Chapter 1

Clinical Epidemiology*

James B. Kirkbride and Annie Jeffery

Introduction

Epidemiology is typically defined as the study of the *frequency*, *distribution* and *determinants* (causes) of health-related states and events in a defined population. These may include disease, disorder, symptoms, wellbeing, causes of death, behaviours and the provision and utilisation of health services.[1] Unlike most other branches of medical science, it is chiefly concerned with understanding and improving the health and disease status of populations rather than individuals, though public health interventions to prevent disease or promote wellbeing may be targeted at a variety of levels, including the individual (e.g. smoking cessation programmes, early detection services for people at risk of psychosis), familial (e.g. parenting interventions for mental, emotional and behavioural problems in children and young people) or societal levels (fluoridation of water supplies to reduce dental caries, folic acid fortification in non-wholemeal wheat flour to reduce birth defects in children).

The concept of *epidemics* – from the Greek *epi*', meaning 'upon'; *demos*, meaning 'people' and *ic*, meaning 'pertaining to' (literally 'pertaining to what is upon the people') – dates back at least to Hippocrates's writings in 400 BCE,[2] who described the relation of the seasons to various diseases occurring in the population at the time. However, the study or discourse – *logos*, in Greek – of epidemics – that is, epidemi*ology* – first arose in the nineteenth century following the identification of bacteria and subsequent observations that epidemics were strongly associated with infectious diseases. Indeed, the study widely credited to be the first epidemiological inquiry of its kind – *On the Mode of Communication of Cholera* – famously saw Dr John Snow remove the pump handle from the Broad Street pump in Soho, London, on 8 September 1854, following a groundbreaking investigation that helped prove cholera was transmitted via contaminated water and not through the air – the prevailing theory at the time.[3]

It soon became clear that many of the methods used for tracking infectious diseases, such as accurate case identification and determining precisely when and where cases had occurred, as well as their frequency in different settings, had a far wider application across population health, extending to our understanding of non-communicable diseases including mental health problems.

Cooper and Morgan[4] provide a brief overview of the history of psychiatric epidemiology, and they credit Émile Durkheim, the French sociologist, as among the first to apply epidemiological methods in psychiatry, in his studies of suicide. Durkheim examined successive five-year average suicide rates in different European countries and showed these were remarkably constant within each country but differed widely between countries, with the Protestant North European countries having rates that were three to four times higher than the Mediterranean and presumably Catholic countries such as Italy. To test his hypothesis further, Durkheim investigated how suicide rates varied within just one country, Germany, where some provinces were strongly Catholic while others were predominantly Protestant. He showed that the Protestant provinces (less than 50% Catholic) had a relatively high mean suicide rate, of 192 per 100,000 population, while the rate for provinces with 90% or more Catholics had rates less than half this, at 75 per 100,000. Those provinces that were 50–90% Catholic fell between these values, with 135 suicide deaths per 100,000 population. Durkheim conducted similar analyses comparing suicides rates between married and divorced people or between those who were fertile against those who were childless and, even without the help of modern statistical tests, found large differences between these different social groups. This led him to conclude that suicide, as a phenomenon, was a collective act, in that it was related to societal forces, and that the Catholic religion in some way appeared to offer a degree of protection.

Although the epidemiology of suicidality is complex and multifaceted,[5] (for a comprehensive introduction), recent evidence confirms that suicide rates are influenced by societal and cultural factors. For example, in Sweden, Hollander et al.[6] have observed that rates of suicides amongst first-generation migrants were over 60% lower than in the Swedish-born comparison population, after taking into account differences in age, natal sex and family income. This suggests that migrants import a range of protective factors that lower their risk of death by suicide, including sociocultural and religious beliefs, behaviours and customs and attitudes to suicide. Most strikingly, however, in this study, rates of suicide in migrants were dependent on the length of time lived in Sweden; no deaths by suicide were reported in migrants living in Sweden

* We are grateful to Matthew Hotopf for allowing us to revise a previous edition of this chapter.

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

for less than five years, while rates then began to increase in a dose-response manner, with no differences in suicide rates observed for those who had lived in Sweden for over 21 years. These findings lend further evidence to suggest that societal forces to which people are exposed can influence risk of suicide (and potentially other adverse health outcomes[7]), as Durkheim first suspected in the nineteenth century.

In the early twentieth century, in the southern United States, there was an alarming rise in the prevalence of pellagra, a debilitating neuropsychiatric disease presenting with neurasthenic symptoms, occasionally psychoses and dementia, as well as skin rashes. It was thought the cause was a specific communicable disease, possibly because of its known association with unsanitary conditions. In 1914, the US public health authority appointed Joseph Goldberger to investigate the cause of pellagra. Goldberger first observed that, in institutions where pellagra was rife, all the cases seemed to occur only among the inmates, and none of the staff were affected. He wrote that 'this pattern seemed to be no more comprehensible on the basis of an infection than is the absolute immunity of the asylum employees'.[8] Furthermore, new cases seemed to occur among inmates who had been there for a long time and who had little contact with the outside world rather than amongst new arrivals who had recent contact with the outside world. In a more detailed survey of an orphanage in Jackson, Mississippi, Goldberger found that the pellagra cases seemed to be confined to those aged 6–12 years. He noted that the younger children (below 6 years old) received a daily ration of fresh milk, while most of those aged 12 years or over were sent out to work on the farms, where they received supplementary food. Meanwhile, those aged 6–12 years subsisted only on the orphanage diet. To confirm his hypothesis that a dietary deficiency was responsible, Goldberger then conducted a dietary survey of households in seven villages in South Carolina, where the prevalence of pellagra was known to be very high. There were no cases of pellagra in households consuming more than 19 quarts of fresh milk per fortnight, but there was a 22.5% rate among households consuming less than one quart per fortnight. A similar pattern was found for the consumption of fresh meat.

This simple but well-designed survey, based only on good case identification and the ascertainment of the age and occupational distribution of cases and non-cases followed by a basic dietary survey, led to the identification of the probable cause of pellagra as a specific dietary deficiency. The disease was then easily prevented by ensuring an adequate supply of fresh milk and meat protein, and all this was clarified long before laboratory scientists had isolated vitamin B6 and identified its deficiency as the definitive biochemical cause of pellagra.

There are two main branches of epidemiology. The first branch provides a framework to *describe* diseases (or, more correctly for psychiatry, disorders, syndromes or dimensions) as they arise in the population. This branch encompasses studies that characterise the *frequency* and *distribution* of disorders such as psychotic disorders, anorexia or depression, or suicide rates as in Durkheim's studies. It is important to know whether disorders are on the increase or in decline and whether they vary dramatically between countries or regions. Having this knowledge allows services to be planned but also helps develop hypotheses about possible causes. Further, it is especially important for patients and their families that their clinical team is able to describe the prognosis of disorders. How many people with first-episode psychosis make a full recovery and never require psychiatric treatment again? How many will develop severe symptoms and require psychiatric care for the rest of their lives?

The second main branch of epidemiology deals with identifying and establishing the *determinants* of a disorder, using *analytic* study methods. It is centrally concerned with establishing whether a putative risk (or protective) factor is causally related to changes in the risk of experiencing a disorder or disease characteristic under study at the population level. Does removal or prevention of exposure to a given risk factor, such as high-potency cannabis, reduce the risk of a disorder, such as psychosis? The studies by Goldberger on pellagra described earlier are one early example of analytic studies in epidemiology. Such analytic studies test hypotheses that exposures (or risk factors) cause disorders or, once the disorder is established, examine whether the exposure (such as different forms of health care or treatment interventions) causes better or worse outcomes. As such, randomised controlled trials, which primarily assess whether an intervention (typically a therapeutic intervention but sometimes extending to social interventions) improve health outcomes, are a special type of analytic study design used in epidemiology. These *experimental* study designs (see 'Randomised Controlled Trials (RCTs)' later in the chapter) are differentiated from *observational* studies in epidemiology based on how the exposure is assigned to the population under study. In experimental designs, the investigator assigns the exposure (often randomly); in all other observational designs, the exposure is not assigned by the investigator, who instead observes what has occurred (or will occur) in the population under study. Inferring causal effects from observational studies requires great care, because of hidden differences that are often present between those who are, and are not exposed to a given risk factor under study. We will explore this critical issue in greater detail throughout this chapter.

As common to many scientific disciplines, analytic epidemiology is centrally concerned with establishing whether an association between two measured variables (typically referred to as exposures and outcomes) is causal. As in all quantitative disciplines, such associations are estimated statistically, but as the old adage goes, correlation does not imply causation, and special *causal inference* techniques are required to evaluate the likelihood that any given relationship is causal. While a vital issue for all analytic studies, causal inference is a particular challenge in observational epidemiology due to the inherent limitations of different study designs along with the (often hidden) roles played by various *biases*, which can nullify or even reverse apparently causal relationships between a risk factor and disorder. Later in this chapter, we provide an

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

overview of both *traditional* and *contemporary* causal inference methods in epidemiology that can be used to investigate causality. The last two decades have seen an explosion in the development and application of contemporary causal inference methods (for an excellent primer, see for example, Hernan & Robins[9]), which – under certain (strong) assumptions – can be applied to observational data to strengthen the plausibility that a given association between an exposure and outcome is causal (see 'Causation', later in this chapter).

Exposures and Outcomes

In most studies, investigators measure three main things:

- Exposures
- Outcomes
- Potential confounders, which are other factors that may influence both the exposure and the outcome

The term 'exposure' encompasses a wide range of different factors that might be important in the aetiology of a disorder. These can include simple demographic variables such as age and gender; biological entities such as genotype, intra-uterine infection and brain abnormalities; psychological variables such as experiences of parenting; or social factors such as life events, deprivation and income inequality. Clearly, these exposures may be measured in many different ways, but the methodological principles behind linking exposure to outcome are essentially similar.

The term 'outcome' is also used broadly – to psychiatrists, the most obvious outcomes are diagnostic categories such as schizophrenia, depression or anorexia nervosa. While some researchers may choose to 'split' psychiatric categories into diagnostic groups as defined in ICD-11[10] or DSM-V, others may 'lump' together broad categories (e.g. 'psychotic disorders', 'common mental disorders' or 'eating disorders'). Increasingly, it is common in both clinical practice and in research to investigate the *dimensions* underlying different presentations, recognising that there are continua of experiences in the population (from no mental health symptoms to mild, moderate or severe symptoms) and that there is often phenomenological overlap in symptom dimensions across traditional categorical diagnostic boundaries. Further, in some countries, clinical practice increasingly seeks to avoid formal diagnoses in the early stages of mental illness to avoid stigma (particularly as most psychiatric conditions begin in adolescence) and allow a clear clinical presentation to unfold. The latest iteration of the *Diagnostic and Statistical Manual*, DSM-V[11], explicitly recognises dimensional approaches to mental illness. Thus, depending on the research question, investigators may choose to study clinical disorders, sets of psychiatric conditions or dimensions of psychopathology.

Potential *confounders* are described in more depth later but are essentially any variable that may present alternative explanations for the observed relationship between exposure and outcome; in causal language, they are referred to as common causes of the exposure and outcome.

Development of Measures: Reliability and Validity

All quantitative research involves the measurement of variables, which may be outcomes or exposures. In physical science, there are often objective criteria on which to base measurement (weight, length, electrical resistance, etc). In psychiatry (and much of medicine besides), such objective, external measures are lacking, and our measurement is therefore particularly prone to error. In developing questionnaires, rating scales or diagnostic interviews, it is necessary to assess their reliability and validity.

Reliability

There are two main types of reliability: inter-rater reliability and test–retest reliability. The term is also used, though, to describe the 'internal' integrity of an instrument – that is, inter-item reliability.

Inter-rater Reliability

Inter-rater reliability indicates whether two or more researchers using the same measure on the same subject will gain similar answers. The measurement of inter-rater reliability depends on the type of variable generated by the questionnaire. If it generates a binary outcome, such as the presence or absence of a specific diagnosis, reliability could be described as the *percentage agreement* between the two researchers. However, this would not take into account agreements that happened just by chance. Instead, *Cohen's kappa* takes into account that some of the observed agreements would be expected by chance. Kappa can vary anywhere between −1 and +1, where positive values indicate above-chance agreement (1 indicates perfect agreement) and negative values indicate below-chance agreement.

If the measure generates an ordered categorical outcome – for example, levels of certainty about the presence of a diagnosis (definite, probable, possible, absent) – a *weighted kappa* can be used. This gives more emphasis to serious levels of disagreement between raters than to trivial ones.

If the measure is a continuous variable, such as a symptom score, the *intraclass correlation coefficient* may be used, which will take a value between 0 and 1, with 1 again indicating perfect agreement.

Test–Retest Reliability

Test–retest reliability involves the same rater using the same measure to assess the same subject twice over an interval of time. The same parameters can be used as for inter-rater reliability. Test–retest reliability is important for measures that assess stable psychological traits, such as personality or intelligence, but is less useful for gauging the reliability of psychological symptoms, as these fluctuate over time.

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

Inter-item Reliability

Split-half reliability describes the integrity or coherence of a questionnaire and assesses whether the questions assess the same underlying construct. It can be measured by calculating a correlation between the scores of the first and second half of the questionnaire or between odd-numbered versus even-numbered questions. Alternatively, Cronbach's α can be used, which provides the average correlation between all possible ways of splitting the items.

Validity

Validity refers to the extent to which an instrument (which in this context usually means a questionnaire or interview) *actually* measures what it sets out to measure. There are three main types of validity:

- *Content validity* (which includes 'face validity') refers to the degree to which the measure covers what it is meant to cover – for example, one would expect a measure of depression to include items on low mood, anhedonia and fatigue.
- *Construct validity* is a more abstract term meaning the degree to which results from a measure fit with underlying theoretical constructs pertaining to that measure. For example, if the phenomenon under study changes with age, one would expect the results of the test to reflect this.
- *Criterion-related validity* (*concurrent* or *predictive*) is the degree to which the measure compares with an alternative criterion. In concurrent validity, the measure is compared with a 'gold standard', and the results are summarised as the sensitivity and specificity of the measure (these are discussed further in the chapter). Predictive validity is assessed by how well the measure is able to *predict* a subsequent outcome that fits into the construct being examined – for example, an IQ test used in children should go some way to predict future academic performance, or a measure of suicidal ideas should be able to predict future suicide attempts to some extent.

