or disease classification is incorrect for unequal proportions of participants in the compared groups:

* Differential misclassification of outcome refers to errors in classification of outcome that are related to the participants’ exposure status.
* Differential misclassification of exposure refers to errors in classification of exposure that are related to the participants’ outcome status.

Differential misclassification can bias the observed estimate either towards or away from the null.

There are two main types of differential misclassification, responder bias and observer bias.

**Responder bias** occurs when the way in which study participants supply information about exposure differs according to outcome status, or the way in which study participants supply information about outcome differs according to exposure status. For example, researchers recruited a group of women with breast cancer, and another group of other women who were healthy, and asked both groups about their history of taking oral contraception (See Lecture 7 on Case-control studies). If the women with breast cancer are more likely to remember to have ever used oral contraceptives than the healthy women, a spurious association between oral contraceptives and breast cancer will result. This is a particular type of responder bias called **recall bias**.

Responder bias can be minimised by keeping the participants unaware of the hypotheses under study and by ensuring that study participants have similar incentives to provide accurate information.

**Observer bias**. Similarly, observers who know the exposure status of an individual may be consciously or unconsciously predisposed to assess outcome variables according to the hypothesis under study. This type of bias is known as observer bias. For instance, in a trial of an exercise intervention to lower blood pressure, observers may underestimate the blood pressure of those in the exercise group or overestimate the blood pressure in those in the control group. (See Lecture 5 on Intervention studies)

One way of minimising observer bias is to keep the exposure status, or the disease status (depending on the study design), of the individual concealed from the observers (ensure observers are masked/blind**)**. But sometimes masking is not feasible as it may be obvious whether a study participant has the exposure, or the disease, or not.

## **2.3. How to identify bias in epidemiologic studies**

Bias is a consequence of defects in the design or execution of an epidemiologic study. Bias cannot be controlled in the analysis of a study and it cannot be eliminated by increasing the sample size. Table 1 highlights some of the questions that help to identify bias in epidemiologic studies.

Besides being able to identify potential sources of bias in a particular study, it is also important to be able to estimate their most likely direction and magnitude. Some strategies may be introduced deliberately into the study to assess the effect of a potential bias. For instance, in a mortality study the vital status of people who were lost to follow-up may be ascertained through routine vital statistics and their mortality compared with that of the people who did participate in the study to ascertain whether selection bias occurred. But often this type of analysis is not possible, and it is always better to strive to minimise bias from the very beginning of the study design process.

## **2.4. Checklist for bias in epidemiologic studies**

**Selection bias (all study designs):**

1. Was the study population clearly defined?
2. What were the inclusion and exclusion criteria?
3. Were non-responses, refusals, losses to follow-up, kept to a minimum?

**In cohort and intervention studies:**

1. Is the follow-up adequate? Is it similar for all groups?

**In case-control studies:**

1. Do the controls represent the population from which the cases arise?
2. Was the identification and selection of cases and controls influenced by the exposure status?

**Information bias (all study designs):**

1. Were the exposures/outcomes of interest clearly defined using standard criteria?
2. Were the measurements as objective as possible?
3. Was the study blinded/masked as much as possible?
4. Were the observers/interviewers rigorously trained?
5. Were clearly written protocols used to standardise procedures in data collection?
6. Were the study participants randomised to observers/interviewers?
7. Was information provided by the patient validated against any existing records?
8. Were any strategies built into the study design to allow one to assess the likely magnitude and direction of the bias?

# **References**

Webb & Bain, *Essential Epidemiology*, Chapter 7. Cambridge, UK: 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, *Epidemiology in Medicine*, Chapters 3, 10-12. Boston, Massachusetts: Little, Brown and Company, 1987

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

von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *PLoS Med*. 2007 Oct 16;4(10):e296.