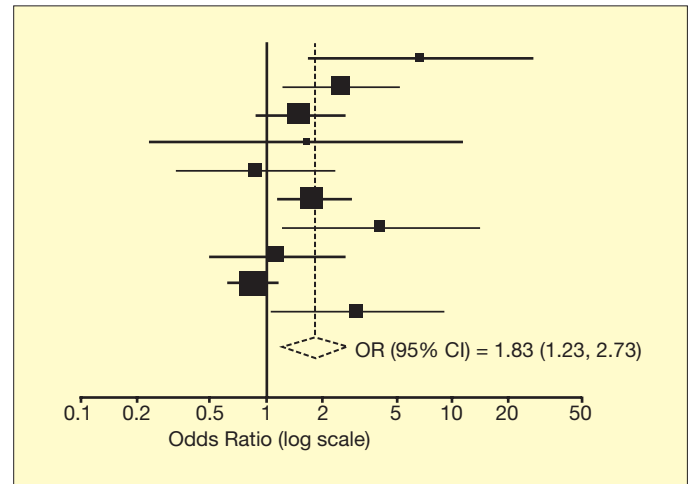


Interpreting data

— assessing study results and reports

By Sarah L. Vowler, MSc

Interpreting the results of clinical trials is often difficult, and pharmacists should be able to establish whether the techniques used are appropriate for the data presented. This article, the second in a special feature, provides guidance on interpreting some of the methods commonly seen in scientific papers



Forest plots are used to summarise the results of meta-analysis

Interpreting the results of studies and clinical trials is often difficult, but health care professionals need to understand why results are presented in different ways and what to look out for. This article provides guidance on assessing the quality of trial data reports, and interpreting the types of analysis that are commonly used in scientific papers.

Assessing trial quality

When presented with data in any form, be it a scientific paper, information from the pharmaceutical industry or work performed by colleagues, it is important to look carefully at the study design, analysis and report. A healthy scepticism will often prove useful as the results are assessed. Attention should be given to the following points.¹

Randomisation The gold standard of study design is the randomised controlled trial (RCT). Randomisation aims to assign participants randomly (and usually equally) between treatment arms to ensure a balance in observed and unobserved patient characteristics across the arms. Randomisation should be included in the study design

wherever possible. It is important to make sure that the randomisation process is adequate — methods such as allocating treatment by date of birth or hospital number or by alternation are not truly random. If the method of randomisation is not reported in a trial, the trial cannot be assumed to be truly random.¹

Eligibility criteria Eligibility criteria are important since they determine the population that the study results generalise to. The recruitment method used in the trial should be described. For example, were patients self-selected via an advertisement or referred to the study by their GP? Inclusion and exclusion criteria for participants should be listed in the report.

It should be remembered that the results only apply to the population included in the study and should not be extrapolated beyond this.¹

Sample size Sample size determination (also called power calculations) should be reported. This requires stating what difference would be clinically significant, the statistical power used (ie, the probability of declaring statistical significance when there is an effect) and some measure of variability for continuous data. It should be possible for a reader to replicate the calculation. The sample size should also be inflated for dropout, so that the sample will still be large enough to detect the stated clinically signifi-

cant difference, despite some patients withdrawing from the trial. If a non-inferiority or equivalence trial is used, the sample size calculation should reflect this. A non-inferiority trial is designed to show that a new treatment is not worse than a current treatment. An equivalence trial shows that it is the same. Both types of trial usually require a larger sample size than a superiority trial. If a non-inferiority or equivalence trial is used, it is important that it is analysed as such and not as a superiority trial.

If the sample size determination is not reported, any negative findings should be treated with caution and should not be taken to mean that, for example, two treatments are the same (a formal equivalence or non-inferiority study should be carried out to establish this).