Concurrent Validity: Sensitivity and Specificity

Table 1.1 gives the overall framework for calculating a range of common parameters for assessing the concurrent validity of an instrument against a gold standard, including *sensitivity* and *specificity*.

The formula for these measures are given below:

$$Sensitivity = \frac{a}{a+c}$$

$$Specificity = \frac{d}{b+d}$$

$$Positive\ predictive\ value = \frac{a}{a+b}$$

$$Negative\ predictive\ vale = \frac{d}{c+d}$$

$$Likelihood\ ratio\ (LR)\ of\ positive\ result = \frac{sensitivity}{(1-specificity)}$$

$$Pretest\ odds\ of\ disorder = \frac{a+c}{b+d}$$

$$Post\ test\ odds\ of\ disorder = \frac{a+c}{b+d} \cdot LR$$

$$Post\ test\ probability\ of\ disorder = \frac{Post\ test\ odds}{(1+post\ test\ odds)}$$

It will be easiest to define and discuss sensitivity and specificity in relation to an example and some actual numbers. Say a general practitioner (GP) decided to screen all attenders with the 12-item General Health Questionnaire (GHQ-12) to improve their detection of common mental disorders. It would be important to know the concurrent validity of the questionnaire – in other words, how it performs against a 'gold standard' psychiatric interview. The GP might therefore compare the results of the GHQ-12 with those on the 'gold standard' Revised Clinical Interview Schedule (CIS-R), which is a structured diagnostic interview. It is then possible to give the sensitivity and specificity of the GHQ-12 (in relation to the CIS-R). Say the doctor uses both measures on 49 patients, and the results are as shown in Table 1.2.

Note, first, that the frequency of psychiatric disorders rated on the CIS-R is high (nearly half the patients score positive). Note also that the frequency of patients who are positive on the GHQ-12 is higher still – this is usually the case when a questionnaire is being used to detect possible cases and indicates that at least some of the 'positives' on the questionnaire are false positives. *Sensitivity* is a measure of the ability of an instrument to pick up genuine cases – in this instance, the sensitivity is close to one (0.96, see below), indicating that the GHQ-12 identifies nearly all those who are true cases.

Table 1.1 Definitions of sensitivity and specificity

		Gold standard		
		Positive	**Negative**	**Total**
Our instrument	**Positive**	a	b	a + b
	Negative	c	d	c + d
	Total	a + c	b + d	a + b + c + d

Table 1.2 Example calculations of sensitivity and specificity for a sample of 49 patients

		CIS-R (Gold standard)		
		Positive	**Negative**	**Total**
GHQ-12	**Positive**	23	9	32
	Negative	1	16	17
	Total	24	25	49

Specificity is a measure of the ability of an instrument to identify correctly those who are free from the disorder. Here the specificity is much lower (0.64), indicating that the GHQ-12 was performing less well. There is a play-off between sensitivity and specificity: the more sensitive a measure is, the more likely it is to also pick up false positives, and *vice versa*. The positive predictive value (0.72) describes the chances that an individual scoring positive on the test will actually have the disorder when the gold standard is applied. Similarly, the negative predictive value (0.94) is the chance that an individual who tests negative will be free from the disorder. Note that the positive and negative predictive values are sensitive to the frequency of the disorder under study. If the disorder is very rare, it is likely that a higher proportion of those who test positive will not have the disorder compared with when it is very common.

$$Sensitivity = \frac{23}{24} = 0.96$$

$$Specificity = \frac{16}{25} = 0.64$$

$$Positive\ predictive\ value = \frac{23}{32} = 0.72$$

$$Negative\ predictivevale = \frac{16}{17} = 0.94$$

The Odds, the Likelihood Ratio and Proportion

The GP knows from past experience that a high proportion (in fact, 49 per cent) of his patients have a psychiatric disorder. How much of a difference does the test make? The likelihood ratio of a positive value gives us an idea of the 'added value' that the test makes, but to use it, we also have to calculate the *odds* of a patient having a disorder. As per the formulae above, this leads to the following values:

$$Likelihood\ ratio\ (LR)\ of\ positive\ result = \frac{0.96}{0.36} = 2.67$$

$$Pretest\ odds\ of\ disorder = \frac{24}{25} = 0.96$$

$$Post\ test\ odds\ of\ disorder = 0.96 \cdot 2.67 = 2.56$$

Note that the odds are different from the probability, and the odds are calculated as the proportion with the disorder divided by the proportion without a disorder (here, 24/25=0.96). The *likelihood ratio* of a positive test is defined as the amount by which a positive test result increases the odds of a patient having the disorder – in this case, 2.67. If a patient scores positive on the GHQ-12, the odds that they have a disorder now increases by 2.67-fold to 2.56. What does this mean in terms of proportions? As above, we now use the formula for the post-test probability of disorder, given as:

$$Post\ test\ probability\ of\ disorder = \frac{2.56}{3.56} = 0.72$$

Hence, the positive test result on the GHQ-12 has changed the probability that the patient has a disorder from 49% to 72%.

Measures of Disorder Frequency: Prevalence and Incidence

One of the basic functions of epidemiology is to describe the frequency of disorders in the population. Knowledge about the burden of disorders in a given population should be the founding principle on which clinical and public health resources are based. There are two main measures of frequency: *prevalence* and *incidence*.

Prevalence

Prevalence is the total number of individuals with the disorder divided by the population from which they are drawn:

$$Prevalence = \frac{Total\ cases}{Total\ population}$$

Prevalence estimates will include some patients who have had the disorder for many years and others who have only just developed it. Prevalence is therefore a function of the number of new cases developing the disorder over a given time period (i.e. the incidence rate) and the average chronicity of the disorder (i.e. its average duration). It is worth noting, therefore, that the prevalence of the disorder will be affected by both recovery and death rates as a result of the disorder – two pertinent and pernicious issues in psychiatry; a higher recovery rate (fewer cases) would reduce prevalence as, paradoxically, would a higher death rate as a result of the disorder (fewer cases).

Two subtypes of prevalence exist: *point prevalence*, which is the proportion of the population who have the disease at the point in time when it is measured, and *period prevalence*, which is the proportion of the population who have experienced the disorder over a defined interval. In psychiatry, there are advantages to using period prevalence as many disorders relapse and remit, and a point prevalence may not reflect the true proportion of the population who have been affected by the condition under study. The two most common timescales for estimating period prevalence in psychiatric epidemiology are annual and lifetime prevalence.

Lifetime prevalence is the proportion of people in the total population who have ever experienced a disorder in their lifetime. There has been considerable controversy over the accuracy of lifetime prevalence estimates when obtained from psychiatric interviews. The problem with such estimates is that they depend on the recall of clusters of symptoms (e.g. for depression: low mood, anhedonia, sleep disturbance) many years before. Recall of such complex information is likely to be very inaccurate. Alternative sources – such as prospectively recorded cases in case registers – may be free from issues of recall bias (see 'Bias') but may still lead to underestimates of lifetime prevalence if case identification is based purely on clinical contact and diagnosis.

Lifetime prevalence is frequently confused with morbid risk of a disorder. Lifetime prevalence is an estimate of the total proportion of people alive at a given point in time (or at a

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

given age) who have ever experienced the disorder of interest (that is, it is dependent on survivorship to that point in time or age). Morbid risk, by contrast, is an estimate of the proportion of disease-free people at a given point in time or age who will go on to experience the disorder of interest over a certain time period or by a certain age.

Finally, a common error in reporting prevalence estimates in the literature is to describe them as prevalence rates; any estimate of prevalence is a proportion in a fixed population (denominator), such as the percentage of people surveyed today who have ever experienced depression.

Incidence, Incidence Rates and Cumulative Incidence

In epidemiology, the term '*incidence*' strictly describes the *number* of individuals in an initially disease-free population who develop the disorder of interest for the first time within a specific time period. For example, there were 80 new cases of schizophrenia in the at-risk population of 400,000 people in 2022. Colloquially, however, *incidence* is used synonymously with the term *incidence rate*, which estimates the rate at which new cases occur within a population:

$$Incidence\ rate = \frac{Number\ of\ new\ cases}{Population\ at\ risk * time\ at\ risk}$$

For example, in the aforementioned population, the incidence rate was 20 new cases of schizophrenia per 100,000 people at risk in 2022. Note here the important concept of the 'population at risk', which includes only individuals who have never had the disorder. It excludes people who have previously had the disorder or those who would not be at risk of developing the disorder. The latter issue may seem trivial, but if someone with an organic brain disorder were to develop psychosis symptoms in the earlier example, there may be a high probability that those symptoms were caused by the organic disorder. Since that person would not meet diagnostic criteria for non-organic psychotic disorders, they could never have been 'at risk', and they should not be included in the estimation of either the numerator (new cases) or denominator (population at risk) when estimating incidence rates.

Cumulative incidence, sometimes referred to as incidence risk, estimates the proportion of new cases in an initially disease-free population at risk over a given length of time. Unlike an incidence rate, the denominator for cumulative incidence is the initial disease-free population at risk, ignoring the time at risk:

$$Cumulative\ incidence = \frac{Number\ of\ new\ cases}{Initial\ population\ at\ risk}$$

Conceptually, cumulative incidence is similar to morbid risk. To illustrate the difference between these measures, we use a simplistic example of 10 individuals in a population, as depicted in Figure 1.1. These individuals are followed for one year to determine who develops a disorder. Three possible outcomes are possible for each individual: remain well, develop the condition under study, or be censored – in other words, stop contributing to the study because of death, emigration or other loss to follow-up. When calculating cumulative incidence, the problem of censoring is ignored. The numerator is all new cases of the disorder, and the denominator is the population at risk. In the study illustrated in Figure 1.1, we would state that two cases (4 and 7) developed the disorder, so the risk is 2/10 or 0.2.

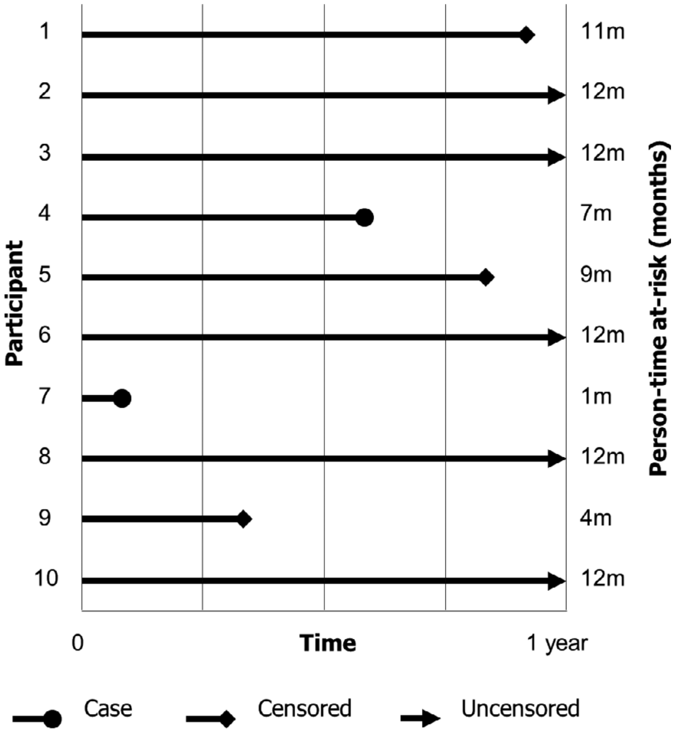

Figure 1.1 Calculation of incidence measures (see text for explanation).

When calculating the rate, a more precise estimate is made to take into account the differing amounts of time each individual spends 'at risk' of the outcome. Individuals who become ill can no longer contribute to 'time at risk', nor can individuals who die or who are otherwise censored. The denominator for the rate is the total 'person-time' at risk in the study. From Figure 1.1, only five individuals (2, 3, 6, 8 and 10) contribute an entire year of time at risk. Case 7 hardly contributes any time, becoming ill after 1 month. Case 4 becomes ill at 7 months, and cases 1, 5 and 9 are all censored, at 11, 9 and 4 months, respectively. The total person-time at risk here is 7.7 years, and the rate is 2 per 7.7 person-years at risk. Typically, we then re-express rates on a common person-years scale, such as per 100 person-years, per 1,000 person-years or even per 100,000 person-years to make it easier to compare between studies or groups. The scale is arbitrary and a function of the frequency of the disorder in the population. Here, we could express 2.0 per 7.7 person-years as 26.0 per 100 person-years, but for rarer disorders, such as psychotic disorders, we typically express incidence rates on a larger scale (i.e. a recent meta-analysis of the incidence of psychotic disorders placed this to

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

be 26.6 per 100,000 person-years).[12] Because it is a more accurate measure, this rate is the preferred expression of incidence.

Finally, in order to precisely estimate incidence rates or prevalence, note that it is important to have accurate information about the denominator. Denominator error occurs if the investigator attempts to define a population using some routinely available data (e.g. the census or electoral register), but these are inaccurate because not everyone provided information in the census or signed up to the electoral register. This may lead to an over-estimate of incidence or prevalence, even if the numerator is being accurately recorded. The strength and effect of this *bias* (see further in this chapter) will depend on the accuracy of the denominator source, whether such bias was differential or non-differential across subgroups and the absolute rarity of the disorder.

