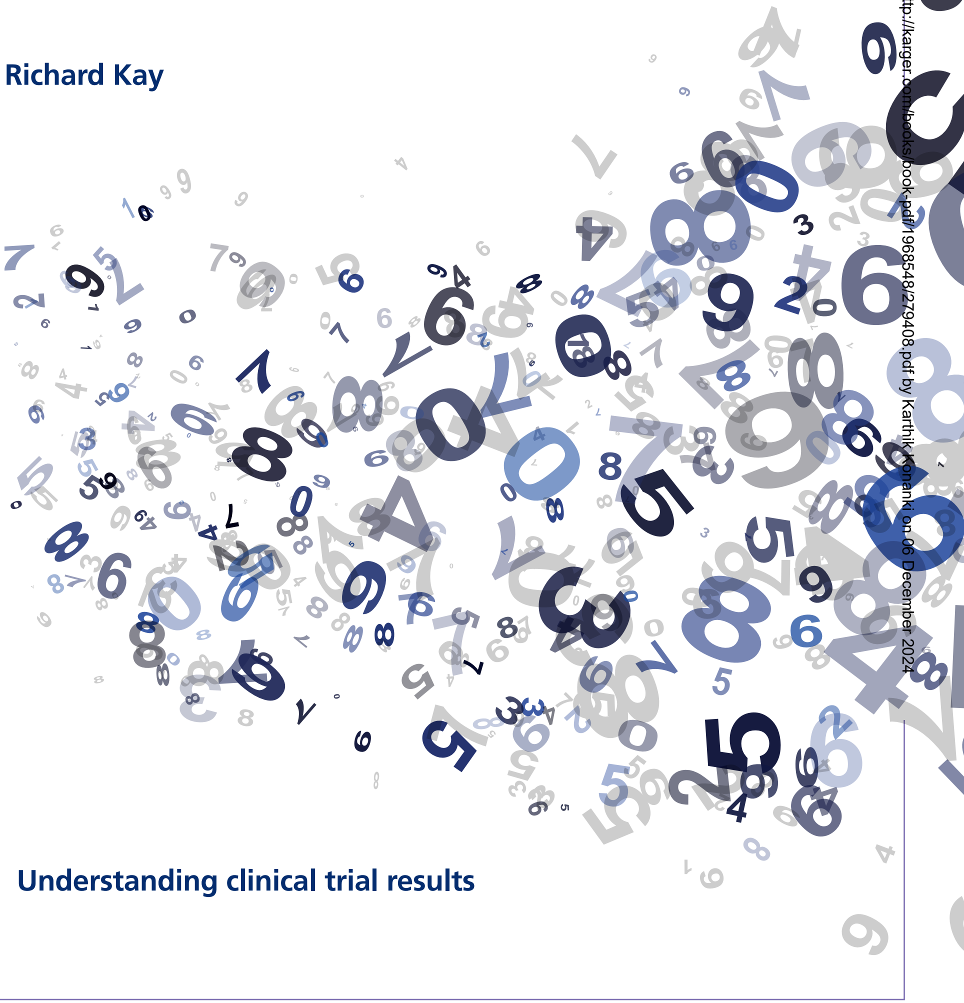
FAST FACTS

Medical Statistics

Richard Kay



Understanding clinical trial results

FAST FACTS

Medical Statistics

Richard Kay PhD
RK Statistics Ltd, Bakewell, UK
Visiting Professor, School of Pharmacy
and Pharmaceutical Medicine
Cardiff University
Wales, UK

Declaration of Independence
This book is as balanced and as practical as we can make it.
Ideas for improvement are always welcome: fastfacts@karger.com

KARGER

Fast Facts: Medical Statistics
First published 2020

S. Karger Publishers Limited, Elizabeth House, Queen Street,
Abingdon, Oxford OX14 3LN, UK
Tel: +44 (0)1235 523233

Book orders can be placed by telephone or email, or via the website.
Please telephone +41 61 306 1440 or email orders@karger.com
To order via the website, please go to karger.com

The publisher and the authors have made every effort to ensure the accuracy of this book but cannot accept responsibility for any errors or omissions.

For all drugs, please consult the product labeling approved in your country for prescribing information.

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

ISBN 978-1-912-77667-2

Kay R (Richard)
Fast Facts: Medical Statistics/
Richard Kay

Typesetting by Thomas Bohm, User Design, Illustration
and Typesetting, UK.
Printed in the UK with Xpedient Print.

List of abbreviations

ANCOVA: analysis of covariance

ANOVA: analysis of variance

AUC: area under the curve

CI: confidence interval

CMH: Cochran–Mantel–Haenszel

dBP: diastolic blood pressure

DFS: disease-free survival

ECOG: Eastern Cooperative Oncology Group

FDA: (US) Food and Drug Administration

FWER: family-wise error rate

H_0: null hypothesis

H_1: alternative hypothesis

HER2: human epidermal growth factor receptor 2

HR: hazard ratio

IQR: interquartile range

LLN: lower limit of normal

NMA: network meta-analysis

NNT: number needed to treat

NNTH: number needed to harm

NRI: non-response imputation

NS: not significant

OR: odds ratio

OS: overall survival

PFS: progression-free survival

RECIST: response evaluation criteria in solid tumors

RMST: restrictive mean survival time

RR: relative risk

RRR: relative risk reduction

SD: standard deviation

SE: standard error

Glossary

Censored measurements or observations: those that occur outside the study period (e.g. a patient who drops out of a study or, in survival analysis, a patient who is still alive after a fixed follow-up period)

Conditional power: the probability that the final outcome will be statistically significant, based on interim data

Confidence interval: a range of values within which the true population value lies

Endpoint: outcome value that is measured for each subject in a trial

Hazard rate: the conditional probability of the event occurring through time

Hazard ratio: the ratio of the hazard rate (see above) in the experimental group divided by the hazard rate in the control group. If there is no difference in risk between the two groups, HR = 1

Hypothesis: a statement about an unknown parameter. A **null hypothesis** assumes there is no difference between two groups, while the **alternative hypothesis** assumes there is. The null and alternative hypotheses are either accepted or rejected as a result of the actual trial data

Interim analysis: the analysis of data while the trial is ongoing and the data are still accumulating

Mean: the sum of values, divided by the number of values

Median: the middle value when all values are ranked from smallest to largest

Mode: the most common value

Multiplicity: the use of multiple significance tests (e.g. for several different endpoints or several different pairwise treatment comparisons), thus increasing the potential for a significant *p*-value to occur purely by chance

Non-parametric (test): a test that is not dependent on the distribution of the data

Normal distribution: symmetrically distributed data (a bell-shaped histogram of a special form)

Number needed to treat/harm: the number of patients that need to be treated for one patient to benefit/be harmed

Odds: the number of times an event happens divided by the number of times it does not happen in a group of patients

Odds ratio: the odds (see above) of an event happening in the experimental group divided by the odds of it happening in the control group

Overall survival: the length of time to death from the point of randomization

Population: a group of people who satisfy a defined set of inclusion/exclusion criteria, i.e. the complete set of subjects from which a sample is drawn

Predictive (factor): a baseline factor that influences the magnitude of difference the treatment makes to the endpoint

Prognostic (factor): a baseline factor that affects the endpoint

Progression-free survival: the length of time to disease recurrence or death, whichever occurs first

***p*-value:** the probability of observed differences having happened by chance

Relative risk: the risk of an event happening in the experimental group divided by the risk of it happening in the control group

Relative risk reduction: the degree to which an intervention reduces the risk of an event happening

Restricted mean survival time: the mean survival time up to a certain point of follow-up

Sample: a group of individuals who participate in a study, drawn from the broader population

Significance level: the probability level at which statistical significance is declared

Standard deviation: a measure of patient-to-patient variability

Standard error: a measure of mean-to-mean variation when samples are repeatedly taken from a population

Statistical inference: the analysis of data to draw conclusions about specific parameters associated with a population

Type I error: a false positive – the null hypothesis is true but the data give a significant *p*-value, suggesting a treatment difference when there is none

Type II error: a false negative – the alternative hypothesis is true but the data give a non-significant value, suggesting no treatment difference when one exists

Introduction

When comparing the efficacy or safety of two treatments, could the difference have occurred by chance or are the data strong enough for us to confidently conclude that the treatments are truly different? This resource is designed for all health professionals and pharmaceutical personnel who want, or need, to understand the medical statistics that help us answer this question.

Using real examples from oncology trials, but keeping it simple, *Fast Facts: Medical Statistics* explains the basic principles of hypothesis testing, calculating the power of a study, and the stopping rules for trials in terms of efficacy or futility. It is essential reading for anyone who wants a better understanding of the statistical terms and methods used both in the planning of study design and the analysis of clinical trial data.

If you find the thought of statistics daunting or have ever wanted to know what a type I error is, how an odds ratio is calculated and interpreted or what a Kaplan–Meier curve is really all about, read on!

1 Statistical inference

In clinical trials, an endpoint is an outcome value that is measured for each individual subject in the trial. The values of these endpoints are then summarized for a group of subjects (summary statistics): for example, a summary value for all subjects who received a particular treatment. This summary value can then be compared with the summary value of other groups in the trial. These results and the conclusions drawn from them are then extrapolated to the population as a whole.

Endpoint types

The type of statistical method used to analyze trial data depends on the type of endpoint that is measured. Common endpoint types are:

- continuous
- score
- count
- binary
- ordered categorical
- time to event.

Continuous endpoints are measured on a continuum of possible values over time: for example, a change in bone mineral density or blood pressure.

Score endpoints are endpoints that arise from scales constructed to capture a discrete value for something that cannot be measured in a continuous way. Examples include scales that provide a score for quality of life or severity of depression.

Count endpoints measure the number of items or events during a specified period: for example, the number of migraine headaches over 28 days.

Binary endpoints represent a dichotomy: a 'yes or no' type of measurement (e.g. success/failure, progression/no progression, survival/death). A common example in oncology is responder versus non-responder, where a responder is a patient whose best response is a complete or partial response, and where a non-responder is a patient whose best response is stable disease or disease progression.

Ordered categorical endpoints occur when an outcome is measured in categories and the categories have an implicit order. In oncology, the response evaluation criteria in solid tumors (RECIST) – complete response, partial response, stable disease or disease progression – are an example. However, when the endpoint is collapsed (e.g. responder versus non-responder), the ordered categorical endpoint is reduced to a binary endpoint.

Time to event. Many endpoints in oncology measure the time to event (see Chapter 2). Examples include overall survival, which measures time to death, and progression-free survival, which measures the time to death or progression, whichever occurs first. Time-to-event endpoints are usually measured from the point of randomization, but not always. Duration of response is an endpoint that is measured from the time of first response to progression.

Summary statistics

Summary statistics provide a quick and simple description of a set of data values. Usually, the sample's average (mean), middle (median) or most common (mode) value is used.

Example 1.1

Five women are diagnosed with breast cancer at ages 45, 50, 52, 54 and 54. To calculate the mean age of diagnosis:

$$\bar{x} = \frac{\sum x}{n}$$

where x is a value, $\bar{x}$ is the mean, $\sum x$ is the sum of all x values and n is the number of values.

$$\bar{x} = \frac{255}{5} = 51$$

The mean age of diagnosis in this sample is 51.

Mean. The mean is the arithmetic average, i.e. the sum of all values divided by the number of values. It is denoted by $\bar{x}$.

The mean is a good measure of comparison to use for sample groups with continuous and score endpoints, provided the endpoint data have a symmetric distribution (Figure 1.1a).

The mean is also a good measure of comparison for samples with count endpoints, but these are usually calculated as means per unit of time to account for different observation periods for different patients.

Median. When the data distribution is skewed (see Figure 1.1b,c), i.e. some of the values are a lot smaller or larger than the others, the mean is not usually the best measure of average. In these cases, the median is often the preferred measurement. The median is the middle value when the data values are placed in order from smallest to largest. It is sometimes denoted by $\tilde{x}$.

Mode. The mode is the most common value (see Figure 1.1). It is used to describe the most frequently occurring outcome, but in general it is of limited value in clinical trials.

Proportions. Comparing the means of binary endpoints makes no sense. Instead, proportions are compared; in this book the symbol *r* is used to denote a proportion.

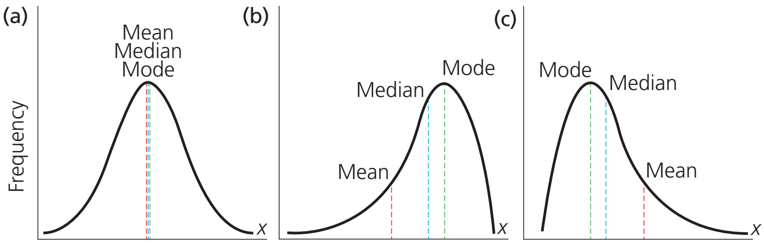


Figure 1.1 Position of the mean, median and mode, depending on the distribution of data. (a) The mean is a good measure to use when the spread of data is similar on each side of the mid-point (symmetric distribution). A common example of this is normal (or Gaussian) distribution. When the data are (b) negatively or (c) positively skewed, the median is the preferred measurement for average. The mode is rarely used.

Proportions for ordered categorical endpoints are also compared between treatment groups, but in such cases the order needs to be taken into account; the statistical procedures that are used take account of the order.

Kaplan–Meier curves are used to compare time-to-event endpoints (see Chapter 2).

Standard deviation (SD) is a measure of patient-to-patient variability. It is particularly important in the analysis of continuous and score endpoints. The SD is the average distance of all data values from the mean. It is not the simple average but a weighted average that gives rather more weight to the points well away from the mean.

SD is most frequently used for data with a symmetric distribution (Figure 1.2).

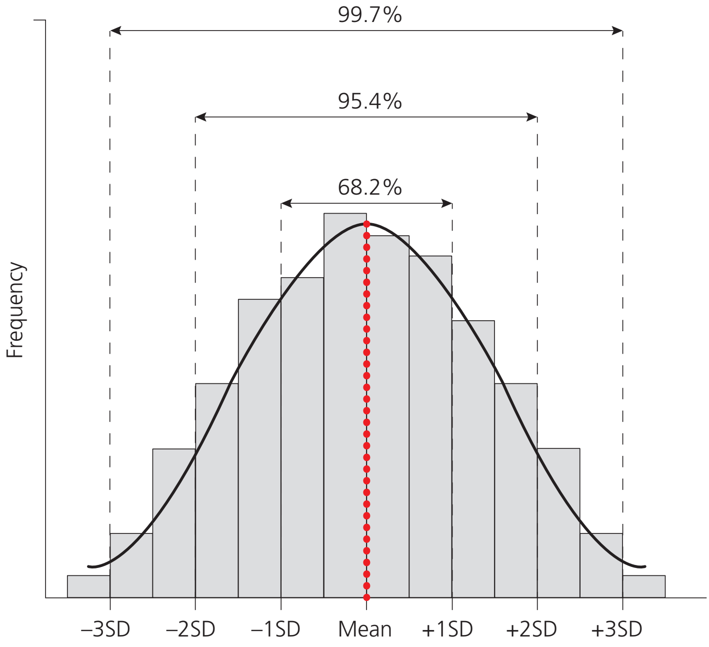


Figure 1.2 For data with a normal distribution, a range of one standard deviation (SD) above and below the mean (± 1SD) includes 68.2% of the values. ± 2SD includes 95.4% of the data. ± 3SD includes 99.7% of the data.

Sample and population

The sample of patients recruited into a clinical trial will be drawn from the population of interest. This population is defined by the inclusion and exclusion criteria, i.e. a set of characteristics, such as stage of disease or previous treatments, that are used to define a subject's eligibility (or ineligibility) for the study.

Inference. Statisticians talk about making inferences about the population based on data from a sample of patients drawn from that population.

Example 1.2

Of 100 patients chosen from the population, 50 are randomized to receive treatment A and 50 are randomized to receive treatment B.

The primary endpoint is the fall in diastolic blood pressure (dBP). The mean values for the fall in dBP are:

$\bar{x}_1$ = 8.6 mmHg for treatment A

$\bar{x}_2$ = 3.9 mmHg for treatment B

On average, treatment A has produced a better result than treatment B.

In general, there are three possible conclusions to be drawn from the results of this type of trial.

- Treatment A is better than treatment B.
- Treatment B is better than treatment A.
- There is no evidence for a difference between treatments.

In Example 1.2, the conclusion drawn is that treatment A is better than treatment B. This conclusion is not just a conclusion about what is happening in the sample, it is a broader conclusion for the complete population. By extrapolating from the sample to the population, it can be concluded that treating the population with treatment A will produce a better outcome, on average, than treating the population with treatment B. This is statistical inference – we draw conclusions about the population based on data from a sample that has been drawn from the population.

Different symbols for the statistics of interest are used when considering the complete population:

- mean = μ
- SD = σ
- proportion = θ

These parameters represent the true values in the population as a whole.

Sampling variation. Patients respond differently to treatments. In Example 1.2, the treatment difference was 4.7 mmHg. If the trial was repeated with the same protocol and same investigators but with a new set of patients, there would be a different numerical value for the difference between the means because of patient-to-patient variability in the outcome. This is referred to as sampling variation and it is ever present in clinical trials.

The observed treatment difference in a study is only ever an estimate of the true difference: in no sense is the result the true difference that would be obtained if the complete population was studied.

Standard error (SE) is a measure of the extent of the sampling variation. If the trial above was conducted 100 times, 100 values for the difference in the treatment means would be obtained. The SD associated with those 100 values is known as the SE. It is a measure of the inherent variability in the statistic being calculated.

If the 100 observed treatment differences are highly variable, the SE will be large, indicating that the outcome is subject to a lot of variability. If the observed treatment differences are similar over 100 trials, the SE will be small, indicating little variability in the outcome. In practice, of course, the trial cannot be repeated multiple times. Instead, formulas are used to calculate the SE based on the data from a single trial.

The SE is an important measure, as it is used to calculate confidence intervals and *p*-values.

Confidence intervals. As discussed above, the precise value for treatment difference can never be obtained, only an estimate of its value. However, an interval (a range of values) can be constructed around that estimate within which the true population value is likely to lie. The 95% confidence interval (CI) contains the true population value 95% of the time.

The 95% CI for a treatment difference is given by:

Observed difference – 2 × SE *to* Observed difference + 2 × SE

To construct a 99% CI, to be 99% confident that the true value has been captured, 2.6 is used as the multiplier in place of 2 in this formula. These formulas for the CIs are a good approximation in all settings except those with small sample sizes.

Example 1.3

The difference between the two mean falls in dBP between treatment A and treatment B is 4.7 mmHg with a 95% CI of 0.5 mmHg to 8.9 mmHg. This is written as:

4.7 mmHg (95% CI 0.5–8.9 mmHg).

Interpretation: We can be 95% confident that the true difference between the treatment means (for the population as a whole) is between 0.5 mmHg and 8.9 mmHg.

Hypothesis testing

In statistics, hypotheses are formulated to gain answers to questions of interest. To test for superiority, i.e. to demonstrate that treatments are different, the null hypothesis (H_0) is that the treatment effects (e.g. the mean effects) are equal. The alternative hypothesis (H_1) is that they are not equal.

Null hypothesis (H_0): $\mu_1 = \mu_2$
Alternative hypothesis (H_1): $\mu_1 \neq \mu_2$
where μ_1 is the mean value for treatment A and μ_2 is the mean value for treatment B.

The *p*-value is a number between 0 and 1 that is a measure of how much evidence there is to accept or reject the null hypothesis (Table 1.1).

TABLE 1.1

A quick guide to *p*-values

p-value	Definition	Level of statistical significance
> 0.05	No evidence against the null hypothesis. The probability of the treatment difference having happened by chance is higher than 0.05 (1 in 20, or 5%)	Not significant
≤ 0.05	Evidence against the null hypothesis. The probability of the treatment difference having happened by chance is 0.05 (1 in 20 or 5%) or lower	Significant
≤ 0.01	Strong evidence against the null hypothesis. The probability of the treatment difference having happened by chance is 0.01 (1 in 100 or 1%) or lower	Highly significant
≤ 0.001	Extremely strong evidence against the null hypothesis. The probability of the treatment difference having happened by chance is 0.001 (1 in 1000 or 0.1%) or lower	Very highly significant

Significance level. The conventional cut-off at which the *p*-value is deemed to be statistically significant is 0.05 or 5%. This is an entirely arbitrary cut-off that operationally seems to be a good compromise in terms of how strong the evidence needs to be before deciding in favor of differences. The cut-off at 5% is termed the significance level. If the *p*-value is above 0.05 then the difference is deemed to be non-significant and the null hypothesis is accepted.

Two-sided and one-sided *p*-values. The *p*-values discussed above are two-sided *p*-values as they detect differences in either direction, i.e. treatment A is better than treatment B or treatment B is better than treatment A. This is the way the hypotheses are set up; they ask the question, are the treatment means the same or are they different? In some cases, however, researchers may only be interested in detecting differences in favor of the experimental treatment.

In oncology, for example, a new drug A may be added to an existing chemotherapeutic combination B+C to see if it increases the proportion of responders. In this scenario, the hypotheses may be set up as follows.

Null hypothesis (H_0): $\theta_1 \leq \theta_2$
Alternative hypothesis (H_1): $\theta_1 > \theta_2$
where θ_1 and θ_2 are the true response proportions for treatment A+B+C and treatment B+C, respectively.

The null hypothesis states that A+B+C is not better than B+C, while the alternative hypothesis says that A+B+C is better than B+C. When the hypothesis is constructed in this one-sided way, a one-sided *p*-value is calculated that only achieves statistical significance if the treatment effect is in favor of the experimental treatment (A+B+C compared to B+C).

Example 1.4

The mean values for the fall in dBP are:

$\bar{x}_1$ = 8.6 mmHg for treatment A

$\bar{x}_2$= 3.9 mmHg for treatment B

The difference between the treatment means is 4.7 mmHg (p = 0.024). This two-sided *p*-value is statistically significant (see Table 1.1) – the null hypothesis is rejected and the alternative hypothesis is accepted.

There is a 2.4% probability of observing a treatment difference as large as 4.7 mmHg by chance, i.e. seeing a difference as large as 4.7 mmHg is very unlikely if the true treatment means were truly identical.

The one-sided *p*-value in favor of treatment A is 0.012, exactly one half of the previously calculated two-sided *p*-value.

There is a 1.2% probability of seeing a difference as large as 4.7 mmHg in favor of the experimental treatment by chance – the one-sided null hypothesis is rejected.

