Data types

Population

The entire number of individuals of which the sample aims to be representative.

Sample

A group taken from the wider population. A sample aims to be representative of the population from which it is taken.

As samples are smaller, they are easier to collect and to analyse statistically. However, as they do not contain all of the values in the population, they can misrepresent it. Statistical analysis is often used to decide whether samples of data come from the same or from different populations. Populations are described by parameters and samples by statistics.

Categorical (qualitative) data

Nominal

Data that have no numerically significant order, such as blood groups.

Ordinal

Data that have an implicit order of magnitude, such as ASA score.

Numerical (quantitative) data

Discrete

Data that have finite values, such as number of children.

Continuous

Data that can take any numerical value including fractional values. Examples include weight or height.

Ratio

Any data series that has zero as its baseline value, for example blood pressure or the Kelvin temperature scale.

Interval

Any data series that includes zero as a point on a larger scale, for example the centigrade temperature scale.

There is a hierarchy of usefulness of data, according to how well it can be statistically manipulated. The accepted order is continuous data > ordinal data > nominal data.

Indices of central tendency and variability

Describing data

Once data have been collected, the values will be distributed around a central point or points. Various terms are used to describe both the measure of central tendency and the spread of data points around it.

Measures of central tendency

Mean

> The average value: the sum of the data values divided by the number of data points. Denoted by the symbol $\bar{x}$ when describing a sample mean and μ when describing a population mean.

The mean is always used when describing the normal distribution and, therefore, it is the most important measure with regards to the examination.

Median

> The middle value of a data series, having 50% of the data points above it and 50% below.

If there are an even number of data points, the median value is assumed to be the average of the middle two values.

Mode

> The most frequently occurring value in a set of data points.

The data can be plotted on a graph to demonstrate the distribution of the values. The individual values are plotted on the x axis with the frequency with which they occur on the y axis.

Measures of spread

Variance

> A measure of the spread of data around a central point. Described by the following equation.

$$\text{Var} = \frac{\Sigma(\bar{x} - x)^2}{n - 1}$$

Standard deviation

A measure of the spread of data around a central point. Described by the following equation (σ for population, SD for sample):

$$\text{SD} = \sqrt{\frac{\Sigma(\bar{x} - x)^2}{n - 1}}$$

Begin by finding the mean value ($\bar{x}$) of the distribution and then subtract each data point from it to find the difference between the values

$$\bar{x} - x$$

Square the results to ensure that all values are positive numbers:

$$(\bar{x} - x)^2$$

Sum the results:

$$\Sigma(\bar{x} - x)^2$$

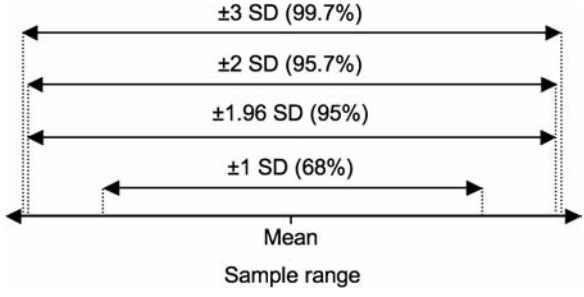
Next divide the result by the number of observations (minus 1 for statistical reasons) to give the mean spread or variance

$$\frac{\Sigma(\bar{x} - x)^2}{n - 1}$$

The units for variance are, therefore, squared, which can cause difficulties. If the observations are measuring time for instance, the variance may be given in seconds squared (s^2), which is meaningless. The square root of the variance is, therefore, used to return to the original units. This is the SD.

$$\text{SD} = \sqrt{\frac{\Sigma(\bar{x} - x)^2}{n - 1}}$$

The spread of data is often described by quoting the percentage of the sample or population that will fall within a certain range. For the normal distribution, 1SD either side of the mean will contain 68% of all data points, 1.96SD 95%, 2SD 95.7% and 3SD 99.7%.



