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.
emotional/cognitive stimulation
when the measure of effect is the difference in the average of a
the rate of smoking cessation.
includes 1, we cannot be confident that the intervention changes
cessation rate (ie, no difference). Because the confidence interval
a reduction in cessation (the part of the range
0.47 to 1.28). In other words, the plausible range included both
by between 2.15 and 4.80 times compared with standard care.
increase in aspirin use (odds ratio 3.22, 95% CI 2.15 to 4.80).
improve secondary prevention of heart disease and reported an
hypothesis test (see below).
This means that we are 95% certain that nurse led clinics
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 −9.87 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.

### 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 situa-
tion 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 defend-
ant, 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 alterna-
tive 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
that we will accept the alternative hypothesis of no effect
and accept instead that the treatment makes a difference (the

### Table 1 Larger sample sizes give more precise confidence intervals

<table>
<thead>
<tr>
<th>Sample size (each group)</th>
<th>Estimated reduction in blood pressure</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>6 mm Hg</td>
<td>−1.0 to 13.0 mm Hg</td>
</tr>
<tr>
<td>100</td>
<td>6 mm Hg</td>
<td>1.1 to 10.9 mm Hg</td>
</tr>
<tr>
<td>200</td>
<td>6 mm Hg</td>
<td>2.5 to 9.5 mm Hg</td>
</tr>
<tr>
<td>1000</td>
<td>6 mm Hg</td>
<td>5.2 to 6.8 mm Hg</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Nurse visited group mean</th>
<th>Comparison group mean</th>
<th>Mean difference</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth weight, g</td>
<td>3032.2</td>
<td>3050.4</td>
<td>−18.2</td>
<td>−98.7 to 62.4</td>
</tr>
<tr>
<td>Emotional/cognitive stimulation (score)</td>
<td>32.3</td>
<td>30.9</td>
<td>1.3</td>
<td>0.4 to 2.2</td>
</tr>
</tbody>
</table>
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. 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 $\chi^2$ (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 (i.e., 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 (e.g., 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


---

**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**


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


**Citation for material taken from a commentary to an article**