Estimating treatment effects: real or the result of chance?

The Notebook in the January 2000 issue of *Evidence-Based Nursing* described how the outcomes of clinical trials are measured and summarised before analysis. We now discuss how we can tell, by using and interpreting statistical tests, if treatments have a real effect on health or if the apparent effects of treatments under trial are a result of chance.

When critically reading a report of a clinical trial, one of the things we are interested in is whether the results of the study provide an accurate estimate of the true treatment effect in the type of patients included in the study.

Sampling error

Even if a study has been carried out in a methodologically sound (unbiased) way, a study result such as "5% more wounds healed in the treatment compared with the control group" does not necessarily mean that this is a *true treatment effect*. This finding could be a chance occurrence even when there is no true effect. To illustrate this, imagine that you are playing a game with dice. We know that, on average, each of the 6 numbers should come up an equal number of times in unbiased dice. However, when your friend throws 2 or even 3 sixes in a row, you are unlikely (depending on the friend) to infer that the dice are loaded (biased) or that he or she is cheating. Instead, you would probably conclude that this was just luck. This example shows us that even if there is no true effect (ie, the dice are not loaded), we can observe events that look like there is an effect, simply because of chance (*sampling error*). This is particularly the case when there are small numbers of observations. For example, if the number six came up in 2 out of 4 throws (ie, 50% of the time), we would assume it was because of chance. However, if it occurred in 100 out of 200 throws, then we would tend to reject the idea that this was just because of chance and instead accept an explanation that the dice were loaded.

Exactly the same logic can be applied to the results of evaluations of clinical interventions. It is possible that a study result showing benefit or harm for an intervention is because of chance, particularly if the study has a small size. Therefore, when we analyse the results of a study, we want to see the extent to which they are likely to have occurred by chance. If the results are highly unlikely to have occurred by chance, we accept that the findings reflect a real treatment effect.

Unlike the dice, which we can check by throwing them repeatedly to see if our run of sixes was a chance finding or because of some bias, we can't easily repeat every clinical trial many times to check if our finding is real. We have to make do with our one study. Of course, replication of the results of studies is an important part of the scientific process, and we would be more confident if the results were confirmed in several studies.

Statistical theory tells us that if we could repeat an experiment several hundred times with different samples of the same number of patients, the result (eg, the mean difference, difference in proportion, or relative risk) would not always be the same. If we plotted the results of these experiments on a graph, the shape of the curve (ie, the distribution of the results) would be approximately normal, or bell shaped (fig 1). On average, the results of the studies would give us a correct estimate of the true treatment effect. However, any one study result could vary from this "true" effect by chance. The degree of variation from the "true" effect is given by the measure of spread or standard deviation of this distribution which, because it indicates the amount of random sampling error that is likely, is called the *standard error* (SE). The bigger the standard error, the more individual study results will vary away from the true effect.

Confidence intervals

From any one study, we cannot be certain which is the true value of the treatment effect because it may vary from the true value by chance. However, because of the shape of the sampling distribution (fig 1) we know that 95% of all possible studies give an estimate that lies 1.96 SE on either side of the true value. Thus, we can say that in 95% of experiments, the true value will lie within 1.96 SE on either side of the estimate of effect found in our single study. In other words, there is a 95% probability that this range (1.96 SE on either side of the study effect size) includes the true value. This is called a 95% *confidence interval* and is a plausible range within which we are 95% confident the true value will lie (fig 1).

If we want to be more confident that our interval includes the true value, we can use a 99% confidence interval which lies 2.58 SE on either side of the estimate from our study. In this case there is only a 1 in 100 chance that the true value falls outside of this range.

The wider the confidence interval, the less *precise* is our estimate of the treatment effect. This precision depends on the size of the SE. This is a measure of the spread of the sampling distribution, which in turn depends on the sample size. The fewer the number of patients in a trial or number of events observed (eg, deaths), the greater will be the sampling error (fig 2). The greater the sampling error, the more likely it is that any one experiment will differ by chance from the true or average value and so the wider will be the 95% confidence interval. On the other hand, if we increase the size of the study so that the distribution becomes less spread out, individual study results will fall much closer to the true or average result and the SE and the width of the confidence interval is reduced (table 1). This is the same as saying that we are more confident in the results of throwing the die lots of times than when we throw it only a few times.