In practice, it makes no difference whether you calculate a one- or two-sided *p*-value, as the conventional cut-off for statistical significance with one-sided *p*-values is 0.025 or 2.5%, exactly one-half of the cut-off for two-sided *p*-values and the *p*-value itself is also divided by 2.

Odds ratio and relative risk

As discussed above, treatment effects can be expressed as differences between two means or differences between two proportions. Treatment differences can also be expressed in terms of ratios; for example, differences for time-to-event endpoints are expressed in terms of hazard ratios (see Chapter 2).

Odds ratios (ORs) and relative risks (RRs) are used to express treatment differences when dealing with binary endpoints. The OR concept also extends to ordered categorical endpoints. An OR shows the odds of the binary outcome occurring after exposure to the experimental treatment compared with the corresponding odds of the binary outcome occurring on the control treatment.

The RR is a similar measure that compares the risks of the binary outcome, rather than the odds.

Example 1.5[1]

Of 800 patients with chronic lymphocytic leukemia, 396 were randomized to treatment with fludarabine and cyclophosphamide (FC) and 404 were randomized to treatment with FC plus rituximab (FCR). The binary endpoint is the occurrence (or not) of grade 3/4 neutropenia.

	Neutropenia	No neutropenia	Total
FCR	136	268	404
FC	83	313	396
Total	219	581	800

The data show that 136 of 404 patients who received FCR had neutropenia, compared with 83 of 396 patients who received FC. The two event proportions can be calculated as:

$$r_1 = \frac{136}{404} = 0.34 \text{ and } r_2 = \frac{83}{396} = 0.21$$

(CONTINUED)

Example 1.5[1] (CONTINUED)

The observed difference between these two proportions (0.34 – 0.21) is 0.13 and the two-sided *p*-value is 0.00006, so the difference is highly significant. Alternative measures of the treatment difference are the odds ratio (OR) and the relative risk (RR).

Odds ratio. First, the odds for the event (the number of patients in whom the event is observed divided by the number of patients in whom the event does not happen) is calculated separately for each treatment group.

$$\text{For FCR: } \frac{136}{268} = 0.507$$

$$\text{For FC: } \frac{83}{313} = 0.265$$

The OR is the ratio of these two odds, with the convention that the odds for the experimental group (numerator) is divided by the odds for the control group (denominator).

$$\text{OR} = \frac{0.507}{0.265} = 1.91$$

Relative risk. First, the risk for the event in each group (the number who experience the event divided by the number in the group) is calculated separately.

$$\text{For FCR: } \frac{136}{404} = 0.34$$

$$\text{For FC: } \frac{83}{396} = 0.21$$

The RR is the ratio of these two risks, again with the convention that the risk for the experimental group (numerator) is divided by the risk for the control group (denominator).

$$\text{RR} = \frac{0.34}{0.21} = 1.62$$

Interpreting OR and RR values. An OR of 1 corresponds to equal treatment effects. If the OR is greater than 1, the odds for the event are higher in the experimental group, and vice versa. In Example 1.5, an OR of 1.91 shows that the odds for neutropenia occurring in the experimental group (FCR) are 91% higher than the corresponding odds in the control group (FC).

Likewise, an RR of 1 corresponds to equal treatment effects. If the RR is greater than 1, the risk is higher in the experimental group, and vice versa. In Example 1.5, an RR of 1.62 corresponds to a 62% increase in the risk of neutropenia in the FCR group.

Relative risk reduction and numbers needed to treat

Relative risk reduction (RRR) shows the degree to which the treatment has reduced the event. This can be calculated when the RR is less than 1. The RRR is 1 minus the RR. It is usually expressed as a percentage.

Example 1.6

Of 200 patients who had surgery for prostate cancer, 100 patients were randomized to further drug treatment (active group) and 100 patients were randomized to no further treatment (control group). 25% of the active group experienced disease recurrence compared with 45% of the control group.

The risk of disease recurrence in the active group was lower than that in the control group:

$$RR = 0.25/0.45 = 0.56$$

$$RRR = 1 - 0.56 = 44\%$$

Numbers needed to treat/harm (NNT/NNTH) is an additional measure that can sometimes be useful. NNT is the number of patients who must be treated for one to gain a benefit. When talking about a harm rather than a benefit, this measure is labeled as NNTH.

Example 1.7

For every 100 patients treated with FCR, 34 are ‘harmed’ (grade 3/4 neutropenia).

For every 100 patients treated with FC, 21 are ‘harmed’.

So, for every 100 patients, an additional 13 patients are harmed by FCR treatment compared with FC treatment. This corresponds to 1 in every 100/13 = 7.7.

The NNTH is 7.7. In practice, this is usually rounded up to the nearest whole number, in this case 8 patients.

Statistical testing

When statisticians talk about undertaking a statistical test, they are referring to the method they are using to obtain the *p*-value.

For comparison of two treatment means for continuous or score endpoints, the appropriate test is the two-sample *t*-test. For comparison of two proportions for a binary endpoint, it is the chi-square test. When comparing two hazard rates (by assessing whether the hazard ratio is equal to one), it is the logrank test (see Chapter 2).

Normal versus non-normal data distribution. The *t*-test is used when the data are normally distributed, or approximately so, i.e. a bell-shaped distribution that is symmetric around the mean (see Figures 1.1a, 1.2). If the data distribution is substantially non-symmetric, strictly, the *t*-test does not apply and using it can result in an incorrect *p*-value.

Alternatively, an attempt to recover normality may be made through a data transformation. Several transformations are possible, such as the log transformation in which the log of each data point is used in the analysis using the *t*-test.

When normality cannot be recovered through a data transformation, a non-parametric test can be used to obtain the *p*-value. Instead of comparing the values of the raw data, either on the original scale or on a transformed scale, the data values are ranked and the ranks are compared. The non-parametric equivalent of the two-sample *t*-test is the Mann–Whitney *U*-test. The Mann–Whitney *U*-test takes the ranks of the data points in the two groups combined and compares the mean rank in treatment group 1 with the mean rank in treatment group 2.

Non-parametric tests usually lack power (see Chapter 5) and so should only be used in preference to the *t*-test (either on the original scale or on the transformed scale) as a last resort.

Key points – statistical inference

- The type of statistical method used to analyze trial data depends on the type of endpoint being measured. Common endpoint types are continuous, score, binary, ordered categorical, count and time to event.
- Summary statistics provide a simple descriptive value for a data sample, which enables comparison of data sets. The mean (arithmetic average) is the most common measure used for continuous and score data with a normal or symmetric distribution. The median is the preferred measure for data with a skewed distribution. Proportions are compared for binary and ordered categorical endpoints.
- The standard deviation is a measure of patient-to-patient variability.
- The standard error is a measure of the extent of sampling variation.
- A 95% confidence interval is a range of values within which the true population value lies 95% of the time.
- The null hypothesis usually states that treatment effects are equal. The alternative hypothesis is that they are not equal. These hypotheses give two-sided *p*-values. One-sided *p*-values can be obtained when looking for differences only in one direction.
- The *p*-value is a number between 0 and 1 that indicates whether to accept the null hypothesis or the alternative hypothesis. For two-sided testing, a *p*-value ≤ 0.05 shows that there is statistically significant evidence against the null hypothesis, i.e. there is a statistically significant difference between the treatments.
- The odds ratio (OR) and relative risk (RR) are used to express the difference between binary endpoints. OR > 1 or RR > 1 indicates that the odds or risk of an event occurring, respectively, are higher in the treatment group.
- The two-sample *t*-test is used to analyze data with a normal or approximately normal distribution. Skewed data sets can be transformed (e.g. by log) before analysis with the *t*-test, or analyzed using a non-parametric Mann–Whitney *U*-test.

Reference

1. Hallek M, Fischer K, Fingerle-Rowson G et al. Addition of rituximab to fludarabine and cyclophosphamide in patients with chronic lymphocytic leukaemia: a randomised, open-label, phase 3 trial. *Lancet* 2010;376:1164–74.

2 Analysis of time-to-event endpoints

Time to-event endpoints

Many endpoints in oncology are time-to-event endpoints, i.e. the time until a specified event of interest occurs, such as death or the occurrence of disease. Typical examples are:

- overall survival (OS)
- progression-free survival (PFS)
- disease-free survival (DFS)
- duration of response.

Often these endpoints are measured from the time the participants in a study are randomized into treatment groups. This is the case for OS, PFS and DFS, but duration of response is measured from the time point at which partial or complete response is achieved to disease progression.

Time-to-event analyses include information from both censored and uncensored observations (see below).

Censoring is a common feature of all time-to-event endpoints. A subject's time to event is said to be censored if the event of interest (death, disease progression etc.) has not occurred in that patient by the end of the follow-up period. A censored observation can also occur when information on a subject is known for a limited duration only (e.g. if a person drops out of a study before the end and is lost to follow-up).

Example 1. In a study with a fixed 24-month follow-up, a subject who is still alive at month 24 will provide a censored value for OS. The OS for such a subject is considered to be at least as long as the duration of the study.

Example 2. At the time of an interim analysis, some subjects may not have been in the study for very long. These subjects will provide censored values if the event of interest has not occurred during the limited follow-up.

Kaplan–Meier curves

Kaplan–Meier curves provide a way of plotting the distribution of a time-to-event endpoint. In Figure 2.1 the dashed line shows that at 24 months' follow-up the estimated survival probability is 54%.

Kaplan–Meier survival curves are often used to compare the data between two groups of subjects. Figure 2.2 shows Kaplan–Meier curves for OS in a randomized study of patients with human epidermal growth factor receptor 2 (HER2)-positive metastatic breast cancer treated either with or without trastuzumab. The Kaplan–Meier curve steps down at time points at which deaths occur, while censored observations are denoted by notches on the curve. In this study, the follow-up period ranged from 3 months to 74 months.

The Kaplan–Meier curve plots the probability of being event free over time, with these probabilities being estimated from the data in the study. Note that the curve for patients who received trastuzumab is consistently above the curve for those who did not receive trastuzumab, indicating a higher survival probability in that group.

For the rest of this chapter, for simplicity, let us assume that the event of interest is death.

Median overall survival. The Kaplan–Meier curves can be used to obtain median survival times (Figure 2.3). The median survival time

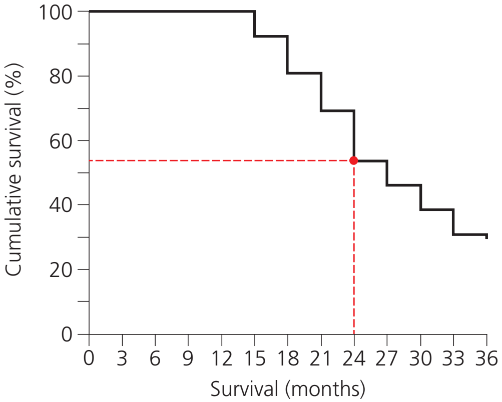


Figure 2.1 Kaplan–Meier survival plot of a cohort of patients with metastatic prostate cancer. At 24 months' follow-up, the estimated survival probability is 54%.

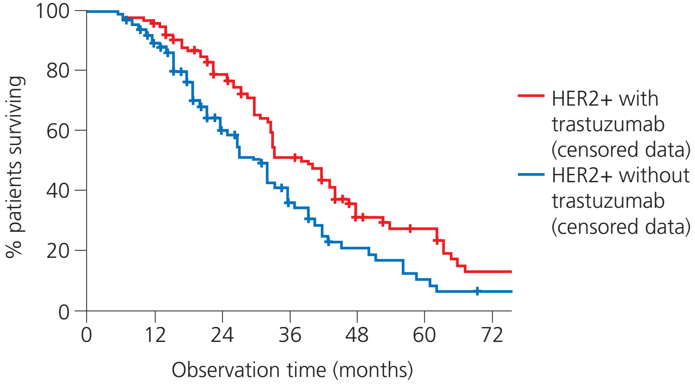


Figure 2.2 Overall survival among patients with HER2-positive metastatic breast cancer treated either with or without trastuzumab. The Kaplan–Meier curves step down at time points at which deaths occur, while censored observations are denoted by notches on the curve. The positions of the curves show a higher survival probability in the treated group. Adapted from Lv et al. 2018.[1]

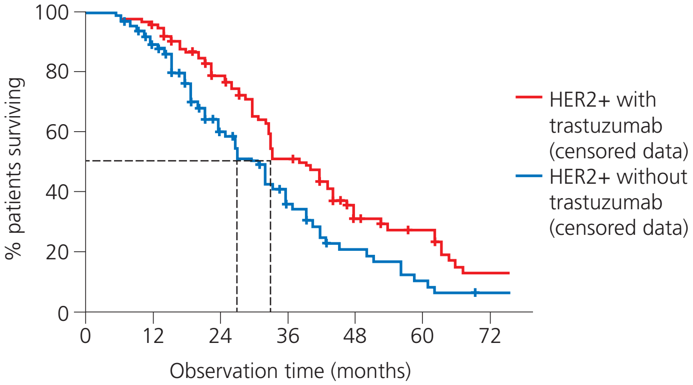


Figure 2.3 The median survival time is the time point at which 50% of patients are estimated to be alive. In this study, the median survival for the patients who received trastuzumab is 33 months compared with 26 months for those who did not, an average survival advantage of 7 months. Adapted from Lv et al. 2018.[1]

is the time point at which 50% of patients are estimated to be alive, and this can be found by drawing a horizontal line at the 50% line on the *y*-axis.

Hazard ratio

Treatment differences are usually expressed as a hazard ratio (HR). In order to understand what an HR is, it is necessary to understand what a hazard rate is.

Hazard rate is the probability that a subject will die within a given time interval among the subjects alive in that treatment group at the start of that interval. In a group of 1000 patients, suppose 10 die in month 1, 15 die in month 2 and 12 die in month 3.

- The hazard rate for month 1 = 10/1000.
- The hazard rate for month 2 = 15/990.
- The hazard rate for month 3 = 12/975 and so on.

In practice, hazard rates can be calculated for smaller time periods and the hazard rate can be considered as a continuous measure.

Figure 2.4 shows two hypothetical hazard rate curves that correspond to active (treatment) and control (placebo) groups. Although hazard rates are never constant over time, in many cases (at least approximately) the ratio of the hazard rates, i.e. the hazard rate for an event in the active group divided by the hazard rate for an event in the control group, will be approximately constant. When this is the case it is assigned a single value, the HR.

What do hazard ratios tell us? By convention, the hazard rate for the active group is divided by the hazard rate for the control group.

- HR = 1: the hazard rate in the active group is identical to that in the control group.
- HR < 1: the hazard rate in the active group is, on average, lower than that in the control group
- HR > 1: the hazard rate in the active group is, on average, higher than that in the control group.

It should be noted that the HR is a relative measure of effect between the treatment arms and does not give any information on the performance of the active treatment in absolute terms.

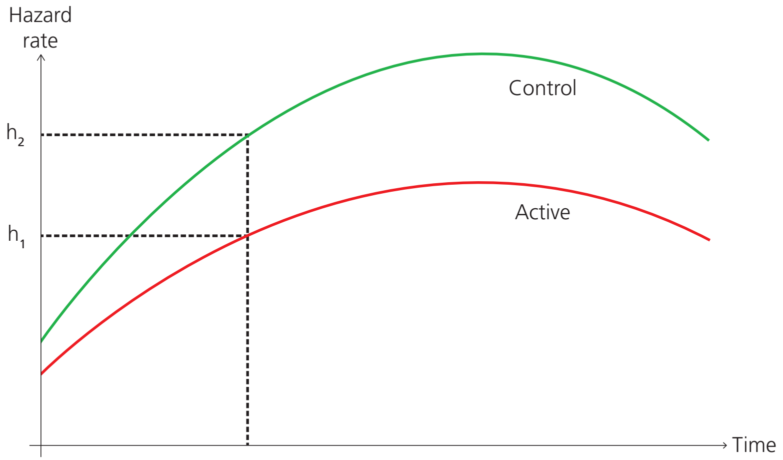


Figure 2.4 Hazard rate curves for active (treatment) and control (placebo) groups. In this case the ratio of the hazard rates, i.e. the hazard of an event in the active group (h_1) divided by the hazard of an event in the control group (h_2) remains constant.

The HR for OS in the previous breast cancer example is 0.502 with a 95% confidence interval (CI) of 0.390 to 0.755. This means that the hazard rate for death for the patients who received trastuzumab was reduced by 49.8%, on average, over time compared with those who did not receive trastuzumab.

HRs are usually given in conjunction with a 95% CI, i.e. the range of values that is likely to include the true population value (see Chapter 1). This range can be used to measure the precision of the HR; the narrower the CI the more precise the HR estimate.

Logrank test

While it is possible to compare the survival of patients in two different study groups at any given point in time on the Kaplan–Meier curves, this does not provide a comparison of the total survival experience of the two groups. The logrank test provides a *p*-value for the total survival experience of the treatment group compared with the control group by comparing the two Kaplan–Meier curves. It takes the whole follow-up period into account.

A significant *p*-value (probabilities lower than 0.05 are usually considered significant, see Chapter 1) shows that there is sufficient evidence to declare treatment differences.

> The logrank test in fact compares the HR to 1.

The logrank test implicitly assumes that the HR is fairly constant. This will not always be the case and when the HR is not constant the logrank test does not provide an appropriate comparison of the Kaplan–Meier curves. Note that a constant HR manifests itself as Kaplan–Meier curves that start together at time zero but then separate out gradually with time (see Figure 2.2 for an example where this is the case).

Restricted mean survival time

The restricted mean survival time (RMST) is the mean survival time up to a certain point of follow-up, for example 10 years. If the RMST is 7 years, this means that the mean survival time in the 10-year period from the point of randomization is estimated to be 7 years. It is a straightforward statistic to calculate as it is the area under the curve (AUC) up to year 10 (Figure 2.5).

Life expectancy difference. If the HR does not provide a suitable measure of treatment effect, for example if the HR is not constant, then a treatment difference can be expressed as the difference between two RMST values (the active RMST minus the control RMST). This is calculated as the area between the two Kaplan–Meier curves (Figure 2.6).

Adjusting for baseline imbalances

The principle of randomizing participants to two (or sometimes more) groups in a study is that, in theory, the two treatment groups are balanced in terms of factors that may influence the outcome. Any differences in outcomes between the two groups are then attributed to the difference in treatments. However, there are numerous baseline variables that may have a bearing on the outcome (e.g. patient characteristics such as age, sex and ethnicity, or stage or grade of tumor). It is therefore usually of value to adjust the analysis to account for any such differences in baseline factors.

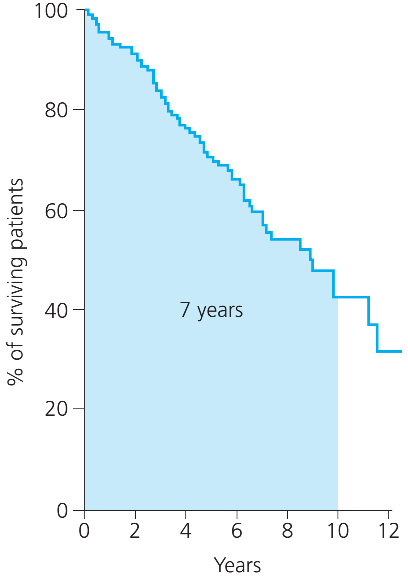


Figure 2.5 Restricted mean survival time is calculated as the area under the (Kaplan–Meier) curve (AUC) up to a specified point in time. Here, the average survival time during 10 years of follow-up is 7 years.

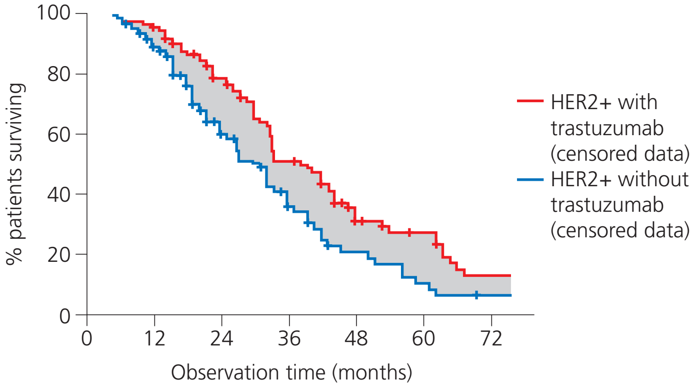


Figure 2.6 The area between two Kaplan–Meier curves represents the difference between the restricted mean survival time (RMST) in the active group and the RMST in the control group. This can be used to show the total difference between two groups when the hazard ratio is not constant.

There are two ways to make such adjustments for time-to-event endpoints. The stratified logrank test can adjust for a small number of factors, while the Cox proportional hazards model is a modeling approach that can, in theory, adjust for many more factors simultaneously.

Prognostic and predictive factors. It is becoming fairly standard practice in the oncology setting to use the term *prognostic* to refer to a baseline factor that affects the endpoint in some way. A prognostic factor can provide information on clinical outcome independent of therapy. In Table 2.1, the Eastern Cooperative Oncology Group (ECOG) performance status is prognostic for death – the greater the ECOG score, the poorer the prognosis.

The term *predictive* is used for a baseline factor that influences the magnitude of difference the treatment makes to the endpoint. In Table 2.2, ECOG performance status is both prognostic and predictive. As in Table 2.1, the higher the ECOG score, the poorer the prognosis. In addition, the lower the ECOG score, the greater the benefit of the treatment over placebo.

In the HER2-positive breast cancer study discussed previously, an analysis was carried out to see if the location of metastases had any influence on the outcome (Figure 2.7). In this example, visceral metastases is clearly a predictive factor for survival.

TABLE 2.1

ECOG status as a prognostic factor

	6-month death rate (%)			
	Control (placebo)	Active (treatment)	Average	Treatment difference
ECOG 0/1	15.2	10.0	12.6	5.2
ECOG 2/3	21.8	16.1	19.0	5.7

ECOG performance status is prognostic for death. The higher the ECOG score, the poorer the prognosis.

TABLE 2.2

ECOG status as a predictive factor

	6-month death rate (%)			
	Control (placebo)	Active (treatment)	Average	Treatment difference
ECOG 0/1	17.4	11.1	14.3	6.3
ECOG 2/3	21.2	20.0	20.6	1.2

ECOG performance status is both prognostic and predictive. The higher the ECOG score, the poorer the prognosis. In addition, the lower the ECOG score, the greater the benefit of the treatment over placebo.

