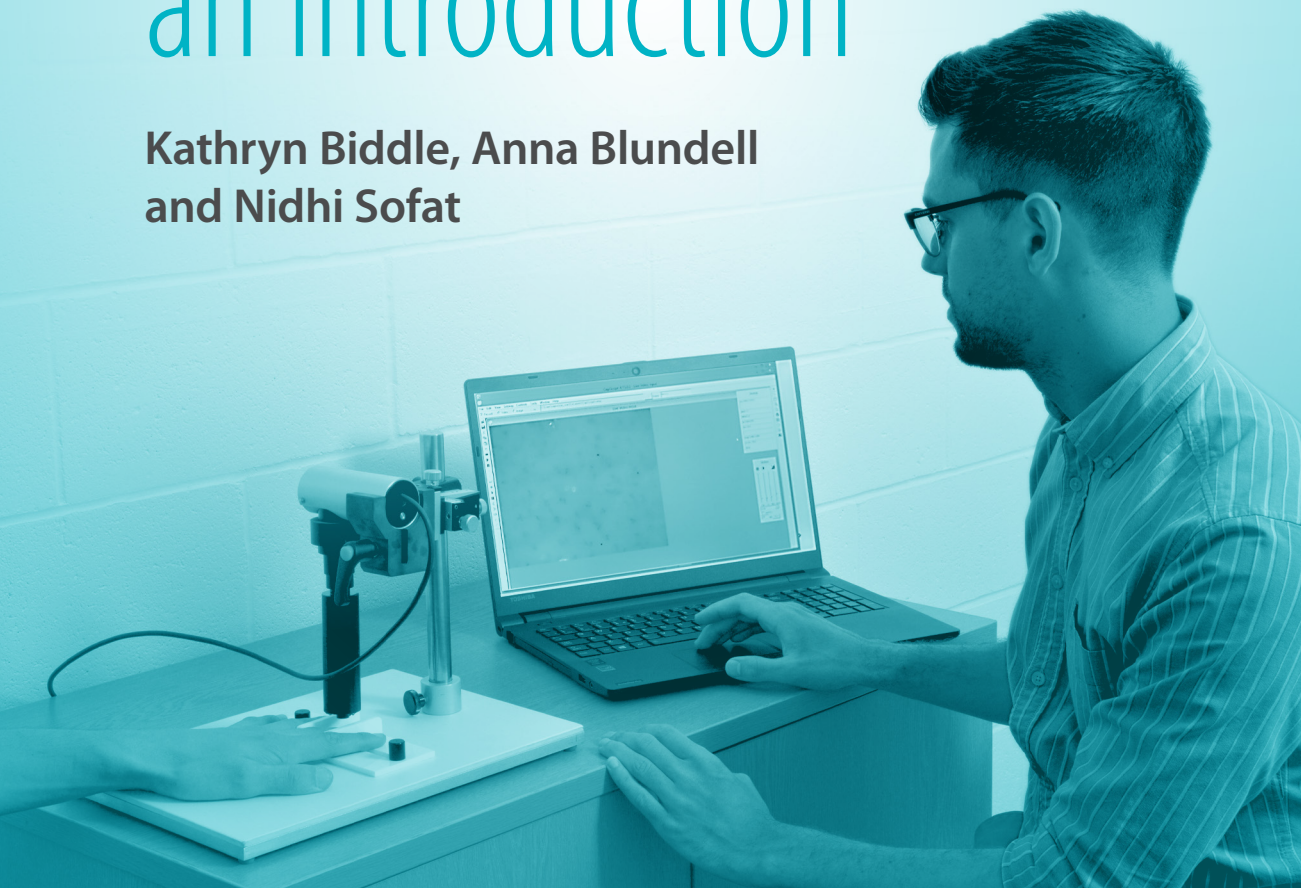
Understanding Clinical Research

an introduction

Kathryn Biddle, Anna Blundell and Nidhi Sofat



Understanding Clinical Research

Understanding Clinical Research
an introduction

Kathryn Biddle, MA, MB BChir, MRCP
Academic Clinical Fellow at St George's, University of London

Anna Blundell, MSci
Research Assistant in Rheumatology, St George's, University of London

Nidhi Sofat, BSc, MBBS, PhD, FRCP, PGCert, FHEA
Professor of Rheumatology, St George's, University of London
Consultant Rheumatologist, St George's University Hospitals NHS Trust
George's Academic Training (GAT) Lead
Co-Director, PGCert in Research Skills and Methods

Scion

First published 2023

A CIP catalogue record for this book is available from the British Library.

ISBN 9781914961267

Scion Publishing Limited
The Old Hayloft, Vantage Business Park, Bloxham Road, Banbury OX16 9UX, UK

www.scionpublishing.com

Important Note from the Publisher

The information contained within this book was obtained by Scion Publishing Ltd from sources believed by us to be reliable. However, while every effort has been made to ensure its accuracy, no responsibility for loss or injury whatsoever occasioned to any person acting or refraining from action as a result of information contained herein can be accepted by the authors or publishers.

Readers are reminded that medicine is a constantly evolving science and while the authors and publishers have ensured that all dosages, applications and practices are based on current indications, there may be specific practices which differ between communities. You should always follow the guidelines laid down by the manufacturers of specific products and the relevant authorities in the country in which you are practising.

Although every effort has been made to ensure that all owners of copyright material have been acknowledged in this publication, we would be pleased to acknowledge in subsequent reprints or editions any omissions brought to our attention.

Registered names, trademarks, etc. used in this book, even when not marked as such, are not to be considered unprotected by law.

Cover design by Andrew Magee Design
Typeset by Evolution Design & Digital Ltd (Kent)
Printed in the UK
Last digit is the print number: 10 9 8 7 6 5 4 3 2 1

Contents

Foreword

It gives me great pleasure to introduce this important book written by colleagues who exemplify what is achievable in a successful clinical academic career. For medical graduates starting out today, it can be daunting to establish where to start and how to successfully combine medical training with academic activity. Although there are many research opportunities available to those early in their career, it is important for each individual to investigate which specific areas interest them and what skills they will need to develop their research career. This book breaks down the many considerations to support your decision-making and support successfully undertaking research. These include how to get started, how to design and appraise clinical studies, understanding statistical tools needed for specific research questions, ethical considerations, public and patient involvement, qualitative research and the vital dissemination of research findings.

I remain completely convinced that academic activities, alongside clinical practice, are a great route to sustain a long, varied and interesting career. Although at times it may feel daunting, or more complicated, the extra effort is so worth it. This book segments areas of knowledge and demonstrates how things have developed – in research techniques and more broadly. The focus on ethics; the importance of involvement of public and patients in research – not as people to be 'done to' but fully involved in co-designing research areas and outcomes that are important and relevant to them – is welcome.

Of course, clinical academic activity is much broader than conducting research projects and as your experiences develop, you may well focus on a variety of areas, all of which should help your progression. There are contributions to education, knowledge exchange, enterprise; other areas of leadership and citizenship, which may become relevant and allow flexibility in your personal academic pathway. My own career has been very varied and leant on different aspects of academic activity over the years, but cumulatively has helped my progression to a leadership level I could never have imagined at the point of my graduation from medical school. So, be ambitious and go for it in your own way!

Professor Jenny Higham
Vice-Chancellor
St George's, University of London

Preface

Over the last few years clinical research has been a crucial driving force behind significant developments in new treatments in medicine, surgery and primary care. Whilst welcoming these advances in treatments and practice, clinicians and researchers may not always be equipped to assess studies and their methodologies in busy clinical or research environments. This book is aimed at budding researchers who are starting out in research and require further information on the established principles of clinical research. It will also be of interest to the practising clinician and researcher who needs to appraise and consider these developments in evaluating best practice.

There are seven chapters in the book, which cover key topics in initiating research, obtaining funding, design, planning and carrying out of research projects. We also summarise case histories, providing information about how recent medical and science graduates can identify research areas they are interested in. There are chapters on designing and appraising clinical studies, and on types of study, such as expert opinion, case reports, cross-sectional studies, case–control, cohort studies, randomised controlled trials, systematic reviews and meta-analysis. There are also chapters on statistics, data acquisition, analysis and research methodology. A chapter dedicated to ethical considerations and governance is also provided. There is a chapter on qualitative research, mixed methods study design and a dedicated chapter on public and patient involvement, which are important considerations for many studies. The final chapter discusses presenting data as oral or abstract presentations, considerations for publishing and selecting appropriate journals for scientific research.

This book is designed to provide an introduction to clinical research. We focus on the 'why' and the 'how' and discuss the rationale for developing clinical research studies based on the questions that a researcher wants to ask. With the recent Covid-19 worldwide pandemic, many clinicians and researchers were asked to contribute to clinical studies and trials which led to the rapid development of new therapies and vaccines for combating the pandemic. Such an international effort required rapid upskilling by the workforce to equip them with the skills required in conducting and reporting clinical research in a time-restricted environment. Many of the published studies for Covid-19 are used as case histories in the book, with worked examples on the types of study, statistical analyses and reporting outcomes. Our examples demonstrate how evidence-based practice is developed through research.

Our book embodies the Postgraduate Certificate in Research Skills and Methods curriculum at St George's, University of London and can also be used as an accompanying text for other PGCert and Masters courses in clinical research. It will also be helpful to those who are embarking on MD/PhD studies.

Kathryn Biddle
Anna Blundell
Nidhi Sofat

Acknowledgements

The authors would like to thank Dr Mathew John Paul for useful discussions and Ms Yvonne Forde for her support in proofreading.

We would also like to thank all our families for their patience and support during the writing of this book – we couldn't have done it without you!

About the authors

Dr Kathryn Biddle trained in Medicine at the University of Cambridge. Following her Foundation year training, she was awarded an Academic Clinical Fellowship from the NIHR (National Institute for Health and Care Research) which allowed her to continue her clinical training in Rheumatology combined with training in academic research. Kathryn is continuing her training as a Clinical Academic.

Anna Blundell completed her Master in Science degree at the University of Bath. Following this she worked as a Research Assistant on clinical trials and translational studies at St George's, University of London. She is currently pursuing her PhD studies.

Professor Nidhi Sofat studied Medicine at University College London. She then trained in Rheumatology at Imperial College Healthcare NHS Trust and was awarded her PhD at the Kennedy Institute for Rheumatology, funded by a Clinical Research Training Fellowship from the Wellcome Trust. Nidhi works as a Consultant Rheumatologist at St George's University Hospitals and also leads her own research group in Translational Medicine at St George's, University of London, where she is the National Institute for Health and Care Research (NIHR) Integrated Academic Training programme lead for Clinical Academic trainees.

Abbreviations

ACF	Academic Clinical Fellow
AE	adverse event
AR	adverse reaction
ARR	absolute risk reduction
AUC	area under the curve
BHF	British Heart Foundation
CAQDAS	computer-assisted qualitative data analysis software
CCT	Certificate of Completion of Training
CI	confidence interval; chief investigator
CRF	case report form
CTIMP	clinical trial of an IMP
df	degrees of freedom
EBM	evidence-based medicine
GCP	Good Clinical Practice
HR	hazard ratio
HRA	Health Research Authority
IMP	investigational medicinal product
IQR	interquartile range
IRAS	Integrated Research Application System
ISF	investigator site file
LR	likelihood ratio
MHRA	Medicines and Healthcare products Regulatory Agency
MRC	Medical Research Council
NIHR	National Institute for Health and Care Research
NNT	number needed to treat
NPV	negative predictive value
OR	odds ratio
PI	principal investigator
PiiAF	Public Involvement Impact Assessment
PIL	patient information leaflet
PPI	public and patient involvement
PPV	positive predictive value
RCT	randomised controlled trial
REC	research ethics committee
ROC	receiver operating characteristic
RR	relative risk
RSI	Reference Safety Information
SAE	serious adverse event
SD	standard deviation
SEM	standard error of the mean
SFP	Specialised Foundation Programme

03

CHAPTER 3

Statistics

3.1 Introduction

Statistics is the science of collecting, analysing, interpreting and presenting data. This chapter will provide a concise overview of some of the key concepts involved in statistics along with worked examples highlighting the application of different statistical methods.

3.2 Obtaining and describing data

3.2.1 Types of data

Different types of statistical methods are used to analyse different types of data. The main types of data are summarised below.

Numerical data

This is quantitative data. The two main types include continuous and discrete data.

- **Continuous data**
  - Continuous data can take any numerical value and can be meaningfully subdivided into finer levels. Continuous data is usually measured on a scale or a continuum.
  - Measurements such as height and weight fall into this category, e.g. 1.54m and 53.4kg.
- **Discrete data**
  - Discrete data can only take certain numerical values, usually integers. The discrete values cannot be subdivided and therefore only a limited number of values is possible.
  - Examples include number of people or number of hospital visits. In these examples, it is not possible to subdivide integers into smaller increments such as half a person or half a hospital visit.

Categorical data

This is data that has been grouped into categories on the basis of qualitative features. The types include nominal, ordinal and binary data.

- **Nominal data**
  - Nominal data is grouped into categories that cannot be ordered.
  - Examples include blood group or ethnicity.
- **Ordinal data**
  - Ordinal data is grouped into categories that can be ordered.
  - Examples include tumour stage.
- **Binary/dichotomous data**
  - Binary data refers to data where there are only two possible values (e.g. 0 or 1) or two possible categories (e.g. dead or alive).