Measures of the Strength of Associations: Risk Difference, Risk Ratios, Rate Ratios and Odds Ratios

In analytical studies, an attempt is usually made to describe the *strength* of an association between an *exposure* (or risk factor) and an *outcome* (or disorder). These are referred to as *measures of effect* or *measures of association*. In cohort studies, the incidence of a disorder is compared in two groups – one exposed to a risk factor, the other not exposed. The study estimates incidence risks or rates for each group. The risk or rate difference (sometimes *excess risk* or *rate*) is the difference in risks (or rates) between the exposed and unexposed group, expressed as:

$$Risk\ difference = Risk_{Exposed} - Risk_{unexposed}$$

and

$$Rate\ difference = Rate_{Exposed} - Rate_{unexposed}$$

Risk (rate) ratios are ratios of the risk (rate) in the exposed population divided by the risk (rate) in the unexposed population, as follows:

$$Risk\ ratio = \frac{Risk_{Exposed}}{Risk_{unexposed}}$$

and

$$Rate\ ratio = \frac{Rate_{Exposed}}{Rate_{unexposed}}$$

A risk ratio of 3, for example, indicates that individuals with the exposure are three times more likely to have the outcome as those unexposed.

Use of ratio measures is more common in clinical epidemiology than difference measures, although the choice will depend on the intended goal of the researcher when setting the research question and designing the study. Risk (or rate) differences can be useful in quantifying the absolute excess incidence or risk of disorder in one group compared with another, providing valuable clinical or public health information. Difference measures are particularly useful when reporting the results from RCTs, since they provide valuable information about the absolute magnitude of benefit or harm of the treatment in those who received the intervention. These should be reported alongside ratio-based measures of effect.[13]

Another widely used measure of impact is the odds ratio, which is used especially in case-control studies. The relationship between the odds ratio and risk or rate ratios is described in detail elsewhere,[14] but it can be illustrated in the following example.

Imagine that we are interested in determining the effect of unemployment on suicide rates in men of working age. We might identify a population of 1 million men for whom we know their employment status. Assume that 5 per cent of the population are unemployed and we follow the population for one year assessing suicide rates to obtain the figures shown in Table 1.3.

From these figures, it is possible to calculate the rate ratio:

$$Rate\ ratio = \frac{36\ per\ 100{,}000}{12\ per\ 100{,}000}$$

$$Rate\ ratio = 3$$

Now let us assume that it was impossible to identify the employment status of the entire population at the start of the study, and instead a case-control design was used. In the case-control study, the exposure status for cases (i.e. people who die by suicide) and controls (i.e. people who do not die by suicide) are compared. Assuming that it was possible to identify all 132 suicides in the population and compare them with a randomly selected sample of individuals who did not commit suicide, and assuming that the rate of unemployment in this randomly selected group of controls was similar to that of the general population, we could compare the odds of exposure in the cases with that in the controls. This might generate a table like Table 1.4.

Table 1.3 Illustration of rate of suicide by employment status

	Employed	Unemployed
Number of suicides	114	18
Denominator	950,000	50,000
Suicide rate	12 per 100,000 per year	36 per 100,000 per year

Table 1.4 Illustration of rate of unemployment by suicide status in a case-control study

	Cases of suicide	Controls	Total
Total	132	264	396
Employed	114	251	365
Unemployed	18	13	31
Odds of exposure	0.158	0.052	-

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

From this, it is possible to calculate the odds ratio, where the odds ratio is defined by the odds of exposure in cases divided by the odds of exposure in controls:

$$Odds\ ratio = \frac{Odds_{Cases}}{Odds_{Controls}}$$

$$Odds\ ratio = \frac{0.158}{0.052}$$

$$Odds\ ratio = 3.05$$

In the above example, we calculated the 'exposure' odds ratio, that is, the increased relative odds of exposure (being unemployed) in cases relative to controls. In a case-control study, we could have equally calculated the 'disease' odds ratio, that is, the increased relative odds of suicide in those who were unemployed (the exposed) relative to those who were employed. This yields the same odds ratio ([18/13]/[114/251] = 3.05).

The odds ratio in this example is a close approximation to the rate ratio estimated in Table 1.3; however, the two are not identical. The odds ratio approximates to the rate or risk ratios where the outcome under study is rare, as is often the case for many psychiatric disorders. When it is not rare, the odds ratio is higher than the risk ratio. See Box 1.1 for a technical note as to why we should not estimate risk (or rate) ratios in case-control studies. An exception to this rule exists for nested case-control studies in which controls are sampled by a method called incidence density sampling; here, provided conditional logistic regression is used, the odds ratios will be equivalent to incidence rate ratios (for further introduction to this advanced issue, see Lubin and Gail).[15]

Box 1.1 Use of odds ratios and not risk ratios in case-control studies

In the example in Table 1.4, we saw how the exposure odds ratio and disease odds ratio yielded the same effect size of 3.05.

The same property does not hold if one were to estimate the 'disease' risk ratio and 'exposure' risk ratio using case-control data. For example, in Table 1.4, the risk of suicide in the unemployed group is 18 of 31 (risk = 0.581) while the risk of suicide in the employed group is 114 of 365 (risk = 0.312), leading to a 'disease' risk ratio of 0.581/0.312=1.86. However, the 'exposure' risk ratio would be estimated as (18/132)/(13/264) = 2.77.

This situation arises in case-control studies because the researcher artificially constrains the number of controls in the study by design (for example, often one control per case). Because of this, risk of disease in both the unexposed and exposed group will change as a function of the proportion of controls to cases.

Suppose now we decide to sample 10 times as many controls for our study, which, under consistent sampling to the data generated in Table 1.4, would yield 130 unemployed and 2,510 employed controls. Now, the risk of suicide in the unemployed group would be 18/148 = 0.122, and the risk of suicide in the employed group would be 114/2624 = 0.043, yielding a 'disease' risk ratio of 2.80, compared with 1.86 previously. Thus, calculation of the risk ratio is a function of the number of controls, which is decided in advance by the researcher in case-control studies, while calculation of the odds ratio remains unchanged ([18/130]/[114/2510] = 3.05).

Assuming that the proportion of exposed to unexposed controls remains consistent as the number of controls increases, the odds of exposure in controls will be unaffected by the total sample size (which is constrained by design), leading to a valid measure of effect in case-control studies.

Measures of Impact

A key question for preventive medicine is determining how much impact a risk factor has on the overall rate of a disorder. Thus, *measures of impact* provide a useful way of understanding how much disease, disorder or burden could – theoretically – be prevented in the population, if a given risk factor or exposure could be removed (e.g. if we could stop everybody from being exposed to bullying in childhood, what proportion of psychiatric disorders in the population would we prevent?). Note, that *measures of impact* are predicated on several assumptions, including that there is a *causal* association between the exposure and outcome, the exposure can be prevented and that removal of the exposure would lead to removal of the outcome in a given population. The extent to which these assumptions are valid is of considerable debate, particularly given the multi-factorial causal structure of most psychiatric conditions, where any single risk factor may be neither sufficient nor necessary to cause morbidity. We return to issues of *causation* later in this chapter.

Returning to the unemployment and suicide example in Table 1.3, we might want to know how much unemployment contributes to the total suicide rate and whether removing the exposure (i.e. providing conditions of full employment) would have a sizeable impact on suicide rates. The population attributable risk (PAR) gives an estimate of this:

$$PAR\% = \left[\frac{P_e(I_e - I_u)}{P_t I_t}\right] \cdot 100$$

where:

P_e = Number of persons exposed = 50,000
P_t = Total population = 1,000,000
I_e = Incidence in the exposed = 36 per 100,000 per year
I_u = Incidence in the unexposed = 12 per 100,000 per year
I_t = Incidence in the total population = 13.2 per 100,000 per year

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

Leading to a PAR estimate of:

$$PAR\% = \left[\frac{50{,}000 \cdot (36 - 12)}{(1{,}000{,}000 \cdot 13.2)}\right] \cdot 100$$

$$PAR\% = \left[\frac{1{,}200{,}000}{13{,}200{,}000}\right] \cdot 100$$

$$PAR\% = 9.1\%$$

In other words, this shows us that, of the 132 suicides that occurred in the year, 9.1% were attributable to unemployment. This implies that if unemployment was removed as a risk factor, the suicide rate would fall by this amount. Thus, the population attributable risk can be defined as the proportion of a population's experience of a disorder that can be explained by the presence of a risk factor. As noted earlier, PAR has major practical limitations. In this example, it may not be feasible to remove unemployment entirely from the population, or the risk factor itself may not be a cause of suicide. For example, people may be unemployed because of other underlying health conditions, including mental health issues such as depression, and these issues may remain even if someone returned to work.

For a more comprehensive overview of the strengths and limitations of such measures of impact and to learn more about various methods for estimating the PAR (and other measures of impact) that exist, we refer the reader elsewhere.[14]

Study Designs

Ecological Studies

The ecological study design looks for population-level associations between the rates of a disease or disease outcome and the rates of a given exposure. This type of study design can be used to compare associations between different geographical locations, across time or between different groups such as migrant groups or social class. This approach requires routinely available estimates of prevalence or incidence as well as data on exposure. The problem in psychiatry, and the reason that ecological studies are not a common design, is that there are relatively few reliable estimates of prevalence or incidence that apply between many populations. Two key exceptions are suicide rates and hospital admissions. An example of an ecological study assessing suicide is that by Helbich et al.,[16] who assessed the relationship between suicide rates and the proportion of green space across different municipalities in the Netherlands. The authors found that municipalities with more green space had lower rates of suicide compared to municipalities with less green space. Another example of an ecological study assessing time trends in involuntary psychiatric hospital admissions is that by Keown et al.[17] The authors assessed the relationship between annual changes in the state provision of mental illness beds in the United Kingdom and involuntary admission rates. They found that reductions in mental illness beds were associated with increases in involuntary psychiatric admissions.

Ecological studies can be useful to generate hypotheses on the aetiology of a disorder at a relatively low cost. However, because they do not measure the exposure or the outcome at the level of the individual, it is not possible to use them to link exposures and outcome at the level of the individual. Therefore, they only provide weak evidence of causal relationships. This is referred to as the *ecological fallacy*. Another problem with ecological studies is *ecological bias*. There are two ways in which ecological bias may occur. Firstly, ecological bias may occur when associations described in ecological studies can be explained by factors that might link the exposure and outcome (confounding). For example, if it was found that suicide rates were highest in areas with the most developed mental health services, a naive interpretation would be that mental health services are bad for mental health and have caused this excess. An alternative explanation is that there are unmeasured confounders, such as social deprivation or urban environments, which are associated with both suicide and the extent of local mental health services.

The second way in which ecological bias may occur is when the effect of the exposure is modified by another factor that varies between populations (effect modification). For example, if a study found that suicide rates were lowest in areas with more green space, this could be modified by the level of perceived safety – areas with high perceived safety may benefit from more green space, whereas this may not be the case in areas with low perceived safety where green space is not utilised.

Despite these concerns, ecological studies can reveal important trends in psychiatric outcomes at a population or group level and across time. This information can be valuable for the planning of health services and public health initiatives, as well as hypothesis generation.

Cross-sectional Studies

Cross-sectional studies examine health outcomes within a defined population at a particular point in time. They are usually survey-based and are conducted on individuals, rather than at the group level like ecological studies. They can be used to assess the frequency of disease occurrence (prevalence) and the distribution of disease occurrence (e.g. by sex, age, ethnicity or social class). There are several important examples of large cross-sectional studies in psychiatry, such as the UK Adults Psychiatric Morbidity Survey[18] and the WHO World Mental Health Survey.[19]

The first step in the design of a cross-sectional study is the identification of a population. For the purposes of most studies, population means individuals living within a defined geographical area. However, it can be any group of individuals of interest to the researchers, as long as that group can be defined in a reproducible way. Thus, cross-sectional studies may be carried out within specific settings, such as primary care or general hospital outpatient departments, and specific populations in these settings, such as among employees of a firm or pupils within a school. In some circumstances, the researcher may be interested in defining a population of individuals with a disorder – such as patients with schizophrenia – and

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

measuring the prevalence of another disorder – such as tardive dyskinesia – within this group.

As it is not always possible to survey an entire population, cross-sectional surveys are typically conducted using samples from an accessible subset. The key to making valid inferences using a study sample is to ensure that it is representative of the *target population*. Thus, if a cross-sectional study of school refusal was carried out, it would clearly be important not to limit the interviews to those children attending school as the group of most interest are those least likely to be there! Another example might be a cross-sectional study that interviewed individuals within their own home. If the survey was performed during working hours, it is likely that the healthiest members of the community would be at work, and the survey would exaggerate rates of illness as a consequence. Relevant groups (e.g. children who do not attend school or household members in full-time employment) should be identified in advance so that efforts can be made to ensure their inclusion. Random sampling is then preferable to maximise the representativeness of the sample.

Although cross-sectional studies can be used to measure associations between risk factors and disease outcomes, because both exposure and outcomes are measured at the same time point, the direction of causation may not be clear. However, there are a number of important examples where cross-sectional studies have been repeated with the same participants over time (e.g. the UK Household Longitudinal Study ('Understanding Society'), the English Longitudinal Study of Ageing). These are termed *panel studies* and are in essence, a hybrid form of cross-sectional and cohort study.

Cohort Studies

Cohort studies examine the relationship between exposures and subsequent health outcomes. In a cohort study, the sample is defined according to its *exposure status* and followed up over time to determine who develops the disorder(s) of interest. The key strength of the cohort study is its longitudinal design, which means that participants are assessed for the exposure before the onset of the disorder. Thus, cohort studies can usually give an insight into the direction of causation (see later) and are not susceptible to recall bias. Cohort studies allow rare exposures to be studied and can assess the effect of such exposures on multiple outcomes. In psychiatric epidemiology, cohort studies identify groups of people exposed to risk factors (such as childhood maltreatment, substance abuse, workplace stress or a history of depression) and compare the incidence of mental health outcomes with that among a non-exposed group. The analysis of a cohort study then involves the calculation of a risk ratio or rate ratio (see previous discussion).