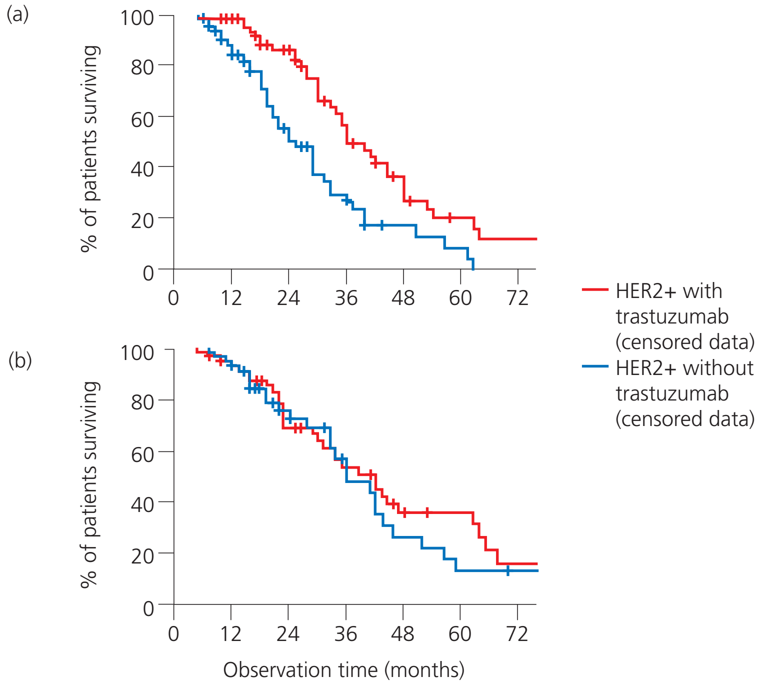


Figure 2.7 Survival analysis by location of metastases. (a) Of the patients with visceral metastases (lung, liver, ovary), those treated with trastuzumab had significantly better survival than those who did not receive trastuzumab ($p < 0.001$). (b) Of the patients with non-visceral metastases (bone, chest wall, locoregional lymph nodes), those who received trastuzumab had no real evidence of survival benefit compared with those who did not receive trastuzumab ($p = 0.128$).

Key points – analysis of time-to-event endpoints

- Typical examples of time-to-event analyses include overall survival, progression-free survival, disease-free survival and duration of response.
- Censoring is a common feature of all time-to-event endpoints. Observations are said to be censored when the event of interest has not happened by the end of the follow-up period.
- Kaplan–Meier curves plot the probability of being event free over time. The curves from different treatment groups can be plotted against each other to show the differences in outcome.
- Treatment differences are often expressed as a hazard ratio (HR). An HR of 1 means the risk in the two groups is the same. An HR > 1 means that the risk in one group is higher than the other, and vice versa.
- The logrank test compares the two Kaplan–Meier curves to provide a *p*-value for the difference in total survival experience of the two groups. This test can only be used if there is a constant HR, i.e. the Kaplan–Meier curves start together at time zero and then separate out gradually with time.
- If the HR is not constant, a treatment difference can be expressed as the difference between two restricted mean survival times (which is equal to the area between two Kaplan–Meier curves).
- The stratified logrank and the Cox proportional hazards model are two ways of adjusting for imbalances in baseline factors.
- Baseline variables that are important prognostic factors (e.g. age and stage of disease) should be taken into consideration in all statistical analyses, even in randomized studies.

Reference

1. Lv S, Wang Y, Sun T et al. Overall survival benefit from trastuzumab-based treatment in HER2-positive metastatic breast cancer: a retrospective analysis. *Oncol Res Treat* 2018;41:450–5.

3 Power and sample size

One of the most important issues at the trial design stage is to choose the appropriate number of subjects to be randomized so that treatment differences – should they exist – can be detected. This sample size calculation is based on the concept of power and the assumptions made about the magnitude of the treatment effect. To understand power, first consider the two types of error that can occur with the null and alternative hypotheses.

Type I and II errors

As discussed in Chapter 1, hypothesis testing involves the statement of a null hypothesis (e.g. that the frequencies of an event with two treatments are, in truth, equal) and the selection of a level of significance that indicates whether to reject or accept the null hypothesis (see page 16). In general, researchers are looking to reject the null hypothesis of equality in order to conclude in favor of treatment differences. The *p*-value framework, in which differences are significant when $p \leq 0.05$ and non-significant when $p > 0.05$, is unfortunately not perfect in terms of ascertaining the truth. Sometimes the data through the *p*-value can be misleading.

The two possible mistakes are the false positive and the false negative. The false positive occurs when the null hypothesis is true (i.e. there is no treatment difference) but the data give a significant *p*-value and a treatment difference is falsely declared. The false negative occurs when the alternative hypothesis is true (i.e. there is a treatment difference) but the data give a non-significant *p*-value and the difference between treatments is not detected.

- The false positive is more formally referred to as the type I error.
- The false negative is referred to as the type II error.

Consider the two-sample *t*-test for comparing two treatment means, μ_1 and μ_2. In truth, the means will either be equal or they will not. Data from the clinical trial comparing those means will provide either a significant difference or a non-significant difference.

		Truth	
		H_0: HR = 1	H_1: HR ≠ 1
Data	p = NS (cannot conclude differences)	✓	✗
	$p \leq 0.05$ (conclude in favor of differences)	✗	✓

Figure 3.1 Type I and type II errors. H_0, null hypothesis; H_1, alternative hypothesis; HR, hazard ratio; NS, non-significant.

Figure 3.1 sets out the various possible outcomes in terms of the truth and the *p*-value. Suppose the endpoint of interest is overall survival (OS) and that treatment differences are being assessed in terms of a hazard ratio (HR). Suppose further that the null hypothesis (H_0) is true and that HR = 1. If the data provide a non-significant *p*-value, the null hypothesis will not be rejected – the correct conclusion. However, from time to time a significant *p*-value ($p \leq 0.05$) will occur and the null hypothesis will be falsely rejected (lower left box). This is the false positive, the type I error (alpha [α] error).

In contrast, suppose that the alternative hypothesis (H_1) is true and that HR ≠ 1. If the data provide a significant *p*-value, the correct conclusion can be drawn, confirming a treatment difference. Again, though, this will not always occur and from time to time a non-significant *p*-value will indicate no treatment difference (top right box). This is the false negative, the type II error (beta [β] error).

Controlling type I and II errors

For a given sample size, it is not possible to eliminate these potential mistakes by, for example, modifying the level at which a significant difference is declared. By reducing the potential for one of these errors, the potential for the other error increases. In practice, there has to be a compromise and each of them needs to be controlled at an acceptably low level.

Controlling the type I error. The probability of the type I error (false positive) is controlled generally at 5%, as determined by the level at which differences are declared statistically significant. For example, if the treatment means are truly identical, then there is a 5%

probability of getting $p \leq 0.05$. This is a consequence of the *p*-value definition: 5% of the time when there is no treatment effect, there will be a numerical difference in the sample means that is extreme enough to give $p \leq 0.05$ by chance.

Controlling the type II error. The type II error is more difficult to control. The potential for that false negative is controlled through a related quantity called power.

$$\text{Power} = 100\% - \text{type II error}$$

As examples, if the probability of the type II error is 10% then the power is 90%; if the trial is powered at 80%, then the probability of a type II error is 20%. The type II error is missing a true difference while the power is capturing a true difference; if there is a 10% chance of missing the bus, there is a 90% chance of catching the bus!

Example 3.1

Consider an oncology trial with OS as its primary endpoint. Suppose the trial is powered at 80% to detect a difference characterized by HR = 0.75, i.e. if the true HR is 0.75 and the trial is run ten times, then on eight of those occasions, on average, a significant *p*-value ($p \leq 0.05$) will be produced, while on two of the ten occasions a non-significant *p*-value will be produced. This is the 20% type II error.

HR, hazard ratio; OS, overall survival.

Real world implications. When considering type I and type II errors, medicines regulators are generally most concerned with the type I error, the false positive; they do not want to be registering drugs that do not work. They are less concerned with the false negative, the type II error. The type II error is more the sponsor's problem – a sponsor does not want to fail to demonstrate a treatment difference when, in truth, there is a treatment benefit for the experimental treatment.

Regulators recommend that later-phase trials have at least 80% power. In a Phase III pivotal study, however, it must be remembered that 80% power means 20% type II error – a 20% potential to end up

Excerpt from the International Conference on Harmonisation (ICH) E9 Guideline (1998)[1]

Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations; …. The probability of type II error is conventionally set at 10% to 20%; it is in the sponsor's interest to keep this figure as low as feasible especially in the case of trials that are difficult or impossible to repeat.

with a negative study that fails to detect statistical significance even if a true treatment effect exists. Many researchers would say that 80% power in a Phase III trial is not enough and 90%, at least, is more appropriate at this pivotal trial stage.

Calculating sample size

One good thing at the trial design stage is that the power of a study can be calculated in advance of running the trial by speculating about what might happen. Consider a cholesterol-lowering placebo-controlled trial with 50 patients per group, in which the primary endpoint is the change from baseline to month 6 (baseline – month 6). Assume that the Phase II data have shown a standard deviation (subject-to-subject variability) associated with the primary endpoint of 1.1 mmol/L and a 5% level test is being used. Table 3.1 shows the calculated power for a range of values for the true treatment

TABLE 3.1

Power for varying levels of true treatment effect, *n* = 50 patients per group

True treatment difference (active – placebo) (mmol/L)	Power (%)
0.25	20.6
0.50	62.3
0.75	92.6
1.00	99.5

difference. Note that when comparing means, the subject-to-subject variation in the endpoint impacts the power. If the variability in the endpoint is large, it is more difficult to ‘see’ differences in the means and vice versa when the variability is small.

From these data, if the true treatment difference is 0.50 mmol/L, the trial has a power of 62.3% (see Table 3.1). This means that if the trial was run on 100 occasions and 0.50 mmol/L was the true treatment difference, then a statistically significant *p*-value ($p < 0.05$) would be achieved for 62 of those trials, while 38 of the trials would have a non-significant *p*-value and fail to detect differences; this is the type II error, equal to 37.7% in this case.

If a difference in the means of 0.50 mmol/L is viewed as being clinically important, a trial with only 62.3% power to detect this level of treatment benefit would be unsatisfactory; this level of power is simply too low. The only way to increase the power is to increase the sample size.

TABLE 3.2

Power for varying levels of true treatment effect, *n* = 100 patients per group

True treatment difference (active – placebo) (mmol/L)	Power (%)
0.25	36.2
0.50	89.5
0.75	99.8
1.00	100.0

Table 3.2 provides the values for power at the same specified levels of effect as in Table 3.1 after doubling the sample size to 100 patients per treatment group. All the values for power have increased. In particular, the power has increased from 62.3% to 89.5% at a treatment difference of 0.50 mmol/L. This is much more acceptable and is close to the generally acceptable level of 90% power for a Phase III pivotal trial. A sample size of 102 patients per treatment

group would give the trial 90% power, while a sample size of 76 per treatment group would give the trial 80% power.

This is the basis of sample size calculation. A level of effect of interest is specified and the sample size required to give, say, 90% power is then calculated. Formulas exist that can be used to calculate sample size for all the commonly occurring settings. To undertake the calculation, several parameters that impact sample size need to be specified.

- Significance level, which is the level at which statistical significance is to be declared (also known as the type I error or α error) – usually set at 5% unless there is a requirement to adjust the significance level to account for multiple testing (see Chapter 4).
- Power, which should – according to regulators – be at least 80% (see Excerpt on page 37), although for Phase III pivotal trials, 80% power is generally considered to be too low and 90%, at least, is preferred.

The remaining parameters that need to be specified depend on the endpoint type.

Binary endpoints, such as overall response, require:

- the response rate expected in the control group – usually based on data from the literature
- the expected difference in response rates with the specified power – this treatment difference could also be expressed in terms of a relative risk reduction or an odds ratio.

Continuous endpoints, such as change from baseline in a quality of life measure, require:

- the standard deviation of the endpoint, which usually comes from the literature, although there may also be some Phase II data based on the same endpoint that could be useful
- the targeted difference between the two means.

Time-to-event endpoints, such as OS, require:

- the median OS anticipated in the control group
- the treatment difference to be targeted, expressed as either an HR or an increase in the median OS.

Missing data. In each of these cases, allowance could also be made for the number of patients expected to be excluded from the analysis because of, for example, dropout, withdrawal of consent or major protocol deviations. If 5% of patients are expected to be non-evaluable for the primary analysis, then the required sample size would be increased by 5% to account for these non-evaluable patients without compromising the power.

It is important, when looking at the literature, to make sure that the planned primary analysis aligns with the analysis presented. Take, for example, an oncology study with overall response as the primary endpoint in which all patients with missing response data are categorized as 'non-responders', even if the reason for the missing data is, for example, dropout (non-response imputation [NRI]). The study from which the control group's response rate is being estimated should also have used NRI for any missing data. If it has not, and perhaps has simply omitted those patients from the analysis, then some adjustment will need to be made for the assumed control group response rate in the sample size calculation.

Event-driven trials

In many cases, the primary endpoint in an oncology trial is a time-to-event endpoint, such as OS or progression-free survival (PFS). The power of a study based on a time-to-event endpoint is determined not directly by the number of patients, but by the number of patients who experience the event being studied. For example, a trial in which 100 out of 1000 patients experienced the event has the same power as a trial in which 100 out of 300 patients had the event. Many oncology studies with a time-to-event primary endpoint are therefore event driven and the trial data are analyzed when the required number of events are observed in the two treatment groups combined.

Calculating the sample size for a trial based on a time-to-event endpoint has an intermediate step as a consequence. First, calculate how many patients with events are needed for the required power. Then, considering the design, calculate how many patients need to be recruited to 'deliver' those events. The key elements of the design in this case are the rate of recruitment over the recruitment period and the length of follow-up.

Example 3.2

MONALEESA-7 was a placebo-controlled randomized Phase III trial comparing ribociclib plus endocrine therapy with endocrine therapy alone in premenopausal women with hormone-receptor-positive advanced breast cancer.

The primary endpoint was investigator-assessed PFS, defined as the time from randomization to either the first documented disease progression per RECIST or death from any cause.

The primary efficacy analysis was the comparison of PFS between the ribociclib and control groups using a logrank test stratified by the criteria used for the stratified randomization. The trialists estimated that 329 events of disease progression or death would provide 95% power to detect an HR of 0.67 in a two-sided 5% level test, assuming a median PFS of 9 months in the control group.

The plan was to recruit 594 patients (33 patients per month over an 18-month period). The data cut-off of 20 August 2017 was specified by the sponsor as the projected date by which 329 events of disease progression or death would be reached.

HR, hazard ratio; PFS, progression-free survival; RECIST, response evaluation criteria in solid tumors.
Tripathy et al. 2018.[2]

Sample size re-evaluation

Blinded re-evaluation. As discussed, the power of a study when the primary endpoint is a time-to-event endpoint is driven by the number of patients who have the event. If the proportion of patients who experience the event is lower than expected, the study may need to be extended to allow more time for the events to occur or to increase the sample size to achieve the required number of events in an acceptable timeframe.

Based on the assumed median for the time-to-event endpoint in the control group and the HR (or, equivalently, both medians), it will be possible to calculate the proportion of patients with events, over various periods of time, that align with those assumptions; these proportions can be monitored as the trial progresses.

In MONALEESA-7, the assumed median PFS in the control group was 9 months and, with an HR of 0.67, the assumed median in the experimental group was set at 13.5 months. Note that, at least approximately, the HR is equal to the ratio of the medians. However, the proportion of PFS events occurring over time in the two groups combined was found to be lower than expected under these assumptions. Indeed, the observed medians at the end of the study were 13.0 months in the control group and 23.8 months in the experimental group.

The sponsor took the decision to increase the sample size (from 594), with 672 participants eventually being randomized. At the planned cut-off date for the final analysis, the number of PFS events observed was 318. Of course, it would have been possible to delay the cut-off date to achieve the required 329 events, but the sponsor decided against this course of action. The reduction in number of events led to a small reduction in the power of the study, but as the power had been set at a high level (95%) at the design stage it remained acceptable.

Unblinded re-evaluation and adaptive designs. Undertaking a sample size re-evaluation based on blinded data is acceptable from a statistical perspective and does not cause any inflation in the false-positive potential for the trial (type I error). The US Food and Drug Administration (FDA) makes a clear distinction between blinded data looks and unblinded looks.

Excerpt from the FDA's Draft Guidance for Industry. Adaptive Design Clinical Trials for Drugs and Biologics, 2010[3]

There is a critical distinction between adaptations based on an interim analysis of unblinded results of the controlled trial … and adaptations based on interim noncomparative analyses of blinded data … revisions based on blinded interim evaluations of data (e.g. aggregate event rates, variance, discontinuation rates, baseline characteristics) do not introduce statistical bias to the study or into subsequent study revisions made by the same personnel.

FDA, Food and Drug Administration.

It is possible to re-evaluate a sample size on the basis of unblinded data, perhaps having observed a smaller than expected treatment difference, but this would constitute an adaptive design and would affect the way in which the data could be analyzed in order to preserve the overall type I error for the comparison. Under these circumstances, it is important to plan for such an evaluation at the outset and consider carefully the implications for the statistical analysis once the trial eventually completes.

Key points – power and sample size

- Every trial has the potential to yield a false-positive or a false-negative result – the false positive is known as the type I error and the false negative the type II error.
- Potential for the false positive is controlled at 5%, the significance level, while the false negative is controlled via the power.
- Power is equal to 100% – type II error. The type II error is the false negative, missing a real difference, while the power is the true positive, capturing a real difference.
- Power can be calculated at the planning stage, in advance of running the trial, by making various assumptions about, in particular, the magnitude of the treatment effect.
- The sample size calculation is based on the requirement to detect a prespecified treatment difference, often referred to as the clinically relevant difference, with a certain level of power, usually 80%, or at least 90% in a Phase III pivotal trial.
- As many oncology studies are powered on a primary time-to-event endpoint – power being determined by the number of patients with events – many oncology studies are event driven, with the statistical analysis taking place once the required number of patients with events have been reported.
- It is possible to increase the sample size if the event rate is lower than anticipated – if this is done in a blinded way, there is no statistical penalty or implication for the statistical method to be used for analysis.

References

1. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH Topic E 9: Statistical Principles for Clinical Trials*. www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials, last accessed 26 March 2020.

2. Tripathy D, Im S-A, Colleoni M et al. Ribociclib plus endocrine therapy for premenopausal women with hormone-receptor-positive, advanced breast cancer (MONALEESA-7): a randomised phase 3 trial. *J Clin Oncol* 2018;36: 904–15.

3. US Food and Drug Administration. *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics – Draft Guidance* 2010.

Further reading

Kay R. *Statistical Thinking for Non-Statisticians in Drug Regulation*, 2nd edn. Chichester: Wiley, 2014.

4 Multiplicity

As we saw in the last chapter, every statistical test has the potential to give a false-positive result – a type I error. This can happen because of the way the *p*-value is defined: the *p*-value is the probability of seeing a difference of a certain order of magnitude purely by chance.

Often, several simultaneous statistical tests are conducted (multiple testing or multiplicity): for example, for several different endpoints or for several different time points, for the overall sample and in several subgroups. By doing this, the potential for a significant *p*-value to occur purely by chance increases.

The overall type I error rate, known as the family-wise error rate (FWER) therefore needs to be controlled in every trial at 5%.

Golden rule

The family-wise error rate (FWER) must be controlled at 5%.

Numerous methods and safeguards are in place to prevent the cherry-picking of significant *p*-values only, or to enable cherry-picking of significant results at a lower significance level. This chapter explains these method.

Note that, in general, these considerations apply to confirmatory studies and efficacy. In general, *p*-values are not calculated in routine safety studies. In safety studies, false negatives are more of a concern than false positives, i.e. that a safety signal is missed. However, if a specific safety issue is being addressed, with a clearly defined safety endpoint (e.g. the occurrence of grade 3/4 neutropenia), then that endpoint could be built into the confirmatory testing for the trial and the same considerations would apply as for efficacy.

If there was just one endpoint of interest, at one time point, with just two treatment groups then only one *p*-value would be calculated to evaluate the efficacy of the experimental treatment

TABLE 4.1

Examples of studies in which multiplicity may arise

- Evaluation of several endpoints
- Evaluation of more than two treatment groups
- Investigation of three dose levels of the experimental treatment and placebo
- Evaluation of several key subgroups
- Interim analysis in addition to the primary analysis at the end of the study

and multiplicity would not arise. This, however, is very rarely the case (Table 4.1).

The testing scenarios shown in Table 4.1 will lead to the calculation of many *p*-values. Every additional *p*-value increases the potential for one or more false positives.

Reducing type I errors

Suppose several different endpoints are being assessed; for example, objective response (partial response [PR] or complete response [CR]), clinical benefit (stable disease [SD], PR or CR), progression-free survival (PFS) and overall survival (OS). A general approach to control the type I error ($\alpha = 0.05$) would be to divide it across the endpoints that are being considered.

The Bonferroni correction divides the 0.05 equally across all endpoints. For the four endpoints mentioned above, the significance level for each endpoint would be $0.05/4 = 0.0125$. Paying a price on the alpha in this way and using a reduced significance level enables researchers to draw confirmatory conclusions for the endpoints that are significant. For example, if two of the four endpoints are significant ($p \leq 0.0125$) and two are non-significant ($p > 0.0125$) confirmatory conclusions can only be drawn about the two that are significant.

For m tests, adjusted significant level = α/m

Making this correction will affect the power of the study (see Chapter 3 for a discussion on power). Achieving a significance level of 0.0125 is much more difficult than achieving a significance level of 0.05 and a larger sample size will be required.

Non-equal division of the alpha. In certain situations, it may not make sense to divide the alpha equally. For example, in a study of patients with advanced prostate cancer, researchers may want the flexibility to draw confirmatory conclusions for both PFS and OS as co-primary endpoints. The Bonferroni correction would divide the 0.05 by 2 and assign a significance level of 0.025 to each endpoint. However, PFS is potentially the easier endpoint in terms of achieving statistical significance, possibly because there is likely to be a larger treatment effect on that endpoint or because, by definition, there will be more PFS events than OS events. In this case, it may make sense to assign a significance level of 0.01 for PFS and 0.04 for OS. Of course, researchers must decide how to divide the alpha at the trial design stage; it is not appropriate to decide on the alpha split once the data are available!

The Hochberg approach. The Bonferroni correction is not the only approach to splitting the alpha; the Hochberg method is also in fairly common use. Along with the Holm approach (not discussed here), the Hochberg approach involves obtaining the *p*-values and then working through them in a specific order.