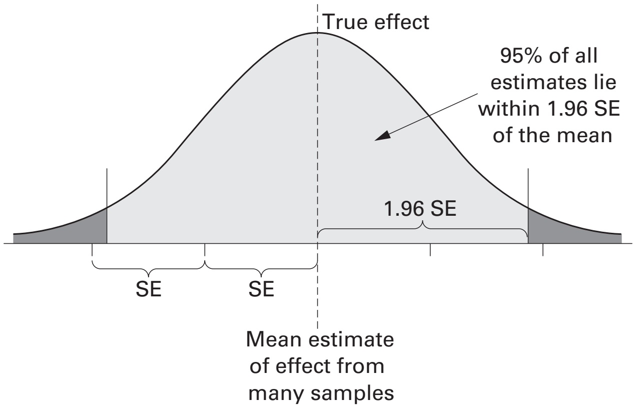

Figure 1 Sampling distribution showing effect of sampling error (SE).

Table 1 Larger sample sizes give more precise confidence intervals

Sample size (each group)	*Estimated reduction in blood pressure*	*95% confidence interval*
50	6 mm Hg	–1.0 to 13.0 mm Hg
100	6 mm Hg	1.1 to 10.9 mm Hg
200	6 mm Hg	2.5 to 9.5 mm Hg
1000	6 mm Hg	5.2 to 6.8 mm Hg

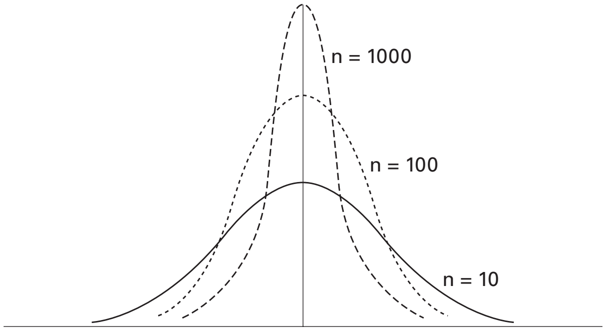

Figure 2 The larger the sample (n), the smaller the sampling error.

Presenting the uncertainty about the treatment effect using confidence intervals is now the method preferred by many good healthcare journals. Confidence intervals can also be used to answer the question of whether or not a treatment has any effect. If the confidence interval of the difference in mean blood pressure reduction or the difference in the proportion of wounds healed includes the value zero (ie, no difference in average or proportions), then we can say that we are not confident that the treatment is better than the control. Similarly, if the confidence interval of an odds ratio or relative risk includes the value 1 (same odds or risk in treated and untreated groups), we cannot be confident that there is a difference between the effects of treatment and control. This is equivalent to carrying out a hypothesis test (see below).

We can illustrate the above with 2 examples. Campbell assessed the value of nurse led clinics in primary care to improve secondary prevention of heart disease and reported an increase in aspirin use (odds ratio 3.22, 95% CI 2.15 to 4.80).[1] This means that we are 95% certain that nurse led clinics increased the odds of patients with heart disease using aspirin by between 2.15 and 4.80 times compared with standard care. However, the odds ratio for smoking cessation was 0.78 (95% CI 0.47 to 1.28). In other words, the plausible range included both a reduction in cessation (the part of the range less than 1 down to 0.47) and an increase in cessation (the part of the range between 1 and 1.28). An odds ratio of 1 means that the nurse led clinic group and the standard care group had the same cessation rate (ie, no difference). Because the confidence interval includes 1, we cannot be confident that the intervention changes the rate of smoking cessation.