The cohort study is best suited to situations where the outcome is common. For rarer outcomes (such as suicide and schizophrenia), cohort studies, unless very large, may have more limited utility. To illustrate this, suppose that a research team designs a cohort study to determine the effect of birth asphyxia on schizophrenia. They may identify babies with birth asphyxia (the 'exposed' group) and babies without such a history (the 'unexposed' group). They then have to follow the babies until adulthood in order to see whether any of them have developed schizophrenia. Assuming that by 25 years of age, the risk of schizophrenia is 0.5 per cent in those without birth asphyxia, the team would have had to follow (on average) 200 babies for each individual with schizophrenia in the unexposed cohort. In order to have a reasonable chance of detecting a twofold increase in the risk of schizophrenia over the course of the study, they would have had to follow over 10,000 individuals for 25 years. This example illustrates that cohort studies can be very expensive and time-consuming, especially if the outcome is rare. Cohort studies need to follow-up as many of the original sample as possible. *Non-response bias* (see 'Non-response Bias' in this chapter) is therefore a major concern in psychiatric cohort studies, as the individuals who cannot be traced may be the ones of most interest. For example, individuals with schizophrenia frequently become homeless or may not be cooperative with requests to participate in research. In the reporting of these studies, the investigators should describe the characteristics of those who could not be traced and how they differ from those who were traced.

Two approaches can be used to overcome some of these difficulties. The first is the use of large population-based cohort studies. In the UK, there are several large birth cohort studies that follow individuals born in a certain year over the course of their lives.[20–22] There is also, for example, the English Longitudinal Study of Ageing, which identified a sample of 11,391 people over the age of 50 in the year 2002 and continues to follow-up these individuals every two years. These cohort studies have looked at many different aspects of health, and because of their size and inclusion of relevant exposures, they have provided important data for psychiatric epidemiologists.[23,24] The second common approach is the retrospective cohort study. To return to the example of birth asphyxia and schizophrenia, instead of following babies born now, the investigators could examine the hospital records of babies born 25 years ago, and – provided sufficient information on asphyxia was available – could then trace the babies to identify individuals who had developed schizophrenia. This is a cheaper approach because the long follow-up time is not required. Population registers, such as those in the Nordic countries,[25] contain health information for all citizens stored under a unique personal identity number – these registers enable easier tracing of whole populations over long periods of time and are ideal data sources from which to conduct retrospective cohort studies. For more detail about the design, strengths and limitations of cohort studies, see Chapter 3.2 on the 'Causes of Depression' by Lewis, Lewis and Srinivasan.

Prognostic Studies

Studies on prognosis essentially use a cohort design in which the participants are patients with a disorder who are followed over time. There is usually no comparison group, as such

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

studies are essentially descriptive – giving insights into the natural history of the disorder rather than its cause. The main methodological consideration is ensuring that an *inception cohort* is defined, meaning that to be included, patients must be as close as possible to the start of their first episode of illness. Most psychiatric disorders have a fluctuating course, with relapses and remissions. If a study assessing the prognosis of psychotic illness gathered a sample of individuals at different stages of their illness, it would tend to give an overly pessimistic view of prognosis because it would preferentially include individuals whose illness had an established chronic course. Determining that the cohort of individuals are all in their first episode ensures that those who get better quickly and never suffer further symptoms are included.

Another consideration with such studies is that the sample should be truly representative of the general population. If patients are recruited from specialist centres, there may be important *referral biases*, where more unusual cases are included, perhaps with a poorer outcome. For example, many of the earlier prognostic studies in the UK, for example of depression, were conducted from the Maudsley Hospital, which is not only a tertiary referral centre but also has an inner-city catchment area, both factors that may skew the outcome in a negative direction.

Case-Control Studies

Like cohort studies, case-control studies examine the relationship between exposures and health outcomes. Unlike the cohort study, where the sample is defined according to its exposure status, in a case-control study, the sample is defined according to its *outcome status*.

Cases with a disorder or outcome of interest are compared with individuals who are free from the disorder or outcome. An example of a case-control study in psychiatric epidemiology is that of Jongsma et al.,[26] where the authors recruited 1,130 cases with schizophrenia and 1,497 controls without schizophrenia, then compared a range of exposures between these groups, including ethnicity and social disadvantage. Unlike cohort studies, case-control studies are useful for rare disorders, and it is possible to determine the relationship between many different exposures and the disorder under study. Case-control studies are usually quicker and cheaper to perform than cohort studies because the disorder has already occurred, and it is not necessary to follow individuals over many years. Unless very large, case-control studies are not useful for rare exposures because insufficient cases and controls will have experienced them to make useful comparisons. The analysis of the case-control study involves a comparison of the odds of exposure in the cases compared with the controls – and is expressed as the odds ratio (see previous discussion).

The most important issue in case-control studies is the selection of both cases and controls. The key problem is *selection bias*, which occurs when the risk factor under study has an effect on the likelihood that the individual will be recruited to the study. This can work for both cases and controls. For example, some neuroimaging studies in psychiatry involve selecting patients with severe chronic psychotic illness from 'centres of excellence' and comparing them with controls who may be PhD students from the same centres. For both cases and controls, equal and opposite selection factors may generate misleading results. Cases may be unlikely to give a true representation of psychotic illness because those most readily available tend to be those with chronic symptoms (an instance of *prevalence bias*, discussed later). The controls are unlikely to represent the typical 'normal' brain because they have been drawn from a highly educated sample. For this reason, much emphasis is placed on attempting to select as representative a sample of cases as possible. The key to the selection of controls is that they should be drawn from a similar population and be similar to the cases in all respects apart from the disorder under study.

Depending on how and when the exposure is assessed, case-control studies may be unable to determine the direction of causality and may be susceptible to recall bias. However, this may not be the case when there is a clear temporal sequence (e.g. exposure to childhood maltreatment and the outcome of substance abuse in adolescence or the exposure to domestic violence and the outcome of suicide). Recall bias may also be overcome if exposures are identified through, for example, medical records.

Randomised Controlled Trials

In the randomised controlled trial (RCT), interventions to treat (or sometimes prevent) a disorder are compared to a placebo or to one or more other active treatments. RCTs can be used to evaluate intervention efficacy, acceptability and adverse effects. RCTs randomly assign participants to an intervention as part of the trial. To perform an RCT, the investigator should demonstrate that there is no evidence to suggest that a treatment is better than placebo or another active intervention. If one treatment was already known to be far superior to another, it would not be ethical to randomise. Unlike studies of risk factors, where it would be unethical for the investigator to assign individuals to receive a potentially harmful exposure, RCTs are ethical because the intervention is expected to do good.

Appropriately designed RCTs are the most robust research method for determining causal relationships between an intervention and outcome (for a more detailed discussion of this, see the 'Causation' section later in this chapter). The key methodological feature of the RCT is randomisation with concealed allocation. The rationale behind randomisation is that each participant has an identical chance of receiving each treatment. Then, if the trial is sufficiently large, potential confounders will be evenly distributed between the groups; this process should theoretically remove *confounding* by both observed and unobserved variables, as well as avoid *selection bias*.

In *simple randomisation*, the participants are assigned to groups in sequence according to a randomly generated number. The problem with this method is that the random

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

Simple randomisation of 48 subjects

abbaaaababbaaabbaaaaaabbabaabaabbabbbabbaaaabaab

Allocated to a = 28

Allocated to b = 20

Balanced randomisation: block size 8

abbaaabb

Figure 1.2 Distinction between simple randomisation and balanced randomisation. In the simple randomisation, the total in each group is unlikely to be balanced. In balanced randomisation, the investigator has decided to randomise within blocks of eight. In each block of eight, there must be four participants on each treatment.

groups may not be balanced: it is possible that, simply by chance, the two groups are of different size, and this is statistically inefficient. *Balanced randomisation* overcomes this by allocating participants in blocks. A typical block size would be eight, and the investigator would arrange that within this block, four participants would receive the intervention and four would be the control condition (see Figure 1.2).

Randomisation usually ensures an even distribution of important confounders between groups. However, in smaller trials, this cannot be guaranteed – by chance, there may be big differences in the distribution of confounders. To get around this, the investigator can perform a *stratified randomisation*, where the sample is divided according to the presence of the variable. For instance, in a trial of sertraline to treat depression, the baseline severity of depression was considered to be a key variable, and so, the investigators stratified the randomisation on this.[27] In *minimisation*, this process is taken a step further, and a wide range of key variables are identified; participants are then effectively matched on each of these variable to ensure that they are as similar as possible.

Concealment of allocation refers to the degree to which it is predictable to the researcher which treatment the patient will receive. If the investigator had considerable faith in a new treatment, they might consciously or unconsciously manipulate the randomisation process in order to ensure that patients with a good prognosis were assigned to the experimental treatment. Thus, the trial would be more likely to report 'positive' results. The best method is to have randomisation performed by an independent third party who is not aware of the study questions.

RCTs usually have a list of inclusion and exclusion criteria to ensure that patients entered are similar. The rationale for exclusion criteria may be to prevent the following groups from participating:

- People with contraindications for the treatments
- Clinical subtypes with particular profiles that might confuse the results (e.g. having psychiatric comorbidities with similar symptoms)
- Certain groups who are considered 'high risk' (e.g. patients with suicidal ideation – this may prevent embarrassment of the investigators and sponsors, but it is not useful to clinicians, who see such patients all the time)
- Those who might be considered to have difficulties consenting or following trial protocol (e.g. individuals with learning disabilities or cognitive impairment)

It is good practice for trials to report the number of individuals approached, the number who refused to participate or were excluded, and the number randomised. The CONSORT statement provides a widely endorsed checklist of standard reporting items for RCTs.[25]

As with cohort studies, dropouts from RCTs are a major problem. The investigators should attempt to follow everyone up, including patients who drop out of treatment. Many trials simply compare those in the two groups who have completed the trial according to protocol. This may mean that a sizeable proportion of those randomised (one-third in average antidepressant trials) are left out. This is misleading and can be a source of potential bias. A better approach is to use *intention to treat* analysis, where all randomised participants, no matter how long they were on treatment, are included in the analysis.

The analysis of RCTs depends on the nature of the outcome. Many RCTs describe results in terms of change of scores on symptom-rating scales. In these cases, it is preferable to present results as differences in the changed scores from the baseline. For categorical outcomes (e.g. recovery or admission to hospital), the approach will be similar to cohort studies, and a relative risk or rate ratio may be calculated. The number needed to treat, which expresses the number of individuals whose recovery can be attributed to the intervention, can also be calculated. This is a clinically useful measure that describes the number of individuals who would have to be placed on a treatment in order to produce one good outcome. For example, in a meta-analysis (see Box 1.2) of the antidepressant fluoxetine versus a placebo to treat depression, 45.8% of those treated with fluoxetine were considered to be in remission after six weeks, compared with 30.2% of those treated with the placebo.[25] The number needed to treat is then the inverse of the risk difference (see Box 1.2). In other words, a doctor would have to prescribe antidepressants to more than six patients (at least seven, in fact) in order for one to meet criteria for remission.

Systematic Reviews and Meta-analysis

Reviews aim to synthesise evidence on a topic of interest and are an important source of information for policy makers,

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

Box 1.2 Example calculation of number needed to treat

Risk of recovery on antidepressant	45.8/100
Risk of recovery on placebo	30.2/100
Risk difference	15.6/100
Number needed to treat (NNT)	6.41

clinicians and researchers. While narrative reviews are sometimes used to summarise a body of knowledge, this approach has been criticised because the methods are not reproducible: important articles may be missed, and the reviewer may overemphasise results that confirm his or her point of view. *Systematic reviews*, on the other hand, involve a systematic effort to identify all relevant literature; inclusion and exclusion criteria are then applied to that literature, and results are extracted in a systematic way.

Just as with primary research, systematic reviews should aim to answer a specific question and state their aims and objectives explicitly. Systematic reviews also include a 'methods' section that describes the search strategy. The reviewer performs a literature search, which will usually involve a combination of searching databases of published research (e.g. MEDLINE), tracing other articles in the reference lists of identified studies, searching clinical trial databases for unpublished trial results and so on. The results section should include information on the number of studies identified from the literature search, the numbers excluded and included, and the characteristics of the studies included. Several reporting guidelines now exist for systematic reviewing and meta-analyses, including the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; www.prisma-statement.org) and the Meta-Analyses of Observation Studies in Epidemiology (MOOSE) reporting guidelines (https://jamanetwork.com/journals/jama/fullarticle/192614). Like other forms of research, prospectively registering systematic reviews, on databases such as PROSPERO (https://www.crd.york.ac.uk/PROSPERO/), is also encouraged.

Meta-analysis is a statistical synthesis of the main results from the studies identified by a systematic review. Because randomised trials in psychiatry are often too small to give reliable information, pooling the results of many similar studies will improve the precision of the effect size. When the effect size varies between studies, meta-analysis can also be used to identify the reason for the variation. For example, different trials of the same intervention may take place in different settings (outpatient, inpatient, primary care), with disorders of differing severity or chronicity. For pharmacological treatments, the drug prescribed in different trials may have been identical, but the dosage may have been different. For non-pharmacological treatments, such as psychotherapy or trials of the way in which community care is delivered, the treatment may differ radically between trials. It is possible to use a statistical test of heterogeneity to assess whether all the trials included in a meta-analysis are 'pulling the same way'. If this test indicates that significant heterogeneity between trials exists, the researchers should investigate why this might be.

Publication Bias

An important problem with meta-analysis is *publication bias*. It is a fact of life that researchers and journal editors like to have 'positive' results. There is considerable evidence that papers that show that one treatment has a clear advantage over another are more likely to be published than those that do not. Substantial publication bias could radically alter the conclusions of a meta-analysis. Publication bias is best avoided by a comprehensive search strategy – unpublished results may be publicly available in clinical trial databases or databases of 'grey literature' (e.g. OpenGrey).

The role of publication bias can then be assessed using a funnel plot.[28] If researchers complete a large RCT, they are likely to want to see it published even if the result is negative because of the effort involved. If publication bias does exist, it is most likely to be due to small negative trials not being published. The funnel plot is a graphical representation of the size of trials plotted against the effect size they report. As the size of trials increases, they are likely to converge around the true, underlying effect size. For the large trials, one would expect to see an even scattering of trials on either side of this true, underlying effect. When publication bias occurs, one expects an *asymmetry* in the scatter of small studies, with more studies showing a positive result than those showing a negative result.