3.2.2 Obtaining data, i.e. sampling

Clinical researchers are usually interested in populations. Populations are defined as groups of individuals who share a common characteristic or condition, usually a disease. For example, a researcher investigating rheumatoid arthritis is interested in all patients who have been diagnosed with this condition. In clinical research, it is generally not feasible to study an entire population, and therefore a subset or sample of the population is recruited to the study. The study reports the results obtained from the

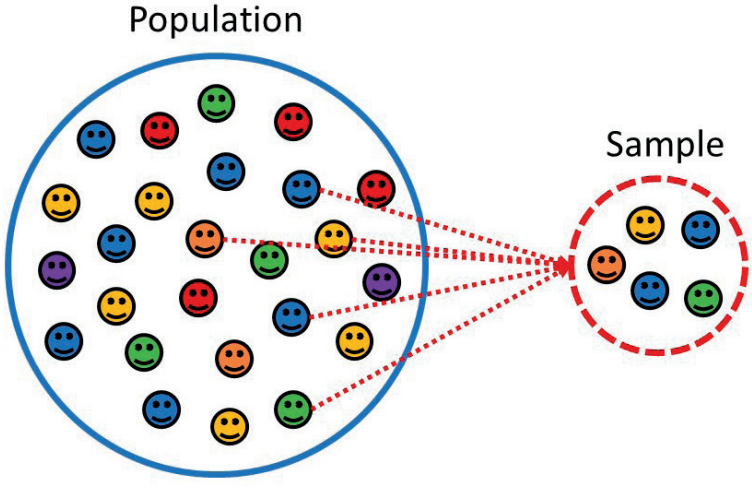


Figure 3.1. Obtaining a sample from a population. A sample is a subset of the population that is included in a research study. The sample should be representative of the population in order for researchers to infer conclusions regarding the wider population.

sample (the sample estimate) and this is used to estimate the value in the population (the population parameter). In order for the sample estimate to accurately reflect the population parameter, the individuals in the sample must be representative of the individuals in the population (see the sampling section below for more details on how this is achieved). When the sample is not representative of the general population, the sample estimate does not equal the population parameter. This is sampling error. When significant sampling error occurs, the results of the study cannot be generalised to the overall population, thus limiting the external validity of the study.

Sampling methods

Different methods are used to select a sample for inclusion into a clinical study. Three examples of sampling method are summarised below.

- **Simple random sampling:** in this scenario, every member of the population has an equal probability of selection into the sample. Theoretically, a researcher may have a list of all of the patients diagnosed with Goodpasture's syndrome. In this case, the list is called the "sampling frame". In order to select a sample, individuals are randomly drawn from the list using methods such as a random number generator. In reality, the entire population of patients with a defined disease, such as Goodpasture's syndrome, is not usually available for possible recruitment and this type of sampling is generally not feasible.
- **Stratified random sampling:** this is a modification of random sampling. In this scenario, the whole population is divided into homogenous strata according to demographic or clinical factors (examples include gender, ethnicity and comorbidity). After the population are divided into strata, the researcher selects a random sample of individuals from each stratum to be included into the clinical trial.

  Stratified random sampling is a widely-used sampling method in clinical trials. It allows researchers to study effect sizes between different groups and allows sampling from under-represented categories.
- **Convenience sampling:** in this method, participants are recruited on the basis of availability and ease of access. For example, in a study investigating patients with rheumatoid arthritis, individuals attending outpatient rheumatology clinics within the study period are recruited. Convenience sampling is a widely used method for subject recruitment as it is easy, cheap and quick. On the downside, convenience sampling may introduce an element of bias into subject recruitment. For example, individuals attending clinic may be more compliant with medical treatments than the individuals who do not attend their appointments.

Accuracy versus precision

When obtaining a sample estimate, it is important to consider its accuracy and its precision.

Precision: a measure of how close measured values are to each other
Accuracy: a measure of how close the sample estimate is expected to be to the population parameter

3.2.3 Measures of central location

Measures of central location represent the average values in a dataset. The most commonly used measures include the mean, median and mode.

The **arithmetic mean** is the best-known average value. It is calculated by summing all of the values in a dataset and dividing by the total number of values. The mean is easy to calculate and convenient to use in many contexts. By considering all values in a dataset, the mean is the most sensitive method to measure an average. The main disadvantage to using the mean is that it is highly influenced by extreme values (or outliers).

The **median** lies at the midpoint of all values in a dataset when they are ordered numerically. Therefore, 50% of values in a dataset are greater than the median value and 50% of values are lower. The median is not influenced by extreme values and is preferable to the mean when outliers are present.

The **mode** is the value that occurs most frequently in a dataset. The mode is not commonly used as a measure of average as it is not generally representative of the data. It is, however, useful to know whether a dataset has one or two modal values. When a dataset has one modal value, it is described as unimodal.

3.2.4 Measures of spread

Common measures of spread include the range, interquartile range, variance and standard deviation.

The **range** is the difference between the largest and smallest value in a dataset. It is very simple to calculate but may not be representative of the dataset, particularly when outliers are present.

The **interquartile range (IQR)** is calculated by ordering the dataset, dividing it into quartiles and calculating the difference between the bottom and top quartile. The IQR therefore indicates where the middle 50% of the data lie. The IQR is not influenced by outliers and therefore is a useful measure of spread when data is not symmetrically distributed.

A **box and whisker plot** is a common method to display the median, IQR and the range of a dataset. Its interpretation is outlined in *Figure 3.2*.

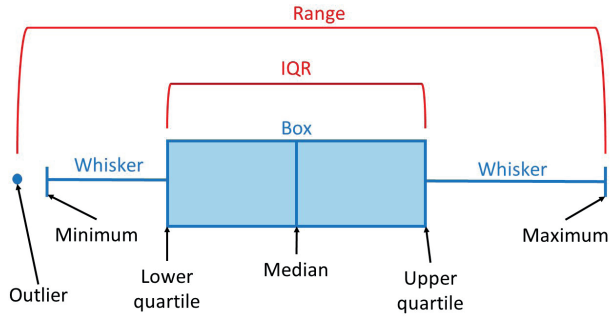


Figure 3.2. The box and whisker plot displays the median, IQR and range of a dataset. Sometimes, outliers are represented as points outside of the whiskers.

The symmetry of the box and whisker plot gives useful information regarding the distribution of the data. This is discussed in more detail in *Section 3.3.1*.

The **variance** measures the degree to which individual values in a dataset deviate from the mean. The larger the variance, the larger the spread of the data.

The variance is calculated using the following steps:

1. Subtract the mean from each value in the dataset
2. Square each of the differences and add all of the squares together
3. Divide the sum of the squares by the number of values in the dataset minus 1

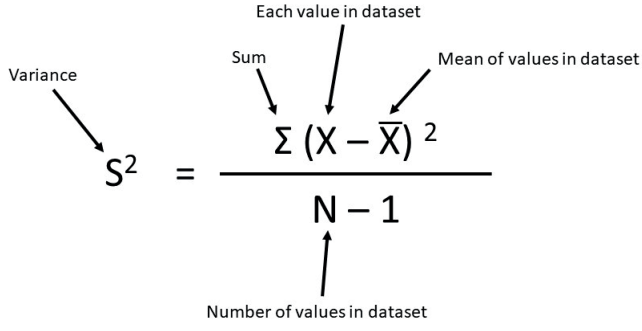


Figure 3.3. Equation to calculate the variance of values in a dataset.

The **standard deviation (SD)** is derived from the variance and is a very commonly reported measure of spread. It is calculated by performing the square root of the variance. Therefore, the larger the standard deviation, the larger the spread of the data. Both the variance and the SD are calculated using computer programs, such as SPSS or GraphPad.

WORKED EXAMPLE
Sphingosine-1-phosphate and CRP as potential combination biomarkers in discrimination of COPD with community-acquired pneumonia and acute exacerbation of COPD

Hsu *et al.* (2022) *Resp Res, 23*: 63, doi.org/10.1186/s12931-022-01991-1

Study aim

This study evaluated the use of the blood marker sphingosine-1-phosphate (S1P) to differentiate between community-acquired pneumonia and acute exacerbation in patients with COPD.

Results

The following box and whisker plots show the S1P readings in COPD patients with acute exacerbation (AE) compared to those with pneumonia (Pn).

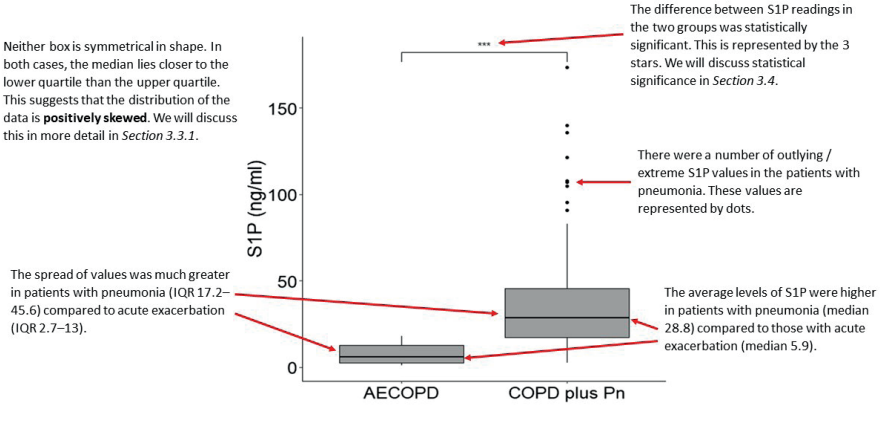


Figure 3.4. A worked example of the interpretation of two box and whisker plots. Image reproduced under a CC BY 4.0 licence.

Comments

- This example nicely illustrates the importance of choosing the correct measure of average and spread to describe your study data. In this case, the median and IQR were used because the data was skewed and because there were a number of outlying data points. In this example, it would have been inappropriate to use the mean and SD, which are highly influenced by outlying values.

3.3 Distribution, probability and confidence intervals

3.3.1 Types of distribution

Throughout this book, we will consider two main types of distribution:

1. **Probability distribution**: this is a mathematical distribution that gives us the predicted probability of an outcome occurring. Probability distributions have important applications in medical statistics, including in the calculation of confidence intervals and in hypothesis testing.
2. **Frequency distribution**: this gives us the observed frequency of a particular data point in a study or experiment. A frequency distribution is plotted on a histogram after data has been collected. In clinical research, quantitative data can follow a variety of different frequency distributions. These are important to consider because they influence hypothesis testing and statistical analysis (discussed in *Section 3.5*).

Probability and probability distributions

Probability is an important concept in statistics. It is defined as a measure of uncertainty, i.e. how likely something is to occur. Numerically, it is usually expressed as a value between 0 and 1, where 0 is impossible and 1 is certain.

Probability distributions are theoretical distributions that show the probability of all of the possible values of a random variable. For example, imagine the probability distribution when rolling two dice. Each die has a 1 in 6 (0.17) probability of rolling any number, one to six. When rolling two dice, the sum of the rolled values on the two dice will form the probability distribution illustrated below.

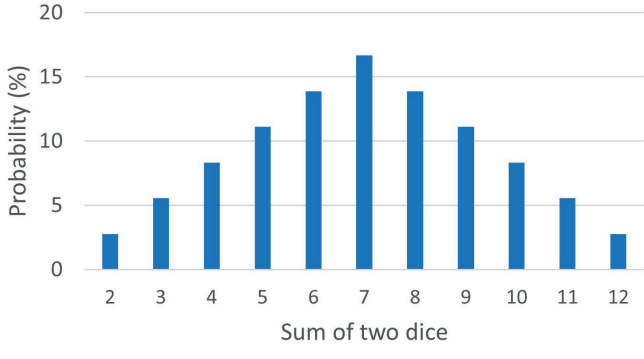


Figure 3.5. Probability distribution for sum of rolling two dice.

As you can see, seven is the most likely number to roll and occurs in 6 out of 36 rolls (17% of rolls). Conversely, rolling a two is much less likely and will only occur in 1 out of 36 rolls (3% of rolls). In this example, the probability distribution gives us the predicted outcome of rolling each number. If an experiment was performed where two dice

were rolled 50 times and the results were plotted, the frequency distribution would differ from the probability distribution due to the effects of chance. As the number of rolls increases, the frequency distribution approaches the shape of the probability distribution.

The normal distribution

Probability and frequency distributions can follow a **normal distribution**. This is a symmetrical distribution that follows a bell-shaped curve with a single peak. Mathematically, the normal distribution follows a Gaussian curve and is symmetrical around the mean value. When data is normally distributed, the mean, median and mode are equal. The width of the curve depends on the variance; as the variance increases, the curve becomes wider.

When data is distributed normally:

- 68% of values fall in the range: mean –1SD to mean +1SD
- 96% of values fall in the range: mean –2SD to mean +2SD
- 99% of values fall in the range: mean –3SD to mean +3SD

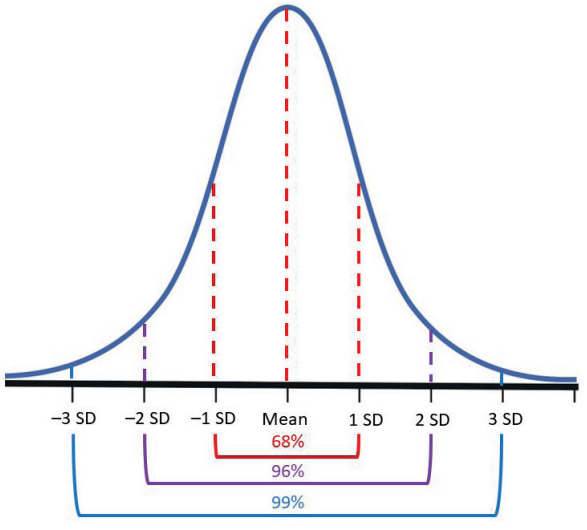


Figure 3.6. Normal distribution curve.

Pragmatically, 95% of data is considered to fall within 2 SD of the mean (rounded from 96%). This distribution range is commonly used to arbitrarily represent the 'normal range'. When values fall outside of this range, they are reported as abnormal. This can be useful in the interpretation of test results and in establishing reference ranges, for example, in the measurement of blood biochemical markers.