Standard error of the mean

The standard deviation of a group of sample means taken from the same population (SEM):

$$\text{SEM} = \sigma/\sqrt{(n-1)}$$

where σ is the SD of the population and n is the number in the samples.

In practice, the population SD is unlikely to be known and so the sample SD is used instead, giving

$$\text{SEM} = \text{SD}/\sqrt{(n-1)}$$

In the same way as the SD is used as a measure of spread around a mean, the SEM is used as a measure of the spread of a group of sample means around the true population mean. It is used to predict how closely the sample mean reflects the population mean.

As the sample size increases, SEM becomes smaller. For this reason, the SEM is sometimes quoted in study results rather than the SD in order to make the data look better.

Degrees of freedom

Statistics frequently involve calculations of the mean of a sample. In order to be able to calculate a mean, there must be at least two values present. For this reason, when describing sample size, the term $n-1$ is often used instead of the actual number. One of the sample points *must* be present in order that each of the other points can be used in the mean calculation. In other words, the size of the freely chosen sample must always be one less than are actually present.

For large sample sizes, the correction factor makes no difference to the calculation, but for small sample sizes it can be quite important. It is, therefore, best always to describe the sample size in this way.

Confidence intervals

The range of values that will contain the true population mean with a stated percentage confidence. Used in parametric tests.

A 95% confidence interval is ± 1.96SD and is the most frequently quoted. There is a 95% certainty that this range of values around the mean will contain the population mean.

Quartile

Any one of the three values that divide a given range of data into four equal parts.

In order to tear a piece of paper into four equally wide strips, three tears must be made. One to tear the original paper in half and the other two to tear those halves in half again. A quartile is the mathematical equivalent of this to a range of ordered data. You should realize that the middle quartile (Q_2) is, in effect, the median for the range. Similarly, the first quartile (Q_1) is effectively the median of the lower half of the dataset and the third quartile (Q_3) the median of the upper half. In the same way as for the median calculation, a quartile should be represented as the mean of two data points if it lies between them.

Interquartile range

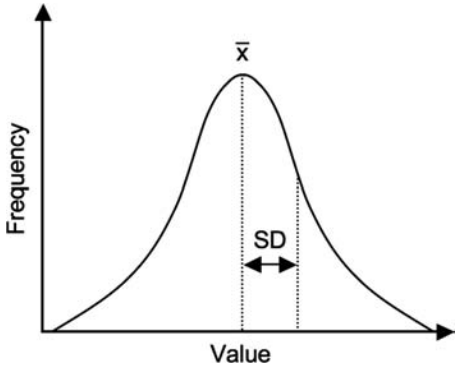
The range of values that lie between the first and third quartiles and, therefore, represent 50% of the data points. Used in non-parametric tests.

Calculating quartiles and using the interquartile range is useful in order to negate the effect of extreme values in a dataset, which tend to create a less stable statistic.

Types of distribution

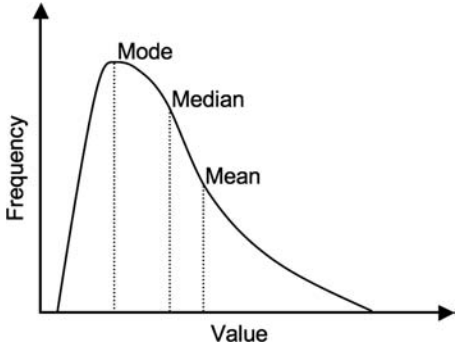
The normal distribution

A bell-shaped distribution in which the mean, median and mode all have the same value, with defined SD distribution as above.



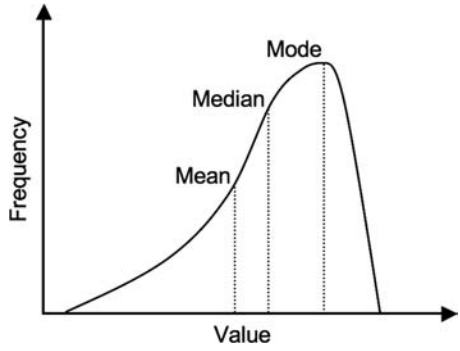
The curve is symmetrical around the mean, which is numerically identical to the median and mode. The SD should be indicated; 1SD lies approximately one third of the way between $\bar{x}$ and the end of the curve.