Choosing a Study Design

The choice of a study design depends on the type of question being asked, the nature of the disorder/outcome and the exposure, and the time and resources available. The first question a researcher should ask is whether the question has been answered already, and the step before any serious research project should be to identify systematic reviews on the topic or to perform one. For some types of questions, the study design may be obvious. Studies on treatment efficacy are usually best answered by an RCT or a systematic review and meta-analysis of RCTs. When the researcher wants to describe the prevalence of a disorder, cross-sectional studies provide the obvious solution. However, it is more difficult to settle a question about the aetiology of a disorder, or the potential harmful effect of an exposure (including exposure to different treatments), and the study design will often be a trade-off between methodological considerations and resources.

The question next to ask is whether there are existing sources of data. Previous research studies may have collected the data necessary to answer the question. Kandola et al.[29] were able to use data from the Avon Longitudinal Study of Parents and Children to determine whether sedentary

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

behaviour between the ages of 12 to 16 was associated with depressive symptoms at age 18, which the study showed to be the case. This was an extremely economical way of answering a well-focused question, which would otherwise have required major resources. Sometimes routinely collected data exist that are not part of a research study but which still allow the question to be answered. For rare side effects of drugs, large databases such as the UK Clinical Practice Research Datalink are an ideal resource.

If existing data do not exist, the choice of whether to use a case-control, cohort or cross-sectional study will depend on the relative frequency of the outcome and exposure and how easy they are to measure. Case-control studies manipulate the frequency of the outcome (by sampling according to participants' disorder status), and cohort studies manipulate the frequency of the exposure (by sampling according to the participants' exposure status). Thus, case-control studies are best for rare disorders and cohort studies for rare exposures.

Causation

The observational (e.g. cohort, case-control) and experimental (e.g. trial) study designs in epidemiology described earlier share the common goal of identifying whether an association between two variables is causal. This fundamental tenet of epidemiology then forms and informs the basis of effective clinical and public health intervention and policy. In observational studies, the researcher attempts to understand whether an association between exposure (a risk or protective factor) and disorder is causal; in experimental studies, the researcher attempts to understand whether an association between an intervention or treatment (a protective factor) and effect is causal. In this section, we provide a theoretical overview to help the reader understand important conceptual issues around causation and how they apply particularly to studies in psychiatric epidemiology and psychiatry more generally. In other chapters of this book, for example, Chapter 3.2 on the "Causes of Depression" by Lewis, Lewis and Srinivasan, more direct application of causal theory to specific issues is given.

As would be expected of this cornerstone issue, causal inference in epidemiology has received substantive theoretical and empirical attention, particularly given the controversies and harms that potentially arise from incorrect inferences; one of the most infamous (and since debunked[30] and retracted[31]) recent examples in psychiatric epidemiology was the erroneous conclusion – based on a very weak study design and (as it later turned out) falsified data and unethical procedures – that a combined measles, mumps and rubella (MMR) vaccine caused an increased risk of autism in children. Subsequent research has demonstrated the profound impact on public health this had, increasing both measles susceptibility in young children[32] in the years after publication until the partial retraction (1998–2004) and in increases in vaccine hesitancy in the population.[33]

Causal inference has been central to the development and evolution of epidemiology as a discipline. In his seminal President's Address to the Royal Society of Medicine in 1965, Sir Austin Bradford Hill outlined nine criteria (Box 1.3) of any exposure-outcome association that we should 'especially consider before deciding that the most likely interpretation of it is causation'.[34] These *traditional causal inference* criteria remain useful today, while recognising that establishing causation requires careful triangulation of a range of evidence across a variety of settings, study designs and methodological disciplines. For example, recent randomised trial evidence that the drug lecanemab can delay cognitive impairment and lead to reductions in amyloid burden in those with early Alzheimer's disease over an 18-month period[35] builds on decades of biomedical, neuroscientific and other observational and experimental research that has identified the agglomeration of amyloid beta (Aβ) in plaques as one of the core features of the pathology of Alzheimer's disease. Indeed, several influential epidemiologists have proposed modern *triangulation* criteria to strengthen *causal inference* in aetiological epidemiology,[36,37] which seek to incorporate and assess evidence generated by different methodological approaches that – although not free from bias – will likely contain different sources of biases that may (or may not) counteract each other to strengthen (or weaken) the plausibility of a causal association.

Box 1.3 Bradford Hill criteria for causation

1. *Strength* – Stronger associations are more likely to be causal
2. *Consistency* – The finding replicates across different studies in different samples by different researchers
3. *Specificity* – Evidence that a single risk factor has a specific effect on one disorder but not others may increase the likelihood of causality (though Hill also recognised that most disorders would have multiple causes, or the so-called *multifactorial aetiology*)
4. *Temporality* – The exposure should precede the outcome
5. *Dose-response* – The greater the level of exposure, the greater the risk of disorder
6. *Plausibility* – The finding agrees with accepted biological understanding
7. *Coherence* – Triangulation of findings across different designs and disciplines. Here, epidemiological evidence would cohere evidence from other disciplines such as neurobiology, psychology and animal evidence
8. *Experimental evidence* – Observational findings are supported by RCT evidence and natural experiments
9. *Analogy* – Analogous exposures and outcomes show similar effects. For example, if low socioeconomic status (SES) was a determinant of schizophrenia, we would expect to see this association across validated measures of education, income, occupation and social class

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

Complementing these approaches, a set of more statistically based *contemporary causal inference* methods in epidemiology have also been developed over the past two decades to strengthen the plausibility that associations between exposure and outcome from observational epidemiology are causal. A causal effect would be established if we could prove that an individual exposed to a risk factor for disease developed the disease following their exposure (i.e. the *factual* scenario) but would not have developed the disease had they not been exposed (i.e. the *counterfactual* scenario). In other words, if we could observe the outcome status that a single individual would experience if they were both exposed or unexposed to a given risk factor or treatment, we could estimate the individual causal effect of the exposure on the outcome. This is an example of counterfactual reasoning, where any individual has two potential outcomes: the outcome they would have received if exposed versus the outcome they would have received if they had remained unexposed. The difficulty here, however, is that we can never simultaneously observe the factual and counterfactual outcomes for a single individual, meaning that individual causal effects are not identifiable. As discussed previously, the great advantage of experimental epidemiology study designs, such as randomised controlled trials, is that provided certain assumptions are satisfied, the process of randomisation ensures that both measured and unmeasured confounders are similarly distributed in both the intervention and control arms of the trial. Given this, the two groups become exchangeable such that the average outcome experienced in the intervention arm (i.e. the factual scenario) would be the same as the average outcome experienced in the control arm, had the control arm been the intervention arm (i.e. the counterfactual scenario), and vice versa. Thus, at the population or group level, it is possible to estimate the average causal effect in an RCT design under certain assumptions; for example, provided that the trial achieves a sufficient sample size, has true randomisation and is free from attrition bias, the effect size (i.e. the odds ratio) becomes equivalent to the causal odds ratio.

Unfortunately, establishing counterfactual effects under a potential outcomes framework from observational study designs is much more difficult. This is particularly problematic for applied research in mental health (as most disciplines), where randomised controlled trials are often infeasible or unethical for testing exposure to putatively harmful effects. *Contemporary causal inference* methods, including genetically informed studies (e.g. twin and sibling designs, Mendelian randomisation (MR)), the broader class of instrumental variable approaches of which MR is a special case, inverse probability weighting and propensity scoring have been developed as statistical techniques to mimic the fundamental concept of exchangeability that is achieved in an RCT through randomisation. Thus, these methods – under certain strong assumptions – recover the *average causal effect* between exposure and outcome in an observational study design. It is beyond the scope of this chapter or book to provide a detailed introduction to this class of causal inference methods in epidemiology, but for excellent introductions on theory and critique of causal inference, see Rothman and Greenland;[38] causal inference methods, see Hernan and Robins;[9] and on contemporary approaches to triangulation, see Lawlor et al.[36] and Munafò et al.[37] In this book, Lewis, Lewis and Srinivasan also provide more details about Mendelian randomisation, its advantages and limitations, and how it has been used in depression research in Chapter 3.2 on the ‘Causes of Depression’.

In addition to these approaches, the use of causal diagrams provide a further *contemporary causal inference* technique to aid transparent identification of causal effects from observational data. Causal diagrams, such as Directed Acyclic Graphs (DAGs), have been developed in parallel to more statistically based approaches to estimating causal effects under a potential outcomes framework using observational (or experimental) data in epidemiology. DAGs provide a useful and transparent tool to declare the theoretical model and assumptions underlying any causal effect of interest to be estimated.[39] Since correct causal inference requires the identification and removal of all potential *threats to validity*, including the

Box 1.4 Directed Acyclic Graphs (DAGs)

DAGs are causal diagrams that can help researchers identify and declare the hypothesised causal data structure underpinning the association between an exposure and outcome. These graphical tools can help researchers identify a minimal set of confounders that would need to be controlled for in the design or analysis of a study to estimate the causal effect of exposure on outcome, as well as to help identify any potential biases that may be introduced in the design or analysis of the study. In this way, DAGs provide a useful conceptual tool to design, conduct and report transparent and reproducible research. All potential variables relevant to the causal model should be included, regardless of whether they can be (or have been) collected.

A DAG includes *nodes* (variables) and *edges* (arrows). They are *directed* because they must indicate the assumed causal direction from one variable to another, and they are *acyclic* because a variable cannot cause itself; no node should have a path via edges that points back to itself.

In a DAG, we are usually interested in estimating the *direct causal effect* of the exposure, A, on the outcome, Y. Many other *paths* between A and Y may exist, via other nodes, including a set of confounders, L. These are so-called *biasing paths* because they are alternate, non-causal paths through which the association between A and Y exists. Failure to account for these biasing paths will result in biased estimation of the direct effect of A on Y. Biasing paths are said to be *open* when a confounder is not controlled for, as depicted in Figure 1.3, and *closed* or *blocked* when a confounder is controlled for – or *conditioned on* – in some way. Note that a potential biasing path is any path between A and Y, regardless of the directionality of the arrows (e.g. A←L→Y is a biasing path between

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

Box 1.4 (*cont.*)

A→Y in Figure 1.3). Conditioning on L, as shown in Figure 1.4, blocks the biasing path of this confounder.

Some variables in the assumed causal model may not be confounders but so-called colliders, C – that is, variables that are common effects (or common descendants) of two other variables, as depicted in Figure 1.7. Here, unlike with confounders, conditioning on a collider will *open* a biasing path, while leaving a collider as unconditioned will block that path.

The open access software 'DAGitty' (www.dagitty.net) provides a helpful tool for researchers to build their hypothesised causal model and understand the potential biasing paths that need to be blocked in the design and analysis of their study.

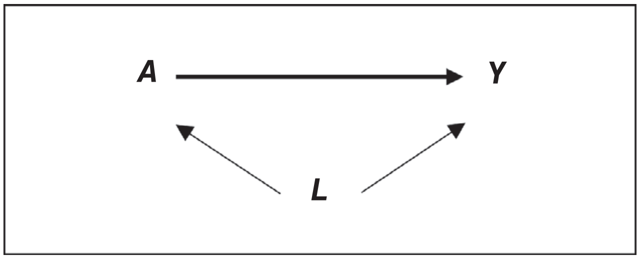

Figure 1.3 Basic confounding structure, represented in a causal diagram. The potentially causal association between the putative exposure, *A*, and the outcome, *Y*, may not be causal in the presence of a confounder, *L*, which is not taken into account during the design or analysis phase of a study.

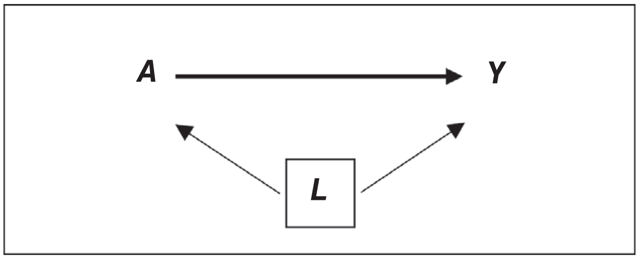

Figure 1.4 Controlling for a confounder, *L*, blocks the potential alternate causal path between *A* and *Y* that travels from $A \leftarrow L \rightarrow Y$, allowing the direct causal effect of $A \rightarrow Y$ to be estimated.

critical issues of *confounding*, *bias* and *chance* (discussed in detail later), we use simplified DAGs (see Box 1.4 for an elementary introduction) in the remainder of this section to highlight how these issues can affect our ability to infer causation from estimated associations of measures of effect (e.g. odds ratios, rate ratios) in observational epidemiology. For a comprehensive introduction to causal diagrams, see both Hernan and Robins[9] and Tennant et al.[39]

Confounding

Confounders are variables that are common causes of both the exposure and the outcome and can lead to a spurious association or eliminate a real one (Figure 1.3). Importantly, this implies that a confounder must temporarily precede the occurrence of both the exposure and the outcome. For example, we might be interested in understanding whether cannabis use (*A*, in the causal diagram in Figure 1.3) was causally associated with the risk of developing psychosis (*Y*, in Figure 1.3). A common cause of both cannabis use and psychosis (i.e. a potential confounder) may be (lower) socioeconomic status (SES) (*L*, in Figure 1.3). Since cannabis use may, theoretically, also change your subsequent SES, any study investigating the potentially causal association between cannabis use and psychosis would need to include methods to control for SES that was measured *prior* to the measurement of cannabis use; this may mean taking measurements of SES in childhood or at birth, for example, parental SES. Note that confounding is a reflection of the relationship between variables in real life – unlike bias (see later in this chapter), confounding is not a result of error in the design or analysis of studies.