Suppose the *p*-values for three different endpoints have been calculated and ordered from largest to smallest. If the largest *p*-value is ≤ 0.05, then all endpoints have *p*-values ≤ 0.05 and confirmatory conclusions can be drawn for all three endpoints (Case 1 in Example 4.1). If, however, the largest *p*-value is > 0.05, statistical significance cannot be declared for that endpoint. Instead, look to the endpoint with the second largest *p*-value. If that *p*-value is ≤ 0.025 (0.05/2) then that endpoint and the endpoint with the smallest *p*-value are statistically significant (Case 2 in Example 4.1). Finally, if the largest *p*-value is > 0.05 and the second largest *p*-value is > 0.025 (0.05/2), look to the endpoint with the smallest value. If that *p*-value is ≤ 0.017 (0.05/3) then that endpoint is statistically significant (Case 3 in Example 4.1).

Hochberg versus Bonferroni. Note that if the Bonferroni correction is applied to Example 4.1, with an adjusted significance level of 0.017

Example 4.1

Consider three endpoints, E_1, E_2 and E_3, with *p*-values ordered from largest to smallest.

	α	Case 1	Case 2	Case 3
E_1	≤ 0.05	0.042 ✓	0.170 ×	0.072 ×
E_2	≤ 0.025 (0.05/2)	0.029 ✓	0.024 ✓	0.033 ×
E_3	≤ 0.017 (0.05/3)	0.019 ✓	0.013 ✓	0.009 ✓

- In Case 1, the largest *p*-value is 0.042, so all p-values are ≤ 0.05 and confirmatory conclusions can be made for all endpoints.
- In Case 2, the largest *p*-value is > 0.05 but the second largest *p*-value is ≤ 0.025 and confirmatory conclusions can be drawn for endpoints E_1 and E_2.
- In Case 3, the largest *p*-value is > 0.05, the second largest is > 0.025, but the third largest is ≤ 0.017 and a confirmatory claim can be made for endpoint E_3.

for all three endpoints, the only statistically significant endpoint would be E_3 in cases 2 and 3. In this sense, Hochberg always performs better than Bonferroni; it is more sensitive in terms of being able to pick up significant differences.

So why not use Hochberg all the time? To use Hochberg, all *p*-values need to be available at the same time to put them in order, and that will not always be the case. For example, the primary analysis for PFS may be triggered once a certain number of PFS events have been observed across all the treatment groups, while the primary analysis of OS may be triggered at a later point in time when the requisite number of OS events have occurred.

Hierarchical testing

In the approaches to multiple testing discussed above, paying a price on the alpha gives researchers the flexibility to cherry-pick significant *p*-values at a lower significance level. Hierarchical testing, or closed testing as it is sometimes called, places an alternative constraint on the testing to control the FWER at 5%. In hierarchical testing, the endpoints of interest are ranked from most important to least important before the study begins (i.e. the order of importance is prespecified in the study protocol). The endpoints are then tested in that order using 0.05 as the significance level, but the statistical significance can only be claimed down to the first non-significant result.

Hierarchical testing rule

Confirmatory conclusions can only be drawn down to the first non-significant endpoint ($p > 0.05$).

Example 4.2

Three endpoints – PFS, OS and objective response – are placed in order of importance as shown, where objective response is the most important.

	Case 1	Case 2	Case 3
Objective response	0.044 ✓	0.017 ✓	0.072 ✗
PFS	0.036 ✓	0.082 ✗	0.033 ✗
OS	0.048 ✓	0.013 ✗	0.059 ✗

Working through the endpoints of each case in the prespecified order:

- In Case 1, each of the *p*-values is ≤ 0.05; all the endpoints are statistically significant.
- In Case 2, objective response is statistically significant ($p = 0.017$) but the *p*-value for PFS is non-significant ($p = 0.082$) at the 5% level, so PFS is not statistically significant and no confirmatory claim for OS can be considered, irrespective of the *p*-value for that endpoint.
- In Case 3, the *p*-value for objective response is non-significant ($p = 0.072$) so no confirmatory conclusion can be made for that endpoint. Furthermore, the formal testing for confirmatory claims must stop regardless of the *p*-values for PFS and OS.

In Example 4.2, OS in Case 2 ($p = 0.013$) and PFS in Case 3 ($p = 0.033$) both have *p*-values < 0.05 but these constitute exploratory findings only and are sometimes labeled as ‘nominally significant’.

Regulatory authorities would not allow these nominally significant *p*-values to be reported in the label. Estimated hazard ratios could be reported, but to avoid any misinterpretation of the statistics the *p*-values and confidence intervals would not be included.

The hierarchical order. Clearly, it is important to get the hierarchical order correct, both in terms of clinical relevance and other practicalities. In general, it is more difficult to achieve statistical significance for OS than PFS, largely because there will be fewer OS events than PFS events at the time of the analysis. For this reason, PFS is usually placed higher up the hierarchy than OS, so that a non-significant result for OS does not preclude looking at PFS for statistical significance (Example 4.3).

Example 4.3

The efficacy of idelalisib and ofatumumab versus ofatumumab was studied in previously treated patients with chronic lymphocytic leukemia.[1]

Prior to the study, the endpoint hierarchy was determined, as shown below.

	HR/OR	*p*-value
PFS	HR = 0.27	$p < 0.001$
Objective response	OR = 15.9	$p < 0.001$
Lymph node response	OR = 487	$p < 0.001$
OS	HR = 0.74	$p = 0.27$
PFS in the del(17p)/*TP53* subgroup	HR = 0.32	$p < 0.001$

The first three endpoints were all statistically significant ($p < 0.001$), but the *p*-value for overall survival (OS) was non-significant ($p = 0.27$). Although the *p*-value for progression-free survival (PFS) in the del(17p)/*TP53* subgroup, if taken in isolation, was significant ($p < 0.001$), given the hierarchy applied in this study, the result for this subgroup can only be considered as an exploratory finding. If a confirmatory claim was applied to this result, it would destroy the control of the type I error at 5%.

HR, hazard ratio; OR, odds ratio.

Hierarchical testing is a good way to think about primary and secondary endpoints when designing the study. Secondary endpoints are simply primary endpoints that are less important than the designated primary endpoint.

Combining approaches

It is also possible to mix and match the various approaches. For example, the primary endpoint may be placed at the top of several endpoints for hierarchical testing, with three key secondary endpoints placed together in position 2 to be tested with a Bonferroni or possibly a Hochberg correction.

Subgroup evaluation

It is not uncommon for a study to show an impressive treatment difference in a key subgroup but fail in terms of the primary endpoint. In this situation, it is generally not possible to refocus attention on the subgroup to make a confirmatory claim because of the problem of multiplicity. Treatment differences within subgroups will occur by chance and, especially when there are a lot of subgroups, they are almost inevitable. Remember that 1 in 40 *p*-values will be statistically significant in favor of the experimental treatment by chance. An 'interesting' result in a subgroup would therefore only constitute an exploratory finding.

If a subgroup is particularly important, perhaps because of a good clinical argument that supports a strong effect in that subgroup, then that subgroup should be built into the confirmatory testing strategy from the beginning, for example by putting it within the prespecified hierarchy.

Forest plots. If there is a treatment difference in favor of the experimental treatment, it is routine practice to produce a forest plot that displays the result overall and in subgroups (Figure 4.1).

The main reason for looking at subgroups in this way is to assess the consistency of an observed treatment effect. The Committee for Medicinal Products for Human Use (CHMP) has provided a draft guideline on subgroup evaluation in which it explains how to interpret forest plots.[2] If all of the point estimates (the hazard ratios shown in Figure 4.1) fall within the confidence interval (CI) for the

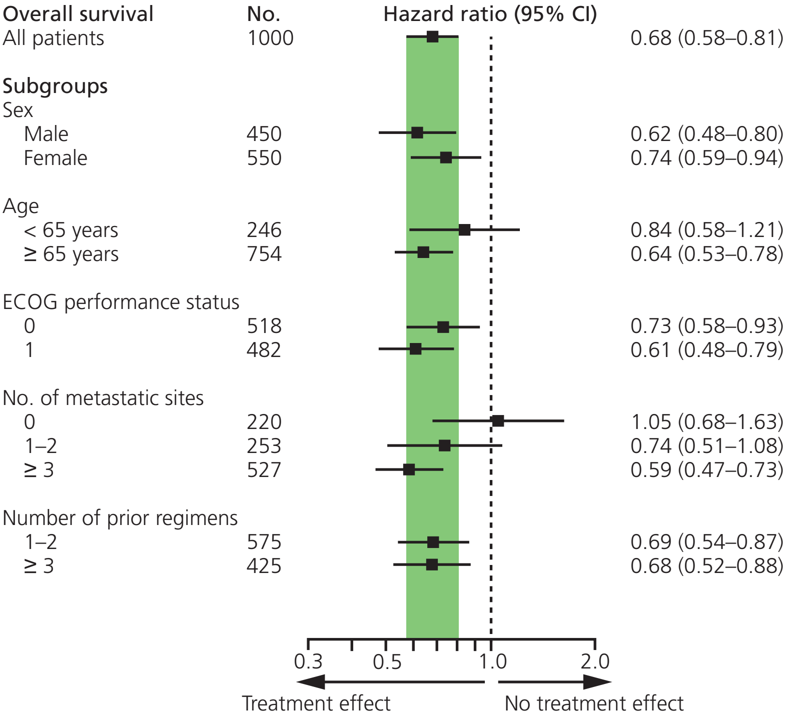


Figure 4.1: Example of a forest plot used to graphically display subgroup analyses in an oncology study. The plot shows the overall survival (OS) for all patients as well as for 11 subgroups. CI, confidence interval; ECOG, Eastern Cooperative Oncology Group.

overall treatment effect, then there is no evidence for heterogeneity. If a point estimate strays outside of the overall CI:

- homogeneity of treatment effect is still supported if the CI for that subgroup substantially overlaps the overall CI
- concerns regarding heterogeneity should be raised when the point estimate moves further away from the overall CI, particularly when the subgroup CI is separate from the overall CI.

To make these judgments, it is useful to mark a band from the overall CI down through the plot (see Figure 4.1).

It is important to note that the CIs for small subgroups will be wide, and even though the point estimate may be well outside the overall CI there may still be considerable overlap of the CIs; for small subgroups there is just not enough information to be able to draw a conclusion of heterogeneity.

In Figure 4.1, the hazard ratio for the subgroup with no metastases is higher than 1 and the CI is starting to separate from the overall confidence interval. However, this is the smallest subgroup and it may also be that there are relatively few deaths in that subgroup, making a firm conclusion regarding heterogeneity problematic.

Finally, it is not appropriate to discount subgroups in terms of a treatment benefit when the CI crosses the 'no effect' line. In Figure 4.1 the treatment effect in the subgroup of patients under 65 years in isolation is non-significant. This does not mean that the treatment does not work in that subgroup. The trial is powered to detect a treatment effect overall, the trial is not powered to detect treatment effects in much smaller subgroups. This is often a fundamental misunderstanding and a mistake that you should not make.

Key points – multiplicity

- As several simultaneous statistical tests are usually conducted (e.g. for several different endpoints or time points, for the overall sample and several subgroups), the potential for a significant *p*-value to occur purely by chance increases.
- The overall type I error rate, known as the family-wise error rate must be controlled in every trial at 5%.
- The Bonferroni correction divides the 0.05 equally across all endpoints; for example, for four endpoints, the significance level for each endpoint would be 0.05/4 = 0.0125. The decision to do this must be made in the study protocol.
- The Hochberg approach places the *p*-values in order, from largest to smallest, and then assigns a different significance level to each; for example, ≤ 0.05 to the largest, ≤ 0.025 (0.05/2) to the next and 0.017 (0.05/3) to the next, and so on.
- In hierarchical testing (or closed testing), the endpoints of interest are ranked from most important to least important before the study begins. The endpoints are then tested in that order using 0.05 as the significance level. The statistical significance can only be claimed down to the first non-significant result.

(CONTINUED)

Key points – multiplicity (CONTINUED)

- Attention should not be refocused on strong subgroup results because of the problem of multiplicity. If there is a good clinical argument that supports a strong effect in a subgroup, then that subgroup should be built into the confirmatory testing strategy from the beginning of the study.
- Forest plots are a good way of graphically displaying the results of subgroup analyses.

References

1. Jones JA, Robak T, Brown JR et al. Efficacy and safety of idelalisib in combination with ofatumumab for previously treated chronic lymphocytic leukaemia: an open-label, randomised phase 3 trial. *Lancet Haematol* 2017;4:e114–26.

2. Committee for Medicinal Products for Human Use. *Guideline on the Investigation of Subgroups in Confirmatory Clinical Trials – Draft.* www.ema.europa.eu/documents/scientific-guideline/draft-guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf, last accessed 28 April 2020.

5 Interim analysis

It is quite common in large Phase III clinical trials in oncology to introduce one or more interim analyses to assess the data for efficacy between treatment arms while the trial is ongoing and the data are accumulating. It is also possible to introduce interim analyses to assess efficacy or safety data in earlier phases of development.

These analyses have predefined stopping rules for what constitutes overwhelming evidence for efficacy of the experimental treatment. A trial may also be stopped for futility. Sponsor companies do not want to continue to invest in a new drug that does not seem to be offering any advantages in terms of efficacy over an existing treatment.

Advantages of interim analyses

The main advantage of an interim analysis is to shorten the duration of the trial. Stopping a Phase II trial early would accelerate the drug development process. Stopping a Phase III trial early based on overwhelming evidence of efficacy would enable marketing authorization to be submitted sooner than would have been the case were the trial to run to completion. As well as commercial advantages, this has clinical and ethical benefits in that the new treatment can be made available to patients more quickly.

Conversely, if, at the interim stage, the drug demonstrates only limited efficacy potential, it may be prudent to stop the study and use the remaining budget and resources elsewhere in the company on a project that has greater potential for success.

Disadvantages of interim analyses

There are several potential disadvantages of stopping trials early that must be considered at the trial design stage. Stopping a Phase II trial early could limit the amount of data available for other key aspects of the process such as the choice of, or justification for, the optimum dose. Stopping a Phase III trial early based on the evidence for the primary endpoint may limit the amount of data collected from secondary endpoints and safety, which ultimately may delay, or even prevent, approval for the new treatment.

Stopping rules for efficacy

Interim analyses for efficacy have stopping rules that trigger a conclusion of overwhelming efficacy for the experimental treatment. At each interim analysis, a *p*-value calculation compares the treatment groups for the primary endpoint. The *p*-value needs to be sufficiently small to draw a positive efficacy conclusion.

This is an example of multiplicity in which repeated significance tests, i.e. interim analyses in addition to the primary analysis at the end of the study, have the potential to inflate the type I error and thus increase the chances of a false-positive claim (see Chapter 4). For example, there may be three *p*-value calculations for two interim analyses and a final analysis, each of which has the potential to be a false positive. As explained in Chapter 4, 0.05 cannot be used as the cut-off for statistical significance for each analysis if the family-wise error rate (FWER) is to be controlled at 5%. Instead, the type I error rate ($\alpha = 0.05$) must be divided across the three analyses.

Several schemes are used to break down the type I error rate for interim analyses. Provided the scheme is prespecified in the protocol, any of the following splits can be imposed on the design.

O'Brien–Fleming, which is the most common scheme applied to interim analyses, recognizes two key issues.[1]

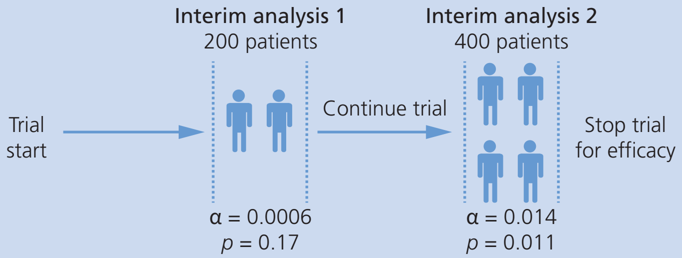
- Trials should not be stopped early for evidence of efficacy unless the evidence is overwhelming.
- The most important analysis is the final analysis, so a substantial amount of the 0.05 should be left over for that final analysis.

For two interim analyses and a final analysis, the division of the alpha ($\alpha = 0.05$) is 0.0006, 0.014 and 0.045. Therefore, at the first interim analysis the adjusted significance level is 0.0006. This must be achieved to trigger a stopping decision for overwhelming evidence for efficacy. At the second interim analysis, the adjusted significance level is 0.014, which leaves a significance level of 0.045 for the final analysis.

Example 5.1

Consider a trial with 600 patients in which interim analyses are conducted when the primary endpoint is reported in 200 and 400 patients.

No. patients	O'Brien–Fleming split	*p*-value
200 (Interim analysis 1)	$\alpha = 0.0006$	$p = 0.17$ Continue trial
400 (Interim analysis 2)	$\alpha = 0.014$	$p = 0.011$ Stop trial
600 (Final analysis)	$\alpha = 0.045$	Analysis not required

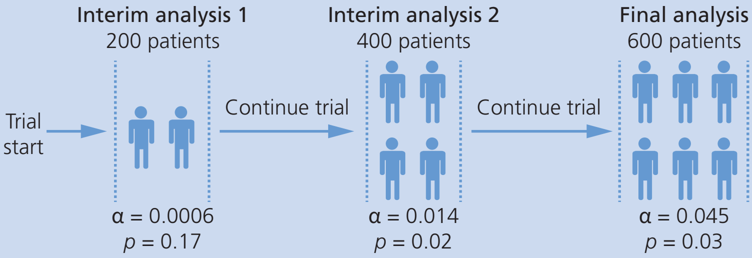


- The *p*-value calculated for the first interim analysis is 0.17. Based on the O'Brien–Fleming split, statistical significance is not achieved according to the required significance level of 0.0006 and the trial therefore continues.
- At the second interim analysis, the *p*-value is 0.011. Statistical significance is achieved according to the required significance level of 0.014 ($p < 0.014$), so the recommendation is to stop the trial.
- The final analysis is not required in this example.

You will probably have recognized that the three adjusted significance levels used in the O'Brien–Fleming scheme do not add up to 0.05. In this situation, one plus one does not equal two – the adjusted α levels are not additive. This is because the analyses are correlated, i.e. the data for the first 200 patients is a subset of the data for the first 400 patients and the data for the first 400 patients is a subset of the data for the total 600 patients. The adjusted significance levels control the overall potential for the false positive (the FWER) at 0.05.

It is important to follow the rules. If the *p*-value in the second interim analysis in Example 5.1 had been 0.02 it would have been statistically significant at a conventional significance level of 0.05. Under the O'Brien–Fleming split, however, it is not. Stopping the trial at $p = 0.02$ would have been a major mistake as it would have destroyed the control of the type I error and destroyed the trial in the eyes of regulators and journal editors. The trial would therefore have continued through to the final analysis (as shown in Example 5.2).

Example 5.2



- Under the O'Brien–Fleming split, the second interim analysis is not statistically significant and the trial therefore continues through to the final analysis.
- In this example, the final analysis is statistically significant.

Other schemes. Another common scheme is Haybittle–Peto, which adjusts the significance levels to 0.001 at each of the interim analyses, again leaving most of the 0.05 for the final analysis.[2,3] Under this scheme, after two interim analyses the final analysis would be conducted at $\alpha = 0.049$.

Other schemes were proposed in Chapter 4 to control the type I error across multiple analyses, namely the Bonferroni correction, the Hochberg correction and hierarchical testing. To recap, the Bonferroni correction divides the 0.05 equally across all analyses. Given that the interim analyses in Example 5.1 are correlated, the Bonferroni significance levels (0.017 for each of three analyses) would be too harsh. In fact, to have the same adjusted level at each analysis, the correct adjusted significance level would be 0.022 (and not 0.017).[4] As the Hochberg scheme would involve obtaining all three *p*-values at the same time and working through them in a specific order, the scheme cannot be applied to Example 5.1, as the *p*-values only become available sequentially over time. Hierarchical testing is also not appropriate for essentially the same reason.

Alpha-spending functions. Schemes such as the O'Brien–Fleming split result in the correct adjusted significance levels for a predetermined number of sequential analyses that are equally spaced (in Example 5.1, the analyses were planned to take place after the primary endpoint had been reported for 200, 400 and 600 patients). However, the trial design may be organized in such a way that the analyses are not equally spaced, in which case the adjusted significance levels change.

For example, it may be undesirable to consider stopping after 200 patients because this would limit the amount of safety data obtained. In this case, there may be only one interim analysis at 400 patients planned, with a final analysis when the primary endpoint is reported in 600 patients. In this setting, the analyses are no longer equally spaced and the adjusted alpha levels would need to be recalculated to control the overall alpha level. This is achieved with computer software using so-called alpha-spending functions to determine the levels needed to control the FWER at 0.05.

Event-driven analyses. With time-to-event endpoints in oncology trials, the power of the study is driven by the number of events rather than the number of patients. Consequently, the final analysis is timed to occur when a certain number of events has been reached. Under these circumstances, the interim analyses will also be event driven.

Example 5.3

In the PRIMA trial, patients with a high tumor burden of follicular lymphoma who had responded to rituximab plus chemotherapy were randomized to receive a further 2 years of rituximab maintenance therapy or no maintenance therapy.[5] The primary endpoint was progression-free survival (PFS).

Two event-driven interim analyses were planned, one at 50% (n = 172) and one at 75% (n = 258) of the total number of anticipated events (n = 344).

> *The alpha-spending function with the O'Brien–Fleming boundary was applied for the interim analysis to maintain the overall two-sided type I error of 0.05.*

These interim analyses:

- are not equally spaced
- may not take place precisely as planned.

Even though the analyses are planned to take place after 172, 258 and 344 PFS events, for practical reasons such as confirming progression and planning for the various analyses, they may, for example, happen at 175, 256 and 352 events. Alpha-spending functions enable the correct adjusted significance levels to be used that are consistent with the preplanned and prespecified O'Brien–Fleming scheme.

Stopping rules for futility

The objective here is to stop the study because it has no real chance of achieving the required level of statistical significance for the primary endpoint(s) if it were to run to completion. In general, a futility assessment takes place reasonably early in the trial, as stopping a study for futility potentially saves a sponsor's resources. There is usually only one interim analysis for futility, which for practical reasons may coincide with the first interim analysis for efficacy if such an analysis has been built into the study design, although this is not a requirement.

Usually futility is based on the data for the primary endpoint, although a surrogate endpoint that matures more quickly than the primary endpoint can also be used.

Conditional power. The most common approach to futility utilizes the concept of conditional power, i.e. the probability that the final analysis will be statistically significant given the data observed in the interim analysis. Power measures the ability to detect differences between treatments should they exist.