Positively skewed distribution



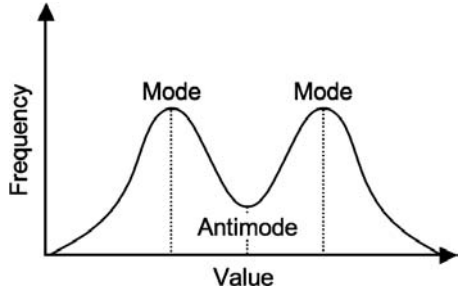
The curve is asymmetrical with a longer tail stretching off towards the more positive values. The mean, median and mode are now separated so that $\bar{x}$ is nearest the tail of the curve; the mode is at the peak frequency and the median is in between the two. This type of distribution can sometimes be made normal by logarithmic transformation of the data.

Negatively skewed distribution



The curve is asymmetrical with a longer tail stretching off towards the more negative values. The mean, median and mode are now separated in the other direction, with $\bar{x}$ remaining closest to the tail. This type of distribution can sometimes be made normal by performing a power transformation (squaring or cubing the data).

Bimodal distribution



The curve need not be symmetrical nor have two modes of exactly the same height but the above curve demonstrates the principle well. The low point between the modes is known as the antimode. This curve could represent the heights of the population, with one mode for men and one for women.

Methods of data analysis

When performing a study, the first step is to pose a question. The question is formulated as a hypothesis that must be proved or disproved. This question is known as the null hypothesis.

The null hypothesis

The hypothesis states that there is no difference between the sample groups; that is, they both are from the same population (H_0).

The study then examines whether this is true. The amount of data needed to prove a difference between the samples depends on the size of the difference that is to be detected. Enough data must be collected to minimize the risk of a false-positive or false-negative result. This is determined by a power calculation.

Power

The ability of a statistical test to reveal a difference of a certain magnitude (%):

$$1 - \beta$$

where β is the β error (type II error).

Acceptable power is 80–90%, which equates to a β value of 10–20%. In effect, this means a 10–20% chance of a false-negative result.

The *p* value

The likelihood of the observed value being a result of chance alone.

Conventionally a p (probability) value of < 0.05 is taken to mean statistical significance. This means that if $p = 0.05$ then the observed difference could occur by chance on 1 in 20 (5%) of occasions. In effect, this means a 5% chance of a false-positive result.

Number needed to treat

The number of patients that have to be treated to prevent one outcome event occurring.

Absolute risk reduction

The numerical difference between the risk of an occurrence in the control and treatment groups.

$$(\text{Incidence in treatment group}) - (\text{Incidence in control group})$$

Relative risk reduction

The ratio of the absolute risk reduction to the control group incidence (%):

$$\frac{(\text{Absolute risk reduction})}{(\text{Control incidence})}$$

Relative risk

The ratio of the risk of an occurrence in the treatment group to that in the control group:

$$\frac{(\text{Incidence in treatment group})}{(\text{Incidence in control group})}$$

If the control incidence is low, this can lead to an overestimation of the treatment effect.

Odds ratio

Ratio of the odds of outcome in the treatment group to the odds of outcome in the control group.

Unpaired test

Different patients are studied in each of the intervention groups.

Paired test

The same patient is studied for each intervention, thereby acting as their own control. Matched patients can also be used.

Student's *t*-test

A parametric test for comparison of sample means where

$$t = \frac{\text{Difference between sample means}}{\text{Estimated SE of the difference}}$$

Once a value for t is obtained, it is read from a table to see if it represents a statistically significant difference at the level of probability required, for example $p < 0.05$.