When planning a study, we must identify, measure and decide on methods to deal with (control for) all potential confounders that may stop us from concluding that there is a direct causal effect of $A \rightarrow Y$. In causal diagrams, we represent a confounder that has been controlled for (or in statistical terms 'conditioned on') by placing a box around the confounder (Figure 1.4). This indicates that the potential alternate causal pathway from *A* to *Y* that travels between $A \leftarrow \boxed{L} \rightarrow Y$ has been blocked. Subject to assumptions (including the perfect measurement of the confounder, no other confounding, and correct specification of the causal model), this would allow estimation of the direct causal effect, $A \rightarrow Y$.

There are five main methods of dealing with confounding:

1. *Restriction* is a method by which individuals with the confounding variable are removed from the study altogether. In the previous example, one could restrict the study to those from the lowest SES group (or highest) and determine whether psychosis is still more common amongst those who smoke more cannabis.
2. *Matching* involves artificially making the two groups similar in terms of the confounding variable. The investigator might ensure that, in a case-control study, each case with psychosis of a given SES was matched with a control of the same SES. Matching in case-control studies is intuitively easy to understand but has some disadvantages in terms of greater sample size requirements as well as difficulty in finding suitable matched participants when matching on several variables. Furthermore, recent epidemiological theory demonstrates that matching alone does not control for the matched factors included in the design and that these still need controlling for via other methods (see the later discussion) at the analysis stage.[40] Older textbooks also suggest that a matched design requires specific statistical methods to take into account the matching, though this is no longer considered necessary and indeed can introduce bias to the results. Our

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

recommendations, following Pearce,[40] are to judiciously use matching in case-control studies as a technique to control for confounders, limited to one or two variables (e.g. age, gender); ensure the matched variables are included in the analysis stage; and use appropriate ('unmatched') analytical methods during the analysis. Matching is also sometimes used in cohort studies, where the object goal is to make the exposed and unexposed more similar to each other on certain confounders. Advanced methods such as *propensity scoring techniques* attempt to match participants on their propensity to be exposed (often, their propensity to receive treatment) in an attempt to improve the conditions under which the assumption of *exchangeability* is satisfied.

3. *Stratification* is a method used at the analysis stage, where instead of lumping all subjects together, the sample is split according to the presence of the confounder – thus those from different SES groups would be analysed separately. It is possible using stratification to calculate a combined estimate of the size of the effect (e.g. the odds ratio) using specific statistical techniques.
4. *Regression adjustment* is a general term for multivariable modelling techniques used at the analysis stage, where the confounders of interest are included as covariates in the regression model to control for their effects on the statistical association between exposure and outcome. Like with stratification, all confounder variables must be identified from theory and empirical evidence before the start of the study and measured appropriately using reliable and valid instruments. Under any of the earlier methods (1–4), failure to perfectly measure a confounder may result in only partial control for the variable, thus only partially blocking the alternate causal path in Figure 1.3; this could lead to *residual confounding* biasing inferences about the direct causal effect of $A \rightarrow Y$.
5. The final approach to confounding is *randomisation*, which is dealt with in the section on RCTs. Under randomisation, all participants in a trial are randomly assigned to receive the intervention or control, meaning that the distribution of the confounding factors, L – whether measured or unmeasured – will be the same in each arm and thus cannot be common causes of the treatment, A, or the outcome, Y, as assumed in Figure 1.5. In practice, one would wish to check whether randomisation achieved balance (or *exchangeability*) of confounders between the two arms of the trial and take additional steps to control for variables in the presence of imbalance. However, randomisation is considered as the strongest method to demonstrate causal effects between exposure and outcome because it will theoretically deal with unknown or unmeasured confounders, which cannot be taken into account by any of the other methods. Nonetheless, because it is not ethical to assign participants in studies on risk factors to receive a potentially hazardous exposure, randomisation is limited to treatment or preventive studies. This means that studies investigating risk factors are generally limited to observational epidemiology, and special causal inference methods have been developed that attempt to strengthen the *counterfactual* strengths that are implicit to unbiased RCTs.

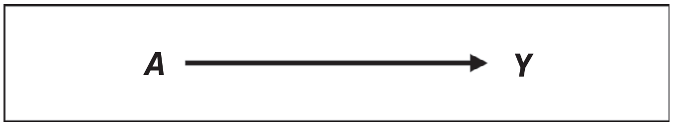

Figure 1.5 Causal diagram of the association between an exposure, *A*, and outcome, *Y*, under the assumption of no confounding, as may arise following randomisation in a randomised controlled trial.

As mentioned earlier, as common causes of exposure and outcome, confounders must temporarily precede the exposure (and outcome). Variables that proceed the exposure but precede the outcome are on the *causal pathway* (Figure 1.6); that is, they are not common causes of the exposure and outcome (i.e. confounders) but potential *mediators* of the relationship. From our example earlier, measuring someone's SES after their cannabis use but before the outcome, psychosis, would make SES a mediator, not a confounder. Inadvertent control for a variable on the causal pathway may induce bias (Figure 1.6) into the results since you are no longer estimating the total *causal effect* of $A \rightarrow Y$.

Typically, more complex confounding structures frequently exist in the causal relationship between an exposure and outcome, including *confounding-by-indication*. For example, in investigating the possible causal role of antidepressants on dementia risk, confounding-by-indication would arise if depression status was an indicator for an antidepressant prescription and if depression was a cause of dementia. For more complex examples of confounding structures in observational data, see Hernan and Robins.[9]

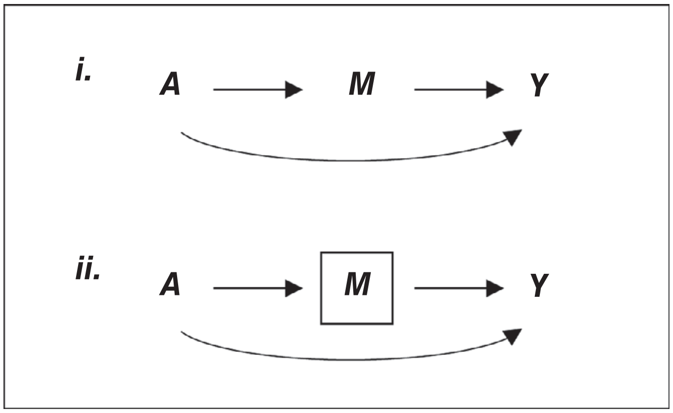

Figure 1.6 ***i.*** When a variable, *M*, lies on the causal pathway, it is not a common cause of the exposure, *A*, and outcome, *Y*. ***ii.*** Inadvertent control for the mediator, *M*, would induce bias in estimation of the total causal effect of $A \rightarrow Y$ by blocking the part of the causal effect that travels via $A \rightarrow M \rightarrow Y$.

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

Bias

Bias refers to systematic errors in the design of a study that may generate misleading results. Unlike confounding, bias comes about as a result of the study design or execution. Bias is classified into selection bias and information bias.

Selection Bias

Selection bias refers to the way in which participants in a study are selected and the impact this may have on the study's results. It occurs when there are systematic differences between those who take part in a study and those who do not. It is a particular problem in case-control studies but is not exclusive to them. An example was given under the 'Case-Control Studies' section. Selection bias tends to be a particular problem when studies identify cases from clinical populations, especially in specialist settings, since these cases may differ in important ways to cases that do not present to services. For example, a case-control study of the relationship between bullying and disordered eating restricted to clinically diagnosed cases may bias the results because those who have presented to services may differ systematically to those cases who do not present to services (but still have an undiagnosed eating disorder) in terms of their exposure or confounding factors.

From a causal inference perspective, *restriction* to clinical cases in this example is a form of conditioning on that variable (clinical presentation) as discussed in the previous section on *confounding*. Conditioning on a common effect of both the exposure, the bullying, as well as the outcome, eating disorders, may induce a spurious – biased – association between the exposure and outcome via a phenomenon called *collider bias* (Figure 1.7).

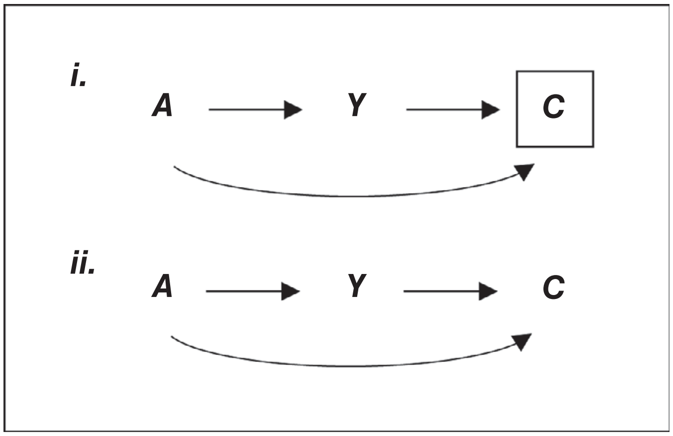

Figure 1.7 ***i.*** Selection bias occurs when those who took part are systematically different to those who did not: the exposure, *A*, and outcome, *Y*, such that selection into the study, *C*, is a common effect of *A* and *Y*. Restriction to those who took part conditions on *C*, inducing a biasing path between $A\rightarrow\boxed{C}\leftarrow Y$, an example of *collider bias*. ***ii.*** If participation is unrelated to exposure or outcome status, there is no conditioning on the common effect, *C*, and the non-causal path $A\rightarrow C\leftarrow Y$ is blocked, allowing correct estimation of the causal effect, $A\rightarrow Y$.

In this example, collider bias occurs because, via restriction, we have conditioned on the common effect (the *collider*) of clinical presentation; in causal inference theory, conditioning on a collider opens an alternate non-causal path between the exposure and outcome via the common effect, *C*, of the form $A\rightarrow\boxed{C}\leftarrow Y$, biasing the true *causal effect* of $A\rightarrow Y$. To estimate the true causal effect of bullying on eating disorders, one would need to obtain a representative sample of cases from the target population, such that there was no conditioning on the collider of clinical presentation. In causal inference theory, the non-causal path between $A\rightarrow C\leftarrow Y$ is blocked when unconditioned, allowing estimation of the causal effect, $A\rightarrow Y$. Various types of *selection bias* exist, including two important ones discussed next.

Non-response Bias

Non-response bias is a form of selection bias of particular importance in cohort studies (but relevant to all study designs), where the individuals of greatest interest may be those who are least likely to participate. This can cause misleading results if the exposure (or outcome) under study also influences participation. In cohort studies, non-response over time is known as loss to follow-up, attrition or censoring.

As a motivating example, consider the long-standing observation that many migrant groups are at increased risk of psychosis.[41] Cohort studies of this association may be affected by differential non-response bias, as depicted in Figure 1.8. In such studies, genetic liability for psychosis is unmeasured (or at best, imperfectly measured), as represented by *U*, but is known to increase both later risk for psychosis[42], *Y*, and drop-out or censoring[43], *C*, in cohort studies. Reasons for this differential loss to follow-up may include greater cognitive impairment or paranoia, *M*, as shown via the mediating path $U\rightarrow M\rightarrow C$. At the same time, migrants, *A*, may also be more likely to be lost to follow-up,

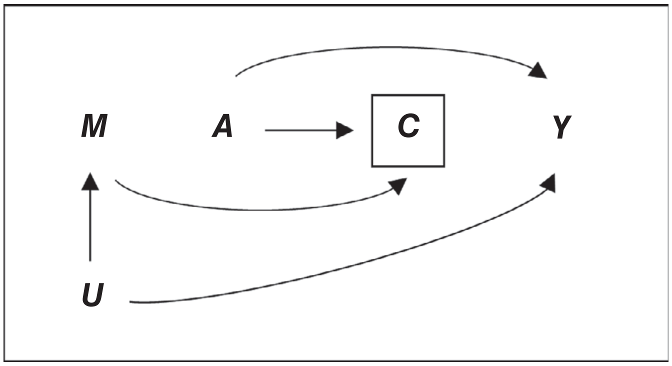

Figure 1.8 Differential non-response bias is induced when censoring (loss to follow-up). C is a common effect of both the exposure, A, and other unmeasured factors, U, which are also related to risk of the outcome, Y. Analyses restricted to those with complete data, indicated by conditioning on censorship, C, would induce a non-causal open path between $A\rightarrow\boxed{C}\leftarrow M\leftarrow U\rightarrow Y$, biasing the estimated association between A→Y. M represents a set of mediating variables that may be caused by unmeasured factors, U, and may influence non-response, C, such as cognitive impairment, paranoia or other symptoms of disorder.Adapted from Hernan and Robins.[9]

C, for a variety of reasons including returning to their home country. If we let drop-out from the study be denoted by C=1, and we restrict the analysis to those who do not drop out from the study (i.e. C=0), then we have conditioned the analysis on the descendant of a common effect, C, which, as in Figure 1.8, induces a non-causal path between $A \rightarrow \boxed{C} \leftarrow M \leftarrow U \rightarrow Y$ resulting in biased estimation of the causal effect of migrant status on psychosis risk, $A \rightarrow Y$.

Various other patterns of selection bias may exist (see Hernan and Robins).[9] Careful consideration of design features of the study should be made before the start of a new study, while methods that attempt to mitigate selection bias at the analysis stage, including inverse probability weighting and multiple imputation, exist though require strong assumptions and theoretical considerations and may not overcome all issues arising from selection bias.

Prevalence Bias

Prevalence bias is a subtype of selection bias that is a problem in case-control and cross-sectional studies where investigators identify prevalent cases, some of whom may have had the disorder for many years. With disorders such as depression, where relapse and remission are the rule, prevalent samples will be biased because they will over-represent those with chronic depression. It is then difficult to determine whether exposures act to cause or maintain the disorder.

Information Bias

Information bias refers to errors made in the gathering of information from participants. There are two main types of information bias – recall bias and observer bias.