A second example illustrates the use of confidence intervals when the measure of effect is the difference in the average of a continuous variable. Kitzman *et al* evaluated the effect of nurse home visits to low income, first time mothers[2] (also see *Evidence-Based Nursing*, 1998 July, p76.) The outcomes included birth weight, number of hospitalisations, number of days hospitalised, and an emotional/cognitive stimulation score. The means, mean differences, and 95% confidence intervals around those mean differences are shown in table 2. The mean difference in birth weight between the treatment and control groups was 18.2 g. However, the 95% confidence interval around this estimate ranged from –98.7 g to 62.4 g. This plausible range for the true treatment effect includes zero difference, and therefore, we cannot infer that there is any effect of home visits on birth weight. We can, however, be more confident that there was a real improvement in the measure of the degree to which the home environment was cognitively and emotionally stimulating. The mean increase in the score of 1.3 had a 95% confidence interval of 0.4 to 2.2. Because this interval does not include zero, we are 95% confident that there is a true treatment effect.

Table 2 Nurse home visits v usual care for infant birth weight and emotional/cognitive stimulation

Dependent variables	*Nurse visited group mean*	*Comparison group mean*	*Mean difference*	*95% confidence interval*
Birth weight, g	3032.2	3050.4	–18.2	–98.7 to 62.4
Emotional/cognitive stimulation (score)	32.3	30.9	1.3	0.4 to 2.2

Hypothesis testing and p values

Instead of trying to estimate a plausible range of values within which the true treatment effect is likely to lie (ie, confidence interval), researchers often begin with a formal assumption that there is no effect (the *null hypothesis*). This is a bit like the situation in a court of law where the person charged with an offence is assumed to be innocent. The aim of the evaluation is similar to that of the prosecution: to gather enough evidence to reject the null hypothesis and to accept instead the alternative hypothesis that the treatment does have an effect (the defendant is guilty). The greater the quantity and quality of evidence that is not compatible with the null hypothesis, the more likely we are to reject this and accept the alternative.

In a court case, the more evidence there is against the defendant, the more likely it is that they are guilty and will be convicted. In a clinical evaluation, the greater the treatment effect (expressed as the number of SEs away from zero), the more likely it is that the null hypothesis of zero effect is not supported and that we will accept the alternative of a true difference between the treatment and control groups. In other words, the number of SEs that the study result is away from the null value, is equivalent in the court case analogy to the amount of evidence against the innocence of the defendant. The SE is regarded as the unit that measures the likelihood that the result is not because of chance. The more SEs the result is away from the null, the less likely it is to have arisen by chance, and the more likely it is to be a true effect.

For example, if the study shows a mean difference in blood pressure of 5 mm Hg when the SE is 5 mm Hg, then the result is only 1 SE above or below the null of no difference. We know that 68% of study results are likely to lie within 1 SE of zero even when there is no treatment difference, simply by chance. In this case, there is insufficient evidence from the study to make us reject the null hypothesis; the result is compatible with chance. However, if the study result is a treatment difference of 15.1 mm Hg (ie, more than 3 SE above or below zero), then we are more likely to reject the null hypothesis because we know that when there is no real treatment effect, 99.7 % of the studies will have a result that is within 3 SE of zero (15 mm Hg). The probability of an experiment giving a result of a 15 mm Hg reduction in blood pressure by chance when there is no real treatment effect is less than 0.3% (3 out of 1000).

How unlikely must it be (ie, how many standard errors away) that a study result is due to chance for us to accept the alternative that a treatment difference really exists? We need to use some rule or criterion to decide when the result will be regarded as compatible with the null hypothesis and when it is so unlikely to be a result of chance that we reject the null hypothesis and accept instead that the treatment makes a difference (the defendant is guilty in the courtroom example). Traditionally, researchers have used probabilities $<5\%$ (1.96 SE) or $<1\%$ (2.56 SE) as cutoffs. In other words, if the treatment effect found in a study falls more than 1.96 SE above or below the null value, then the probability of this or a more extreme result occurring by chance when the null is in fact true is $<5\%$ or 1 in 20; we say it is statistically significant at the 5% level or $p<0.05$. Similarly, if the result is more than 2.58 SE above or below zero, then it is regarded as statistically significant at the 1% level ($p<0.01$). For example, in the study of nurse led clinics, it was found that the mean change in the overall score for 6 secondary prevention components over a year of follow up was 0.59 in patients allocated to the nurse clinics and 0.06 in the control group. The difference between the groups of 0.53 was highly statistically significant at $p<0.001$; that is, there was a probability of less than 1 in 1,000 that this result was because of chance.