Skewed datasets

In some cases, frequency distributions are asymmetrical with a substantially longer tail on one side of the frequency histogram. In these cases, data is termed skewed and the direction of skew depends on the tail.

Positively skewed data has a longer tail on the right. In other words, there are a relatively large number of low values and a lower number of extreme higher values. In positively skewed data, the mean is greater than the median which is greater than the mode.

Negatively skewed data has a longer tail on the left. In other words, there are a relatively large number of high values and a lower number of extreme low values. In negatively skewed data, the mode is greater than the median which is greater than the mean.

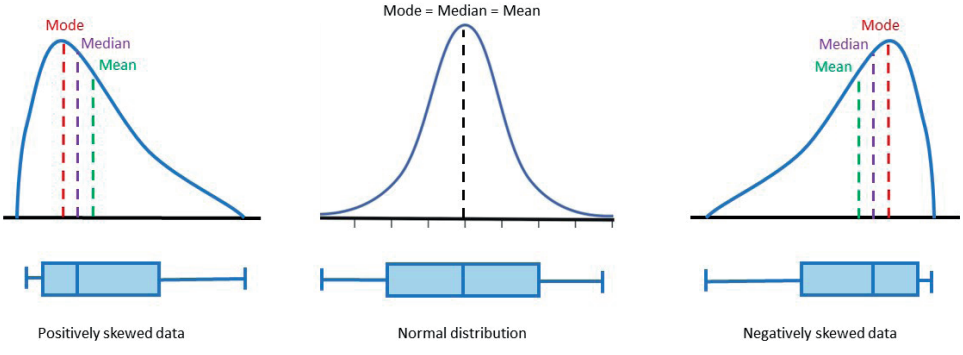


Figure 3.7. Skewed datasets versus the normal distribution.

The t *distribution*

The ***t* distribution** is a probability distribution that is widely used in statistics. It is most commonly used in studies with a small sample size, usually under 30, and when the population standard deviation is unknown. The *t* distribution looks very similar to the normal distribution curve but shorter and wider, reflecting a greater degree of uncertainty. The exact shape of the *t* distribution is influenced by the mean, variance and degrees of freedom (df) of the data, where df equals the sample size –1.

The *t* distribution has two main applications in statistics:

1. Calculation of the confidence interval (discussed in *Section 3.3.3*)

2. Testing hypotheses about one or more means (discussed in *Section 3.4*).

Figure 3.8 illustrates the *t* distribution. As the sample size increases, the *t* distribution approaches the normal distribution curve. When the sample size is greater than 30, the *t* distribution is very similar to the normal distribution.

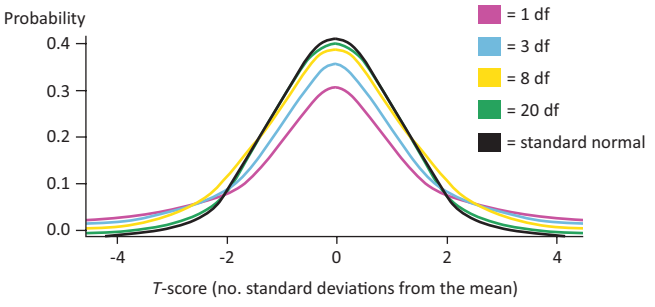


Figure 3.8. The *t* distribution with increasing sample size. As the sample size increases, the *t* distribution approaches the normal distribution.

3.3.2 Standard error of a sample mean

Due to sampling error (discussed in *Section 3.2.2*), the sample estimate varies between different studies and may not always be an accurate representation of the population parameter. For example, imagine that you are interested in estimating the mean HbA1c of all patients with a diagnosis of type 2 diabetes under the care of an endocrine team at a tertiary hospital. In order to do this, you decide to measure the mean HbA1c in a sample of diabetic patients. If this process was repeated 100 times, 100 different sample estimates would be derived and a histogram of these estimates could be plotted. This distribution represents the sampling distribution of the mean. *Figure 3.9* demonstrates the relationship between the frequency distribution for the population parameter and the sampling distribution of the mean in studies with different sample sizes (sample

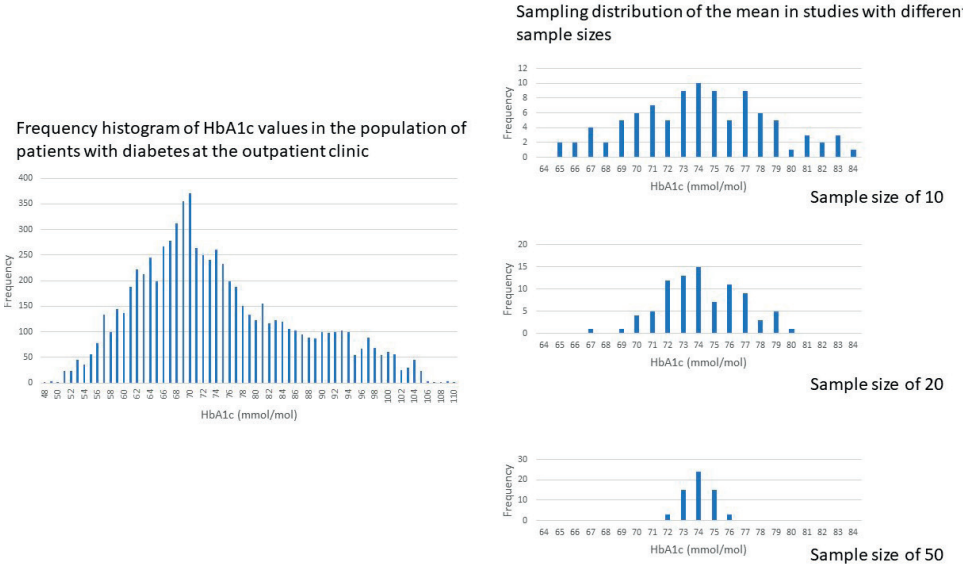


Figure 3.9. Histograms showing the relationship between the frequency distribution of the population parameter and the sampling distribution of the mean. As the sample size increases, the sampling distribution of the mean narrows and the sample estimates become a more precise representation of the population mean.

sizes of 10, 20 or 50). In this example, when the sample size is 10, the range of the sample estimates of HbA1c value is large and ranges from 65mmol/L to 84mmol/L. When the sample size is large, the sample estimates lie closer to the true population mean HbA1c value and the range of estimates decreases. When the sample size is greater than 30, the estimates of the mean follow a normal distribution.

Statistically, the difference between the sample estimate and the population parameter is quantified using the **standard error of the mean (SEM)**. The smaller the SEM, the greater the precision of the sample estimate.

Mathematically, the SEM is calculated using the following formula:

SEM = standard deviation / square root of the sample size

As illustrated in the mathematical formula above, the SEM is influenced by two factors: the standard deviation of the sample estimates and the size of the sample. By increasing the study sample size, the SEM decreases and the sample estimate is more precise. Precision also increases when the sample variance decreases.

3.3.3 Confidence intervals

A research study allows us to calculate a point estimate of the population parameter of interest. Whilst the SEM represents the precision of the estimate, it is not intuitive or easily interpretable by most clinicians. Therefore, the SEM is generally used to estimate the **confidence interval (CI)** for the parameter.

The CI gives a range in which the true population parameter is likely to lie. Most commonly, the 95% confidence interval is used. This is the interval around the sample estimate in which there is a 95% probability that the population parameter lies.

The 95% confidence interval can be calculated using two main methods:

Method 1: Calculation using the normal distribution

As discussed in *Section 3.3.2,* provided that the sample size is large, the sample means follow a normal distribution around the population parametric. We also know that in a normal distribution, approximately 95% of values fall within 1.96 SD of the mean (as discussed in *Section 3.3.1*). When referring to sample estimates in relation to the population parameter, the SD is termed the SEM.

Therefore, in order to calculate the 95% confidence interval, we can apply the following formula:

95% CI = *from* **sample mean – (1.96 × SEM)** *to* **sample mean + (1.96 × SEM)**

Method 2: Calculation using the *t* distribution

Strictly speaking, we should only use Method 1 when the variance in the population is known.

Moreover, Method 1 should only be used when the data is normally distributed. This might not be the case when the sample is small.

If it is not appropriate to use Method 1, we can calculate the confidence interval using the *t* distribution:

$$\textbf{95\% CI} = \textit{from}\ \textbf{sample mean} - (t_{0.05} \times \textbf{SEM})\ \textit{to}\ \textbf{sample mean} + (t_{0.05} \times \textbf{SEM})$$

where $t_{0.05}$ is obtained from a *t* distribution table which can be found online or in a statistics textbook. In order to find the relevant value for $t_{0.05,}$ we simply need to reference the value corresponding to the study's degrees of freedom and the desired significance levels. We will talk about this in more detail in *Section 3.4*.

Interpretation of the confidence interval

Consider the following example. In this study, investigators measured the mean change in blood pressure in study participants prescribed a trial medication versus placebo. *Figure 3.10* visualises their results, with the sample estimate represented by the circles and the 95% confidence interval represented by the horizontal lines.

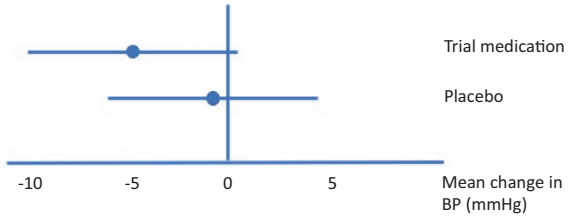


Figure 3.10. A visual representation of sample estimates and confidence intervals. In this example, investigators measured the mean change in blood pressure in study participants prescribed a trial medication versus placebo.

The CI gives us three useful pieces of information:

1. The **width of the confidence interval** represents the precision of the sample estimate. The wider the confidence interval, the less precise and greater the uncertainty of the estimate. In this example, the confidence interval for the trial medication and for the placebo are both around 10mmHg. This suggests that there is a large degree of uncertainty in the estimates for both medications.
2. The **range of the confidence interval** quantifies the magnitude of the effect of interest and enables us to assess the clinical implications of the result. In this example, the effect of the trial medication can be anywhere between a reduction in blood pressure of 10mmHg versus an increase in blood pressure of 1mmHg.
3. The **position of the confidence interval** relative to values of interest, most notably the line of null effect (the value at which there is no association between exposure and outcome or no difference between interventions on outcome). In this example, the confidence interval for the trial medication crosses zero; this suggests that the treatment may not have any effect on blood pressure. This result is commonly extrapolated to suggest that there is no effect of treatment with a significance level of 0.05 (discussed in *Section 3.4.2*).

WORKED EXAMPLE
Hemodynamic effects of sacubitril/valsartan in patients with reduced left ventricular ejection fraction over 24 months: a retrospective study

Abumayyaleh *et al.* (2022) *Am J Cardiovasc Drugs,* **22**: 535, doi.org/10.1007/s40256-022-00525-w

This study aimed to characterise the haemodynamic effects of sacubitril/valsartan in patients with heart failure. In patients who responded to this medication, the risk factors for mortality were investigated. *Table 3.1* summarises the effect of three of the hypothesised predictors of mortality.

Table 3.1: The effect of type 2 diabetes mellitus, congestion at admission and coronary artery disease on mortality

	HR	95% CI	*P* value
Type 2 diabetes mellitus	2.17	0.59–7.92	0.24
Congestion at admission	5.57	1.45–21.48	0.01
Coronary artery disease	3.70	0.47–29.44	0.22

The first column summarises the hazard ratio (HR) corresponding to the three hypothesised predictor variables. As discussed in *Chapter 2*, a HR >1 indicates that the predictor is associated with the outcome. As all three comorbidities have a HR greater than 1, they are all associated with an increased risk of mortality in this cohort. The size of the HR gives us an estimate of the increase in risk. For example, in this study, patients with type 2 diabetes are at 2.17 times greater risk of death than those without. Although this sounds very significant, we need to review the confidence interval before interpreting the significance of these results.

The second column summarises the 95% CI for each of the comorbidities. We can draw the following conclusions from these values:

1. The CI intervals for all three comorbidities are very wide. This indicates low precision of the estimates and a large degree of uncertainty regarding the true HR for the population. To illustrate this further, consider the 95% CI for type 2 diabetes. We can state that we are 95% confident that the true value for HR falls between the range of 0.59 and 7.92. In other words, the true HR could be as low as 0.59 (i.e. protective against death) or as high as 7.92 (i.e. a significant risk factor for death). In this scenario, the large degree of uncertainty makes it difficult to draw any useful conclusions.
2. The CI intervals for type 2 diabetes and coronary artery disease cross the line of null effect (i.e. HR = 1). This is often used as a marker of statistical significance and when this line is crossed, it is often inferred that there is no significant association between the exposure and outcome of interest. Accordingly, in this example, both type 2 diabetes and coronary artery disease are not significantly associated with mortality. This is also represented by the *P* value of greater than 0.05. In contrast, congestion at admission is significantly associated with mortality. This is represented by a 95% confidence interval that does not cross a HR of 1 and a *P* value of 0.01. The interpretation of the *P* value is discussed in detail in *Section 3.4.2*.

3.4 Statistical hypothesis testing and significance levels

Statistical hypothesis testing is a vital concept in medical research. In order to perform statistical hypothesis testing, five main steps need to be performed:

1. Define the **null and alternative hypotheses**.
2. Choose an appropriate **test statistic**.
3. Determine the **critical value of the test statistic**; i.e. at what value do we consider the hypothesis proved or disproved? This is also known as the **significance level**.
4. Perform the statistical test and obtain the *P* value.
5. Interpret the *P* value.