If a non-significant result is found, this may be due to lack of power to detect a clinically significant difference should one exist. In order to be valid, trials should usually report no evidence of a difference, rather than no significant difference. This is because there may still be a clinically significant difference that the sample size was too small to detect.²

Blinding Concealment of treatment allocation protects against possible selection bias. Clinicians enrolling study participants should be unaware of the treatment allocation. This is always possible and the process should be described in the report. Blinding

Sarah Vowler is a medical statistician at the Centre for Applied Medical Statistics, Department of Public Health and Primary Care, University of Cambridge

protects against bias after treatment allocation. It is possible to blind patients, investigators and statisticians to how treatment has been allocated. This is particularly important when the outcomes of the study are subjective. Complete blinding is not always possible, therefore it is important that studies reach their potential degree of blinding. The method of blinding should be explained in the report, as should any reasons for not blinding.¹

Missing data Missing data also cause problems. Any missing data should be reported and accounted for. If a large proportion of data is missing, the results of the study should be treated with caution. Statistical methods exist for handling missing data and should be used where possible. If a number of patients have withdrawn from a clinical trial, it is important to see if there is a differential dropout between the treatment and control arms. Withdrawal may be related to the treatment, for example, one arm might have more side effects, or the treatment in one arm might not work as well.

Hypotheses A null hypothesis is the working hypothesis that is to be disproved by a statistical test in favour of the alternative hypothesis. It is usually a hypothesis of no difference or association, for example: "Increasing the dose of drug x has no effect on cardiovascular side effects." Often, not all hypotheses that have been tested are reported in scientific papers. Studies should have one pre-stated primary hypothesis that the power calculation is based on. This should usually be tested at the 5 per cent significance level (ie, $P < 0.05$). Other hypotheses are known as secondary hypotheses and should be tested at a more stringent significance level or with a method allowing for multiple comparisons, such as the Bonferroni correction.

Analyses that were not pre-planned should be reported as exploratory, and would need validation by a further study. When reading papers it is not always possible to tell whether analyses were primary, secondary or exploratory.

Panel 1 lists some checklists that can be used to help assess the quality of reports.

— Statistical analysis

The type of analysis performed in any study should be appropriate for the type of data that have been collected. The statistical analysis performed should be reported in sufficient detail for the reader to assess appropriateness and to enable replication. If the analysis has not been sufficiently reported, it cannot be assumed to be correct.

Statistical tests assume independence of data. This means that any patient should only appear in the analysis once. If the data

Panel 1: Checklists to assess the quality of reports

When reading papers there are several checklists available to assess the quality of the report. There are different checklists for different study types. The CONSORT (consolidated standards of reporting trials) checklist is used for assessing the quality of randomised controlled trials. There is an extension of this for cluster-randomised trials, and planned extensions for crossover, non-inferiority, factorial, and multi-arm trials.¹

Other checklists include QUOROM (quality of reporting of meta-analyses) for systematic reviews and meta-analyses, MOOSE (meta-analysis of observational studies in epidemiology), STARD (standards for reporting of diagnostic accuracy) and STROBE (strengthening the reporting of observational studies in epidemiology).

Trisha Greenhalgh's "How to read a paper" series, which has been made into a book,³ is another useful guide on the areas to look out for in scientific papers.

include repeated measures (see p43) or matching (in a case-control study where controls are matched to be similar to cases eg, in terms of age, sex and smoking status) the analysis must take this into account.

If a trial is cluster randomised, (eg, it is the GP practices that are randomised rather than individual patients), then the cluster (ie, the GP practice) must be the unit of analysis rather than the individual patients. Similarly, using the example of fertility trials, it would be the women who are the units of analysis, not the fertility cycles or the eggs.

Analysis in a clinical trial should be performed on the basis of "analyse as you randomise", where each patient is analysed in the group they were randomised to, whatever treatment they received. This is called the intention to treat analysis. This should be the primary analysis. "Per protocol" (those treated as specified in the protocol) or "as treated" (subjects analysed by treatment received) analyses may be carried out as secondary analyses.

The next section describes terms commonly used in association with data analysis.⁴

P-values The *P*-value of a test is the probability of seeing a result as extreme or more extreme, given that the null hypothesis is true. *P*-values are often divided up into arbitrary cut off points to indicate significance (usually $P < 0.05$) or not being significant ($P \geq 0.05$).

P-values are often misinterpreted. For example, there is a common misconception that *P*-values describe the probability that an observed difference is due to chance alone.