Type I error - the risk of a false positive result

By now it should be clear that we cannot know with complete certainty what the true treatment effect is, even when there is no study bias, because of the play of chance. All we can do is make probabilistic statements that indicate how likely it is that the true value lies within a range and how likely it is that the result or one greater is a result of chance. This means that there is always the possibility of being wrong. One type of error is to reject the null hypothesis falsely (ie, to say that a treatment effect exists when in fact the null is true or, in our court case example, to falsely convict the defendant). This false positive is called a *type I error*. The risk or probability of this type of error is given by the p value or statistical significance of the treatment effect found. If a treatment difference in blood pressure is 11 mm Hg ($p=0.04$) and we infer that a real treatment effect exists, then we have a 4 in 100 risk of being wrong, because 4% of all experiments of the same type would have reported this or a bigger difference by chance. The lower the p value, the less likely it is to be a false positive, and the lower the risk of a type I error. This is the same as saying that the more evidence we have to support the guilt of the defendant, the less likely it is that an innocent person will be falsely convicted. To an extent, the level of type I error that one accepts depends on the costs and consequences of making an incorrect inference and saying that a treatment works when, in fact, it does not. Because this will vary according to what is evaluated, use of a single cutoff for all research studies is not very sensible.

Type II error - risk of a false negative result

The uncertainty, of course, works the other way as well. Even if insufficient evidence is presented in a trial to convict a defendant, this doesn't necessarily mean that she is innocent—it could be that a guilty person was wrongly acquitted. Similarly, even if we can not exclude chance as the explanation of the result from our study, it does not necessarily mean that the treatment is ineffective. This type of error—a false negative result—where we wrongly accept the null hypothesis of no treatment effect is called a type II error.

The wider the spread of the sampling distribution (larger SE), the harder it is to exclude chance and, therefore, the greater the probability of falsely accepting the null hypothesis of no treatment effect. This is particularly a problem with small studies because they have large SEs. In these cases, even large estimates of treatment effect do not provide sufficient evidence of a true treatment effect (ie, they are not statistically significant) because the SE is so large. We say that the study has a low *power* to detect a treatment effect as statistically significant when there really is one.

If a study is too small, the confidence intervals can be so wide that they cannot really exclude from the range a value indicating no effect. For example, several studies of debriding agents for the treatment of chronic wounds are so small that estimates have large confidence intervals. A study of cadexomer iodine, for example, reported an odds ratio for healed wounds of 5.5 favouring cadexomer iodine compared with dextranomer.[3] However, because only 27 patients were included in the trial, the 95% confidence interval was wide, ranging from 0.88 to 34.48. Thus, the study was too small to be able to exclude an odds ratio of 1 (no treatment difference).

When a study is undertaken, the number of patients should be sufficient to allow the study to have enough power to reject the null hypothesis if a treatment effect of clinical importance exists. Researchers should, therefore, carry out a power or sample size calculation when designing a study to ensure that it has a reasonable chance of correctly rejecting the null hypothesis. This prior power calculation should be reported in the paper.

One of the problems in clinical research is the plethora of studies that are too small and so have insufficient power. In these cases, one cannot interpret a statistically non-significant result to mean that no treatment effect exists.

Because so many studies are too small and have low power, it is possible that important clinical effects are being missed. One approach for dealing with this is to pool or combine the results of similar studies to get an overall and more precise estimate of treatment effect. This approach is called meta-analysis and will be discussed in a future Notebook.

Tests for different types of outcome measures

In the January issue of *Evidence-Based Nursing*, we described different ways of expressing treatment effects depending on the type of outcome measures used. These also affect the type of statistical test used to determine the extent to which the estimate of treatment effect is because of chance.

CONTINUOUS MEASURES

When a trial uses a continuous measure, such as blood pressure, the treatment effect is often calculated by measuring the difference in mean improvement in blood pressure between groups. In these cases (if the data are normally distributed), a *t*-test is commonly used. If, however, the data are skewed (ie, not normally distributed), it is better to test for differences in the median, using non-parametric tests, such as the Mann Whitney U test.