Throughout the next sections of the chapter, we will discuss these steps in more detail.

3.4.1 The null hypothesis

A hypothesis is a proposed explanation for an observation and is the starting point for all clinical research. It is important to define a hypothesis prior to a clinical study taking place. This usually takes the form of the null and alternative hypotheses.

The **null hypothesis (H_0)** states that there is no difference in the outcome of interest between the defined groups.

The **alternative hypothesis (H_1)** states that there is a difference in outcome of interest between groups; this difference can be in either direction (if the hypothesis is two-tailed). One-tailed hypotheses state the direction of effect.

For example, in the ARISTOTLE trial (see *Section 2.10.9*), the null and alternative hypotheses are as follows:

H_0: There is no difference between the risk of ischaemic stroke in patients with AF on warfarin as compared to patients on apixaban.

H_1: There is a difference between the risk of ischaemic stroke in patients with AF on warfarin as compared to patients on apixaban (this could either be increased or decreased risk of stroke).

3.4.2 Significance levels and test statistics

The level of significance should be defined prior to the statistical test being performed. When defined at this stage, it is known as the **alpha value**. The alpha value is the probability of incorrectly rejecting the null hypothesis when it is actually true, i.e. finding a difference due to chance when there is in fact no difference. A value of 0.05 is conventionally chosen. This equates to a 5% chance of incorrectly rejecting the null hypothesis due to the effects of chance.

The ***P* value** is reported after the statistical test has been performed. It is defined as the probability of obtaining the result, or something more extreme, if the null hypothesis is

true. In similarity to the alpha value, a *P* value of 0.05 is conventionally chosen for statistical significance. When the *P* value is less than 0.05, there is a less than 5% probability that the null hypothesis is true, and therefore we reject the null hypothesis and accept the alternative hypothesis. As the *P* value approaches zero, there is decreasing evidence in favour of the null hypothesis.

Test statistics

The test statistic describes how closely the distribution of your data matches the distribution predicted under the null hypothesis you are using. The most commonly used test statistics include the *Z*-score and the *T*-score. Other test statistics include the *f* statistic in ANOVA and the chi-square statistic in chi-squared (χ^2) test.

1. The ***Z*-score** describes the relationship of the mean of the dataset to the mean of the population. It is measured in terms of standard deviation from the population mean. The *Z*-score ranges from –3 to +3. When the *Z*-score is 0, it equals the population mean. When it is 1, it is 1 SD from the population mean. *Z*-scores can be positive or negative depending on whether the value is greater or lower than the mean.

   Mathematically, the *Z*-score is calculated using the following formula:

   $$\boldsymbol{Z = (X - \mu) / (s/\sqrt{n})}$$

   Where:
   X = sample mean
   μ = population mean
   s = SD of the population means
   n = sample size

2. The ***T*-score** is similar to the *Z*-score but is used when the sample size is small and therefore the population mean is not known. The *T*-score is used in hypothesis testing, using the student's *t* test (see *Section 3.5.1*).

   Mathematically, the *T*-score is calculated using the following formula:

   $$\boldsymbol{T = (X - \mu)/ (s/\sqrt{n})}$$

   Where:
   X = sample mean
   μ = population mean
   s = SD of the sample means
   n = sample size

The *T*-score is traditionally referenced from a *t* distribution table. In order to look it up, you need to calculate the **degrees of freedom (df)** for your study. The df is dependent on the sample size of the study; the larger the sample size, the greater the df. Using the df, significance level and number of tails, the *T*-score can be easily referenced in a *t* distribution table – this can be found online or in any statistics textbook. In practice, *T*- and *Z*-scores are always calculated using statistical programs such as SPSS or Excel.

Degrees of freedom (df)

The degrees of freedom of an estimate is the number of independent variables in a dataset. In order to obtain the degrees of freedom for a sample estimate, subtract 1 from the number of measurements. For example, imagine that you are estimating the mean blood pressure reduction with a new medication. If you use 10 people, the df is 9 and if you use 200 people, the df is 199. When calculating df using more than 1 sample from each patient or ANOVA tests, different formulae need to be used.

Once the test statistic has been generated, we can consider its location on the normal or *t* distribution curve. The central region of the curve is the acceptance region. When the test statistic falls within this region, we can infer that there is no statistically significant difference between the sample estimate and the population parameter and we will accept the null hypothesis. The tail(s) of the *t* distribution are the rejection regions. Mostly commonly, both tails are used as rejection regions; this is the case in two-tailed significance tests. Conversely, in one-tailed significance testing, only one tail is used as the rejection area. Hence the hypothesis predicts the direction of effect. In either case, when the sample estimate falls within these areas, we infer that there is a **statistically significant** difference between the sample estimate and population parameter and we reject the null hypothesis. Traditionally, a cut-off value of 1.96 is used for the rejection region; this equates to a significance level of 5%. Statistical tests are used to derive a *P* value from a test statistic. The choice of statistical test is determined by the type of data. This is discussed in *Sections 3.5* and *3.6*.

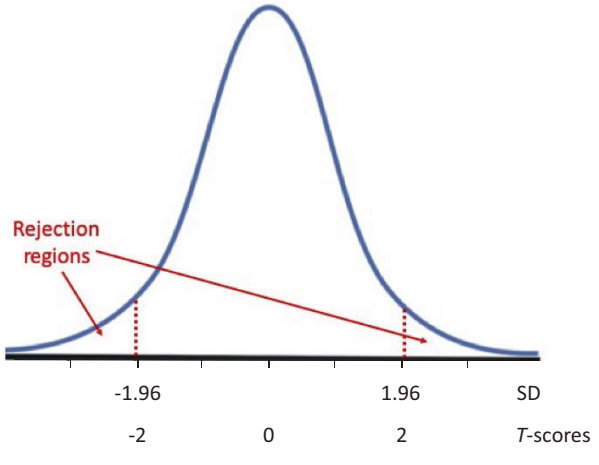


Figure 3.11. Rejection regions on a normal or t distribution curve. In this example, the rejection regions are two-tailed with a significance level of 0.05.

3.4.3 Types of error (type I versus type II)

There are two main types of error that occur in hypothesis testing.

A **type I error** occurs when a significant difference between groups is reported when in reality, one does not exist. This is a false positive result, e.g. reporting that an

antihypertensive is superior to placebo in lowering blood pressure when in reality there is no difference. Mathematically, the alpha value is the probability of obtaining a type I error. A type I error is more likely to occur in studies with a small sample size due to increased effect of confounding.

A **type II error** occurs when it is incorrectly reported that there is no difference between the two groups when in reality one exists. This is a false negative result, e.g. reporting that an antihypertensive medication does not lower blood pressure when in reality it does. Mathematically, the probability of a type II error is denoted as the **beta value**. Type II errors are more likely to occur when the sample size is small and the study is not powered to detect clinically significant differences between the two groups.

3.4.4 Statistical power

Statistical power is defined as the probability of rejecting the null hypothesis when it is false. It is defined mathematically as 1 – beta. Increased statistical power is associated with a reduced chance of incorrectly failing to reject the null hypothesis, i.e. obtaining a false negative result. In most studies, a power of 80–90% is chosen (this represents 10–20% chance of incorrectly rejecting the null hypothesis).

The power of the study is affected by four major components:

1. Sample size

   As sample size increases, power increases.

2. Effect size

   The larger the effect size of the treatment, the easier it will be to detect a difference between treatment arms and the larger the power.

3. The variability of the observations

   As the variability of the observations increases, the power decreases.

4. Significance level

   The larger the significance level, the greater the power of the study but the larger the probability of making a type II error.

In order to increase the power of the study, a larger sample size is generally required. In addition to being more expensive, resource-intensive and time-consuming, recruiting a larger sample may be unethical as more participants are unnecessarily recruited to an experimental study (discussed in *Chapter 4*). Therefore, prior to performing the study, a sample size estimation should be performed. This is an estimation that calculates the sample size required to detect a clinically significant difference. The considerations needed when performing a sample size estimation are summarised in *Section 3.9*.

3.4.5 Multiple hypothesis testing and adjustment

In some studies, multiple hypotheses are tested using the same sample. As the number of hypotheses tested increases, the chance of a type I error increases dramatically. This creates issues with data interpretation and deciding whether a result is truly significant

or not. For example, imagine that you are testing 20 different hypotheses in one sample, each with a significance level of 0.05. Mathematically, the probability of obtaining a significant result through the effects of chance alone is equal to $1 - ((1–0.05)^{20})$. This equates to a **64%** chance of observing at least one statistically significant result.

In order to deal with multiple hypothesis testing, the alpha value (the predetermined significance level) can be adjusted. The Bonferroni correction is the simplest adjustment method and adjusts the alpha value dependent on the number of hypotheses tested. It is calculated using the following formula:

Alpha / number of hypotheses tested

The Bonferroni correction assumes that all hypotheses are independent of each other. In research settings, this is often not the case and, in some circumstances, can be overly conservative and result in a very high rate of false negative results.

WORKED EXAMPLE
Efficacy and safety of tofacitinib monotherapy, tofacitinib with methotrexate, and adalimumab with methotrexate in patients with rheumatoid arthritis (ORAL Strategy): a phase 3b/4, double-blind, head-to-head, randomised controlled trial

Fleischmann *et al*. (2017) *Lancet*, 390: 457, doi.org/10.1016/S0140-6736(17)31618-5

The ORAL Strategy study was a phase 3b/4 RCT investigating the efficacy of tofacitinib (a janus kinase inhibitor) compared to adalimumab (an anti-TNF biologic) for the treatment of patients with RA. In this study, participants were randomised to one of three arms; tofacitinib monotherapy (A), tofacitinib with methotrexate (B) or adalimumab with methotrexate (C).

In a two-armed trial, there is only one comparison (A vs. B). In this study, there were three study arms and hence three comparisons (A vs. B, B vs. C, A vs. C). As the number of comparisons increase, the probability of obtaining a false positive result (i.e. a type I error) increases. Some studies, including this example, use a Bonferroni correction to account for this.

In this example, the study investigators used three study arms and hence three comparisons. Therefore, an alpha value of 0.0167 (0.05/3) was used to preserve the overall type I error rate to 5%.

The Bonferroni correction can be used for studies where multiple comparisons are used. These most commonly include studies with more than two treatment arms and studies with multiple endpoints. There is some debate as to when a Bonferroni correction should be used and concern that it increases the risk of false negative results (type II error).

3.5 Statistical significance tests to compare means

In order to choose an appropriate statistical test to compare the means of samples, we need to ask ourselves the following questions.

1. Is the data continuous?

   This is a necessity for all of the following statistical tests. Categorical data will be discussed in the next section.

2. How many groups do I want to compare?

3. Is the outcome data parametric or non-parametric?

   Parametric statistics are based on assumptions about the population from which the sample was taken. In order to use parametric statistics, the population distribution frequency should follow a normal distribution. Furthermore, the variances in each group should be equal.

   Non-parametric statistics are not based on assumptions and this data can be collected from a sample that does not follow the normal distribution.

4. Are the comparison groups independent or dependent?

   In independent samples, information about subjects in one group does not provide information about the subjects in the other groups. In this scenario, groups contain different subjects and there is no meaningful way to compare them.

   In dependent samples, subjects in one group provide information about other groups. This occurs in two scenarios:

   - Measurements are taken from the same individuals at two different time points; for example, before and after an intervention. This is the most common example.
   - Measurements are taken from different subjects who have been intentionally matched to each other. For example, in case–control studies, cases and controls may have been matched on the basis of demographic or clinical features. Although the matched pairs are different people, the statistical analysis treats them as the same subject because they are intentionally very similar.

3.5.1 Selecting a statistical test to compare means

The following flowchart (*Fig. 3.12*) can be used to identify which statistical test is most appropriate for comparing means.

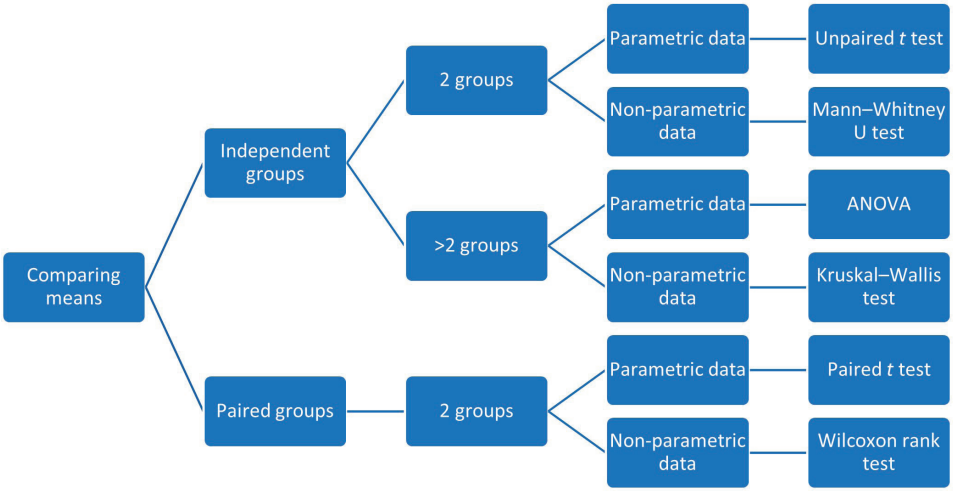


Figure 3.12. Flowchart summarising statistical tests that can be used when comparing means.

After the appropriate test has been chosen, statistical programs such as SPSS or GraphPad can be used to calculate *P* values.

3.6 Statistical significance tests to compare percentages or proportions

In order to choose an appropriate statistical test to compare differences between proportions or percentages between populations, we need to ask ourselves the following:

1. Is the data categorical? (This is a necessity for all of the following statistical tests).
2. How many groups do I want to compare?
3. Are the groups paired or independent?
4. Does the data fulfil the prerequisites required for the statistical test of note?