One of the main problems with *P*-values is that they show what is statistically significant, but this is not necessarily what is clinically significant. The two can be brought into line using a power calculation, but in practice these are often not conducted.

For these reasons, it is preferable to include a 95 per cent confidence interval wherever possible. If the study is carried out a large number of times, 95 per cent of the resulting confidence intervals will contain the true parameter estimate (eg, the mean).

It is easier to tell from a confidence interval whether or not the results are clinically significant — if the clinically significant difference lies inside the confidence interval then the result is clinically significant.

When a *P*-value is reported it should be stated exactly rather than with an arbitrary cut-off, for example, <0.05 or <0.01 , unless *P* is very small (ie, <0.0001).

Absolute risk reduction When looking at results from trials, the outcome is often binary, such as an event that either happened or did not happen (eg, death, heart attack, still-birth). The risk of this event occurring is calculated as the number of events divided by the total number of people. If a trial has two groups, for example, a control group and a treatment group, the risk can be calculated for each group. The absolute risk reduction (ARR) is the difference in risks between the control group and the treatment group. If the ARR is positive then the treatment is beneficial. If the ARR is negative, the treatment is harmful. For example, a risk of 1.70 in the control group and 1.66 in the treatment group gives an ARR of +0.04, meaning that the treatment is beneficial.

Relative risk Risks can be combined into a risk ratio or relative risk (RR), calculated by dividing the risk in the treatment group by the risk in the control group. Relative risks are usually reported in cohort and cross-sectional studies. If this value is greater than one, the risk is higher in the treatment arm than in the control group. When considering the relative risk it is important to also consider the absolute risk; if this is low, the risk to an individual will also be low. The relative risk has a fairly simple interpretation; if there is a relative risk of three, the risk of the event in the treatment group is three times that of the control group. If the relative risk is 0.75, the risk in the treatment group is three quarters of that in the control group.

Odds ratio Odds are obtained by dividing the probability of an event happening by the probability of it not happening. The odds ratio is the ratio of the odds of an event hap-

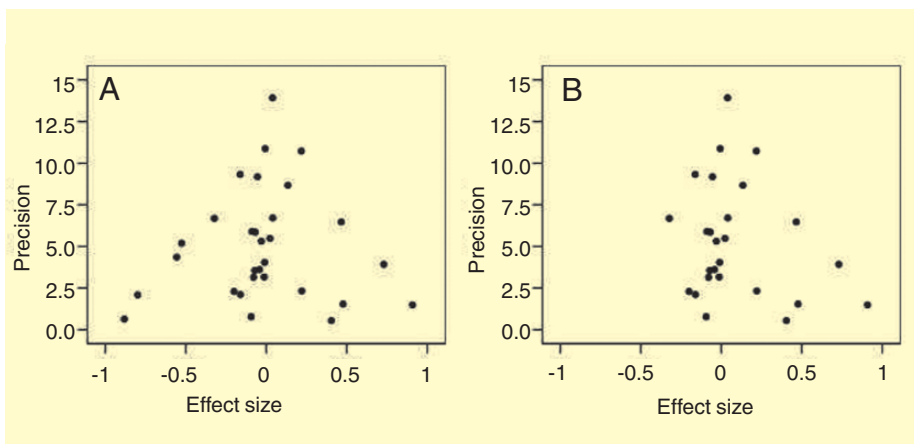


Figure 1: Example of two funnel plots. A symmetric funnel shape is formed if publications bias is not present (A). Plot B illustrates possible publication bias.

pening in two groups. Odds ratios are usually reported in case-control studies. The odds ratio of an event not happening is the inverse of the odds ratio of the event happening. If a disease is rare (eg, if it has a prevalence of <10 per cent) the odds ratio can be used as an approximation of relative risk. Where a disease is not rare, it can be difficult to interpret odds ratios and they are often incorrectly interpreted as relative risks.

Odds ratios are reported where logistic regression is used for a outcome. When the predictor variable is continuous (eg age, weight) the odds ratio applies to each unit increase in the predictor variable. Where the predictor is a binary variable and the odds ratio is greater than one, the odds are greater for the category coded as one (eg, male) compared with that coded as zero (eg, female). This interpretation assumes that all other variables are held constant. To combine the effects of more than one variable, odds ratios need to be multiplied.