CATEGORICAL VARIABLES

When a study measures categorical variables and expresses results as proportions (eg, numbers infected or wounds healed), then a χ^2 (chi-squared) test is used. This tests the extent to which the difference between the observed proportion in the treatment group is different from what would have been expected by chance if there was no real difference between the treatment and control groups. Alternatively, if the odds ratio is used, the standard error of the odds ratio can be calculated and, assuming a normal distribution, 95% confidence intervals can be calculated and hypothesis tests can be done.

PAIRED ANALYSIS

The tests described above apply to situations where independent groups of patients are being compared. There are, however, situations where patients are matched, or where patients are used as their own controls. In these "paired" comparisons, it is not appropriate to use the tests outlined above and instead paired analyses are needed. For normally distributed continuous measures, one can use the *paired t-test*. For skewed continuous paired data, the *Wilcoxon signed-rank test* is available. In the case of categorical outcomes, a number of alternative tests are available, such as McNemar's test. The important thing to remember is that if the design of the comparison is paired or matched, so must be the analysis.

Statistical significance is not clinical significance

Up to now, we have been concentrating on ways of testing whether the estimate of a treatment effect found in a trial is likely to be an accurate reflection of the true effect or whether it is likely to have arisen by chance. Researchers and readers of research often focus excessively on whether a result is statistically significant (ie, not likely the result of chance). However, just because a test shows a treatment effect to be statistically significant, it does not mean that the result is *clinically* important. For example, if a study is very large (and therefore has a small standard error), it is easier to find small and clinically unimportant treatment effects to be statistically significant. A large randomised controlled trial compared rehospitalisations in patients receiving a new heart drug with patients receiving usual care. A 1% reduction in rehospitalisation was reported in the treatment group (49% rehospitalisations *v* 50% in the usual care group). This was highly statistically significant ($p < 0.0001$) mainly because this is a large trial. However, it is unlikely that clinical practice would be changed on the basis of such a small reduction in hospitalisation.

Summary

When reading reports of clinical trials, it is important to remember that apparent differences in the outcomes of patients in different treatment groups may be chance occurrences and not necessarily true treatment effects. Ideally the results of clinical trials (eg, as risk differences, odds ratios, or differences in means) are presented with confidence intervals around them. Confidence intervals represent the degree of uncertainty around the result: the narrower the confidence interval, the more precise the result. When a statistically significant difference is present, it is also important to consider whether the difference is clinically important and large enough to warrant a change in practice.

TREVOR A SHELDON, DSc
Department of Health Studies
University of York,
York, UK

1 Campbell NC, Ritchie LD, Thain J, *et al.* Secondary prevention in coronary heart disease: a randomised trial of nurse led clinics in primary care. *Heart* 1998;**80**:447–52.

2 Kitzman H, Olds DL, Henderson CR Jr, *et al.* Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing. A randomized controlled trial. *JAMA* 1997;**278**:644–52.

3 Tarvainen K. Cadexomer iodine (Iodosorb) compared with dextranomer (Debrisan) in the treatment of chronic leg ulcers. *Acta Chir Scand Suppl* 1988;**544**:57–9.

How to cite material from *Evidence-Based Nursing*

The citation styles given below can be used when citing material that was published in *Evidence-Based Nursing*.

Citation for material from an editorial

- Mulhall A. Nursing, research, and the evidence [editorial]. *Evidence-Based Nursing* 1998 Jan;1:4–6.

Citation for material taken from a structured abstract, written without attribution by a staff member

- Sucrose and commercially available milk reduced crying in newborns during a blood collection procedure [abstract]. *Evidence-Based Nursing* 1998 Jan;1:10. Abstract of: Blass EM. Milk-induced hypoalgesia in human newborns. *Pediatrics* 1997 Jan;99:825–9.

Citation for material taken from a commentary to an article

- Carlisle C. Commentary on "A conceptual model described how adults responded to a simulated literacy assessment." *Evidence-Based Nursing* 1998 Jan;1:29. Comment on: Brez SM, Taylor M. Assessing literacy for patient teaching: perspectives of adults with low literacy skills. *J Adv Nurs* May;25:1040–7.