3.6.1 Selecting a statistical test to compare percentages

The following flowchart (*Fig. 3.13*) can be used to identify which statistical test is most appropriate for comparing percentages.

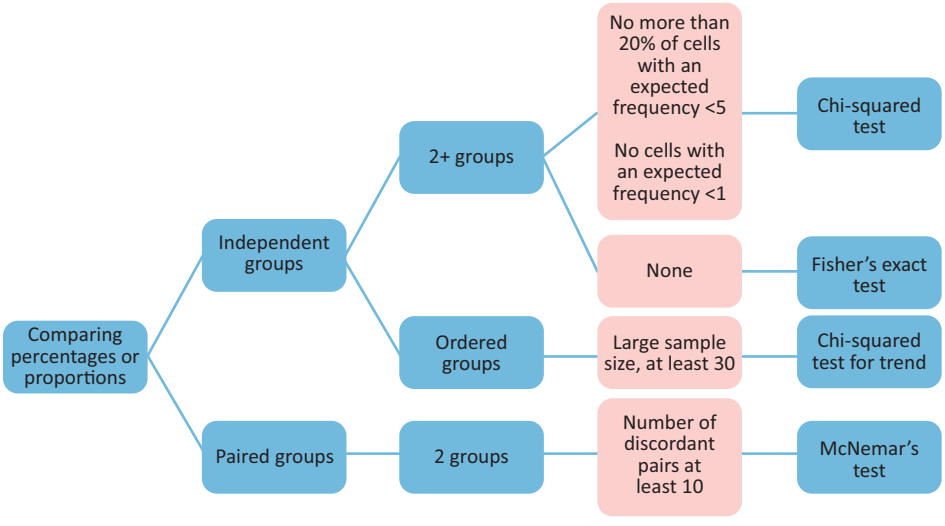


Figure 3.13. Flowchart on statistical tests that can be used when comparing percentages or proportions. Prerequisites are shown in the pink boxes.

WORKED EXAMPLE
Adalimumab reduces extraintestinal manifestations in patients with Crohn's disease: a pooled analysis of 11 clinical studies

Louis *et al*. (2018) *Advances in Therapy*, 35: 563, doi.org/10.1007/s12325-018-0678-0

Background

Extra-intestinal manifestations (EIM) are common in patients with Crohn's disease. This study aimed to investigate the effect of the biological medication adalimumab on EIM in patients with Crohn's disease.

Methods

In order to compare the differences in EIM between the two groups, the authors first described the demographics of recruited subjects and used statistical testing to determine whether there was any difference between the two groups. This is an important part of any study comparing two groups of patients and will be the focus of this worked example.

T tests were used to compare the differences between groups in continuous variables. These included age, disease duration and biochemical markers such as CRP. In this example, the *t* tests are unpaired because the values in one group do not influence the readings in the other. By choosing the *t* test, the authors are assuming that the data is parametric.

Chi-squared tests were used to compare differences between groups in categorical variables including gender and disease activity. By choosing the chi-squared tests, the authors assumed that the data fulfilled the prerequisites outlined in *Figure 3.13*. Data must be independent and must fulfil the expected frequencies requirement when recorded in a two-way table. In this example, a two-way frequency table for gender would appear as follows:

	Male	Female
Adalimumab group	390	747
Placebo group	101	196

In order to perform a chi-squared test, no more than 20% of cells should have an expected frequency of less than five and no cells should have an expected frequency of less than 1.

Results

Table 3.2: A comparison of demographic and disease features in patients in the study on placebo versus adalimumab

	Patients with EIM at baseline		
Characteristic	Placebo ($n = 297$)	Adalimumab ($n = 1137$)	*P* value
Age, years, mean (SD)	38.9 (11.9)	37.5 (12.2)	0.090
Female, *n* (%)	196 (66.0)	747 (65.7)	0.924
Disease duration, years, mean (SD)	11.0 (8.8)	10.3 (8.5)	
Disease activity			0.001
Moderate, *n* (%)	149 (50.2)	686 (60.3)	
Severe, *n* (%)	148 (49.8)	450 (39.6)	
Albumin, g/L, mean (SD)	39.6 (4.8)	39.7 (5.0)	0.701
CRP, mg/dl, mean (SD)	1.7 (2.7)	1.9 (2.7)	0.248

Summary data abstracted from *Advances in Therapy*, 35: 563.

Table 3.2 allows us to compare the two groups and make several conclusions:

1. The patients in the two groups are similar with regard to age, gender, disease duration, albumin level and CRP. This is demonstrated by a *P* value of greater than 0.05, indicating no difference between the two groups.
2. The patients allocated to receive placebo had higher rates of severe disease activity, compared to those allocated to receive adalimumab. This is demonstrated by a *P* value of 0.001.

Discussion

This study investigated the difference in EIM in patients with Crohn's disease on adalimumab versus placebo. In order to draw meaningful conclusions, the authors first described the demographic characteristics in the two groups. They used null hypothesis testing to do so and found that the patients allocated to placebo were significantly more likely to have severe disease. This indicates that the two groups are not well matched with respect to disease activity and this could have influenced the overall results from the study. This example describes the use of null hypothesis statistical testing to describe the differences between two groups and indicates the importance of doing this.

3.7 Measures of risk

3.7.1 Relative risk versus odds ratio

In *Chapter 2*, we discussed the calculation of relative risk and odds ratio. A summary of the differences between the two measures is illustrated in *Table 3.3*.

Table 3.3: A comparison of relative risk versus odds ratio

	Relative risk	Odds ratio
Definition	Risk in exposed / risk in unexposed	Odds in exposed / odds in unexposed
Measure	Risk = the total number of outcomes in a group divided by the number of people in the group	Odds = the number of outcomes in a group divided by the number of people in the group that did not experience the outcome
Use	Used in a variety of studies including observational studies such as cohort studies and interventional studies such as RCTs	Used in case–control studies; in such cases, the relative risk cannot be used
Interpretation	RR/OR >1: the probability of the outcome occurring is greater in the exposed than the unexposed group RR/OR = 1: the probability of an outcome occurring is the same in the exposed and unexposed groups RR/OR <1: the probability of an outcome occurring is less in the exposed than the unexposed group	
Association	OR approximately equal to RR when outcome is rare OR greater than the RR when the outcome is common	

3.7.2 Absolute risk reduction and number needed to treat

The **absolute risk reduction (ARR)** is another method to compare the risk of an outcome in one group to another. It is calculated using the following equation:

ARR = risk in exposed – risk in unexposed

The **number needed to treat (NNT)** is derived from the ARR. It is derived from the following calculation:

NNT = 1/ARR

The NNT is the number of patients needed to treat to prevent one adverse outcome (for example; death, heart attack or stroke). It is a measure that is commonly used to report the findings from RCTs and gives a measure of the benefit obtained from a treatment or intervention. The higher the NNT, the more patients need to receive treatment for any benefit to be seen. This information can be used when considering the risk:benefit

ratio of a treatment for an individual patient. On a wider level, the NNT gives us an idea of the cost-effectiveness of a treatment.

For example, in classical studies comparing thrombolysis to streptokinase for the management of stroke in the 1980s, there was a 1% ARR in patients treated with thrombolysis. Therefore, 100 patients needed to be treated with thrombolysis for a single patient to gain benefit. At a time when thrombolysis was very expensive, this benefit was not cost-effective and therefore, streptokinase remained the treatment of choice until stronger evidence was reported.

WORKED EXAMPLE
Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial

RECOVERY Collaborative Group (2021) *Lancet*, **397**: 1637, doi.org/10.1016/S0140-6736(21)00676-0

The aim of this study was to evaluate the efficacy and safety of tocilizumab therapy in adult inpatients admitted with severe Covid-19. The primary outcome was 28-day mortality.

In this trial, 2022 patients were randomised to receive tocilizumab whilst 2094 patients received standard-of-care therapy. In total, 631 participants randomised to the tocilizumab group died compared to 729 participants in the standard-of-care therapy group.

Let's consider the different comparisons of risk:

First, we need to calculate the risk of death in the two groups. This can be done by using the simple calculation:

number of events (in this case 28-day mortality) / total number of participants in group

Therefore, the risk of death in the tocilizumab group is: 621/2022 = **31%**

And the risk of death in the standard-of-care therapy group is: 729/2094 = **35%**

The relative risk is calculated by dividing the risk of death in the treatment group by the risk of death in the standard-of-care therapy group:

$$RR = 31/35 = \mathbf{0.89}$$

As the RR is less than 1, participants receiving tocilizumab had a lower risk of death than those who didn't. This is a proportional measure of risk reduction, in contrast to the ARR which can be calculated as follows:

$$ARR = 35 - 31 = \mathbf{4\%}$$

Therefore, patients receiving tocilizumab had a 4% lower risk of dying than those who didn't. This means that if 100 patients received tocilizumab, four patients would be prevented from dying. This can also be expressed as the NNT:

$$NNT = 1/ARR = 1/0.04 = \mathbf{25}$$

This means that 25 patients need to be treated with tocilizumab in order to prevent one death.

Overall, this data provided strong evidence for the survival benefit with tocilizumab in hospitalised Covid-19 patients.

3.8 Correlation and regression

3.8.1 Introduction

Correlation and regression are used to characterise the relationship between two variables. Correlation allows us to characterise the strength of the association between the two variables and can be used when neither variable is assumed to predict the other. In contrast, regression analyses are used to predict the effect of an explanatory variable on the outcome variable. These analyses can therefore only be used when one variable is thought to change the other.

3.8.2 Correlation

Correlation is a statistical technique used to measure the strength of the association between two variables. There are two main measures of correlation: Pearson correlation coefficient and Spearman correlation coefficient.

Pearson correlation coefficient (*r*) provides a measure of the correlation between two variables when the relationship between the two is linear. It is a parametric test that can be used when data is normally distributed. *r* can be calculated using most statistical programs with the result ranging from – 1 to + 1:

1. If *r* is positive, an increase in one variable results in an increase in the other.
2. If *r* is negative, an increase in one variable results in a decrease in the other.
3. The magnitude of *r* indicates how closely the data points lie in relation to the line of best fit. When *r* is +1 or –1, there is perfect correlation between the two variables.
4. r^2 represents the proportion of the variability in the dependent variable that can be explained by variability in the explanatory variable.

This is graphically depicted in *Figure 3.14*.

Spearman correlation coefficient is a non-parametric measure of correlation. It should be used if any of the following are true, in which cases, it is not possible to perform Pearson correlation.

- The sample size is small
- The relationship between the two variables is non-linear
- Neither *x* or *y* are normally distributed.

The value for Spearman correlation coefficient can be interpreted in a similar manner to Pearson correlation coefficient, with a value of +1 or –1 indicating perfect correlation and 0 indicating no correlation.

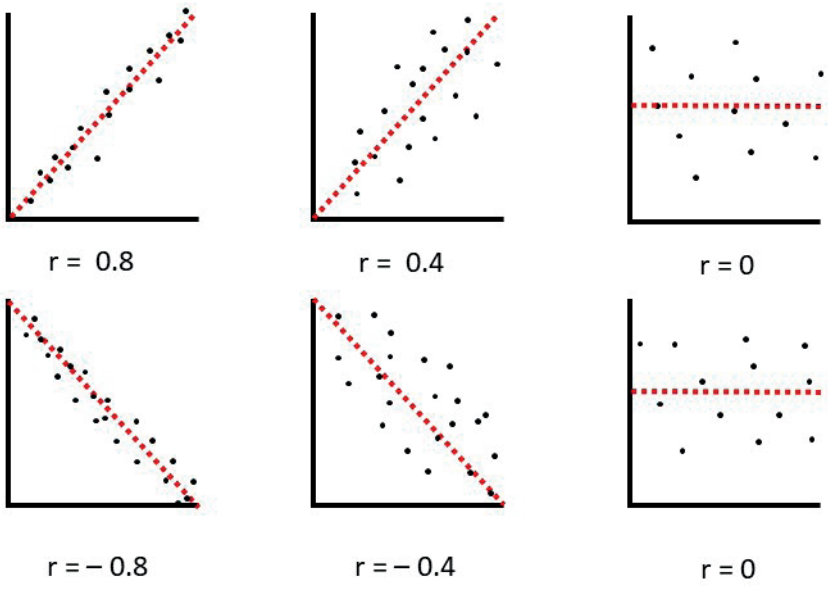


Figure 3.14. Scatter plots demonstrating positive and negative correlation with estimated *r* values.

3.8.3 Linear regression

Linear regression quantifies the linear relationship between two continuous variables; an explanatory or independent variable and a dependent variable. In these models, the **explanatory variable** is used to predict the dependent variable.

Imagine a simple study where investigators wanted to determine whether height (explanatory variable) predicts weight (dependent variable) in a cohort of children. In order to do this, the researchers measure the height and weight of 100 children. The next step would involve plotting the results on a scatter graph where *x* is the explanatory variable (height) and *y* is the dependent variable (weight). When the points form a linear relationship, we can plot the regression line; this is the line that best fits through all of the data points (*Figure 3.15*).

The following equation models the simple linear regression line:

$$y = a + bx$$

x = explanatory or independent variable e.g. height (cm)

y = dependent variable e.g. weight (kg)

a = Y intercept of the line

b = the gradient of the line, i.e. how much *y* increases for every unit increase in *x* and in this example, how much weight increases for every centimetre increase in height.