Recall Bias

Recall bias particularly occurs when a disorder has an impact on the participant's recall. For example, patients with depression, when asked about recent life events, may be more inclined to dwell on negative events and over-look positive ones, as this is a feature of depressive thinking. In schizophrenia research, it is notoriously difficult to gain reliable information on early experiences via retrospective recall, such as obstetric complications, and mothers of people with schizophrenia may be inclined to put a good deal more effort into remembering remote events than mothers of healthy controls. Recall bias is best prevented by using documentary evidence (e.g. clinical or other routine records) or by choosing a study design less prone to recall (e.g. cohort studies). Other strategies are to use a control group of individuals with another disorder not thought to be associated with the risk factor under study, where similar recall effects would be expected, serving as a *negative control outcome*.

Observer Bias

In its most general sense, observer bias arises whenever the way in which something in a study (exposure, outcome, confounder) is measured leads to a systematic departure from the true value of that variable. As such, observer bias comes in many guises.

It can relate to the way in which researchers ask questions of participants in studies. If the researcher is aware of the hypothesis under study and also knows which group the participant is from, they may ask questions in subtly different ways. For example, if the study was assessing the efficacy of cognitive therapy versus standard care for depression, the researcher may probe depressive symptoms in a less persistent way to the group who have had cognitive therapy. 'Blinding' is an important approach to prevent this type of observer bias, but it is not always possible to blind the researcher – in case-control studies it may be very obvious which participants have a psychiatric disorder and which do not. This type of observer bias may be overcome by using highly structured interviews or self-completed questionnaires so that every participant is asked the same question in the same way.

Observer bias may also relate to some other systematic error in data collection that we have already come across; for example, many epidemiological studies that rely on diagnoses made as part of someone's routine care will implicitly include (or ignore) between-clinician variance – or *inter-rater reliability* – as a result of the way that different clinicians formulate and apply the same diagnostic criteria to patients. Without consideration, this could introduce bias into the results. For example, in the Social Epidemiology of Psychoses in East Anglia (SEPEA) study of the epidemiology of first-episode psychosis in a rural part of the east of England,[44] the authors used clinical diagnoses made in routine Early Intervention in Psychosis (EIP) care to identify potential cases before asking a panel of clinicians – all trained in the same way with good inter-rater reliability – to make diagnoses using a standardised research instrument; this ensured both reliable and valid diagnoses were used to define the epidemiology of psychotic disorders in this study as well as minimise possible observer bias.

In other situations, systematic errors in observation may arise during the analysis phase of the study if the measures, techniques or observers introduce incorrect data, have faulty readings or fail to correctly interpret information. Classic examples of this exist, including the so-called dead salmon experiment in which the authors demonstrated that without correction for multiple comparisons, results from (though by no means limited to) functional magnetic resonance imaging (fMRI) experiments would show substantial *post-mortem* neural activation in the brain of an Atlantic salmon in response to socially stressful stimuli (to humans, whether to fish remains unclear).[45]

Ascertainment bias is another form of observer bias, whereby the methods used to detect cases may systematically fail to identify and include relevant cases in the target population from different groups equally. For example, studies in which hospitalised cases of depression were over-represented compared with community cases (who may be harder to find) would underestimate the true incidence or prevalence of

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

depression in a population; moreover, if hospitalised cases differed from those in the community in terms of the frequency of the exposure of interest, this would introduce differential bias, leading to inaccurate estimation of the true effect size (i.e. the odds ratio) between exposure and outcome. For example, if hospitalised cases were more likely to have a family history of depression than community cases, this would lead to an overestimation of the true effect of family history of depression on the depression risk in this example.

Reverse Causation

In reverse causation, the association between the risk factor and the disorder is a valid one, but the interpretation is turned around. For example, a study might find that there is a strong association between job loss and depression, and this might be interpreted as indicating that those who lose their jobs are at greater risk of becoming depressed. However, an alternative hypothesis is that depression is an important cause of job loss – individuals who become depressed perform less well at work and are therefore more at risk of losing their jobs.

From a causal inference perspective,[9] *reverse causation* is effectively a special form of confounding, when an unmeasured factor – say the prodromal symptoms of psychotic disorder – is a common cause of both a decline in SES (because people in the prodromal phases of psychotic disorder can no longer hold down a job due to their symptoms) and a clinical diagnosis of psychotic disorder (Figure 1.9).

In psychosis research, reverse causation is a long-standing problem[46] in understanding whether the higher rates of schizophrenia and other non-affective psychotic disorders seen in city dwellers is due to features of urban life ('social causation'), or whether those suffering from non-affective psychotic disorders are more likely to migrate to the cities ('social drift'). These issues can be partially addressed using longitudinal study designs, such as cohort studies, where the risk factor is measured many years *before* the onset of the disorder to decrease the likelihood that prodromal symptoms could be related to exposure (i.e. effectively removing the arrow between *U* and *A* in Figure 1.9). There is now strong evidence from such studies that an association persists between urbanicity at birth[47] or during upbringing[48] and later schizophrenia risk. Nonetheless, more complex – intergenerational – social drift patterns may still explain such an association if parental genetic liability for schizophrenia (now *U* in Figure 1.9) was a common cause of both child genetic liability for schizophrenia (now *L* in Figure 1.9) and urbanicity at child birth (now *A* in Figure 1.9). Evidence to support this possibility is currently equivocal (see Colodro-Conde et al., Solmi et al. and Paksarian et al.[49–51] for further reading on this issue).

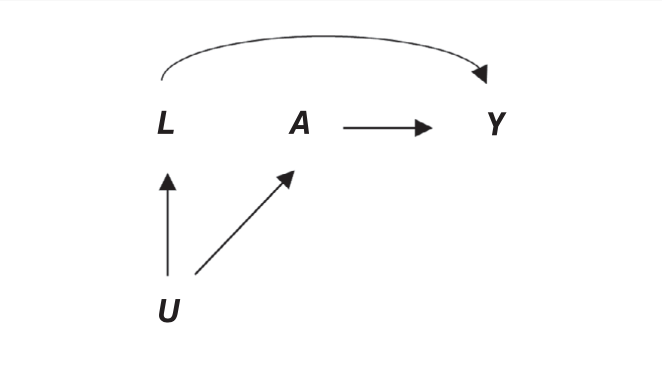

Figure 1.9 *Reverse causation* is effectively a form of confounding, where the putative relationship between an exposure, *A* (for example, SES), and outcome, *Y* (for example, psychotic disorder), is actually due to the prodromal symptoms of psychosis which are unmeasured, *U*, and which may be a common cause of SES (for example, loss of a job due to the prodromal symptoms of psychosis) and which increase the risk of psychotic disorder, *Y*, via some unobserved pathway, *L*, for example, cognitive impairment. Since *U* and *L* are unobserved, reverse causation may provide an alternate non-causal path between *A*→*Y*.

Chance

Type 1 and Type 2 Error

Most studies aim to describe reality by taking a *sample* of the total population. However, the sample will not exactly describe the true underlying population distribution: there is always a degree of *sampling error*. Tossing a coin 10 times will yield different combinations of heads and tails. Statistically the *most likely* result would be five heads and five tails, but any combination of heads and tails is possible. More extreme results (e.g. all tails or all heads) become less probable with increasing numbers of tosses of the coin. In other words, increasing the number of tosses increases the *precision* with which the underlying 'true' situation can be estimated.

In any analytic study, we hope that the results of our study reflect reality. Nevertheless, if 10 identical studies were performed, they would all come up with slightly different results. The size of the difference would depend on the size of the sample in each study. Studies that report an association between two variables may either be describing the true underlying situation or, by chance, have committed a *type 1 error* (see Table 1.5). A type 1 error occurs where a spurious association is detected by chance (a 'false positive'), and the probability that this has occurred is assessed by statistical testing. By convention, the type 1 error rate is often set at the arbitrary level of $P < 0.05$.

Studies that report a 'negative finding' (i.e. do not show an association between two variables) may either be describing the

Table 1.5 The relationship between the results of a study and 'true life'

	'True life'	
Study	**Association exists**	**No association exists**
Association demonstrated	*	Type 1 error ('false positive')
No association demonstrated	Type 2 error ('false negative')	*

* Indicates where the study results represents 'true life'

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

true underlying situation or may, by chance, have committed a type 2 error. A type 2 error occurs when a genuine association is missed by chance (a 'false negative'). In designing a study, the power calculation takes into account an acceptable type 2 error rate, usually set at 10–20%, meaning that most studies accept that there is a 10–20% chance that they will fail to detect a true effect. Statistical power is the converse of the type 2 error rate (and is therefore usually set at 80–90%). While the type 1 error rate can be set as an arbitrary threshold beyond which statistical significance is inferred, the type 2 error rate is determined by the power the study has to detect an effect size of a pre-specified magnitude in a sample of a given size. This means that power and sample size calculations are required before starting a study to understand how big a study needs to be to detect an effect should one of (at least) that size exist in reality. More powerful studies require bigger sample sizes.

Hypothesis Testing, Statistical Significance and Uncertainty

Most epidemiological studies seek to test whether there is an association or effect between a hypothesised exposure and outcome. Implicitly, this is a hypothesis test that the association differs from what would be expected under the null condition – that is, there being no association between the two variables. Conventionally, a test of the 'statistical significance' of this association is made, with the test being appropriate to the type of data and model used. If the estimated P-value is smaller than an arbitrary threshold (often, $P < 0.05$, though smaller in genetic studies due to multiple comparisons), conclusions are drawn that the observed effect differs from the null and is 'unlikely to be due to chance' (type 1 error).

The received wisdom presented in the previous paragraph is, however, a bastardisation of the use of statistics in medical research. There is no P-value that can 'prove' an association is true. The misuse and misinterpretation of statistical testing, P-values and related measures such as confidence intervals are one of the most endemic and enduring issues in medical statistics, and we encourage readers of this chapter to develop a deeper understanding of the correct use and interpretation of statistics in epidemiology and other fields of medicine (see Greenland et al.[52] for an excellent primer on this topic).

Briefly, though, any statistical model we construct defines a set of assumptions we, as researchers, make about the relationship between exposures, confounders and outcomes. We collect data from a (hopefully unbiased and representative) sample of our target population, and we test the extent to which that model provides an accurate representation of the data collected. Effect sizes between the outcome and other variables in our model – including any exposure(s) of interest – are estimated alongside the level of uncertainty around them. This statistical uncertainty codifies the probability or likelihood of the observed data, given the effect size(s) estimated, and is often represented as confidence intervals around the estimated effect size. Models produced from smaller datasets will estimate effect sizes with greater statistical uncertainty (and wider confidence intervals). P-values and confidence intervals are intimately linked in statistics (a type 1 error alpha level of 0.05 corresponds to a 95% confidence interval, or 1 – 0.05), and P-values should more correctly be thought of 'as a statistical summary of the compatibility between the observed data and what we would . . . expect to see if we knew the entire statistical model . . . were correct' (Lewis, Lewis and Srinivasan, p.339).[48]

Table 1.6 Parameters and their null values

Parameter	Null value
Differences in means, risk difference	0
Odds ratio, rate ratio, risk ratio, hazard ratio	1
Number needed to treat	∞

Confidence intervals provide us with a more intuitive measure of the likelihood or probability that the estimated interval from our study sample contains the true effect size in our target population, or the precision of our effect. If one were to repeat our study 100 times in another valid (i.e. comparable) population, on 95 of those 100 occasions we would expect the confidence interval around our effect size to contain the true effect size. Confidence intervals may be calculated for most parameters we estimate. For example, it is possible to calculate a confidence interval around purely descriptive statistics like a mean or a proportion. It is also possible to show a confidence interval around a comparative parameter such as a difference between two means, a relative risk or a number needed to treat. When the 95% confidence interval crosses the null value of a parameter, this indicates that there is no difference at the P=0.05 level between the groups compared. It is important here to know the null value (see Table 1.6).

We recommend de-emphasising the reliance on P-values, arbitrary thresholds of 'significance testing' or the reporting of 'statistically significant' results (or worse, 'significant' results) in favour of interpretation of effect sizes alongside 95% confidence intervals, which tells us about the level of uncertainty in our observed data given the model.

Points to Consider If a Study Reports One or More Positive Associations

Uncertainty around any estimate and the possibility of a type 1 error mean that no single study will provide sufficient evidence to demonstrate a causal effect between an exposure and outcome. This is why researchers seek to replicate one another's results and triangulate evidence from a body of studies around causation. Positive findings may arise for several reasons.

First, it may be that the study is very big, and even differences that are in clinical terms trivial appear 'statistically significant'. Here, it is important to consider the effect size (i.e. how big the odds ratio is) and its potential clinical relevance instead of reliance on the P-value.

Second, the more statistical comparisons made, the greater the likelihood that a 'statistically significant' association will be found by chance. If researchers collected data on 40 possible

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

risk factors for schizophrenia, they could expect two to be associated at $P < 0.05$ *just by chance*.

There is a balance to be reached between wasting data and ‘data trawling’. The best way to overcome this dilemma is to set out with one or two main hypotheses that form the centre of the research protocol and on which the power calculation is based. All additional findings can be labelled ‘secondary analyses’ and be seen as a useful by-product of the main research, potentially for follow-up in other future studies. Further, in keeping with contemporary causal inference methods, defining one or two main hypotheses *a priori* and specifying the theoretical model you assume to be relevant (for example, via construction of a DAG, see previous discussion) also reduces the reliance on data trawling and multiple testing; instead, one constructs a theoretical model for the hypothesised relationship, designs the study, collects the data and tests that model.

An alternative approach to multiple statistical testing is to use a *Bonferroni correction*. This works on the principle that the level of statistical significance set should be adjusted to take account of the number of tests performed. Thus, if we set $P < 0.05$ for a single significance test, this should be reduced to $P < 0.005$ if we perform 10 comparisons. This approach is generally considered too conservative and may lead one to miss significant positive findings. A better approach is to express results in terms of their *precision* (see the previous section on confidence intervals).