One-tailed test

A statistical test in which the values that will allow rejection of the null hypothesis are located only at one end of the distribution curve.

For example, if a study were to investigate the potential of a new antihypertensive drug, a one-tailed test may be used to look for a decrease but not an increase in BP.

Two-tailed test

A statistical test in which the values that will allow rejection of the null hypothesis are located at either end of the distribution curve.

A study investigating the effect of a drug on serum Na^+ levels could use a two-tailed test to identify both an increase and a decrease. In general, unless you are sure that a variable can only move in one direction, it is wise to use a two-tailed test.

Chi-square (χ^2) test

Compares the frequency of observed results against the frequency that would be expected if there were no difference between the groups.

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

where χ^2 is the chi-square statistic, E is the number of expected occurrences and O is the number of observed occurrences.

It is best demonstrated by constructing a simple 3×3 table. You may be provided with a pre-printed table in the examination but be prepared to draw your own.

		Smokers		
		Yes	No	Total
Sex	M	30	60	90
	F	70	20	90
	Total	100	80	180

The numbers in the unshaded portion of the table give you the observed frequency. The expected percentage of smokers if there were no difference between the sexes would be 100/180 (55.6%) smokers and 80/180 (44.4%) non-smokers in each group. To find the actual frequency in each group, this percentage is multiplied by the respective row total.

$$E = \frac{\text{Column total}}{\text{Grand total}} \times \text{Row total}$$

		Smokers		
		Yes	No	Total
Sex	M	30 (50)	60 (40)	90
	F	70 (50)	20 (40)	90
	Total	100	80	180

The table now has an expected frequency in parentheses in each cell along with the observed frequency. The calculation $(O - E)^2/E$ is performed for each cell and the results summed to give the χ^2 statistic.

Degrees of freedom for χ^2

Degrees of freedom for a table are calculated in a similar way to those for distributions.

$$\text{DF} = (\text{No. of rows} - 1) \times (\text{No. of columns} - 1)$$

Therefore for a 2×2 table

$$\text{DF} = (2 - 1) \times (2 - 1)$$
$$\text{DF} = 1 \times 1$$
$$\text{DF} = 1$$

When the χ^2 statistic has been calculated, it is cross-referenced to a table of values together with various degrees of freedom. The table will enable the statistician to see if the groups are statistically different or not.

Fisher's exact test

This is a variation of the χ^2 test that is used when the value for E in any cell is 5 or less.

Correlation

A representation of the degree of association between two variables.

Importantly, this does not identify a cause and effect relationship but simply an association.

Correlation coefficient

A numerical description of how closely the points adhere to the best fit straight line on a correlation plot (r).

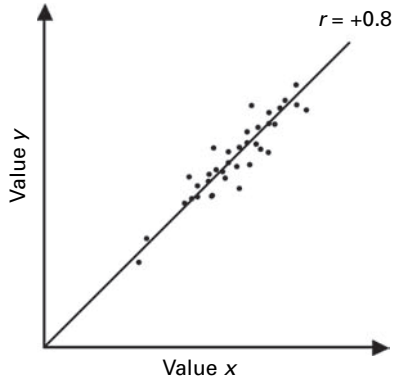
The value of r lies between ± 1. A value of $+1$ indicates a perfect positive correlation and a value of -1 a perfect negative correlation. A value of 0 indicates that there is no correlation between the two variables.

Regression coefficient

A numerical description of the gradient of the line of best fit using linear regression analysis (b).

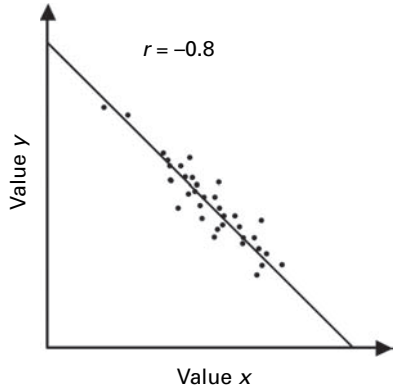
The regression coefficient allows prediction of one value from another. However, it is only useful when the intercept on the y axis is also known, thereby describing the relationship by fixing the position of the line as for the equation $y = bx + a$.