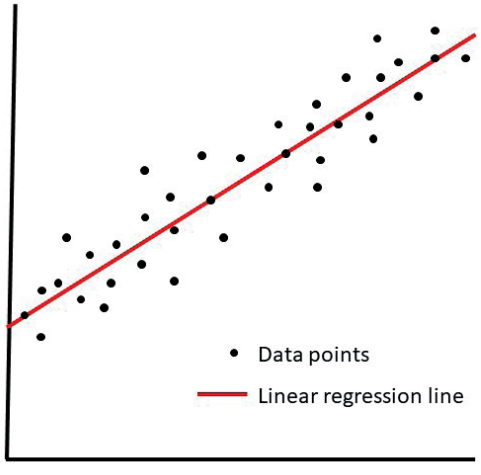


Figure 3.15. Linear regression line.

By plotting the line of best fit, simple linear regression allows us to quantify the values for a and b. This allows us to predict the value of the dependent variable (e.g. weight) for any given value of the explanatory value (e.g. height).

Now that we have discussed the interpretation of the regression line, we can consider methods used for its derivation; most commonly, the method of least squares.

The method of least squares minimises the differences between the observed values and the values predicted by the linear regression line.

Figure 3.16 illustrates the principles of this. Every observed value deviates from the linear regression line. The difference between the observed value and the line is called the residual. The method of least squares minimises the sum of the squares of these residuals, allowing the line to pass through the data points as closely as possible.

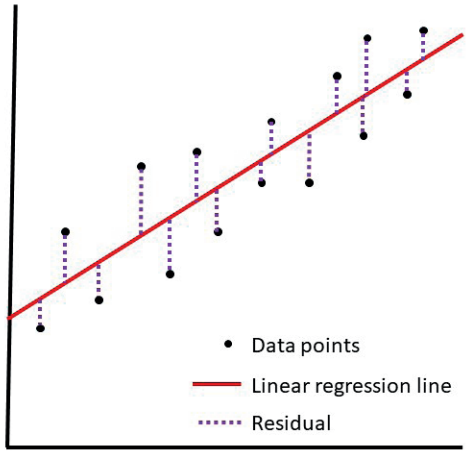


Figure 3.16. Principles of linear regression using the method of least squares.

Assessing goodness of fit

In order to judge how well the line fits the data, we can calculate R^2. This is defined as the percentage of variability of *y* that can be explained by variability in *x*. in general, the closer the points lie to the regression line, the higher the value for R^2.

Linear regression is a useful tool to model the association between two continuous variables; however, it can only be performed when the following assumptions are met:

1. There is a linear relationship between *x* and *y*.
2. The observations in the sample are independent of each other.
3. The residuals are normally distributed.
4. The residuals have the same variability for all of the fitted values of *y*.

3.8.4 Multiple regression

In some cases, investigators are interested in the effect of several explanatory variables on the dependent variable. Multiple linear regression allows us to investigate this.

Multiple regression allows us to identify whether explanatory variables are associated with the dependent variable. By incorporating more than one explanatory variable into the model, multiple regression allows us to determine the extent to which explanatory variables are associated to the dependent variable after adjusting for other variables (often confounding factors). In similarity to linear regression, multiple regression can allow us to predict the value of the dependent variable from the value of the explanatory variables.

Different regression models are used for different types of explanatory variables.

- Multiple linear regression models include more than one continuous explanatory variable.
- Multiple regression or Analysis of Covariance (ANCOVA) are models with more than one categorical explanatory variable, with or without multiple continuous explanatory variables.

3.8.5 Logistic regression

Logistic regression is similar to linear regression except that the dependent variable is a binary outcome; for example, the presence or absence of disease. Like linear regression, logistic regression allows us to characterise which explanatory variables influence the outcome. It can also be used to predict the risk of an outcome in the presence of explanatory variables (usually a risk factor for the development of a disease).

The results from logistic regression are usually presented as odds ratios (OR). In these analyses, both unadjusted and adjusted OR are usually presented. Unadjusted OR represents the association between the explanatory and dependent variable, without taking the other explanatory variables into account. In contrast, the adjusted OR represents the association between the explanatory and dependent variable when all of the explanatory variables are considered. This allows us to consider whether the association between the variables is a true association or whether it is secondary to the presence of confounding variables.

3.8.6 Comparing and contrasting correlation and regression

Table 3.4: Comparing and contrasting correlation and regression

Correlation	Regression
Measures the strength of association between two variables	Describes the effect of the explanatory variable (x) on the dependent variable (y)
The correlation coefficient measures the degree to which two variables move together	Allows us to predict the effect of changes in the explanatory variable on the value of the dependent variables
	Regression analyses quantify the change in the dependent variable for every unit change in the explanatory variable

WORKED EXAMPLE
Dose-response relation between dietary sodium and blood pressure: a meta-regression analysis of 133 randomized controlled trials

Graudal *et al.* (2019) *Am J Clin Nutr*, 109: 1273, doi.org/10.1093/ajcn/nqy384

This study aimed to investigate the effect of reducing sodium intake on blood pressure (BP) measurements in a population of hypertensive individuals. In order to do this, the investigators performed a meta-analysis of 133 RCTs investigating sodium intake on BP. The results from all of the RCTs are plotted in *Figure 3.17* with the size of the circle representing the weight of the individual RCT. In this scatter plot, systolic blood pressure effect is the dependent / *y* variable and sodium reduction is the explanatory / *x* variable. A simple linear regression model has been used to characterise the relationship between sodium reduction and BP improvement in patients with a BP greater than 130mmHg.

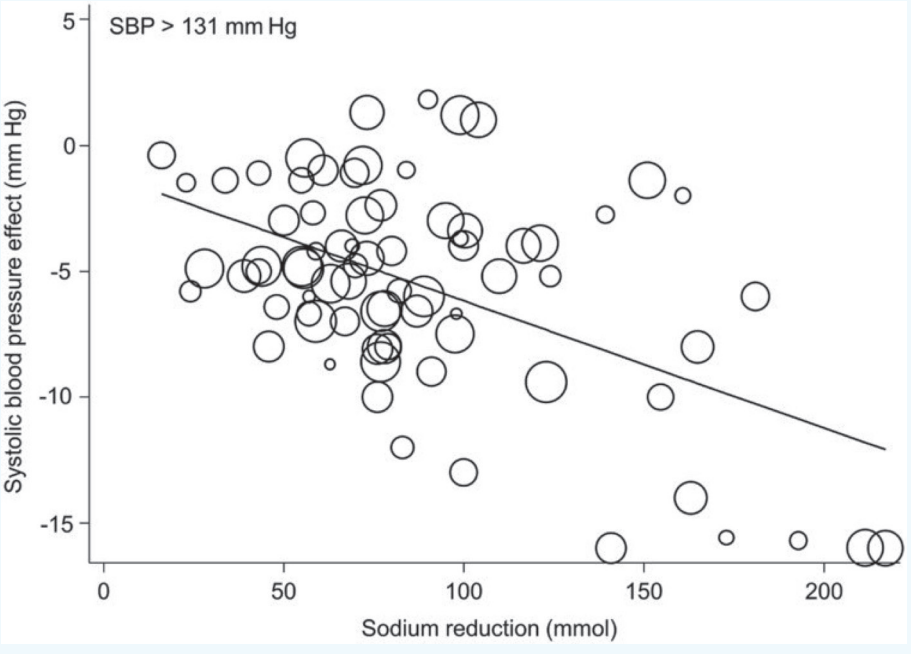


Figure 3.17. Linear regression showing the relationship between systolic blood pressure effect and sodium reduction. Each data point represents a different study, with the size of the circle representative of the size of the study. Reproduced from *Am J Clin Nutr*, 109: 1273 with permission from Oxford University Press.

We can make the following conclusion from these results:

- The gradient of the line represents the improvement in BP for a given reduction in sodium intake. In this case, there is a 5mmHg drop in BP for every 100mmol reduction in sodium intake.

We cannot make the following conclusions:

- Linear regression does not allow us to make predictions about the explanatory variable from the value of the dependent variable. Therefore, we cannot predict the sodium reduction from a patient's SBP.
- We cannot use this model to make predictions for values that fall outside of the measured range. For example, it would be inappropriate to predict the BP effect of reducing sodium intake by 300mmol.
- In this paper, a measure of goodness of fit was not reported. Therefore, we cannot comment on the proportion of variability in BP reduction explained by the sodium reduction.

In this study, the authors went on to perform a multivariable analysis to allow for adjustment of confounding factors. The results from both analyses are summarised in *Table 3.5*.

Table 3.5: Study results from the univariable analysis (simple linear regression model) and the multivariable analysis

	Univariable analysis		Multivariable analysis	
Baseline BP	Effect (95% CI) in mm Hg	*P* value	Effect (95% CI) in mm Hg	*P* value
SBP >131mmHg	−5.0 (−7.1, −3.0)	0.0001	−6.2 (−8.5, −3.9)	0.0001

As previously discussed, in the simple linear model, there was a 5mmHg drop in BP for every 100mmol reduction in sodium intake. This is a statistically significant result with a *P* value of 0.0001. In model 1 of the multivariable analysis, the researchers adjusted for possible confounding factors including baseline BP, age, ethnicity and antihypertensive use. Some of these explanatory variables were continuous (e.g. baseline BP) and others were categorical (e.g. ethnicity). When the multiple regression model was used, a significant impact of sodium reduction on BP improvement remained.

3.9 Determination of sample size

It is essential to perform a sample size calculation during the planning of any confirmatory research study. This determines the number of participants needed to detect a statistically significant difference between the study groups.

3.9.1 Exploratory versus confirmatory research

The differentiation between exploratory and confirmatory research is essential when planning and appraising clinical studies. Broadly speaking, exploratory research aims to generate new hypotheses in areas where little may be known. In contrast, confirmatory research builds upon data from exploratory research in order to test existing hypotheses. The main differences between the two research types are summarised in *Table 3.6*.

Table 3.6: Comparing and contrasting exploratory vs. confirmatory research

Exploratory research	Confirmatory research
Explores unknown research questions	Tests a priori hypotheses
Discovers new knowledge and is not based on previous studies	Generally based on previous studies
Does not offer final and conclusive statements but generates hypotheses to test in confirmatory research	Provides evidence to make inferences from the sample about the population
Less stringent research methods	More stringent research methods
Generates data for sample size calculations in confirmatory research	Sample size calculation should be performed prior to the study
Descriptive statistics and accuracy of sample estimates can be performed	Descriptive statistics and accuracy of sample estimates can be performed
Null hypothesis statistical testing should not be performed	Null hypothesis statistical testing can be performed

3.9.2 Sample size calculations

In order to perform a sample size estimation, the following pieces of information should be considered. It is important to discuss all of these factors with a statistician during the planning of any clinical trial.

- The smallest magnitude of a **clinically significant difference**
  - For example, in a study of antihypertensive medications, we need to consider the fall in blood pressure that would be considered as clinically meaningful. This is based on prior research; for example, the fall in blood pressure required to reduce the risk for cardiovascular outcomes.
- The expected **standard deviation** of observations in each group
  - This is estimated from previous research.
- The **power** that is required of the study
  - Statistical power was discussed in *Section 3.4.4*. The power of the study describes the likelihood that it will detect a clinically significant difference between groups, if one exists in reality.
  - In most studies, the power is set at 80–90%.
- The type of **statistical test** that will be performed with the results
  - When planning a study, it is useful to consider the type of statistical analysis that will be performed with the results. Prior to performing the study, researchers often meet with a statistician to develop a plan for statistical testing that can be added to the study protocol.
- The **critical level of significance** chosen
  - We know that a higher critical level of significance is associated with a decreased risk of type I error (false positive) and increased risk of type II error (false negative).
  - Traditionally, we select a cut significance level of <0.05. At this cut-off level, in 5% of cases, a type II error will be reported.

WORKED EXAMPLE

Efficacy and safety of albendazole and high-dose ivermectin coadministration in school-aged children infected with Trichuris trichiura in Honduras: a randomized controlled trial

Matamoros *et al*. (2021) *Clin Infect Dis,* 73: 1203, doi.org/10.1093/cid/ciab365

The following paragraph has been taken from a study comparing the efficacy of albendazole (ALB) and ivermectin (IVM) in the treatment of parasitic infection in children in Honduras.

Sample size was calculated estimating the efficacy of the different experimental drug or combinations and gathering the individual samples sizes for the study. The sample size was calculated using a 1-tailed test for pairwise comparisons of the expected cure rates for 4 study groups – 17% for single-does ALB, 55% for single-dose ALB-IVM, 85% for 3-dose ALB-IVM, and 60% for 3-dose AL – with an overall significance level of 5% adjusted for multiple tests by Bonferroni correction and 80% power and inflated for 10% loss to follow-up. The estimated sample size was 177 participants, included 39 participants for single-dose ALB (arm 1), 57 for single-dose ALB-IVM (arm 2), 24 for 3-dose ALB (arm 3), and 57 for 3-dose ALB-IVM (arm 4).

As you can see the authors have considered:

1. The statistical test that will be used; in this case, a 1-tailed test for pairwise comparisons.
2. The expected cure rates with the different medications; this data has been taken from the existing literature.
3. The significance level required; the study has followed convention and used a cut-off of 5%.
4. As multiple hypotheses have been tested, the authors corrected the 5% significance level using the Bonferroni correction. This was discussed in *Section 3.4.5*.
5. A power of 80% was selected; this is a common value to use.
6. The number of participants lost to follow-up was estimated to be 10%.

Overall, the sample size calculation estimated that 177 participants would need to be recruited to the study. The authors stuck to this estimation and recruited a total of 176 children to the study. Recruiting more children would be unnecessarily time-consuming and costly.

3.10 Analysis of survival data

Survival analyses characterise the time an individual takes to reach an endpoint of interest, often but not always death.