Another way of getting spurious P-values is by using subgroup analyses. Here, the researcher breaks down the statistical analysis according to certain characteristics of the participants. For example, a researcher may have performed a randomised trial comparing a new atypical antipsychotic with haloperidol in patients with treatment-resistant psychosis. The main results show no overall difference, but the researcher may investigate whether there are any particular subgroups of patients in whom there was a difference. For example, the researcher might hypothesise that patients with pronounced positive symptoms will respond better to the atypical antipsychotic. Such analyses are often reported as showing positive evidence for the new intervention but are best avoided as it is notoriously easy to generate a type 1 error in this way.

Points to Consider If the Study Reports a Negative Finding

The key question to consider when a negative finding is presented is whether the sample size was sufficient to detect a true difference if it really existed – that is, are the results due to a type 2 error? For example, in evaluations of new antidepressants, the new treatment is often pitted against a reference compound. Studies often report no difference in treatment effect between the two drugs, suggesting perhaps that the new treatment is as good as the old one. However, comparisons between two active treatments require large samples. If one assumes that two-thirds of patients treated with imipramine respond within 6 weeks, Table 1.7 indicates the sample size required to detect differing levels of recovery rates for a new treatment at 95% confidence and 80% power. It indicates that a sample size of 82 would be able to detect only a very big difference between the treatments and that to detect a difference of 10 percentage points in recovery rates (which would be a clinically meaningful difference) would require over 650 participants. Thus, an underpowered study that demonstrates no difference between two treatments *does not* indicate that the treatments have similar efficacy!

Table 1.7 Power calculations for different effect sizes: sample size required to detect differing levels of recovery rates for a new treatment at 95% confidence and 80% power

Recovery rate on new antidepressant (compared with 66% improvement on imipramine)	Number required to be randomised
33%	82
40%	128
50%	320
55%	654
60%	2096

Internal versus External Validity

Most of the previous discussion has concentrated on threats to the *internal validity* of research studies. However, a common complaint about research is that participants may be so dissimilar from patients seen in normal clinical practice that it is impossible to generalise from the research findings. This complaint particularly applies to RCTs, which are indeed performed in different settings and with different patient groups to those seen in standard practice. In general, complaints about lack of generalisability of RCTs have been misplaced. There are very few examples of treatments working on one group of patients with a disorder but not on others. In some circumstances, the practicalities of recruiting patients make it difficult to ensure that those entered into the study are similar to those seen in clinical practice; this particularly applies to patients with psychotic or manic illness, where the most severely affected are least likely to give their consent to participate. In other circumstances, RCTs impose unnecessarily long lists of exclusion criteria. It is now more common in psychiatric epidemiology to see pragmatic RCTs, which apply standard trial methodology on representative samples of patients to reduce potential issues affecting external validity.[53,54]

Critical Appraisal

This chapter has discussed a number of aspects of clinical epidemiology, with particular emphasis on developing understanding of the principal study designs and their associated flaws. Critical appraisal is the approach of putting this knowledge into practice when assessing research findings. While some knowledge of study designs and common

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

flaws helps, critical appraisal is a skill that requires practice. This is best gained by:

- Reading new research with a sceptical frame of mind, including assessment of the major 'threats to validity' of the reported findings (chance, bias, confounding)
- Following correspondence about published studies
- Discussing studies with colleagues
- Presenting and attending journal clubs
- Being prepared to ask 'what's your evidence?' when given a 'fact' (even in this book!)

References

1. Last JM. *A Dictionary of Epidemiology*, 4th ed. Oxford: Oxford University Press; 2001.
2. Hippocrates. *Hippocrates, Volume I: Ancient Medicine*, Loeb Classical Library. Cambridge, MA: Harvard University Press; 2022.
3. Snow J. *On the Mode of the Communication of Cholera*, 2nd ed. London: John Churchill, New Burlington Street; 1855 (reprinted New York; 1936).
4. Cooper B, Morgan HG. *Epidemiological Psychiatry*. Springfield, IL: Charles C Thomas Pub Ltd; 1973.
5. Wasserman D (ed.). *Oxford Textbook of Suicidology and Suicide Prevention*, 2nd ed. Oxford: Oxford University Press; 2021.
6. Hollander A-C, Pitman A, Sjöqvist H, et al. Suicide risk among refugees compared with non-refugee migrants and the Swedish-born majority population. *The British Journal of Psychiatry* 2020;217(6):686–92. doi.org/10.1192/bjp.2019.220.
7. Harris S, Dykxhoorn J, Hollander A-C, et al. Substance use disorders in refugee and migrant groups in Sweden: a nationwide cohort study of 1.2 million people. *PLoS Medicine* 2019;16(11): e1002944. doi.org/10.1371/journal.pmed.1002944.
8. Goldberger J, The cause and prevention of Pellagra on JSTOR. *Public Health Report* 1914;29(37):2354.
9. Hernan M, Robins J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.
10. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems* (11th ed.). https://icd.who.int/ Geneva: World Health Organization; 2019.
11. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-V)*, 5th ed. Washington DC: American Psychiatric Association; 2013.
12. Jongsma HE, Turner C, Kirkbride JB, et al. International incidence of psychotic disorders, 2002–17: a systematic review and meta-analysis. *The Lancet Public Health* 2019;4(5): e229–e244. doi.org/10.1016/S2468-2667(19)30056-8.
13. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:869. doi.org/10.1136/BMJ.C869.
14. Lash TL, VanderWeele TJ, Haneuse S, et al. *Modern Epidemiology*, 4th ed. Philadelphia: Lippincott Williams & Wilkins; 2021.
15. Lubin JK, Gail MH. Biased selection of controls for case-control analyses of cohort studies. *Biometrics* 1984;40 (1):63. doi.org/10.2307/2530744.
16. Helbich M, de Beurs D, Kwan MP, et al. Natural environments and suicide mortality in the Netherlands: a cross-sectional, ecological study. *The Lancet Planetary Health* 2018;2 (3):e134–e139. doi.org/10.1016/S2542-5196(18)30033-0.
17. Keown P, Weich S, Bhui KS, et al. Association between provision of mental illness beds and rate of involuntary admissions in the NHS in England 1988–2008: ecological study. *BMJ* 2011;343:d3736. doi.org/10.1136/BMJ.D3736.
18. McManus S, Bebbington PE, Jenkins R, et al. Data resource profile: adult psychiatric morbidity survey (APMS). *International Journal of Epidemiology* 2020;49(2):361–62e. doi.org/10.1093/IJE/DYZ224.
19. Scott KM, de Jonge P, Stein DJ, et al. (eds.). *Mental Disorders around the World: Facts and Figures from the WHO World Mental Health Surveys*. Cambridge: Cambridge University Press; 2018.
20. Wadsworth M, Kuh D, Richards M, et al. Cohort profile: the 1946 national birth cohort (MRC National Survey of Health and Development). *International Journal of Epidemiology* 2006;35 (1):49–54. doi.org/10.1093/IJE/DYI201.
21. Elliott J, Shepherd P. Cohort profile: 1970 British Birth Cohort (BCS70). *International Journal of Epidemiology* 2006;35(4):836–43. doi.org/10.1093/IJE/DYL174.
22. Boyd A, Golding J, Macleod J, et al. Cohort profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* 2013;42(1):111. doi.org/10.1093/IJE/DYS064.
23. Solmi F, Lewis G, Zammit S, et al. Neighborhood characteristics at birth and positive and negative psychotic symptoms in adolescence: findings from the ALSPAC birth cohort. *Schizophrenia Bulletin* 2020;46(3):581–91. doi.org/10.1093/SCHBUL/SBZ049.
24. Geoffroy MC, Arseneault L, Girard A, et al. Association of childhood bullying victimisation with suicide deaths: findings from a 50-year nationwide cohort study. *Psychological Medicine* 2022. doi.org/10.1017/S0033291722000836.
25. Laugesen K, Ludvigsson KF, Schmidt M, et al. Nordic health registry-based research: a review of health care systems and key registries. *Clinical Epidemiology* 2021;13:533–4. doi.org/10.2147/CLEP.S314959.
26. Jongsma HE, Gayer-Anderson C, Tarricone I, et al. Social disadvantage, linguistic distance, ethnic minority status and first-episode psychosis: results from the EU-GEI case-control study. *Psychological Medicine* 2021;51 (9):1536–48. doi.org/10.1017/S003329172000029X.
27. Lewis G, Duffy L, Ades A, et al. The clinical effectiveness of sertraline in primary care and the role of depression severity and duration (PANDA): a pragmatic, double-blind, placebo-controlled randomised trial. *The Lancet Psychiatry* 2019;6 (11):903–14. doi.org/10.1016/S2215-0366(19)30366-9.
28. Egger M, Smith GD, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 1997;315:629–34. doi.org/10.1136/BMJ.315.7109.629.

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press

29. Kandola A, Lewis G, Osborn DPJ, et al. Depressive symptoms and objectively measured physical activity and sedentary behaviour throughout adolescence: a prospective cohort study. *The Lancet Psychiatry* 2020;7 (3):262–71. doi.org/10.1016/S2215-0366(20)30034-1.

30. Godlee F, Smith J, Marcovitch H. Wakefield's article linking MMR vaccine and autism was fraudulent. *BMJ* 2011;342:64–6. doi.org/10.1136/BMJ.C7452.

31. Wakefield AJ, Murch SH, Anthony A, et al. Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet* 1998;351 (9103):637–41. doi.org/10.1016/S0140-6736(97)11096-0.

32. Napier G, Lee D, Robertson C, et al. A model to estimate the impact of changes in MMR vaccine uptake on inequalities in measles susceptibility in Scotland. *Statistical Methods in Medical Research* 2016;25 (4):1185–200. doi.org/10.1177/0962280216660420.

33. Motta M, Stecula D. Quantifying the effect of Wakefield et al. (1998) on skepticism about MMR vaccine safety in the U.S. *PLoS ONE* 2021;16(8): e0256395. doi.org/10.1371/JOURNAL.PONE.0256395.

34. Hill AB. The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine* 1965;58 (5):295–300.

35. van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in early Alzheimer's disease. *The New England Journal of Medicine* 2023;388(1):9–21. doi.org/10.1056/nejmoa2212948.

36. Lawlor DA, Tilling K, Smith GD. Triangulation in aetiological epidemiology. *International Journal of Epidemiology* 2016;45(6):1866–86. doi.org/10.1093/IJE/DYW314.

37. Munafò MR, Higgins JPT, Smith GD. Triangulating evidence through the inclusion of genetically informed designs. *Cold Spring Harbor Perspectives in Medicine* 2020;1:11. doi.org/10.1101/CSHPERSPECT.A040659.

38. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *American Journal of Public Health* 2005;95:S144–S150. doi.org/10.2105/AJPH.2004.059204.

39. Tennant PWG, Murray EJ, Arnold KF, et al. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology* 2021;50 (2):620–32. doi.org/10.1093/IJE/DYAA213.

40. Pearce N. Analysis of matched case-control studies. *BMJ* 2016;352:i969. doi.org/10.1136/bmj.i969.

41. Henssler J, Brandt L, Müller M, et al. Migration and schizophrenia: meta-analysis and explanatory framework. *European Archives of Psychiatry and Clinical Neuroscience* 2020;270:325–35. doi.org/10.1007/S00406-019-01028-7.

42. Pardiñas AF, Holmans P, Pocklington AJ, et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature Genetics* 2018;50:381–9. doi.org/10.1038/s41588-018-0059-2.

43. Martin J, Tilling K, Hubbard L, et al. Association of genetic risk for schizophrenia with nonparticipation over time in a population-based cohort study. *American Journal of Epidemiology* 2016;183(12):1149–58. doi.org/10.1093/aje/kww009.

44. Kirkbride JB, Hameed Y, Ankireddypalli G, et al. The epidemiology of first-episode psychosis in early intervention in psychosis services: findings from the social epidemiology of psychoses in East Anglia [SEPEA] study. *American Journal of Psychiatry* 2017;174 (2):143–53. doi.org/10.1176/appi.ajp.2016.16010103.

45. Bennett C, Miller M, Wolford G. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: an argument for multiple comparisons correction. *NeuroImage* 2009;47:S125. doi.org/10.1016/S1053-8119(09)71202-9.

46. Faris R, Dunham H. *Mental Disorders in Urban Areas: An Ecological Study of Schizophrenia and Other Psychoses.* Chicago/London: The University of Chicago Press; 1939.

47. Lewis G, Dykxhoorn J, Karlsson H, et al. Assessment of the role of IQ in associations between population density and deprivation and nonaffective psychosis. *JAMA Psychiatry* 2020;77 (7):729–36. doi.org/10.1001/jamapsychiatry.2020.0103.

48. Lewis G, David A, Andreasson S, et al. Schizophrenia and city life. *Lancet* 1992;340(8812):137–40. doi.org/10.1016/0140-6736(92)93213-7.

49. Colodro-Conde L, Couvy-Duchesne B, Whitfield JB, et al. Association between population density and genetic risk for schizophrenia. *JAMA Psychiatry* 2018;75(9):901–10. doi.org/10.1001/jamapsychiatry.2018.1581.

50. Solmi F, Lewis G, Zammit S, et al. Neighborhood characteristics at birth and positive and negative psychotic symptoms in adolescence: findings from the ALSPAC birth cohort. *Schizophrenia Bulletin* 2020;46 (3):581–91. doi.org/10.1093/SCHBUL/SBZ049.

51. Paksarian D, Trabjerg BB, Merikangas KR, et al. The role of genetic liability in the association of urbanicity at birth and during upbringing with schizophrenia in Denmark. *Psychological Medicine* 2018;48 (2):305–14. doi.org/10.1017/S0033291717001696.

52. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 2016;31:337–50. doi.org/10.1007/S10654-016-0149-3.

53. Ford I, Norrie J. Pragmatic trials. *New England Journal of Medicine* 2016;375:454–63. www.nejm.org/doi/10.1056/NEJMra1510059.

54. Hotopf M, Churchill R, Lewis G. Pragmatic randomised controlled trials in psychiatry. *The British Journal of Psychiatry* 1999;175(3):217–23. doi.org/10.1192/BJP.175.3.217.

https://doi.org/10.1017/9781911623861.002 Published online by Cambridge University Press