In survival analysis, the data reported are “time to event” data, where the outcome is often death. Survival times are not always fully observed, they may be censored (see p44). This could be due to completion of the study or death from another cause. In these cases special analyses that allow for such censoring are required. Cox regression is a method that allows for the presence of censoring, so is preferable to logistic regression for analysing a study in which censoring is present. The results of Cox regression analysis are reported in terms of hazard ratios (see below) rather than odds ratios.

Hazard ratio The hazard is the probability of experiencing an event at a particular time, given that it has not happened up until that time. Similarly to odds ratios, if the predictor is binary, the hazard ratio compares those coded as one with those coded as zero. For continuous predictors, the hazard ratio is the hazard for each unit increase in the predictor variable. To combine the effects of variables it is necessary to multiply the hazard ratios.

A 95 per cent confidence interval should always be reported along with a risk, odds or hazard ratio. If this confidence interval includes one (eg, 0.77, 1.23) the corresponding risk, odds or hazard ratio is usually not significant.

Number needed to treat The number needed to treat (NNT) may be preferable to a relative risk or an odds ratio for reporting the results of a clinical trial, since it is thought to be more intuitive. It is interpreted as the number of patients needed to be treated with the treatment, rather than the control, to benefit one extra patient. The NNT is calculated as one divided by the ARR. For example, an ARR of +0.04 gives a NNT of 25. If the ARR is negative, the resulting figure is known as the number needed to harm (NNH), ie, the number of patients that would need to be treated for one patient to be harmed. In a screening study, the number of people needed to be screened in order to prevent one serious event is known as the number needed to screen (NNS). The baseline risk needs to be similar to allow a comparison of NNT/H/S between different treatments. If there is a different baseline risk, the impact of the treatments will be different. For example, if the risk of death is low in a particular population then the treatment will have less impact than in a population where the risk of death is higher. It is therefore important to know the absolute risk as well as the NNT/H/S.

Correlation and regression

A correlation coefficient quantifies the degree of association between two variables. It is measured on a scale from 0 to ± 1 , where 0 means there is no association and +1 or -1 means there is perfect association. When the coefficient is positive, one variable increases as the other increases. In negative correlation, one variable increases as the other decreases. The most common correlation coefficient is Pearson’s correlation.

Correlation is another measure that is often incorrectly used or interpreted. Correlation does not imply causation, and a statistically significant association is not necessarily clinically significant. It can be useful to square the correlation coefficient and multiply it by 100. This value is the percentage of variability explained by the association. Since Pearson correlation only measures linear association, it is important to verify graphically that the association is plausibly linear. If it is not plausibly linear then a rank correlation such as Spearman or Kendall should be used.

Correlation cannot be used for the following situations:

- To assess agreement between raters or variables
- To look at measurements taken repeatedly over time or measures repeatedly taken in one individual
- To look at samples with a restricted range eg, age restricted to those up to 20 years
- When subgroups with differing correlations may be present
- To relate a part to the whole, or a change to an initial value eg, relating a change in pulse after exercise to the pulse before exercise
- Pairwise over a large number of variables, reporting only those that are significant (eg if there are 10 variables then 45 correlation coefficients can be produced. Often, only those that prove significant would be reported)

From a multiple linear regression, the coefficient gives the change in outcome variable caused by a unit change in the predictor variable, assuming all other variables remain constant. The regression equation can be used to see the effect of several variables at once. This finds the sum of the effects of the predictor variable to predict the outcome.

The results of any regression are only valid over the range of variables measured in the study. Results should not be extrapolated beyond the range of data in the study. The report of any regression analysis should confirm that the assumptions of the regression were met by the data (eg, that there is a linear relationship between variables). If this is not stated, it should not be assumed the assumptions were met.