3.10.1 Kaplan–Meier survival curves

A survival curve is usually calculated by the Kaplan–Meier method and displays the cumulative probability of an individual remaining free of the endpoint of interest at

any time after baseline. The cumulative probability will only change once an endpoint has occurred. Therefore, the curve is drawn in a series of steps, starting at a survival probability of 100% and falling towards 0% as time increases.

An example of a survival curve is illustrated in *Figure 3.18*:

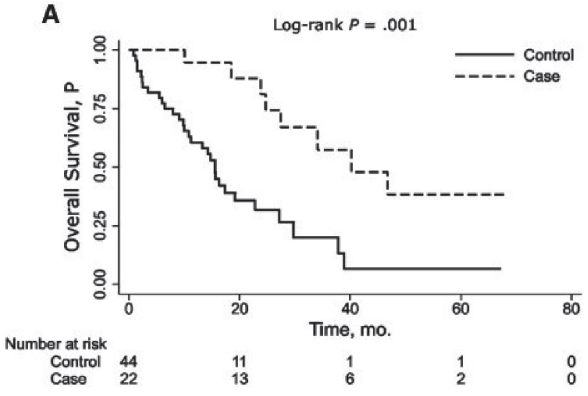


Figure 3.18. Kaplan–Meier curve showing overall survival of patients diagnosed with pancreatic cancer with ATM mutations (cases) and without (controls). Figure reproduced from *JNCI Cancer Spectrum*, 2021: 5: pkaa121, with permission from Oxford University Press.

This study characterised the survival of patients with pancreatic cancer with and without pathogenic mutations in ATM. This Kaplan–Meier curve illustrates survival time following a diagnosis of pancreatic cancer. In this example, the cases did not have pathogenic ATM mutations, whereas the control patients did. As in all survival curves, percentage survival falls with time in a stepwise fashion as events (deaths) occur.

Kaplan–Meier survival curves can be used to generate the following useful pieces of information:

- Survival rates at defined time points
  - This is a common method of explaining survival data, for example, the 1-, 5- or 10-year survival rate.
  - In the above example, the 20-month survival rate is around 85% in cases compared to only around 35% in controls.
- Median survival time (i.e. the time at which 50% of the patients are still alive)
  - In order to determine this, simply note the time at which the survival curve crosses 50%.
  - In the above example, the median survival in cases was 40.2 months compared to 15.5 months in the controls.

In the majority of cases, survival analyses measure an adverse endpoint, usually death. When the endpoint is favourable, such as recovery, the Kaplan–Meier curve is conventionally plotted upwards. In these scenarios, the curve starts at zero and increases towards 100% as time progresses.

3.10.2 Censored data

An important concept in survival analysis is censoring. When a patient's data is censored, this means that we do not know the true survival time for the patient. There are three main reasons for this:

1. The patient does not reach the endpoint by the time that the study has finished.
2. The patient withdraws from the study.
3. The patient is lost to follow-up during the period of the study.

There are two main types of censored data:

Right censored data is the most common. It occurs when the patient does not reach the endpoint during the study. This is either because the patient survives until the end of the study or because they withdraw from the study before they reach the endpoint. In both cases, the exact survival time is not known; however, the true survival time will always be greater than the observed survival time.

In contrast, in **left censored data**, the true survival time is shorter than the observed survival time. This is rare but can occur in some circumstances. For example, imagine measuring the survival time of people infected with hepatitis C. In this scenario, survival time is measured from the date of serological diagnosis. In these cases, it is generally not possible to determine the exact time of infection. As infection precedes diagnosis, observed survival time is shorter than the true survival time.

Censored data is usually plotted as a plus on the survival curve.

3.10.3 Log rank test

Statistical methods can be used to compare differences in the survival times in the two groups studied. The log rank test is one such method. This is a non-parametric test that guides the acceptance or rejection of the null hypothesis. The downside of this test is that it cannot assess the independent roles of more than one factor on the time to endpoint, and therefore cannot correct for confounding factors.

3.10.4 The Cox proportional hazards model and hazard ratios

The Cox model is the most widely used method to analyse time-to-event data. The model uses regression analysis to provide an estimate of the **hazard ratio**. This is the ratio of hazard rate in one group compared to the hazard rate in another group, where the hazard rate describes the probability of an outcome occurring over a defined time period.

The major advantage of the Cox model is that it can test the effects of explanatory variables on the time-to-endpoint. These variables can take the form of continuous, binary or categorical data. For example, imagine a study investigating the survival of patients following a new chemotherapy agent. The Cox model allows us to investigate whether a range of explanatory variables, such as patient age, sex or cancer stage, affect survival time. Clearly, this gives us important clinical information and allows us to determine which patients are most likely to benefit from this new treatment.

The major disadvantage of the Cox model is that it assumes that the hazard ratio is constant over time. When this is not the case, the Cox model should not be used, at least not in its most simple form.

Interpreting hazard ratios from the Cox model

Interpretation of the hazard ratio is similar to interpretation of the relative risk.

HR = 1 → there is no relationship between the explanatory variable and the outcome of interest.

HR <1 → the explanatory variable is protective against developing the hazard.

HR >1 → the explanatory variable is a risk factor for developing the hazard.

WORKED EXAMPLE
Association between bone mineral density at different anatomical sites and both mortality and fracture risk in patients receiving renal replacement therapy: a longitudinal study

Jaques *et al*. (2022) *Clin Kidney J,* 15: 1188, doi.org/10.1093/ckj/sfac034

This study aimed to investigate the effect of bone mineral density (BMD) on survival and fracture risk in patients with end-stage renal disease on renal replacement therapy (RRT).

Figure 3.19 represents a Kaplan–Meier curve demonstrating the risk of fracture (hip and overall) in patients with low BMD (blue line) compared to high BMD (red line).

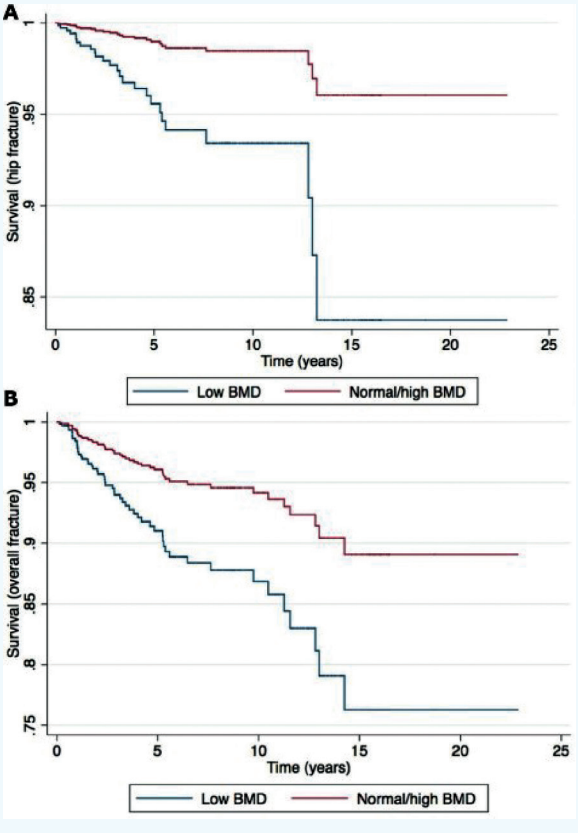


Figure 3.19. Worked example outlining the interpretation of Kaplan–Meier curves. These Kaplan–Meier curves clearly illustrate that patients with a low BMD have increased risk of hip and overall fracture compared to those with normal or high BMD. At 10 years, around 7% of patients with a low BMD have suffered from a hip fracture compared to around 2% of patients with normal or high BMD. Reproduced from *Clin Kidney J,* 15: 1188 with permission from Oxford University Press.

Table 3.7: Table summarising HR (95% CI) and *P* values for hip and any fracture using BMD as a predictor variable (normal/high versus low). In this example, the authors have used three models to assess the risk of BMD on hip and any fracture.

Model	HR (95% CI)	*P* value
Hip fracture		
Univariate model	0.21 (0.10–0.45)	<0.001
Partially adjusted model	0.33 (0.15–0.74)	0.007
Fully adjusted model	0.22 (0.08–0.62)	0.004
Any fracture		
Univariate model	0.30 (0.18–0.50)	<0.001
Partially adjusted model	0.45 (0.26–0.77)	0.004
Fully adjusted model	0.42 (0.21–0.83)	0.013

As you can see, patients with a normal/high BMD have a HR <1 when compared to those with a low BMD. This suggests that normal/high BMD is protective against fractures (hip and total).

The use of different models in this analysis allows for adjustment for confounding factors. The first model (the univariate model) does not correct for confounding factors. The partially adjusted model corrects for the confounders RRT mode, age and gender. The fully adjusted model corrects for the above confounders in addition to BMI, ethnicity, gender, PTH, smoking and CRP. These are all factors that are known to influence fracture risk. In all three models, the effect of BMD on fracture risk remains significant. Therefore we can conclude that low BMD is associated with increased risk of fracture, even when confounding factors are accounted for.

Conclusion

In conclusion, this study demonstrated that patients on RRT with a low BMD were more likely to suffer from fractures than those with a normal or high BMD.

3.11 Meta-analysis

As discussed in *Section 2.11.3*, a meta-analysis is a type of systematic review that combines numerical data from multiple studies.

3.11.1 Forest plots

A **forest plot** is the most common method to display the results from a meta-analysis (see *Fig. 3.22* for an example). A forest plot summarises the estimated effect from individual trials and a summary measure which is derived from pooling the results from all of the studies.

The solid vertical line in a forest plot represents the 'line of no effect'. This corresponds to a relative risk or odds ratio of one.

The results from individual studies are plotted vertically, with each study graphically represented by a box and a line. The location of the box gives us the effect estimate from the study of note. The size of the box corresponds to the weighting of the study to the summary measure. This is usually dependent on the size of the study, with larger studies carrying more weight and hence represented by larger squares. The length of the horizontal line represents the confidence interval of each study.

The summary measure is usually represented by a diamond. This is the weighted average of the effect estimates from all included studies and gives us the effect estimate from the meta-analysis. The horizontal length of the diamond represents the confidence interval of the summary estimate. The longer the diamond, the less certain we are in the summary result.

3.11.2 Publication bias

As discussed in *Chapter 2*, one disadvantage to meta-analysis is publication bias. This describes the tendency for studies with positive results to be published over those with negative results.

To consider whether publication bias is present, we can draw a **funnel plot**. This is a scatter diagram of all published studies with treatment effect on the horizontal axis and a measure of study precision, such as standard error, on the vertical axis.

When there is no publication bias, the funnel plot is symmetrical. If publication bias is present, the funnel is asymmetrical. This is depicted graphically in *Figure 3.20*, where published studies are generally larger or have a larger effect size. By omitting the results from the missing or unpublished studies, the estimated effect size is artificially inflated.

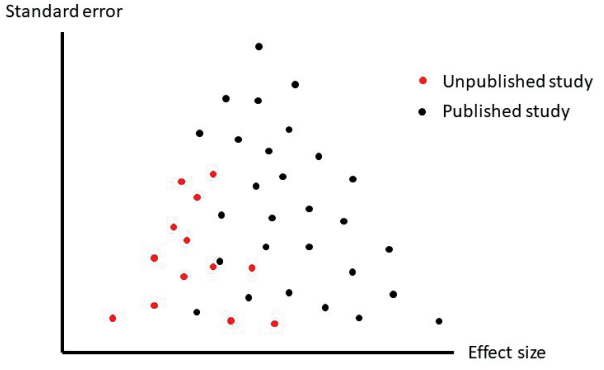


Figure 3.20. Illustration of publication bias. In this funnel plot, published studies have a greater effect size and smaller standard error. The presence of publication bias leads to significant asymmetry in the funnel plot.

If there is significant publication bias, there will be substantial asymmetry on the funnel plot. If the funnel plot appears grossly symmetrical, statistical tests, such as the Begg rank correlation test or Egger linear regression test, can be used to assess for publication bias. These methods should only be used when there are at least ten studies in the meta-analysis, as the power of the test is too low to distinguish chance from real asymmetry. These tests can also be used to adjust for publication bias.

3.11.3 Tests for heterogeneity

In the context of meta-analysis, heterogeneity implies differences between study estimates. Heterogeneity occurs due to differences between study protocol, study design and participant demographics or comorbidities.

Statistical methods can be used to test for heterogeneity in meta-analyses. The I^2 statistic is a common measure of heterogeneity. It provides an estimate of the proportion of the total variability between estimates that can be attributed to heterogeneity between studies. The I^2 statistic ranges between 0 and 100. The higher the I^2, the larger the degree of heterogeneity. Although there is no hard and fast rule, we generally consider heterogeneity to be present when the I^2 is greater than 50%. When the I^2 is very large, the validity of combining study results, and therefore the summary estimate, is called into question.

The presence of heterogeneity has multiple implications:

- Researchers use random effects methods rather than fixed effects methods (this is discussed in *Section 2.11.3)*
- Researchers can explore the treatment effects between groups with the aim of finding groups where homogeneity exists. For instance, a treatment may have a beneficial effect in one subgroup compared to another. By identifying this group, we can target the treatment to the right patient cohort.

WORKED EXAMPLE
Statin use and mortality in COVID-19 patients: updated systematic review and meta-analysis

Kollias *et al.* (2021) *Atherosclerosis,* 330: 114, doi.org/10.1016/j.atherosclerosis.2021.06.911

Background

Statins are lipid-lowering medications that are prescribed to patients with high cardiovascular risk. In previous studies, statins have demonstrated cardioprotective and immunomodulatory effects. In light of these effects, observational studies have been performed to investigate whether statins are associated with improved survival in patients with Covid-19. This study performed a systematic review and meta-analysis of observational studies investigating the relationship between statin use and Covid-19-related mortality.