As an example, imagine a study is powered at 90% to detect a 25% reduction in the hazard for death in the experimental group compared with the control group (i.e. hazard ratio [HR] 0.75). To achieve this, the researchers plan to recruit 700 patients and perform the primary analysis after 520 events. If the true HR is indeed 0.75, then there will be a 90% probability of detecting a statistically significant difference between the treatments.

Now, suppose that an interim analysis for futility is conducted after 173 events (approximately one-third of the way through the trial in terms of the number of events), and the HR is 0.92. Is there any point in continuing the trial? The data are nowhere near the anticipated treatment difference (HR = 0.75), and if this trend were to continue a significant *p*-value at the end of the trial would be highly unlikely.

To judge this, the conditional power of the study is calculated. The conditional power recalculates the power of the study having now seen one-third of the data (173 events) and how the events are split across the two treatment groups. There are two calculations that can be performed here:

- conditional power under the current trend
- conditional power under the original assumption.

Conditional power under the current trend assumes that the data observed so far (HR = 0.92) reflect the truth. If conditional power under the current trend is, for example, < 20%, and if the observed trend in each treatment arm continues for the remainder of the study, it can be argued that the trial is not worth continuing as there is such a small probability of a statistically significant result at trial completion.

Example 5.4

In a study comparing nintedanib/pemetrexed and placebo/pemetrexed in patients with relapsed or refractory advanced non-small-cell lung cancer (LUME-Lung 2 study), an interim analysis for futility was planned based on the conditional power under the current trend of < 20%.[6] The trial was stopped with a conditional power of 10.3%.

Based on the pre-planned futility analysis of investigator-assessed PFS, conducted by an independent data monitoring committee, recruitment was halted on 18 June 2011 after 713 (n = *353 nintedanib/pemetrexed;* n = *360 placebo/pemetrexed) of 1300 planned patients had enrolled.*

Conditional power under the original assumption. Alternatively, a more optimistic view is that the data observed so far do not reflect the truth and that the true HR is still 0.75. This increases the conditional power, but if it remains low (e.g. < 30%), then it could still be argued that it is not worth continuing the trial. In this instance, even with optimistic assumptions about the trends in the data moving forward, the probability for success remains low.

Choice of cut-off for futility, as determined by the conditional power, is entirely at the discretion of the sponsor. Even a trial with a conditional power of 30% still has a 30% probability of success and the sponsor must decide whether this is something that is worth pursuing. Cut-off values for conditional power under the current trend are generally about 20% or 30%, but there are examples of sponsors using values as high as 50%.

Other futility approaches. Futility rules based on conditional power are the most common, but there are other approaches. A futility decision can also be based on a high *p*-value (e.g. > 0.5) or on the 95% confidence interval excluding hazard ratios that would be viewed as clinically important. There are mathematical connections between these rules and the high *p*-value and 95% confidence interval exclusion rules can both be translated into a rule based on conditional power.

There is no alpha price to pay for these futility rules since they are not associated with any kind of inflation of the false-positive potential for the trial.

Pros and cons of stopping for futility

The value of negative data. Although stopping a study for futility is potentially a benefit for the sponsor, some clinicians argue that negative studies are of value as they contribute to the overall body of evidence for a drug or a class of drugs.

Premature trial cessation. Stopping a trial for futility reduces power; a trial with 30% conditional power still has a 30% chance of success, but stopping a study at that point with that conditional power diminishes the probability of a positive trial result. Some trials with low conditional power have been stopped prematurely. Indeed, the LUME-Lung 2 study in Example 5.4 is one such trial. In that trial, the futility rule was built around a surrogate endpoint (investigator-assessed PFS) but subsequent analysis showed that the primary endpoint of improvement in centrally reviewed PFS was met.[7]

The independent data monitoring committee

It is vitally important that interim analyses for efficacy and/or futility do not cause bias. Two forms of bias are of concern: statistical bias and operational bias.

Statistical bias occurs if there is no formal control of the type I error. The schemes discussed above, such as O'Brien–Fleming, must be strictly applied to avoid this kind of bias.

Operational bias occurs when interim results influence changes of behavior; for example, investigators may decide to recruit patients with different characteristics based on especially positive or negative interim results. Control of interim analyses by an independent data monitoring committee (IDMC), which keeps interim results confidential, helps to avoid operational bias. The committee reviews the analyses and makes recommendations back to the sponsor or the clinical trial steering committee regarding the stopping or continuation of the study. Both the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) have issued comprehensive guidelines on the structure and operating procedures for IDMCs.[8,9]

Key points – interim analysis

- Interim analyses assess the data for efficacy or for futility (or both) while the trial is ongoing and data are still accumulating.
- Interim analyses have predefined stopping rules for overwhelming efficacy of the experimental treatment, or futility if the experimental treatment does not appear to offer any efficacy advantage over the existing treatment (or against placebo).
- Multiple analyses have the potential to inflate the type I probability error (multiplicity), increasing the chances of false-positive results. The type I error rate ($\alpha = 0.05$) must therefore be divided across the interim analyses for efficacy.
- In most schemes to divide the alpha, a substantial amount of the 0.05 is left for the final analysis, with the interim analyses for efficacy having very stringent significance levels.
- There is usually only one interim analysis for futility, relatively early in the trial.
- The conditional power recalculates the power of the study based on the interim data under either the current trend or the original assumption. The former assumes the data in the interim analysis reflect the truth, whereas the latter assumes the original assumption about the data is correct.

References

1. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979;35:549–56.

2. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971;44:793–7.

3. Peto R, Pike MC, Armitage P et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br J Cancer* 1976;34:585–612.

4. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 1977;64:191–9.

5. Salles G, Seymour JF, Offner F et al. Rituximab maintenance for 2 years in patients with high tumour burden follicular lymphoma responding to rituximab plus chemotherapy (PRIMA): a phase 3, randomised controlled trial. *Lancet* 2011:377: 42–51.

6. Hanna NH, Kaiser R, Sullivan RN et al. Nintedanib plus pemetrexed versus placebo plus pemetrexed in patients with relapsed or refractory, advanced non-small cell lung cancer (LUME-Lung 2): a randomized, double-blind, phase III trial. *Lung Cancer* 2016;102:65–73.

7. Lesaffre E, Edelman MJ, Hanna NH et al. Statistical controversies in clinical research: futility analyses in oncology–lessons on potential pitfalls from a randomized controlled trial. *Ann Oncol* 2017;28:1419–26.

8. Committee for Medicinal Products for Human Use. *Guideline on Data Monitoring Committees*. www.ema.europa.eu/documents/scientific-guideline/guideline-data-monitoring-committees_en.pdf, last accessed 26 March 2020.

9. US Food and Drug Administration. Guidance for Clinical Trial Sponsors. Establishment and Operation of Clinical Trial Data Monitoring Committees. www.fda.gov/downloads/regulatoryinformation/guidances/ucm127073.pdf, last accessed 26 March 2020.

6 Modeling

Statistical models are mathematical equations that are fit to a sample of data to help researchers understand how baseline factors, patient characteristics and/or treatment group affect outcomes. Statistical modeling enables researchers to:

- ascertain whether outcomes are dependent on baseline variables/ factors
- compare treatments, taking into account baseline imbalances and correcting for multiple baseline variables/factors simultaneously to produce an unbiased comparison
- improve the power of the study
- investigate treatment × factor interactions to evaluate the consistency/homogeneity of the treatment effect.

In Chapter 1 (page 21), the comparison of two treatment group means was based on the two-sample *t*-test. Sometimes, treatment groups are not well balanced for baseline factors that are predictive of outcome. While randomization helps to prevent those imbalances, they can still occur by chance, and when such imbalances are present the resulting simple treatment comparisons are biased.

This chapter begins by explaining how to account for such imbalances in a simple way to produce an unbiased comparison.

It then looks at how to evaluate dependence, i.e how to assess whether an outcome depends on one or more factors measured at baseline.

The chapter then switches back to the comparison of two treatments accounting for baseline imbalances, extending the simple methods developed earlier to allow adjustment for imbalances in several baseline factors simultaneously.

Finally, the concept of a statistical model is introduced to incorporate all of the outlined methods within a single mathematical framework. This provides a general unified approach to the investigation of dependence and the adjustment for baseline imbalances.

Adjusting for baseline imbalances when comparing means

Consider the setting in which an active drug is evaluated for its ability to lower total cholesterol in a placebo-controlled trial. Patients were randomized to receive dietary and lifestyle advice plus placebo or the active drug. Table 6.1 shows summary statistics (where n is the number of subjects and $\bar{x}$ is the mean fall in total cholesterol over the treatment period) for each of three age groups: < 50, 50 to < 60 and ≥ 60 years. Note these data are artificial, chosen to allow a simple explanation of the methodology.

Two observations can be made from the data in Table 6.1.

First, on average, older subjects did less well. In the control (placebo) group, subjects younger than 50 years old had a mean fall in total cholesterol of 0.36 mmol/L, the middle age group had a mean fall of 0.17 mmol/L, while the oldest group had a mean fall of 0.10 mmol/L. A similar pattern was seen in the active treatment group, with the youngest subjects having a mean fall of 0.68 mmol/L and the oldest subjects having a mean fall of only 0.36 mmol/L.

Second, there is a baseline imbalance in age. Table 6.2 presents the age distribution in terms of percentages. Of the 78 patients randomized to the control group, 36% were younger than 50 years old and 42% were aged 60 years or older. In contrast, of the 80 patients randomized to the active treatment group, only 26% were younger than 50 years old and 54% were 60 years or over. In fact, the mean age in the control group was 55.6 years compared with 57.8 years in the

TABLE 6.1

Summary statistics for a cholesterol-lowering trial*

Age (years)	Control (placebo)	Active treatment
< 50	$n = 28, \bar{x} = 0.36$	$n = 21, \bar{x} = 0.68$
50 to < 60	$n = 17, \bar{x} = 0.17$	$n = 16, \bar{x} = 0.39$
≥ 60	$n = 33, \bar{x} = 0.10$	$n = 43, \bar{x} = 0.36$

*Artificial data.

TABLE 6.2

Age distribution for a cholesterol-lowering trial*

Age (years)	Control (placebo)	Active treatment
< 50	*n* = 28 (36%)	*n* = 21 (26%)
50 to < 60	*n* = 17 (22%)	*n* = 16 (20%)
≥ 60	*n* = 33 (42%)	*n* = 43 (54%)
Total	*n* = 78 (100%)	*n* = 80 (100%)

*Artificial data.

active treatment group. This is not a huge difference, but it is a difference that could influence the treatment comparison, given that the efficacy seems to depend quite strongly on the age of the patient.

The raw means in the two groups were 0.21 mmol/L (control) and 0.45 mmol/L (active treatment), with a treatment difference of 0.24 mmol/L. This difference was not statistically significant, with a two-sided *p*-value of 0.072.

The question clearly arises, if there had been no baseline imbalances in age distribution, would the treatment difference have been statistically significant?

Table 6.3 shows the age distribution in the two groups combined (final column), together with the observed mean fall in total cholesterol for each treatment group in each age category.

In order to correct for the baseline imbalances in age distribution, the means are reconstructed to show what they would have been if the age distribution in each of the two treatments groups had been identical, as per the breakdown in the final column of Table 6.3.

In that case, the mean fall in total cholesterol in the control group would have been as follows:

$$\text{Control group mean} = 31\% \times 0.36 + 21\% \times 0.17 + 48\% \times 0.10 = 0.20$$

with 31% of subjects giving a mean of 0.36, 21% of subjects giving a mean of 0.17 and 48% of subjects giving a mean of 0.10.

TABLE 6.3

Mean fall in cholesterol by age

	Control	Active	
Age (years)	(placebo)	treatment	Total
< 50	$\bar{x} = 0.36$	$\bar{x} = 0.68$	$n = 49$ (31%)
50 to < 60	$\bar{x} = 0.17$	$\bar{x} = 0.39$	$n = 33$ (21%)
≥ 60	$\bar{x} = 0.10$	$\bar{x} = 0.36$	$n = 76$ (48%)
			$n = 158$ (100%)

With this breakdown of subjects across the three age categories, there would have been a mean fall in total cholesterol of 0.20 mmol/L in the control group. The raw data gave a mean fall of 0.21 mmol/L.

Similarly, the mean fall in total cholesterol in the active treatment group would have been as follows:

$$\text{Active group mean} = 31\% \times 0.68 + 21\% \times 0.39 + 48\% \times 0.36 = 0.47$$

with 31% of subjects giving a mean of 0.68, 21% of subjects giving a mean of 0.39 and 48% of subjects giving a mean of 0.36.

With this breakdown of subjects across the three age groups, there would have been a mean fall in total cholesterol of 0.47 mmol/L in the active treatment group. In contrast, the raw data gave a mean fall of 0.45 mmol/L.

The difference between these so-called adjusted means is 0.27 mmol/L (0.47 – 0.20) compared with a smaller difference between the raw means of 0.24 mmol/L (0.45 – 0.21). This is as expected. The active treatment group was penalized by having more older subjects and fewer younger subjects than the control group. When this is corrected, there is a larger treatment effect.

This adjusted treatment difference of 0.27 mmol/L is statistically significant, with a two-sided *p*-value of 0.045. Adjusting for the imbalances in age across the treatment groups has produced a statistically significant difference.

The calculation for the adjusted mean difference can also be shown as follows:

$$\begin{aligned}\text{Adjusted difference} &= (31\% \times [0.68 - 0.36]) + (21\% \times [0.39 - 0.17]) \\ &\quad + (48\% \times [0.36 - 0.10]) \\ &= 0.27\end{aligned}$$

This calculation takes the treatment difference in each of the age groups (strata) and combines them by weighting according to the percentage of subjects in each stratum. This corresponds to a weighted average of the treatment differences in each stratum. Had the percentages in each stratum been the same, i.e. equal numbers of subjects in each stratum for both treatment groups combined, then the adjusted difference would have been the simple average of the treatment differences in each stratum.

Is this modified analysis legitimate? Yes, it is! In fact, because of the age imbalances, the unadjusted analysis based on the raw means is incorrect – it is not comparing like for like in terms of the mix of subjects. The adjusted analysis is the correct analysis.

It is routine to adjust for baseline imbalances in factors that influence outcome. Even minor imbalances can affect the comparability of the treatment groups and bias the estimate of the true treatment effect.

Is adjusting the analysis always necessary? If the groups are perfectly balanced, the adjusted analysis will be the same as the unadjusted analysis, so adjustment is not required. However, it is impossible to identify imbalances until the data are unblinded; so, given the emphasis that regulators place on preplanning, it is always preferable to plan for adjustment routinely. Furthermore, the technique outlined above for adjusting the analysis is more sensitive to detecting treatment differences; it increases the power to detect differences if those differences truly exist.

The statistical technique for adjustment is called analysis of variance (ANOVA). Other terms for the technique are adjusted or stratified analysis.

Adjusting the analysis for binary and time-to-event endpoints

The same technique can be used for different types of endpoint.

For a binary endpoint (e.g. responder versus non-responder), the data can be broken down according to categories of the factor that needs adjustment, calculating the proportions of events in each treatment group in each factor category. The treatment difference (response rate on active treatment minus response rate on placebo) is then calculated in each factor category. The adjusted treatment difference is the average of the treatment differences in the factor categories weighted by the percentages of patients overall in those categories. The *p*-value comparing the treatment groups is then based on this adjusted difference.

The adjusted response rates for each of the treatment groups are calculated as weighted averages of the response rates in each factor category, using the weights according to the overall percentages in the factor categories. This technique for adjustment and *p*-value calculation is called the Cochran–Mantel–Haenszel (CMH) test.

For a time-to-event endpoint (e.g. overall survival), the same process is followed. Having broken down the data according to the factor categories, a hazard ratio (HR) is calculated within each category. Note that the HR is already a measure of treatment difference. The HRs are then averaged on the log scale in the same way, by weighting according to the percentage of patients in both treatment groups, to give an adjusted HR and an associated *p*-value. This technique is termed the stratified logrank test.

Adjusting the analysis for multiple baseline factors

These stratified analyses (ANOVA for continuous and score endpoints, CMH test for binary endpoints and stratified logrank test for time-to-event endpoints) are routinely used to adjust for one or two baseline factors; for example, to adjust for two factors (e.g. age in three categories and sex), the data would need to be broken down into six categories according to the six combinations of age and sex.

When adjusting for multiple baseline factors, however, the technique can break down. For example, to adjust for three baseline factors – age (< 50 vs 50 to < 60 vs ≥ 60), sex (male vs female) and baseline risk (ECOG 1 vs ECOG 2) – the data would need to be broken down into 12 (3 × 2 × 2) categories (strata), but there may not be any patients in the active treatment group who are, say, under 50 years, male and ECOG 1. If that is the case, a treatment difference for that category could not be obtained and the methodology described above would break down.

Modeling provides an extension of this methodology which, at least in principle, enables multiple baseline factors to be considered simultaneously (see below).

Assessing the dependence of outcome on baseline factors

To introduce ideas on modeling, the following section moves away from the comparison of two treatments to look at the dependence of an outcome on a baseline factor. In some studies, researchers look to ascertain whether any baseline variables (e.g. the patient's age, sex or ethnicity) have influenced outcome.

Simple linear regression evaluates a single baseline variable (termed a univariable [or univariate] analysis). Multiple regression evaluates several variables simultaneously (termed a multivariable [or multivariate] analysis).

Simple linear regression analysis is a statistical tool used to quantify the dependence of an outcome (the dependent variable) on a baseline factor (the independent variable). For example, in patients with lung cancer, is time to disease progression (the dependent variable) affected by the size of the primary tumor at baseline (the independent variable) (Figure 6.1)? For simplicity, the analysis assumes that there is no censoring for time to disease progression (see page 24 for a discussion on censoring).

Visual inspection of the plot in Figure 6.1 suggests that patients with larger tumors at baseline have a shorter time to disease progression. To investigate the nature of this dependence, a straight line is 'fitted' to the data (Figure 6.2).

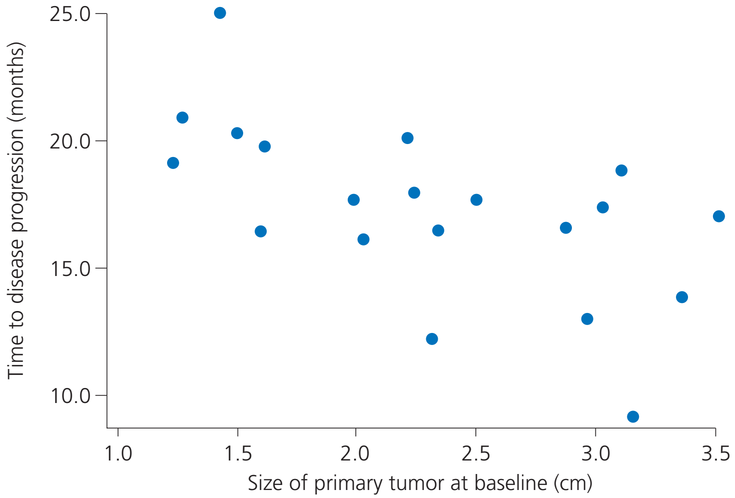


Figure 6.1 Dependence of time to disease progression on size of primary tumor in 20 patients.

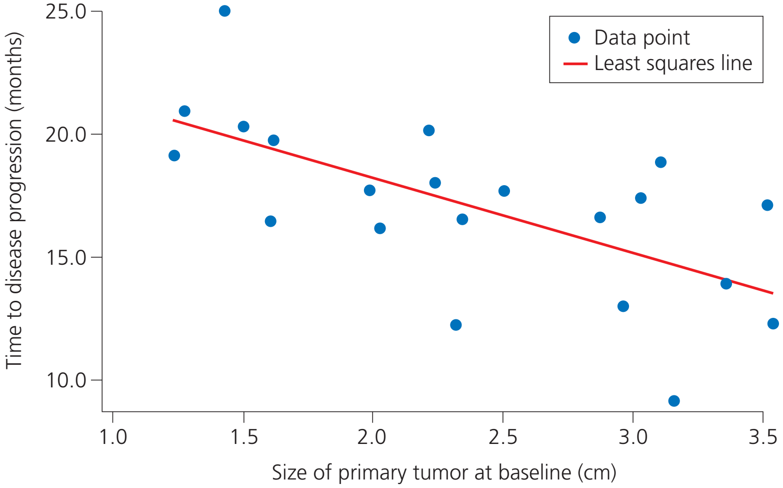


Figure 6.2 A straight line is 'fitted' to the data. This is called the least squares regression line (see text).

What does the line demonstrate? The equation of a straight line is:

$$y = a + bx$$

where 'a' is the point at which the line crosses the y-axis (the intercept) and 'b' is the slope of the line. If the line runs upward from bottom left to top right, b is positive; if the line runs downward from top left to bottom right then b is negative. If the line is horizontal, then $b = 0$ and indicates no dependence (i.e. as x increases, nothing happens to y). To further clarify what is meant by the slope of the line, b is the amount by which y is increased when x is increased by one unit.

Figure 6.2 shows the best fitting line in this case. The equation of this line is:

$$y = 24.3 - 3.05x$$

The method used to obtain this line is discussed later, but for the moment the focus is on the interpretation. If the line were to be tracked back to the y-axis, at $x = 0$ it would cross the y-axis at 24.3 months. The value of b is negative, indicating a negative dependence, i.e. as x increases, y is reduced. More specifically, as the size of the primary tumor (x) increases by one unit (e.g. from 2 cm to 3 cm), the time to disease progression (y) is reduced, on average, by 3.05 months.

How is the line fitted to the data? The best-fitting line is obtained by drawing a vertical line from each data point to the line (as shown in Figure 6.3 for two of the points) and calculating the average distance from all points to the line. The best-fitting line is the line that makes the average distance from all the points to the line as small as possible. However, it is not the simple average that is used as a metric, but a weighted average. The best fitting line is the line that makes the average squared distance of all of the points to the line as small as possible. This process is called least squares and the resulting line is called the least squares regression line.