Random effects models

Observations from trials are frequently grouped in some way and are therefore related and not statistically independent. There might be repeated measures of a particular value over time or a trial might be cluster randomised. If these effects are ignored, the standard error will be underestimated, confidence intervals will be too

narrow and *P*-values will be too small. Random effects models allow for repeats or clustering within the data.^{4,5}

There is some debate over whether effects of variables included in a model should be fixed or random. If the comparison between two generic categories were of interest then fixed effects would be used. For example, comparing the effects in males versus females, or in obese versus non-obese patients may be of interest. However, if, for example, a random selection of pharmacies were included in a study, comparisons between pharmacies may not be of interest, but the model still needs to account for data being from different pharmacies. They are therefore included as random effects. As a rule of thumb, if the study were to be repeated and the same units were to be chosen, eg, males and females, the effect should be fixed.

Random effects are also used in meta-analysis to allow for both between study heterogeneity (eg, differences in eligibility criteria or differences in treatment length) and within study variability.

There are two types of random effects models: cluster-specific models and marginal models. Marginal models are fitted to the data using generalised estimating equations. These are usually easier to fit but require at least 20 and often as many as 40 clusters to be reliable. Cluster-specific models can be fitted using multi-level or mixed models (containing fixed and random effects). These are complex methods and require expert advice.

— Meta-analysis

A systematic review involves searching the literature in a systematic way according to pre-defined protocols and search criteria, in order to find all similar studies (eg, all clinical trials investigating a particular treatment). Evidence from these studies can be formally combined to give an overall treatment effect using meta-analysis. Before

combining the data, it should be determined whether combination makes clinical and biological sense. It can be difficult to decide which studies to include in a meta-analysis, but the studies should involve similar patients, methods, outcomes and treatments. Meta-analyses give an observed effect size, with a 95 per cent confidence interval, to judge clinical significance as well as statistical significance. From meta-analysis it can be difficult to know which population the results generalise to; this depends on the selection criteria for individual studies.⁵

The main advantage of systematic reviews and meta-analyses is that they combine many more subjects, so are more likely to detect an effect if it is present than the individual papers might on their own. However, a meta-analysis cannot replace a single large randomised controlled trial to answer a specific question.

Publication bias Publication bias can be a problem since studies with significant results are more likely to be published by journals and are more likely to be submitted for publication in the first place. Smaller studies may only be published if they are statistically significant and non-significant results are often thought to be unimportant or uninteresting.

A funnel plot, which plots precision (or sample size) against effect size, is often used to demonstrate whether publication bias is present or not. A symmetric funnel shape is formed if publication bias is not present (see Figure 1, p50). A forest plot is used to summarise the results of a meta-analysis (see Figure on p47). A square centred on the point estimate of the relative risk or odds ratio from each study is plotted, with lines to represent the confidence intervals. The area of the box is proportional to the information from the study. A diamond is used to represent the combined results; this is centred on the combined point estimate and the ends of the diamond extend to the confidence interval.

— Conclusion

When reading study results it is important to assess the quality of the study design, the analysis performed and the report. If these are thought to be poor, the results should be treated with caution. If it is not possible to tell from the paper if the study has been designed or analysed correctly, it is best practice to assume that it has not. Statisticians can be approached for help with critical appraisal of papers or meta-analyses of studies.

— References

1. The CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. Available at www.consort-statement.org (accessed 7 December 2006).
2. Altman DG, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. *BMJ* 1995; 311:485.
3. Greenhalgh T. How to read a paper. London: BMJ Books;2001.
4. Campbell, MJ. Statistics at square two. London: BMJ books;2001.
5. Everitt, BS and Palmer, CR (editors). Encyclopaedic companion to medical statistics. London: Hodder Arnold;2005.

— Further reading

1. Altman, DG. Practical statistics for medical research. London: Chapman and Hall;1991.
2. Everitt, BS. The Cambridge dictionary of statistics in the medical sciences. Cambridge: Cambridge University Press;1995.
3. Swinscow, TDV and Campbell, MJ. Statistics at square one, 10th edition. London: BMJ Books;2002.
4. Sutton, AJ, Abrams, KR, Jones, DR, Sheldon, TA, Song, F. Methods for meta-analysis in medical research. Chichester: Wiley;2000.