Methods

In this review, 22 studies fulfilled the inclusion criteria and were included in the meta-analysis. Of these studies, 12 studies reported odds ratios and 10 studies reported hazard ratios. We will focus on the studies reporting odds ratio.

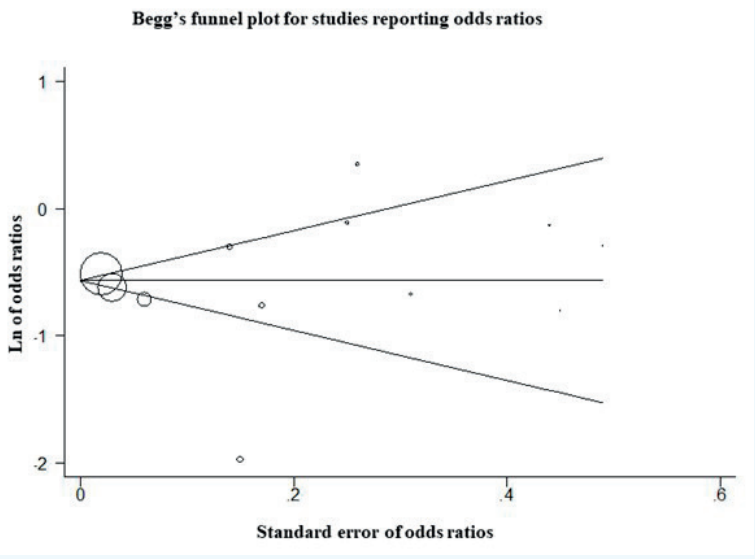


Figure 3.21. A worked example of a funnel plot. Reproduced from *Atherosclerosis*, 330: 114 with permission from Elsevier.

The authors used funnel plots to assess for publication bias. In this example, the ln(odds ratio) is plotted on the vertical axis and the standard error of the odds ratio is plotted on the horizontal axis. Each study is represented by an individual circle, with larger circles representing larger studies. In this example, the funnel appears roughly symmetrical; however, its interpretation is limited by a small number of included studies.

Results

The results from the meta-analysis of studies reporting odds ratios are illustrated in *Figure 3.22*. The 12 included studies are listed vertically. As you can see, there are two large studies (Rosenthal and Mallow) that contribute 50% of the weighting to the pooled estimate. These studies are associated with a OR <1, and hence a protective effect of statins on Covid-19-related mortality. The remainder of the studies are smaller and contribute a smaller weighting to the pooled estimate. These studies are represented by small boxes and wide confidence intervals.

The pooled OR estimate is 0.65 with a 95% confidence interval ranging from 0.55 to 0.78. Therefore, we are 95% certain that statin users have a OR of 0.55 to 0.78 for Covid-19-related mortality compared to non-statin users. This is a significant result with a *P* value <0.01 indicating that statin users are significantly less likely to die from Covid-19 than non-users.

In this meta-analysis, heterogeneity is present ($I^2 = 61\%$). This could be explained by between-study differences in statin dose, statin class, patient characteristics or Covid-19 severity. Further work is needed to investigate the cause of heterogeneity and whether certain patients are more likely to benefit from statin therapy.

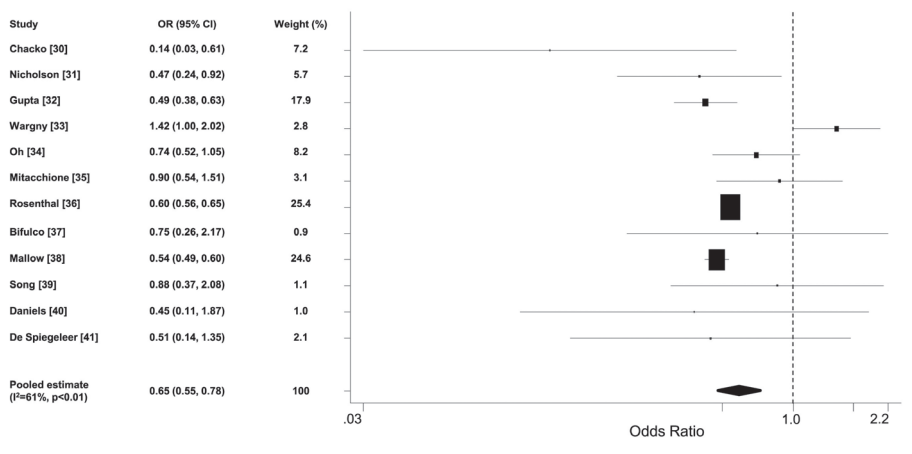


Figure 3.22. A worked example of a forest plot. Reproduced from *Atherosclerosis*, 330: 114 with permission from Elsevier.

Conclusion

In conclusion, this meta-analysis suggests that statins are associated with a lower mortality rate in patients with Covid-19. Limitations to this study include its observational and retrospective nature. Furthermore, there was significant heterogeneity between studies, likely due to differences in study protocol, treatment regimen and patient characteristics.

3.12 Diagnostic tests

Together with history taking and clinical examination, investigations are vital in the diagnosis of many clinical conditions. No diagnostic test is 100% accurate in the detection of disease, and this section will discuss the measures of test validity.

3.12.1 Sensitivity and specificity

When evaluating the diagnostic utility of a test, we commonly consider the sensitivity and the specificity of the test.

In order to understand the definition of sensitivity and specificity, consider the following 2×2 table outlining the frequencies of test results in those with and without the disease of interest.

Test result	Disease	No disease
Positive	a	b
Negative	c	d

Specificity = the proportion of individuals without the disease who test negative using the test
= d / b + d

Sensitivity = the proportion of individuals with the disease who test positive using the test
= a / a + c

In an ideal world, all tests would be 100% sensitive and 100% specific for the diagnosis of disease. In clinical practice, tests often gain sensitivity at the expense of specificity and vice versa. Whether we aim for high sensitivity or high specificity depends on the disease in question. For example, tests with a high sensitivity are preferred when the disease of interest is easily treatable. In this scenario, we want to detect all cases of disease in order to provide treatment.

3.12.2 Predictive value

Whilst the sensitivity and specificity of the test characterise the diagnostic ability of the test, the predictive values indicate how likely it is that the individual has the disease in light of their test result.

Positive predictive value (PPV) = proportion of individuals with a positive test who have the disease
= a / (a + b)

Negative predictive value (NPV) = proportion of individuals with a negative test who do not have the disease
= c / (c + d)

The predictive values depend on the prevalence of the disease in the population of interest. In samples where the disease is common, the PPV is higher than in samples where the disease is rare.

3.12.3 Likelihood ratios and pre- and post-test odds

The **likelihood ratio (LR)** gives another measure of the performance of the test. It is calculated using the following formula:

LR = sensitivity / (1–specificity)

Therefore, when the LR is greater than 1, the test is more likely to give a positive result if the patient has the disease than if they did not. The greater the LR, the greater the discriminatory power of the test.

The pre- and post-test odds are the odds of the patient having the disease before and after the test is performed. Before the test is performed, the odds of the patient having the disease are the same as the general population. This is calculated as:

Pre-test odds = prevalence (1–prevalence)

Following a positive test result, the odds of disease (post-test odds) depend on both the pre-test odds and the LR of the test. The post-test odds are calculated as:

Post-test odds = pre-test odds × LR

3.12.4 Receiver operating characteristic (ROC) curves

A ROC curve is used to determine a cut-off value for a diagnostic test, i.e. the value at which we state that the test is positive. The ROC curve is a plot of sensitivity versus (1 – specificity) across different cut-off values. A ROC curve is usually plotted with a line at an angle of 45°; this represents a test that performs no better than chance. The better the discriminatory capacity of the test, the closer the curve lies to the upper left-hand corner. The area under the curve (AUC) summarises the location of the curve, giving a combined measure of the sensitivity and specificity and hence the validity of the test. The higher the AUC, the higher the validity of the test. Mathematically, the AUC represents that a randomly chosen diseased individual is rated as more likely to have the disease by the test. The maximum AUC is 1, meaning that the test perfectly discriminates between diseased and non-diseased individuals. An AUC of 0.5 indicates that the test performs equally to chance. The AUC provides a useful measure of the test performance and allows us to compare the validity of multiple diagnostic tests.

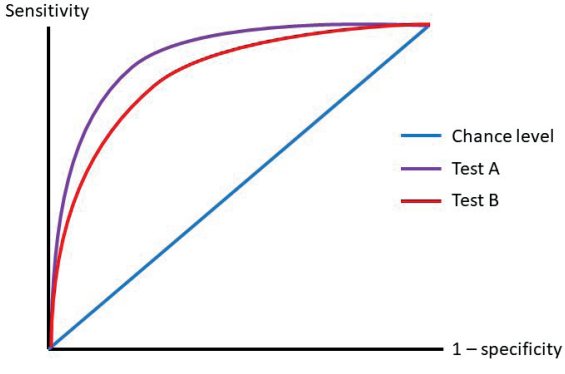


Figure 3.23. Two ROC curves in comparison to chance level. The ROC curve for test A lies closer to the upper left corner of the graph and has a higher AUC and hence validity, as compared to test B.

WORKED EXAMPLE
Diagnostic value of anti-cyclic citrullinated peptide antibody in patients with rheumatoid arthritis

Zeng *et al*. (2003) *J Rheumatol*, 30: 1451, PMID: 12858440.

Background

The detection of anti-cyclic citrullinated peptide (CCP) antibody is a widely used diagnostic test in the investigation of rheumatoid arthritis (RA). This study aimed to describe the validity of a novel ELISA for the measurement of anti-CCP in the diagnosis of RA.

Methods

In this study, the authors described a modified ELISA for the detection of anti-CCP antibodies. In order to describe the sensitivity and specificity of the ELISA, the investigators tested the serum of 191 patients with RA and 230 control subjects (including healthy controls and patients with non-RA rheumatological diagnoses).

Results

Table 3.8: A 2×2 table outlining the frequencies of CCP positivity in those with and without RA

	RA	Control patient
CCP positive	90	6
CCP negative	101	230

Table 3.8 summarises the results from the study. From this data, we can calculate the sensitivity and specificity of the anti-CCP ELISA:

Specificity = the proportion of individuals without the disease who test negative using the test
= 230 / (230 + 6) = 97.4%

As the specificity of the test is high, false positive results are rare (2.6% of all positive results). Therefore, we can conclude that the test is highly specific for RA and a positive result means that the patient is highly likely to have the disease.

Sensitivity = the proportion of individuals with the disease who test positive using the test
= 90 / (90 + 101) = 47.1%

The sensitivity of the test is only 47%. Therefore, only 47% of people with RA test positive for anti-CCP using this ELISA. This means that the test has a high false negative rate and cannot be used to rule out the diagnosis of RA.

The authors calculated the PPV and NPV of the anti-CCP test in the study population. It is important to remember that the PPV and NPV are population-specific. Therefore, these results cannot be extrapolated to the wider population when the prevalence of RA is much lower than the study cohort.

PPV = proportion of individuals with a positive test who have the disease
= 90 / (90 + 6)
= 94%

Therefore, in this population, 94% of patients with a positive test had RA.

NPV = proportion of individuals with a negative test who do not have the disease
= 230 / (230 + 101)
= 69%

Therefore, 69% of patients with a negative test did not have RA.

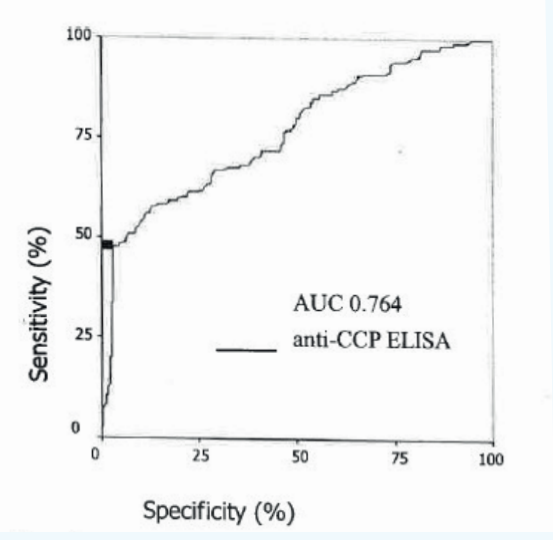


Figure 3.24. A worked example of a ROC curve. Reproduced from the *Journal of Rheumatology* with permission.

The authors plotted a ROC curve to determine the optimal value for the cut-off for a positive result. A cut-off value of 99 units was found to produce the optimal sensitivity and specificity. The AUC equals 0.764.

Conclusion

In this study, the presence of anti-CCP antibodies was highly specific but moderately sensitive for the diagnosis of RA. Therefore, testing for anti-CCP antibody provides a useful adjunct in the diagnosis of RA, but a negative test cannot be used to rule out the diagnosis.

3.13 Chapter summary

Throughout this chapter, we have introduced some of the principles and methods that are commonly used in the statistical analysis of clinical research. Further detail can be found in dedicated statistics textbooks. Given the complexity of the subject, it is always important to include statisticians in the design, undertaking and analysis of research studies.

3.14 Further reading

Statistics at Square One: www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one

Understanding statistics 1: presenting data from clinical trials: https://learning.bmj.com/learning/module-intro/.html?moduleId=5003158

Understanding statistics 2: what is statistical uncertainty?: https://learning.bmj.com/learning/module-intro/.html?moduleId=5001080

Harris, M. and Taylor, G. (2020) *Medical Statistics Made Easy*, 4e. Scion Publishing Ltd.

Peacock, J.L. and Peacock P.J. (2020) *Oxford Handbook of Medical Statistics*, 2e. Oxford University Press.

Petrie, A. and Sabin, C. (2020) *Medical Statistics at a Glance*, 4e. Wiley-Blackwell.