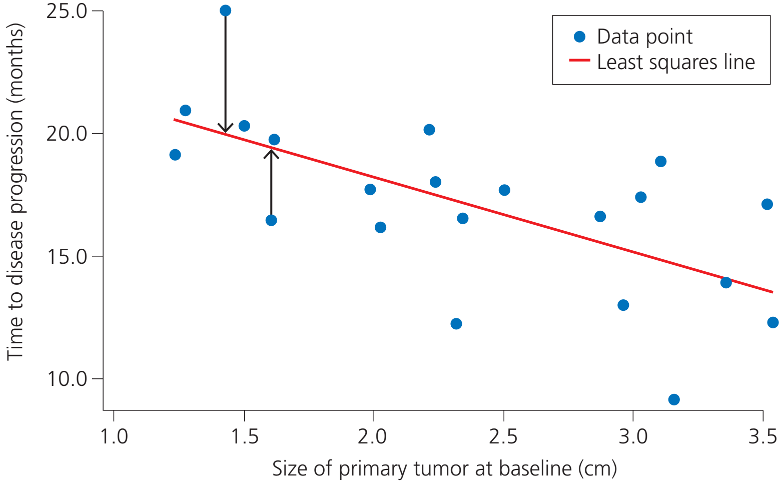


Figure 6.3 Fitting the least squares regression line.

Evaluating dependence. In this example, there is a clear dependence, but in practice it is not always as clear cut as this and the main question will usually be, is there evidence of dependence? To address this question statistically, two hypotheses are formulated:

$$\text{Null hypothesis, } H_0 : b = 0, \text{ no dependence}$$

versus

$$\text{Alternative hypothesis, } H_1 : b \neq 0, \text{ dependence}$$

A p-value is obtained, based on the calculated value for b. In this example, the two-sided p-value is 0.0004, a highly significant difference from zero, so there is strong evidence for dependence.

Multiple regression analysis. Having considered the dependence of an outcome (time to disease progression) on a single baseline factor (size of primary tumor at baseline), this approach can be extended to look simultaneously at the dependence on multiple baseline factors.

Although it is not possible to draw this as in Figures 6.1–6.3, the simple regression equation can be extended to a multiple regression equation by including additional x variables.

For example, evaluating the dependence of time to disease progression (y) on three baseline factors rather than one: size of primary tumor (x_1), age (x_2) and sex (x_3). Note that sex is a binary variable (male/female) while the other two variables are numeric. Binary variables can be incorporated into the modeling framework by defining them with a so-called indicator variable:

$$0 = \text{male},\ 1 = \text{female}$$

To investigate the dependence, the following equation is fitted to the data:

$$y = a + b_1\, x_1 + b_2\, x_2 + b_3\, x_3$$

The coefficients b_1, b_2 and b_3 reflect how each of the baseline factors affect time to disease progression independently. Least squares can again be used to obtain the best-fitting equation, and the effects of each baseline factor on the outcome can be interpreted according to the direction (positive or negative) and magnitude of the b coefficients.

Statistical significance can be assessed by testing hypotheses for each of the b coefficients in turn. Suppose, for example, that the least squares regression line for this example is as follows:

$$y = 22.7 - 3.52x_1 - 0.23x_2 + 1.51x_3$$

with $p = 0.006$ for b_1, $p = 0.37$ for b_2 and $p = 0.04$ for b_3.

The effects of primary tumor size and sex are statistically significant, but age is not.

Why one equation? What are the advantages of fitting these baseline factors into one equation? Why not perform three separate regression analyses on each of the variables? The latter approach could be misleading if the baseline factors are correlated, which they usually will be. For example, age and size of primary tumor may be correlated, with older patients generally presenting with larger tumors. A simple regression analysis, with age as the single baseline factor, could well give a significant *p*-value suggesting a dependence, but this would be misleading. It is not age that drives outcome: age is correlated with the size of primary tumor and it is the size of the primary tumor that determines outcome. The only way to ensure this correlation is accounted for is to include all variables simultaneously in a multiple regression analysis. This technique sorts out the correlation and identifies where the true dependence lies.

Comparing groups while accounting for baseline imbalances

Analysis of covariance for a single covariate. Analysis of covariance (ANCOVA) is used to compare treatment groups while accounting for baseline imbalances. It is an extension of the ANOVA introduced earlier.

Figure 6.4 is a scatter plot showing the relationship between time to disease progression and the size of the primary tumor (the covariate) in two groups (active treatment versus control [placebo]), each comprising 20 subjects.

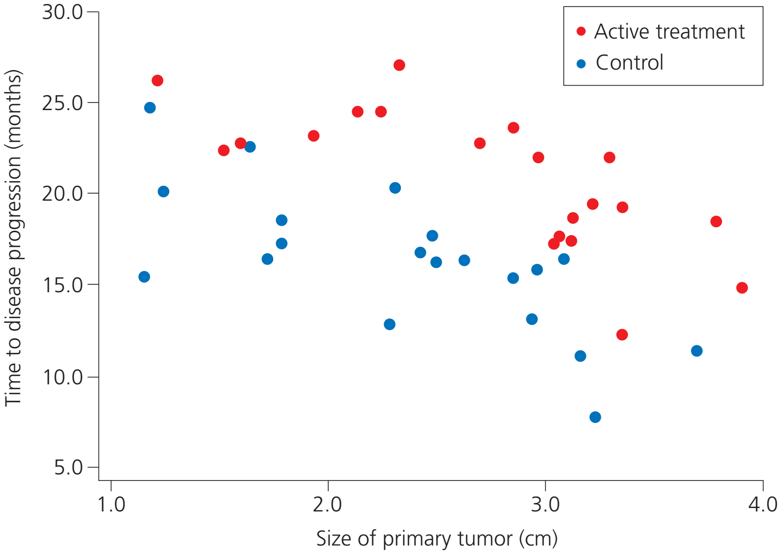


Figure 6.4 Randomized parallel-group study comparing treatment groups for time to disease progression.

In this instance, a simple comparison of mean time to disease progression in the active treatment group versus the mean time to disease progression in the control group, ignoring the role of size of primary tumor, would be a perfectly valid approach. However, it would not be especially sensitive as those means would not be that different. As can be seen in Figure 6.4, there are patients in the active treatment group with large tumors whose outcomes are poor (time to disease progression of 15 months or less). There are also patients in the control group who are doing well (time to disease progression of over 20 months). Therefore, a straight comparison of mean times to disease progression would not be a good reflection of the separation between the two groups that is clearly visible in Figure 6.4.

An alternative way of comparing the two treatment groups would be to exploit the dependence of time to disease progression on size of primary tumor. To do this, two equations are fitted to the data, one for the active treatment group and one for the control group:

$$y = a + bx, \text{ control group}$$

$$y = (a + c) + bx, \text{ active treatment group}$$

Note that it has been assumed that the two lines have the same slope b, i.e. they are parallel. This assumption is discussed later. The control group regression line crosses the y-axis at a while the active treatment group regression line crosses the y-axis at $a + c$. This means that the vertical distance between the two lines is equal to c (Figure 6.5).

The fitted regression lines for these data are:

$$y = 25.18 - 3.77x, \text{ control group}$$

$$y = (25.18 + 5.94) - 3.77x, \text{ active treatment group}$$

If the active treatment was not effective, the data points for that group would be overlapping with the data points for the control group and the two fitted lines would not be separated.

The greater the separation between the two lines, the greater the treatment effect.

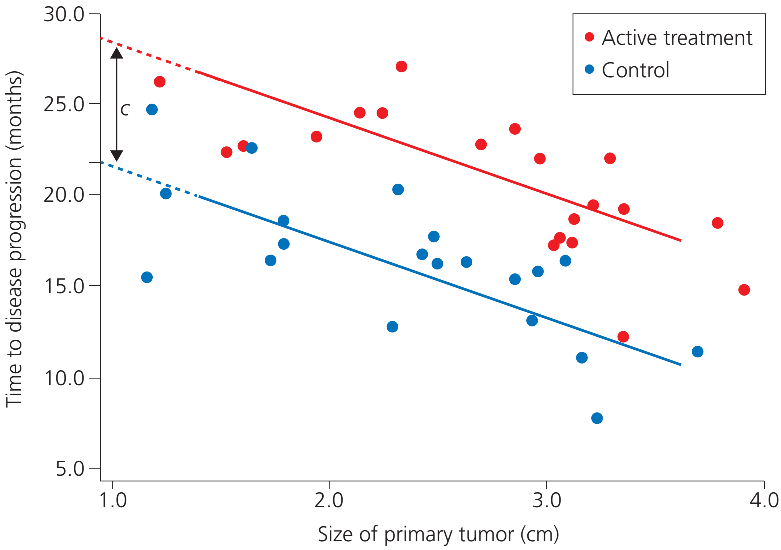


Figure 6.5 Randomized parallel-group study comparing treatment groups for time to disease progression.

The vertical distance between the lines (c) is 5.94, suggesting that, on average and irrespective of baseline tumor size, the active treatment extends time to disease progression by almost 6 months

To investigate whether the distance between the regression lines is significantly different from zero, hypotheses are formulated around the value of c, the distance between the lines.

$$\text{Null hypothesis, } H_0 : c = 0\text{, no treatment effect}$$
$$\text{versus}$$
$$\text{Alternative hypothesis, } H_1 : c \neq 0\text{, treatment effect}$$

As expected by visual inspection of the lines, the p-value is highly statistically significant ($p < 0.001$).

Suppose that size of primary tumor was not well balanced across the control and active treatment groups, with more patients with larger tumors being randomized to the active treatment group and more patients with smaller tumors being randomized to the control group.

This imbalance would penalize the active treatment group. Simply comparing the mean time to disease progression for each group would be misleading and could fail to pick up a treatment effect. ANCOVA corrects for those baseline imbalances.

The method fits the two regression lines to the data and then compares the lines, ignoring any imbalances. The fact that the groups are unbalanced according to size of primary tumor does not affect the comparison of the regression lines. It is also the case that ANCOVA increases sensitivity, providing more power for treatment comparisons.

ANCOVA or ANOVA? How do they differ, and which is preferred? ANOVA works on a simple division of the independent variable, in this case tumor size, so if patients had been subdivided according to the size of primary tumor (< 2 cm, 2–3 cm, > 3cm) then ANOVA could have been used. The *p*-value and the reported treatment difference would have been very similar with both techniques, so the choice of method is largely a matter of personal preference.

There are two aspects, however, that should be considered. First, ANOVA works around a simple division in terms of tumor size, placing the patients in one of three categories in this example. ANCOVA uses the information on tumor size more efficiently by including this variable on a continuous scale; in that respect, it is the preferred technique.

Second, ANCOVA assumes that the effect of tumor size on time to disease progression is linear; i.e. it can be described by a straight-line relationship. This may not always be the case. For example, patients with small tumors may have different outcomes from those with medium or large tumors, while patients with medium or large tumors may have very similar outcomes. This would show as a plateauing of the data and would violate the assumption of linearity. ANOVA, on the other hand, makes no assumptions regarding the underlying relationship between tumor size and time to disease progression and for that reason could be the preferred approach under some circumstances.

The real advantage of ANCOVA, however, is its ability to incorporate multiple covariates, a setting in which ANOVA often breaks down, as discussed earlier. ANCOVA can also be used to investigate treatment × covariate interactions. These two aspects are discussed below.

Analysis of covariance with multiple covariates. ANCOVA can be extended to include several covariates simultaneously by simply adding more baseline factors into the two equations; for example, with size of primary tumor, age and sex as baseline factors (multiple covariates) the equations become:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3, \text{ control group}$$

$$y = (a + c) + b_1x_1 + b_2x_2 + b_3x_3, \text{ active treatment group}$$

This method provides a treatment comparison based on the value of *c* and its associated *p*-value, adjusting for baseline imbalances in size of primary tumor, age and sex.

Statistical interactions

In the example with a single covariate only, the ANCOVA model assumes that the slopes of the two lines are the same, i.e. they are parallel, and, therefore, that the average treatment benefit is constant: the treatment benefit for a patient with a 2-cm tumor will, on average, be the same as the treatment benefit for a patient with a 3-cm tumor. This may of course not be the case (Figure 6.6).

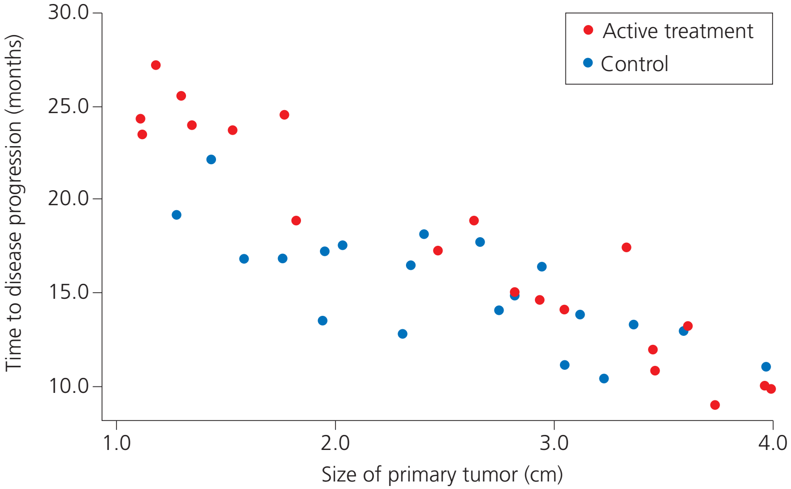


Figure 6.6 Comparing time to disease progression in two treatment groups for which the treatment effect may not be constant.

Consider therefore relaxing the assumption of parallel lines and allowing the two regression lines to have different slopes. Below are the equations, with slopes b_1 and b_2 for the control and active treatment groups, respectively. Note that the intercepts have been labeled a_1 and a_2 for convenience in this case.

$$y = a_1 + b_1x, \text{ control group}$$

$$y = a_2 + b_2x, \text{ active treatment group}$$

For this example, the best-fitting (least squares) regression lines are:

$$y = 25.36 - 3.85x, \text{ control group}$$

$$y = 21.51 - 3.70x, \text{ active treatment group}$$

The lines are displayed in Figure 6.7. The apparent visual difference is very much a function of the scaling for the x- and y-axes, as the actual numeric values of the two slopes, –3.85 and –3.70, are similar.

Comparing b_1 and b_2 to assess whether the data demonstrate a non-constant treatment effect produced a two-sided p-value of 0.90 (non-significant). It is therefore reasonable to assume a homogeneous effect across all values for size of primary tumor.

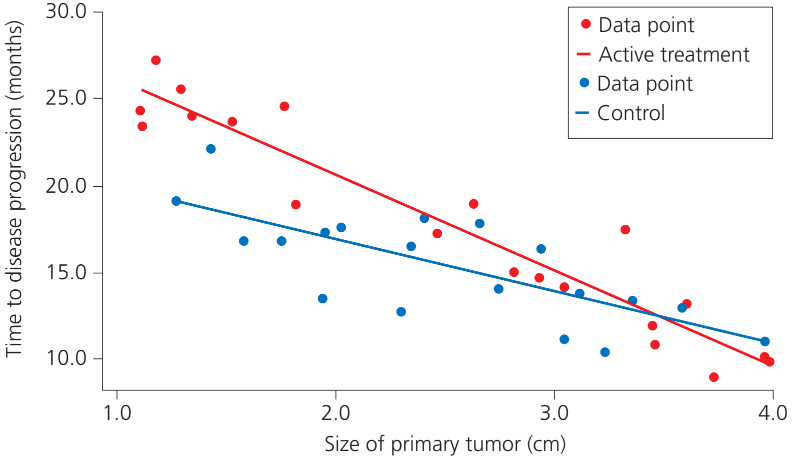


Figure 6.7 Non-parallel regression lines fitted to data from two groups for which the treatment effect may not be constant.

If the differences between the slopes had been significant, there would have been a treatment group by covariate (treatment × covariate) interaction, with evidence that the treatment difference depends on the size of the primary tumor.

However, in our example that is not the case and the next step is to constrain the lines to be parallel, fit the two equations assuming a common slope and proceed as previously. In this example, the common slope is –3.77 and the treatment effect (the difference in the intercepts) is 5.55 months, significantly different from 0 with $p < 0.001$. Thus, the experimental treatment offers, on average, 5.55 months longer to disease progression than the control and this benefit does not depend on the size of the primary tumor at baseline.

Residual variability

Residual variability is an additional element to the modeling framework. With simple linear regression, the straight-line equation ($y = a + bx$) is fitted to the data. It is a model for what is expected to happen on average. According to this model, patients who present with size of primary tumor = x will have mean time to disease progression = $a + bx$. Individual patient values will vary around that mean, and the usual assumption is that the variability is consistent with a normal distribution (Figure 6.8). This takes account of the patient-to-patient variability.

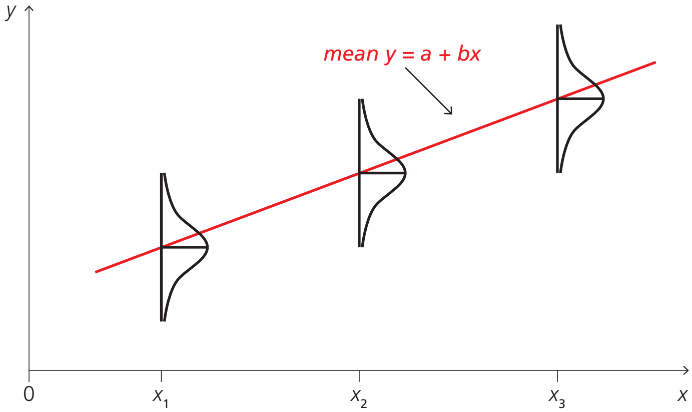


Figure 6.8 Normal distribution residual variability.

Statistical modeling

It is possible to use a statistical model that incorporates all these methods in a single mathematical framework.

Models for continuous and score endpoints with a single covariate. For the different settings outlined above, with a single baseline factor (covariate), the following equations specify the statistical model.

- Regression (studying dependence)

$$\text{mean } y = a + bx$$

- ANCOVA (adjusting for baseline imbalances)

$$\text{mean } y = a + cz + bx$$

where $z = 0$ if the patient is in the control group, $z = 1$ if the patient is in the active treatment group.

This model reduces to:

$$\text{mean } y = a + bx\text{, control group}$$

$$\text{mean } y = (a + c) + bx\text{, active treatment group}$$

- ANCOVA (investigating treatment × covariate interactions)

$$\text{mean } y = a + cz + bx + dx \times z$$

This model reduces to:

$$\text{mean } y = a + bx\text{, control group}$$

$$\text{mean } y = (a + c) + (b + d)x\text{, active treatment group.}$$

Each setting is fundamentally using *mean y* on the left-hand side of the equation and terms that incorporate covariates and treatment on the right-hand side. It is straightforward mathematically to include several covariates simultaneously by adding more x variables to the right-hand side.

Models for binary and time-to-event endpoints. The models discussed above work for continuous and score endpoints where the mean value is being modeled. When the endpoint is binary the corresponding model is called the logistic model.

Logistic regression is the term used when looking at dependence. The model is expressed in terms of the log odds for the event on the left-hand side, while the right-hand side is exactly like the models discussed above:

$$\text{log odds for the event} = a + bx$$

To adjust for baseline imbalances, the model is:

$$\text{log odds for the event} = a + cz + bx$$

For this model it is easy to show that c is directly related to the odds ratio (*OR*) for treatment with natural logarithm (ln) $c = OR$.

Cox regression for time-to-event endpoints (Cox proportional hazards model) is expressed in terms of the log hazard rate. When looking for dependence the model is expressed as:

$$\text{log hazard rate for the event} = a + bx$$

To adjust for baseline imbalances the model is:

$$\text{log hazard rate for the event} = a + cz + bx$$

For this model it is easy to show that ln c = HR where HR is the hazard ratio with treatment.

Key points – modeling

- A stratified analysis (analysis of variance; ANOVA) gives a treatment effect, adjusting for baseline imbalances for continuous and score endpoints.
- The Cochran–Mantel–Haenszel test compares treatments, adjusting for baseline imbalances for a binary endpoint, while the stratified logrank test does the same for a time-to-event endpoint.
- Linear regression (univariate analysis) and multiple regression (multivariate analysis) investigate the dependence of an outcome on baseline factors.
- Analysis of covariance (ANCOVA) is an extension of ANOVA that can adjust for continuous covariates and for several covariates simultaneously when comparing treatments.
- ANCOVA can also be used to investigate treatment x covariate interactions.
- These statistical models can be incorporated into a single modeling framework that models the mean for continuous and score endpoints. Patient-to-patient variation is assumed to follow a normal distribution.
- The logistic model provides a corresponding framework for binary endpoints that models the odds for the event.
- The Cox model provides a corresponding framework for time-to-event endpoints that models the hazard rate.

7 Graphical methods

Graphical methods are becoming increasingly important in the way that efficacy and safety data are presented and evaluated. This is especially true in the area of oncology, where Kaplan–Meier curves are used to display the distribution of the key time-to-event endpoints and forest plots are the basis for assessing homogeneity of treatment effect. More recently, waterfall plots and swimmer plots have been introduced to display patient-level data on tumor shrinkage and duration of response. The interpretation of safety data has traditionally been centered around tables and listings associated with adverse events, laboratory measurements and vital signs data. Plots showing relative risks for adverse events data can aid interpretation, while box and whisker plots can be used to display summary statistics over time for laboratory measurements and vital signs. Trellis plots are used in place of, or at least in support of, shift tables, showing the movement of laboratory parameters in terms of multiples of the normal range.

In this chapter, these techniques are explained and illustrated through relevant examples.

Kaplan–Meier curves

Kaplan–Meier curves are discussed in Chapter 2 in conjunction with the analysis of time-to-event endpoints (see Figures 2.2 and 2.3). In this chapter, Figure 7.1 shows two sets of curves in the open-label ToGA trial, which compared treatment with trastuzumab and chemotherapy with chemotherapy alone in patients with human epidermal growth factor receptor 2 (*HER2*)-positive advanced gastric or gastroesophageal junction cancer.[1] The upper set of curves relate to overall survival (OS) while the lower set relate to progression-free survival (PFS).

Kaplan–Meier curves are conventionally drawn in this way for clinical trials in oncology. The *x*-axis represents time (t) from randomization and the estimated probability of being event free at time t is plotted on the *y*-axis.