Positive correlation



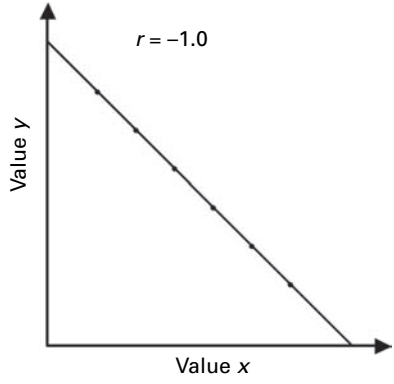
Draw and label the axes. The x axis is traditionally where the independent variable is plotted. Draw a line of best fit surrounded by data points. As the line of best fit has a positive slope, both b and r will be positive. However, r will not be $+1$ as the data points do not lie exactly on the line. In this case r is approximately $+0.8$.

Negative correlation



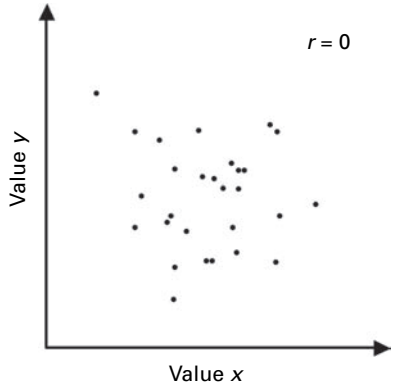
This plot is drawn in exactly the same way but now with a negative slope to the line of best fit. Both b and r will now be negative but, again, r will not be -1 as the data points do not lie exactly on the line. In this case r is approximately -0.8.

Exact negative correlation



This plot is drawn in the same way as the negative plot but now the line of best fit becomes a line of exact fit. Both b and r will now be negative and r will be -1 as the data points lie exactly on the line.

No correlation



Draw and label the axes as before but note that on this plot there is no meaningful line of best fit as the data points are truly random. It is not possible to give a value for *b* as a line of best fit cannot be generated but the value of *r* is 0.

Bland–Altman plot

The Bland–Altman plot is superior to regression/correlation analysis when used to compare two methods of measurement. It is the method of choice when comparing one method to an agreed gold standard.

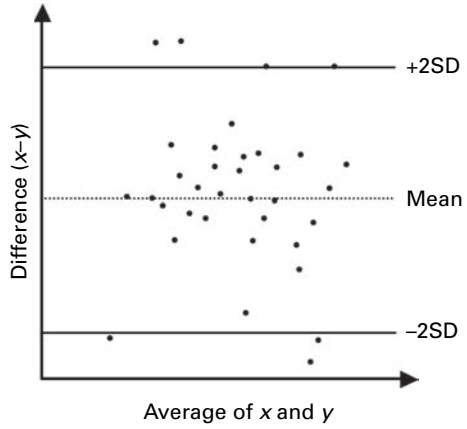
The true value being measured by the two methods is assumed to be the average of their readings. This is then plotted against the difference between the two readings at that point. The level of agreement or disagreement at every value is, therefore, obtained and a mean and SD can be calculated.

Bias

> The extent to which one method varies with respect to another when the two methods are compared.

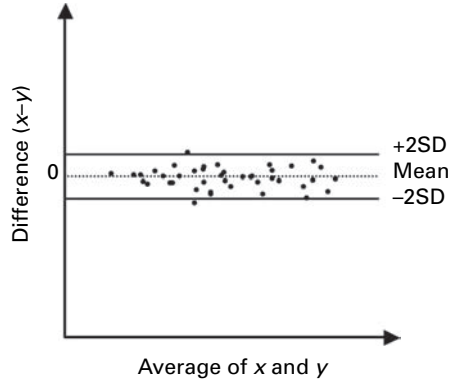
The mean difference between methods should ideally be zero. However, if it is felt that the clinical difference between the methods is not significant, then the mean difference can simply be added to or subtracted from the results of one method in order to bring them into line with the gold standard. The amount by which the mean differs from zero is called the bias.

No agreement



Draw and label the axes as shown. Widely scattered data points as shown suggest no firm comparison between methods *x* and *y*. Demonstrate that ±2SD (95% CI) is wide and the distribution of the points appears arbitrary. Bias can be demonstrated by showing a mean point that does not lie at zero on the *y* axis.

Good agreement



On the same axes draw a tightly packed group of data points centred around a mean difference of zero. The ±2SD should show a narrow range. This plot demonstrates good agreement between the methods used.

Interpretation

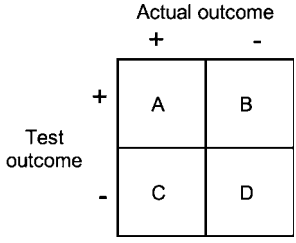
The test does not indicate which method is superior, only the level of agreement between them. It is entirely possible that a method which shows no agreement with a current standard is, in fact, superior to it, although other tests would have to be used to determine its suitability.

Reference table of statistical tests

Type of data	Two groups		More than two groups	
	Unpaired	Paired	Unpaired	Paired
Parametric				
Continuous	Student's unpaired *t*-test	Student's paired *t*-test	ANOVA	Paired ANOVA
Non-parametric				
Nominal	χ^2 with Yates' correction	McNemar's test	χ^2	–
Ordinal or numerical	Mann–Whitney *U* test	Wilcoxon signed rank test	Kruskal–Wallis	Friedman

Error and outcome prediction

In medicine, we often try to predict an outcome based on the result of a test. There are various terms used to describe how useful a test is, which may be best understood by reference to a table such as the one below.



Type I error

The occurrence of a positive test result when the actual value is negative (%).

This type of error equates to box B and is variously described as a type I error, a false-positive error or the α error. A type I error in a study result would lead to the incorrect rejection of the null hypothesis.

Type II error

The occurrence of a negative test result when the actual value is positive (%).

This type of error equates to box C and is variously described as a type II error, a false-negative error or the β error. A type II error in a study result would lead to the incorrect acceptance of the null hypothesis.

Sensitivity

The ability of a test to correctly identify a positive outcome where one exists (%):

$$\frac{\text{The number correctly identified as positive}}{\text{Total number that are actually positive}}$$

or, in the Figure:

$$\mathrm{A}/(\mathrm{A}+\mathrm{C})$$

Specificity

The ability of a test to correctly identify a negative outcome where one exists (%):

$$\frac{\text{The number correctly identified as negative}}{\text{Total number that are actually negative}}$$

or

$$D/(B+D)$$

Positive predictive value

The certainty with which a positive test result correctly predicts a positive value (%):

$$\frac{\text{The number correctly identified as positive}}{\text{Total number with positive outcome}}$$

or

$$A/(A+B)$$

Negative predictive value

The certainty with which a negative test result correctly predicts a negative value (%):

$$\frac{\text{The number correctly identified as negative}}{\text{Total number with negative outcome}}$$

or

$$D/(C+D)$$

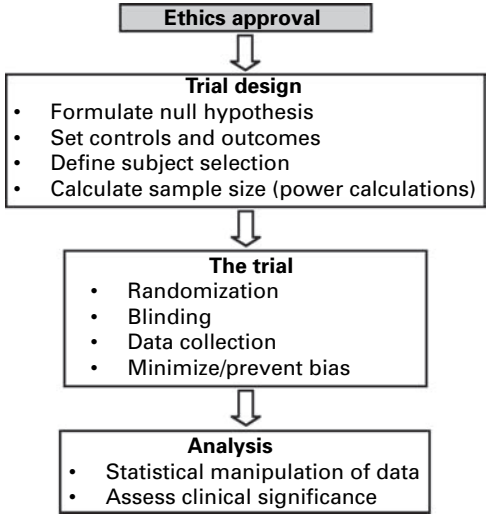
Clinical trials

Phases of clinical trials

Clinical trials will be preceded by in-vitro and animal studies before progressing through the stages shown in the table.