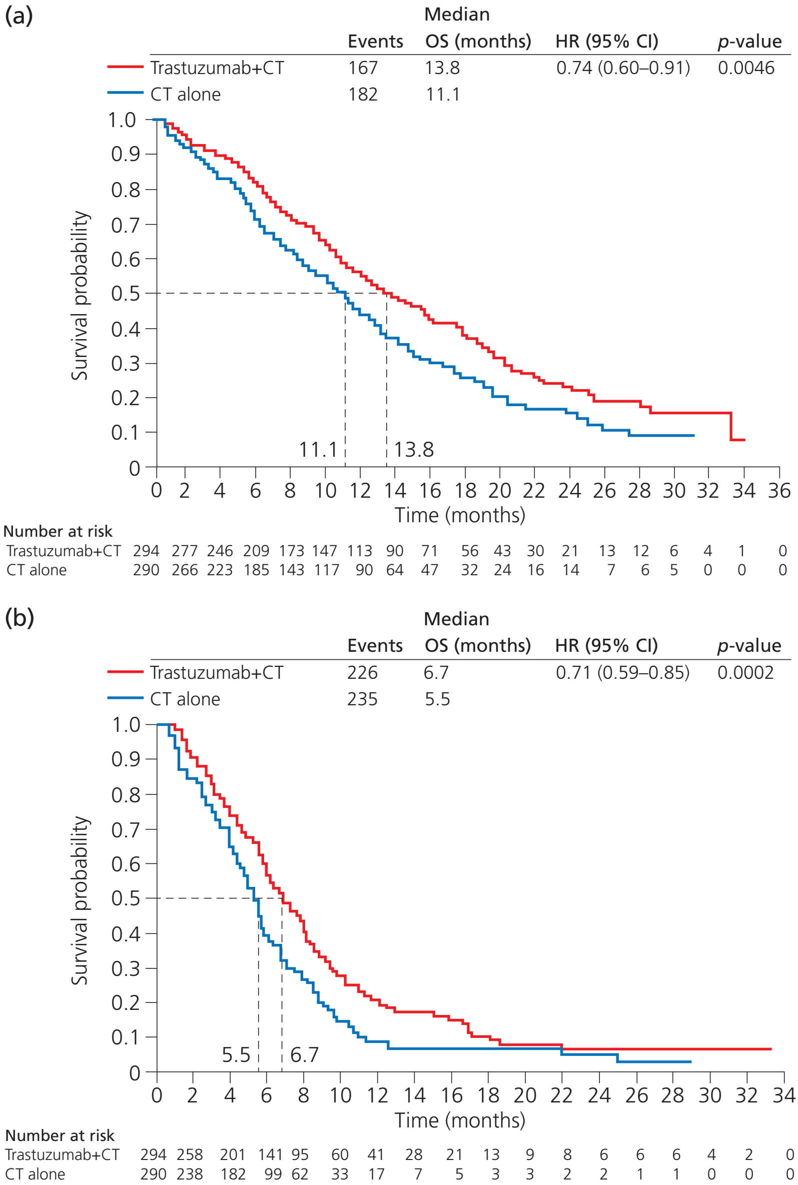


Figure 7.1 Kaplan–Meier curves and medians for (a) overall survival (OS) and (b) progression-free survival (PFS) in patients receiving trastuzumab and chemotherapy (CT) versus CT. CI, confidence interval; HR, hazard ratio. Reproduced with permission from Bang et al. 2010.[1]

In Figures 7.1a and b, the curve for the group that received trastuzumab and chemotherapy is consistently above the corresponding curve for the group that received chemotherapy alone, showing a clear advantage in terms of both OS and PFS. The medians for both OS and PFS, for each of the treatment groups, are clearly identifiable by drawing horizontal lines at probability = 0.5 on the *y*-axis.

When event rates are low, plotting Kaplan–Meier curves in the conventional way can be problematic since the probability of being event free will stay close to 1 through time and there will be little definition regarding possible separation of the curves. Instead, the *y*-axis can be broken at *y* = 0 with the axis starting at a suitable point above zero. Alternatively – to avoid breaking the *y*-axis – the curves can be plotted with the *y*-axis representing the probability of the event occurring by time t rather than the probability of being event free at time t. The Kaplan–Meier curve plotted in this way is simply the inverse of the curve plotted in the conventional way, where:

$$\begin{array}{ccc}\textit{probability of the event} & = & \textit{1 – probability of being}\\ \textit{occurring by time t} & & \textit{event free at time t}\end{array}$$

Cumulative incidence curves

In many settings, a researcher will be interested in a composite event, such as progression or death, and the associated composite endpoint of time to death or progression (PFS). In some circumstances, there may be interest in separating out the components of the composite endpoint. In this regard, it is convenient to think in terms of curves plotted in reverse (see Kaplan–Meier curves above) in order to calculate the probability of the event occurring by time t.

Consider the events ‘death without progression’ and ‘progression’. Cumulative incidence functions allow the components of the composite to be separated by estimating two separate curves:

probability of dying without progression by time t
and
probability of progression by time t

These two cumulative incidence functions can be estimated from data and plotted on the same graph, and it is easy to show the overall reverse Kaplan–Meier probability at time t for the composite PFS as it is the sum of these two cumulative incidence function curves. The cumulative incidence functions for each of the two components can be plotted on separate graphs (as in Figure 7.2), or on the same graph for comparison purposes.

Gray's test is an alternative to the logrank test for comparing treatments in the presence of composite events and is the test of choice in this setting. The hazard ratio (HR) and logrank test *p*-value quoted in Figure 7.2 relate to an alternative analysis based on cause-specific hazard functions; see Dignam and Kocherginsky (2008) for further discussion of the analysis of competing risks data and Gray's test.[2]

Forest plots

Forest plots are routinely used in two distinct settings.

- To display results in subgroups within a trial, enabling an assessment of homogeneity of treatment effect.
- To present results across trials in a meta-analysis.

Judging the homogeneity of effect is discussed in Chapter 4 (see Figure 4.1). In this chapter, Figure 7.3 displays data from the ToGA trial, comparing the effect of trastuzumab and chemotherapy with chemotherapy alone in terms of OS. The overall HR with its 95% confidence interval (CI) is shown, along with the HRs and 95% CIs for each subgroup.

Virtually all the subgroup CIs substantially overlap the overall CI, with one exception – the subgroup of patients presenting with non-measurable disease. Given the small sample size, care must be taken in concluding a differential treatment effect in that subgroup. Nonetheless, this exploratory finding may raise some concerns and, moving forward, it would be important to remain aware of this issue.

As mentioned in Chapter 3, the power for a time-to-event endpoint is driven by the number of patients with events and not directly by the number of patients. It would be of value, therefore, within the forest plot (e.g. in the column headed 'Number of patients' in Figure 7.3), to give not only the number of patients in each subgroup but also the number of patients with events. This would help in understanding how much power is available for specific subgroup evaluations.

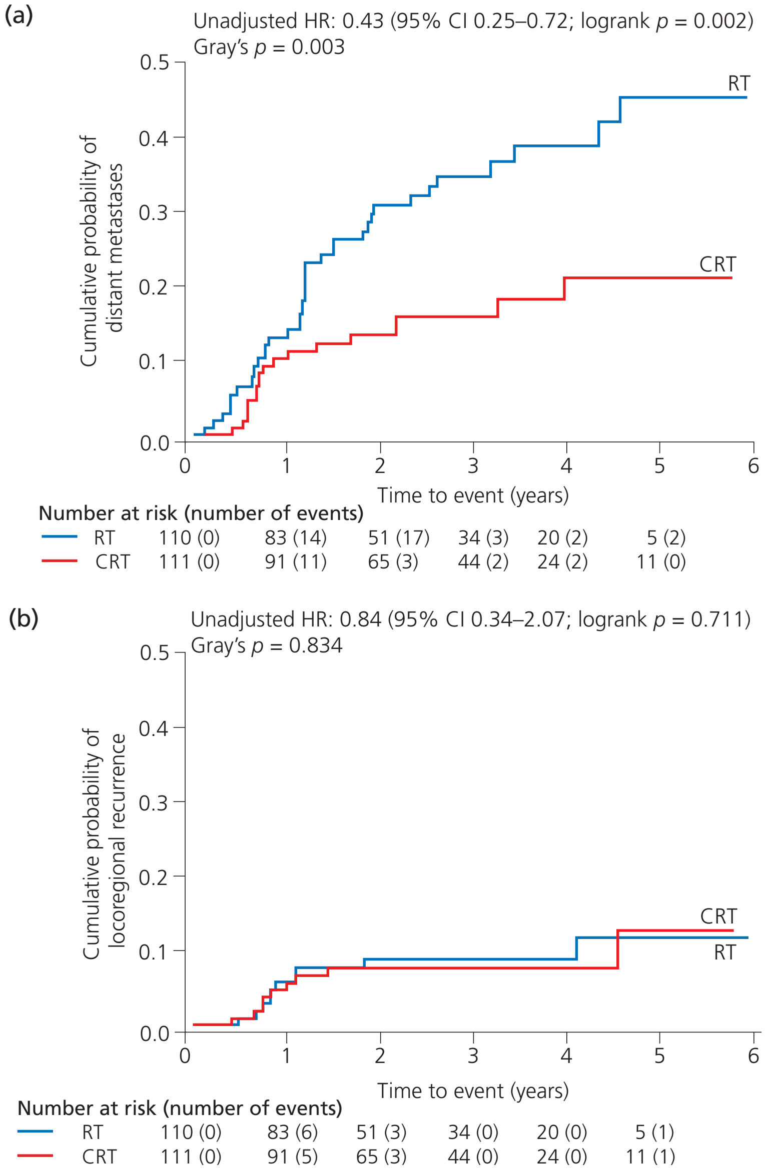
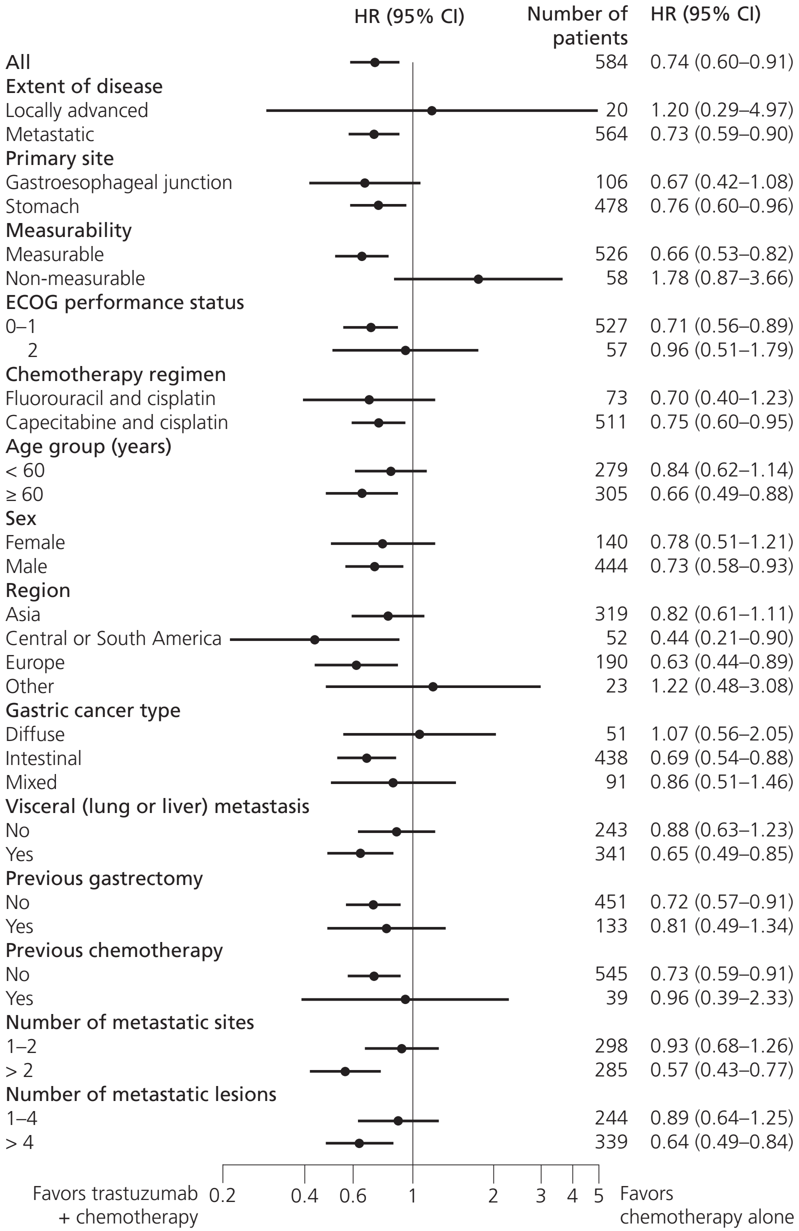


Figure 7.2 Cumulative incidence of relapse in patients treated with radiotherapy (RT) versus chemotherapy and radiotherapy (CRT) where relapse is defined as (a) distant metastasis and (b) locoregional recurrence. HR, hazard ratio; CI, confidence interval. Reproduced from Tai et al. 2011,[3] licensed under CC BY 2.0.



	Number of patients	HR (95% CI)
All	584	0.74 (0.60–0.91)
Extent of disease		
Locally advanced	20	1.20 (0.29–4.97)
Metastatic	564	0.73 (0.59–0.90)
Primary site		
Gastroesophageal junction	106	0.67 (0.42–1.08)
Stomach	478	0.76 (0.60–0.96)
Measurability		
Measurable	526	0.66 (0.53–0.82)
Non-measurable	58	1.78 (0.87–3.66)
ECOG performance status		
0–1	527	0.71 (0.56–0.89)
2	57	0.96 (0.51–1.79)
Chemotherapy regimen		
Fluorouracil and cisplatin	73	0.70 (0.40–1.23)
Capecitabine and cisplatin	511	0.75 (0.60–0.95)
Age group (years)		
< 60	279	0.84 (0.62–1.14)
≥ 60	305	0.66 (0.49–0.88)
Sex		
Female	140	0.78 (0.51–1.21)
Male	444	0.73 (0.58–0.93)
Region		
Asia	319	0.82 (0.61–1.11)
Central or South America	52	0.44 (0.21–0.90)
Europe	190	0.63 (0.44–0.89)
Other	23	1.22 (0.48–3.08)
Gastric cancer type		
Diffuse	51	1.07 (0.56–2.05)
Intestinal	438	0.69 (0.54–0.88)
Mixed	91	0.86 (0.51–1.46)
Visceral (lung or liver) metastasis		
No	243	0.88 (0.63–1.23)
Yes	341	0.65 (0.49–0.85)
Previous gastrectomy		
No	451	0.72 (0.57–0.91)
Yes	133	0.81 (0.49–1.34)
Previous chemotherapy		
No	545	0.73 (0.59–0.91)
Yes	39	0.96 (0.39–2.33)
Number of metastatic sites		
1–2	298	0.93 (0.68–1.26)
> 2	285	0.57 (0.43–0.77)
Number of metastatic lesions		
1–4	244	0.89 (0.64–1.25)
> 4	339	0.64 (0.49–0.84)

Figure 7.3 Hazard ratio (HR) and 95% confidence interval (CI) for overall survival in prespecified subgroups in the ToGA trial. ECOG, Eastern Cooperative Oncology Group (score). Reproduced with permission from Bang et al. 2010.[1]

Another useful feature for the forest plot would be the calculation and inclusion of *p*-values for treatment × covariate interactions. The potential for a differential treatment effect across the subgroups of patients with measurable and non-measurable disease is raised on page 90. A treatment × covariate *p*-value comparing the two HRs (0.66 in the measurable subgroup and 1.78 in the non-measurable subgroup) would help to assess whether there is any statistical evidence for such a differential effect.

An example of a forest plot for presentation of trial data in a meta-analysis is shown in Figure 7.4. Note that lower limit and upper limit refer to the lower and upper limits of the 95% CIs.

	Statistics for each study				Events/total	
	Risk ratio	Lower limit	Upper limit	*p*-value	Bev	Control
Allegra 2008	0.996	0.575	1.725	0.989	25/1326	25/1321
Escudier 2007	3.608	2.458	5.298	0.000	112/337	28/304
Giantonio 2007	8.397	1.140	70.085	0.037	9/287	1/285
Herbst 2007	3.231	0.943	11.073	0.062	9/39	3/42
Hurwitz 2004	1.212	0.530	2.773	0.649	12/393	10/397
Johnson 2004	5.091	1.999	12.963	0.001	42/66	4/32
Kabbinavar 2003	5.224	2.034	13.414	0.001	40/67	4/35
Kabbinavar 2005	1.733	0.425	7.063	0.443	5/100	3/104
Karrison 2007	2.634	1.567	4.427	0.000	33/53	13/55
Kindler 2007	1.519	0.503	4.584	0.458	8/277	5/263
Miles 2008	1.406	0.286	6.916	0.675	6/497	2/233
Miller 2005	2.582	1.682	3.963	0.000	66/229	24/215
Miller 2007	4.740	0.228	98.390	0.315	2/365	0/346
Price 2008	4.155	2.237	7.719	0.000	46/157	11/156
Reck 2009	1.791	1.150	2.787	0.010	83/659	23/327
Rini 2008	4.768	0.560	40.605	0.153	5/366	1/349
Saltz 2008	1.581	0.659	3.789	0.305	13/694	8/675
Sandler 2006	6.526	1.945	21.893	0.002	19/427	3/440
Van Cutsem 2008	1.794	1.399	2.302	0.000	124/296	67/287
Yang 2003	10.000	1.389	72.001	0.022	19/76	1/40
Random	2.479	1.934	3.177	0.000	678/6711	236/5906

Test for heterogeneity: $Q = 40.617$, $p = 0.003$, $I^2 = 53.222$

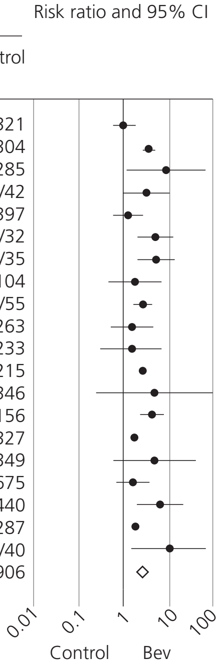


Figure 7.4 Forest plot summarizing meta-analysis data – risk ratios and 95% confidence intervals (CIs) for hemorrhage in patients receiving bevacizumab (Bev), 2.5 mg/kg/week – across 20 trials. Reproduced from Hapani et al. 2010.[4]

Funnel plots

Data for a meta-analysis come from the summary statistics in publications identified as being relevant for a treatment comparison of interest.

One potential source of bias is publication bias, as a study that gives a statistically significant result in favor of the experimental treatment is more likely to be reported, and accepted for publication, than a corresponding study that gives a non-significant result. If the focus is only on published studies, this could result in a biased view of the experimental treatment effect.

The funnel plot helps to detect the presence of publication bias by plotting the treatment effect in each study (e.g. the HR or relative risk [RR]/risk ratio) on the *x*-axis against some function of the precision of the study (e.g. the number of events observed or the standard error of the log HR) on the *y*-axis. Smaller studies will have low precision and more variable results in terms of the observed treatment effect, and these will appear toward the foot of the plot, while larger studies with greater precision will give more consistent results and will appear toward the top of the plot.

The plot with all the studies included should look like an inverted funnel with the narrow part of the funnel at the top and the wide part at the bottom. If, however, there is publication bias then the non-significant studies (very often those with smaller sample sizes) will be under-represented and the lower right-hand part of the plot with HRs of around 1 or above, will be missing (Figure 7.5).

In the data from the meta-analysis depicted in Figure 7.5, the RR was 0.58 (log RR = –0.54); this value is depicted by the vertical line in the funnel plot. If there was no publication bias, there would be a balance of RRs to the left and right of this vertical line from top to bottom of the plot. In this example, there appears to be an absence of both small- and medium-sized studies to the right of the vertical line. Such studies would be ‘negative’, with RRs around and above 1 (log risk ratio around or above 0). This indicates the presence of publication bias.

Visually inspecting the funnel plot in this way is somewhat informal and the technique can be supplemented by a more formal statistical evaluation. See Murad et al. for a discussion of these methods.[5]

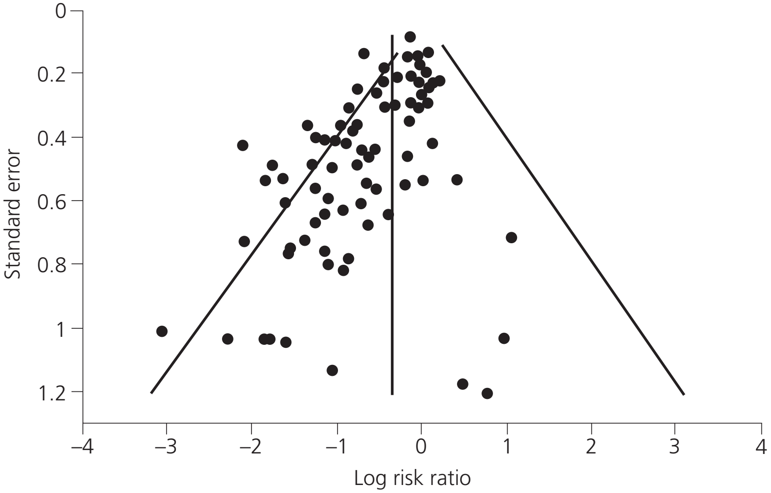


Figure 7.5 Log of the risk ratios plotted against the standard error of the risk ratio of each study to identify asymmetry in the distribution of trials. Gaps in the funnel plot indicate potential publication bias. Reproduced from Ritchie and Romanuk 2012,[6] licensed under CC BY 2.0.

Waterfall plots

Waterfall plots can be used to display the change in tumor burden for each individual patient following treatment. The *y*-axis usually represents the maximum percentage change from baseline in an appropriate measure of tumor burden; for example, the sum of the longest diameters of the target lesions. Positive values correspond to patients whose tumor burden has increased, while negative values are recorded for patients whose tumor burden has reduced.

Patients are usually ordered on the plot: patients with the largest percentage increase at the beginning of the *x*-axis through to those with the largest percentage decrease at the end of the *x*-axis (Figure 7.6).

It is common for individual patients to be color coded according to the response evaluation criteria in solid tumors (RECIST): complete response, partial response, stable disease, progressive disease. Displaying waterfall plots for each treatment group, one on top of the other, or color coded on the same plot, can be a useful aid when comparing treatments.

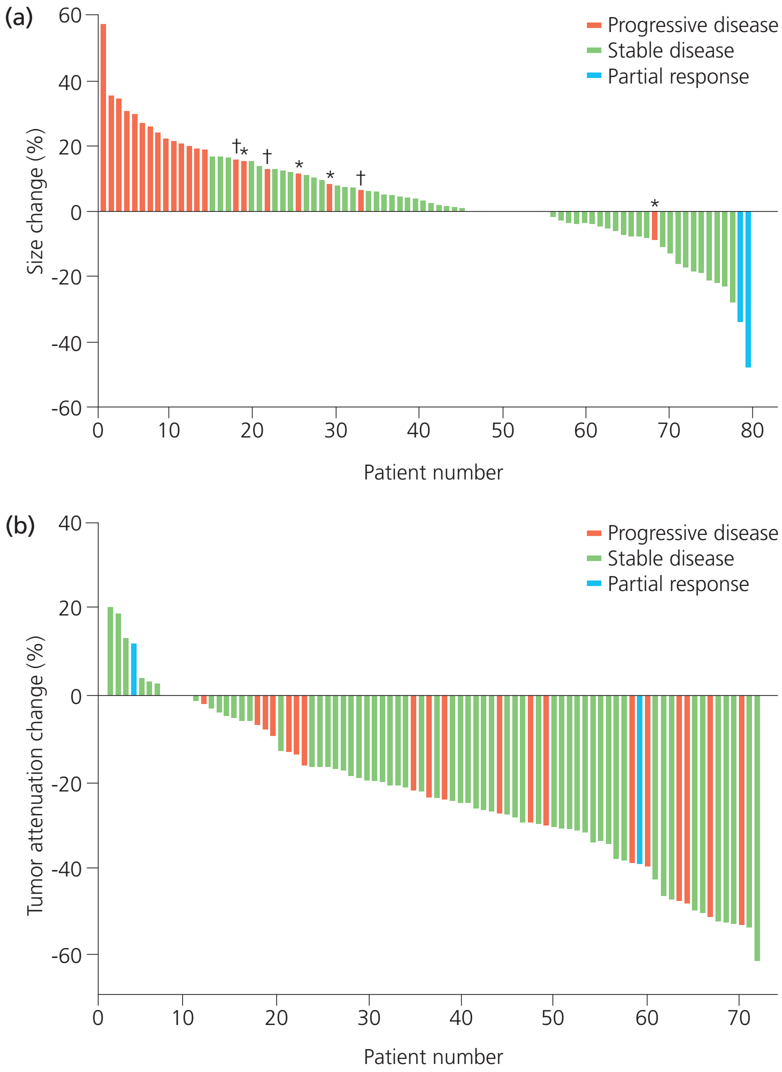


Figure 7.6 Waterfall plots showing changes in (a) tumor size by measurement of change in size of target lesions from baseline, and (b) tumor attenuation by measurement of size changes of target lesions from baseline after 2 cycles of treatment. *Progression due to appearance of new lesion. †Progression of non-target lesion. Reproduced from Lim et al. 2015,[7] licensed under CC BY 4.0.

Swimmer plots

Swimmer plots also focus on individual patient data and display key events during treatment and follow-up, such as start of response, end of response, end of treatment, progression and death (Figure 7.7).

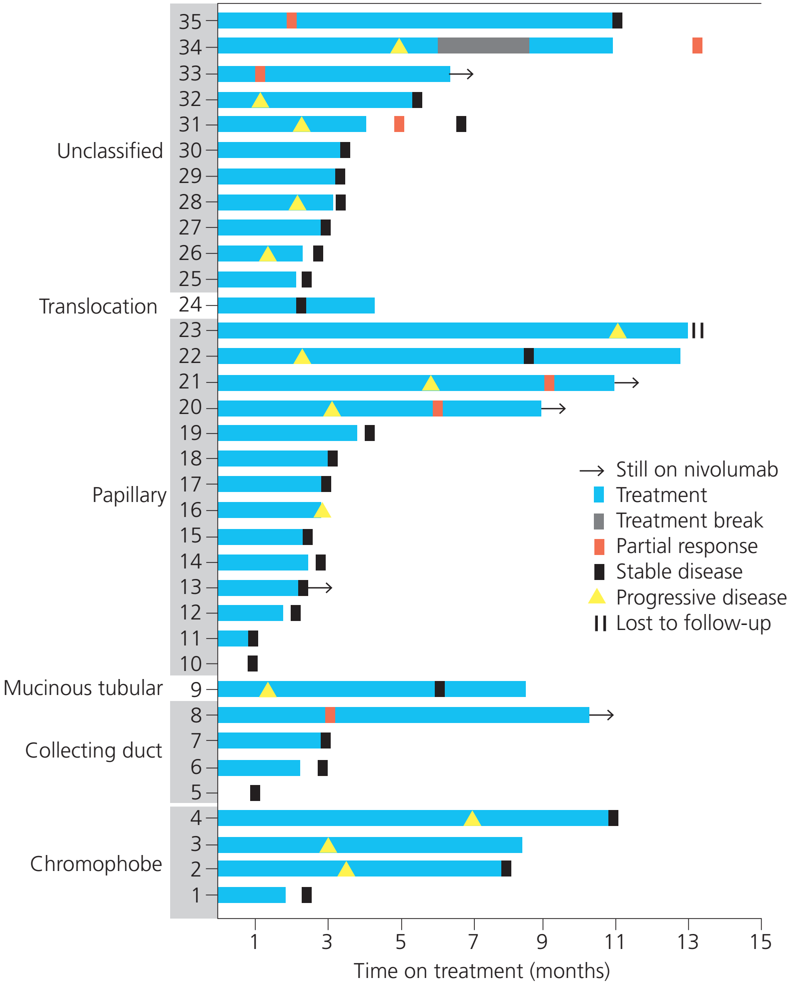


Figure 7.7 Swimmer plot of time on treatment with nivolumab for 35 patients with renal cell carcinoma. The plot shows key events during the time course of treatment. Patients 5 and 10 received a single dose of nivolumab. Patients 31 and 34 had a partial response after discontinuing treatment because of adverse events. Reproduced from Koshkin et al. 2018,[8] licensed under CC BY 4.0.

The length of the bar in a swimmer plot is often the duration of follow-up, with different symbols representing the different events. Patients can be ordered according to the length of follow-up, so that the longest survivors appear at the top of the plot. Bars can also be color coded to differentiate treatment groups or tumor types, or separated appropriately as with waterfall plots (see Figure 7.7).

Box and whisker plots

Box and whisker plots are good ways to represent aspects of the distribution of an outcome. The 'box' in a box and whisker plot displays the median (see Chapter 1), the upper quartile (i.e. the value in the data that cuts off the largest 25% of data values) and the lower quartile (i.e. the value in the data that cuts off the smallest 25% of data values) (Figure 7.8).

The 'whiskers' are obtained by calculating the interquartile range (IQR) (i.e. the numeric difference between the upper and the lower quartiles) and then marking:

- the largest value ≤ the median + 1.5 × IQR (upper whisker), and
- the smallest value ≥ the median – 1.5 × IQR (lower whisker).

Finally, all data values outside of the lower and upper whiskers are marked on the plot and considered as outliers.

Box and whisker plots are a useful way of displaying data over time or across visits. They can be used to compare an outcome across different tumor types or to highlight differences between treatments: for example, for a particular laboratory parameter (Figure 7.9).

The horizontal line drawn within each box is the median value. For a laboratory parameter or vital sign measured over time, these medians can be joined across time within each treatment group to help pick up trends. Trends are much easier to identify in a plot than in a table, especially when comparing treatment groups.

Note that displaying individual values outside of the median ± 1.5 × IQR enables outlying values and patients with values exceeding the standard thresholds to be identified.

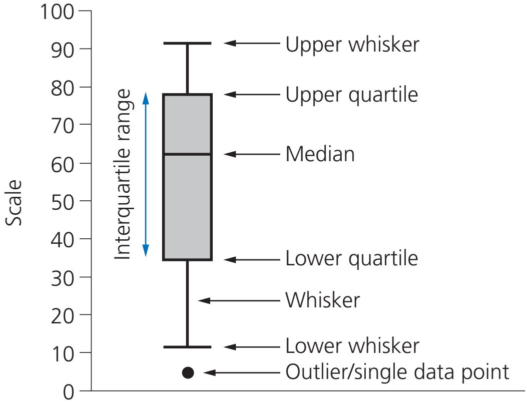


Figure 7.8 Box and whisker plot definitions.

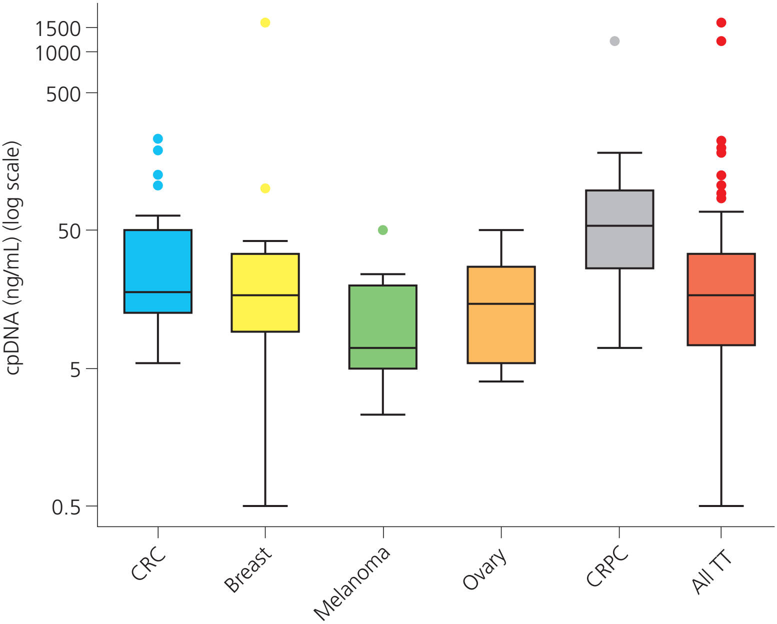


Figure 7.9 DNA concentrations (ng/mL) classified by tumor types. Box and whisker plots showing 25th, 50th and 75th percentiles, upper and lower adjacent values (whiskers) and outliers. For example, the median concentration of circulating plasma DNA (cpDNA) for patients with colorectal cancer (CRC) is 18 ng/mL (range 5–230). All TT, all tumor types; CRPC, castration-resistant prostate cancer. Adapted from Perkins et al. 2012,[9] licensed under CC BY 4.0.

Relative risks for adverse events

The purpose here is to present key adverse event data in a single plot, which highlights where treatment differences lie (Figure 7.10).

The events displayed in Figure 7.10 are the most frequent adverse events in two treatment groups and are potentially those of greatest concern. The circles and triangles on the left-hand side are the percentage incidence rates by treatment group, while the right-hand side shows the RR for each adverse event with 95% CIs. The events are ordered according to the value of the RR, so the adverse events with an increased risk are immediately apparent.

In some situations, it may be preferable to plot the risk difference rather than the RR; for example, if there are zero adverse events of a certain type in the control group. With a zero observed risk in the control group, the RR is not defined. Ordering according to RR or risk difference is likely to be of most interest but ordering by absolute risk in the experimental arm may also be appropriate.

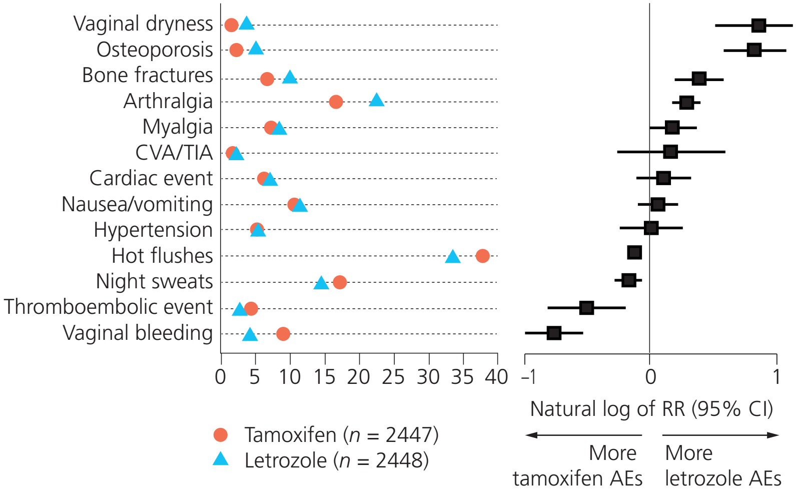


Figure 7.10 Relative risk (RR) of an adverse event (AE), grades 1–5, in 4895 patients receiving either tamoxifen or letrozole monotherapy. CVA, cardiovascular attack; TIA, transient ischemic attack. Adapted from Regan et al. 2011.[10]

Shift and trellis plots for laboratory parameters

Shift tables are often presented for laboratory parameters to show movement above and below the normal range over the course of treatment. A shift plot displaying such data in a graphical way can make it easier to interpret the data, especially when data from both treatment groups are displayed in the same plot. In Figure 7.11, each patient can be represented by a single point, with the treatment groups identified using different colors/symbols in the plot.

If there were separate normal ranges for males and females, for example, then the shift plot would need to be produced by sex. If there were different normal ranges for different centers in the study, then the plot would need to be modified by standardizing the measurements across centers. See Kay (2014) for further discussion of how this could be done.[11]

The shift plot in Figure 7.11 has the laboratory value at the end of treatment on the *y*-axis. Alternatively, it may be the maximum value on treatment that is the most critical value and the *y*-axis can be defined in this way.

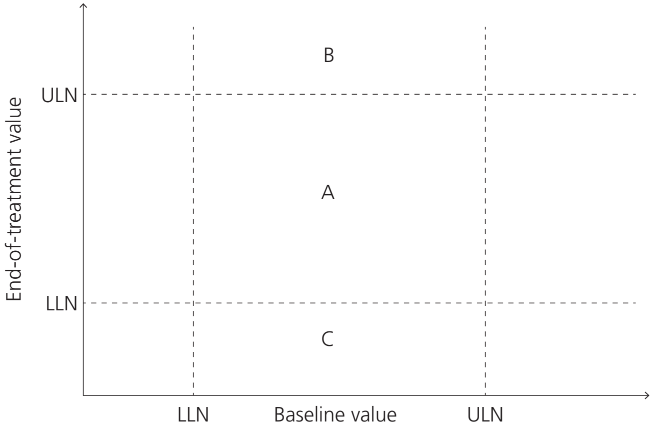


Figure 7.11 Shift plot. A point within region A would correspond to a patient with a value that starts within the normal range and remains in the normal range at the end of treatment. A point in region B would correspond to a patient with a value that starts within the normal range but has a value above the upper limit of normal (ULN) at the end of treatment. A point in region C would be for a patient with a value in the normal range at baseline but below the lower limit of normal (LLN) at the end of treatment.

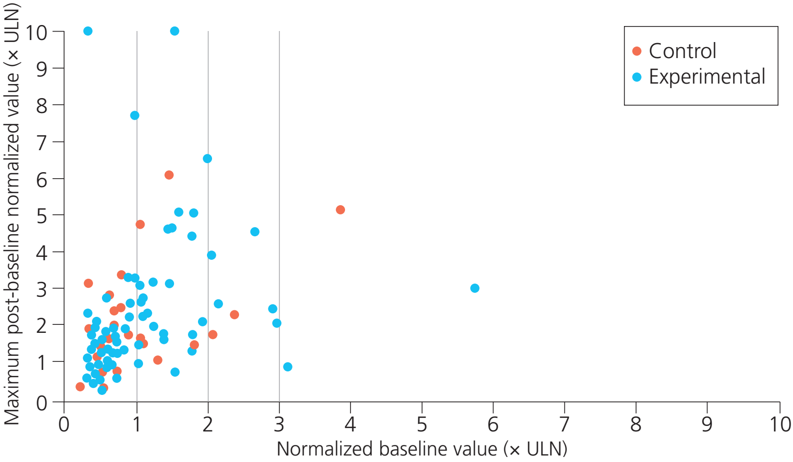


Figure 7.12 Trellis plot of preliminary data for alanine aminotransferase (ALT) levels in patients with pancreatic cancer. The *x*-axis is the normalized value at baseline while the *y*-axis represents the maximum normalized value on treatment. Note that randomization of patients to the experimental and control groups was in a 2:1 ratio. Unpublished data.

Trellis plots are an extension of this idea but with regions of the plot defined by relevant multiples of the normal range (Figure 7.12).

In the trellis plot shown in Figure 7.12, there is some evidence of greater elevations of alanine aminotransferase (ALT) in the experimental group than in the control group: 4 patients who start with values below 2 × the upper limit of normal (ULN) have maximum values on treatment above 6 × the upper limit of normal. Corresponding increases in the control group occur but to a lesser degree.

Network geometry plots (network meta-analysis)

Network meta-analysis (NMA) is an extension of the standard meta-analysis, incorporating indirect evidence for treatment comparisons. Such analyses are frequently used as part of a systematic review. A meta-analysis combines the results from a series of trials comparing two treatments, A and B. The NMA supplements the A versus B evidence with trials that compare A versus C and B versus C, for example. For a critical review of NMAs, see Li et al.[12]

A network geometry plot provides a summary of the studies that make up the network. The area of the circle for each so-called node in the network corresponds to the number of patients who were randomly assigned to receive that treatment across all of the studies. The lines connecting the treatments indicate the available head-to-head randomized comparisons, while the thickness of the lines corresponds to the number of trials (Figure 7.13). These lines can also be annotated by the actual number of head-to-head trials.

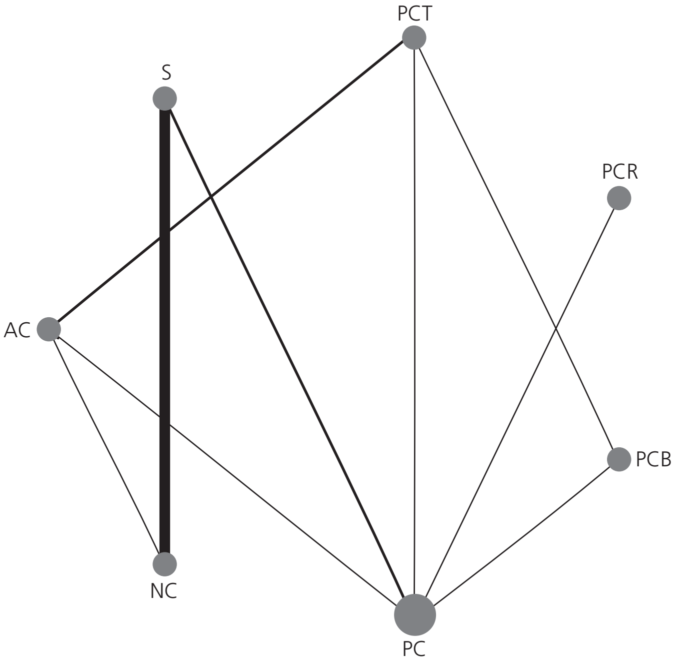


Figure 7.13 Network geometry plot showing all treatments in the network meta-analysis. The size of each node corresponds to the number of patients randomized to the given treatment. The lines connect the treatment regimens that were directly compared in head-to-head randomized controlled trials (RCTs). The thickness of the lines corresponds to the number of RCTs. AC, adjuvant chemotherapy; NC, neoadjuvant chemotherapy; PC, perioperative chemotherapy without a taxane; PCB, perioperative chemotherapy plus bevacizumab; PCR, perioperative chemotherapy with adjuvant chemoradio-therapy; PCT, taxane-based perioperative chemotherapy; S, surgery only. Reproduced from van den Ende et al. 2019,[13] licensed under CC BY 4.0.

Key points – graphical methods

- Kaplan–Meier curves are used to display data on a time-to-event endpoint.
- Cumulative incidence functions are used with a composite endpoint when there is interest in separating out the different components of the composite.
- Forest plots display results for subgroups in a clinical trial, or across a series of trials in a meta-analysis.
- The change in tumor burden from baseline over a treatment period or through to the end of follow-up at the individual patient level can be displayed in a waterfall plot. Swimmer plots also identify key events during treatment at the individual patient level.
- Box and whisker plots look at the distribution of data and can be used for laboratory parameters, vital signs and QTc intervals over time or across different tumor types.
- Relative risks for adverse events can be displayed for the most frequent adverse events to highlight where treatment differences lie.
- Concerning increases in laboratory parameters from baseline can be identified by plotting, relative to the normal range, the baseline value on the *x*-axis and the maximum value over time on the *y*-axis in trellis plots.

References

1. Bang Y-J, Van Cutsem E, Feyereislova A et al. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. *Lancet* 2010;376:687–97.

2. Dignam JJ, Kocherginsky MN. Choice and interpretation of statistical tests used when competing risks are present. *J Clin Oncol* 2008;26:4027–34.

3. Tai BC, Wee J, Machin D. Analysis and design of randomised clinical trials involving competing risks endpoints. *Trials* 2011;12:127.

4. Hapani S, Sher A, Chu D, Wu S. Increased risk of serious hemorrhage with bevacizumab in cancer patients: a meta-analysis. *Oncology* 2010;79:27–38.

5. Murad MH, Chu H, Wang Z. The effect of publication bias magnitude and direction on the certainty in evidence. *BMJ Evid Based Med* 2018;23:84–6.

6. Ritchie ML, Romanuk TN. A meta-analysis of probiotic efficacy for gastrointestinal diseases. *PLoS One* 2012;7:e34938.

7. Lim Y, Han SW, Yoon JH et al. Clinical implication of anti-angiogenic effect of regorafenib in metastatic colorectal cancer. *PLoS One* 2015;10:e0145004.

8. Koshkin VS, Barata PC, Zhang T et al. Clinical activity of nivolumab in patients with non-clear cell renal cell carcinoma. *J Immunother Cancer* 2018;6:9.

9. Perkins G, Yap TA, Pope L et al. Multi-purpose utility of circulating plasma DNA testing in patients with advanced cancer. *PLoS One* 2012;7:e47020.

10. Regan MM, Price KN, Giobbie-Hurder A et al. Interpreting breast international group (BIG) 1-98: a randomized, double-blind, phase III trial comparing letrozole and tamoxifen as adjuvant endocrine therapy for postmenopausal women with hormone receptor-positive, early breast cancer. *Breast Cancer Res* 2011;13:209.

11. Kay R, ed. Section 19.2.3. Laboratory data. In: *Statistical Thinking for Non-Statisticians in Drug Regulation*, 2nd edn. Wiley Blackwell, 2014:284.

12. Li T, Puhan MA, Vedula SS et al. Network meta-analysis-highly attractive but more methodological research is needed. *BMC Med* 2011;9:79.

13. van den Ende T, ter Veer E, Machiels M et al. The efficacy and safety of (neo)adjuvant therapy for gastric cancer: a network meta-analysis. *Cancers* 2019;11:80.

FastTest

You've read the book ... now test yourself with key questions from the authors

- Go to the FastTest for this title ***FREE*** **at karger.com/fastfacts**
- Approximate time **10 minutes**
- For best retention of the key issues, try taking the FastTest before and after reading

Index

Fill the gap in your knowledge, *fast!*

with ***Fast Facts*** – the ultimate medical handbook series

FAST FACTS Medical Statistics

KARGER

ISBN 978-1-912-77667-2

9 781912 776672

FAST FACTS

KARGER.COM