Phase	Description	Numbers
1	Healthy volunteers: pharmacokinetic and pharmacodynamic effects	20–50
2	More pharmacokinetic and dynamic information: different drug doses and frequencies	50–300
3	Randomized controlled trials: comparison with current treatments; assessment of frequent side effects	250–1000 +
PRODUCT LICENCE		
4	Postmarketing surveillance: rare side effects	2000–10 000 +

Trial design flow sheet



Evidence-based medicine

Evidence-based medicine

The use of current best evidence, clinical expertise and patient values to make decisions about the care of individual patients.

Levels of evidence

In this era of evidence-based medicine, there needs to be a method of categorizing the available evidence to indicate how useful it is. The following system is the one used by the UK National Institute for Health and Clinical Excellence (NICE). Other organizations that produce guidelines may use slightly different systems but the hierarchy of usefulness remains the same. The levels of evidence are based on study design, with some systems, such as this one, subdividing the grades further depending on the methodological quality of individual studies.

Level	Evidence description
1a	Systematic review or meta-analysis of one or more randomized controlled trials (RCT)
1b	At least one RCT
2a	At least one well-designed, controlled, non-randomized study
2b	At least one well-designed quasi-experimental study; for example a cohort study
3	Well-designed non-experimental descriptive studies; for example comparative, correlation or case–control studies, or case series
4	Expert opinion

Grade of recommendations

Similarly, the strength of any recommendation made on the basis of the evidence can be categorized. This is an example from NICE.

Grade	Recommendation description
A	Based directly on level 1 evidence
B	Based directly on level 2 evidence or extrapolated from level 1 evidence
C	Based directly on level 3 evidence or extrapolated from level 1 or level 2 evidence
D	Based directly on level 4 evidence or extrapolated from level 1, level 2 or level 3 evidence
GPP	Good practice point based on the view of the Guideline Development Group

An alternative is to think in terms of 'do it' or 'don't do it', based on conclusions drawn from high-quality evidence or 'probably do it' or 'probably don't do it' based on moderate quality evidence. Low-quality evidence leads to uncertainly and inability to make a recommendation.

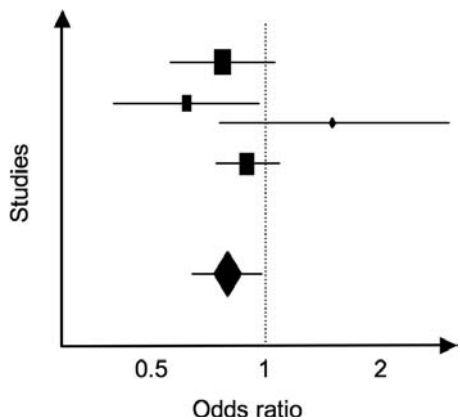
Meta-analysis

A statistical technique that combines the results of several independent studies that address a similar research hypothesis.

Meta-analysis aims to increase the statistical power of the available evidence by combining the results of smaller trials together using specific statistical methods. The validity of the meta-analysis will depend on the quality of the evidence on which it is based and how homogeneous or comparable the samples are. Combining very heterogeneous study populations can lead to bias.

Forest plot

A graphical representation of the results of a meta-analysis.



Begin by drawing and labelling the axes as shown. Draw a vertical line from 1 on the *x* axis. This is the line of no effect. The results of the individual trials are shown as boxes with the size of the box relating to the size of the trial and its position relating to the result of the trial. The lines are usually the 95% confidence intervals. The combined result is shown at the bottom of all the trials as a diamond, the size of which represents the combined numbers from all the trials. The result can be considered statistically significant if the confidence intervals of the combined result do not cross the line of no effect.