The idea of using statistical methods to help reach conclusions about the merits of new therapies is nowadays well established, both in the everyday practices of those involved in clinical research, and in the formal regulations and guidelines of the governmental and other agencies who review the results of such research for the purposes of granting marketing licenses or reimbursement approval. Historically, the use of statistical methods was largely restricted to the application of statistical “tests of significance”, and it became almost mandatory in reporting clinical research to quote the results of such tests, and the “p-values” associated with them. Sadly, in many cases in the literature and elsewhere, the quoting of significance test results and p-values is still considered to be essential, at least by editors, even though in recent years the consensus has developed among medical statisticians that the use of p-values alone is rarely, if ever, sufficient, and is certainly often superfluous.

“...the use of p-values alone is rarely... sufficient, and... often superfluous”

The inadequacy of p-values stems partly from some theoretical questions about their meaning and interpretation (issues that will not be discussed here) but also from the fact that they do not in themselves convey adequate information about a clinical result. What is much more interesting and important than a mere p-value is the size of a treatment difference, and how confident we can be about the true magnitude of a treatment difference in the patient population as a whole. This information is provided by confidence intervals: versatile tools which, properly interpreted, yield all of the information that can be gleaned from a simple comparison of the effects of two treatments.

Confidence intervals
To recap on the meaning of a confidence interval, consider a simple trial in which a test treatment is compared to placebo in terms of effect on lowering systolic blood pressure.

Suppose that in the treated group, the mean reduction in systolic blood pressure is 10.0 mmHg, and in the placebo group that the mean reduction is 5.0 mmHg. In this case the treatment effect would be estimated as 10.0 - 5.0 i.e. we estimate that the test treatment is better than the placebo by 5.0 mmHg.

But this is just an estimate taken from this particular trial, and what we would also like to know is how much better would the test treatment be, on average, if we could actually give it to all patients in our target population. We can’t do that in practice, of course, but from the trial we can calculate a confidence interval. Suppose that this interval ranges from 2.5 mmHg to 7.5 mmHg. What this tells us is that we can be confident that the real effect of the test treatment is somewhere between 2.5 and 7.5 mmHg. If this is a 90% or 95% confidence interval, our degree of confidence that the true value lies between these limits is high.

So the confidence interval can be interpreted as a pair of values which, with a high degree of certainty, contain the “true” value of the treatment effect (this is simplified working definition of a confidence interval: for a more detailed explanation of the meaning and calculation of various types of confidence interval, the text by Altman et al is recommended).

“These uses have... come more into vogue... since their use [was] formalised in FDA guidelines...”
As confidence intervals have become increasingly widely employed, it has become clear that they can be used not only for assessing whether or not a new treatment is superior to an existing one, but also to help make decisions in situations where we simply expect a new treatment to be equivalent to an existing one, or when we wish only to show that a new treatment is not inferior to an existing one. These uses have been well known to medical statisticians for a considerable time, but have come more into vogue in recent times, especially since their use in this way became formalised in FDA guidelines on bioequivalence testing. The following examples from actual clinical trials illustrate the use of confidence intervals in assessing superiority, equivalence, and non-inferiority.

Superiority: Oseltamivir versus placebo in prevention of influenza
Many trials are designed in the expectation that it will be possible to show that a new experimental drug is superior to a control. For example, Welliver et al\(^2\) report a randomised controlled clinical trial in which the neuraminidase inhibitor oseltamivir was compared with placebo for use in prevention of influenza. Of 206 placebo subjects exposed to influenza virus, 26 (12.6%) developed clinical influenza, and in the oseltamivir group, of 209 subjects exposed to influenza virus, only 3 (1.4%) developed clinical influenza.

The conventional way of expressing a treatment effect with this type of endpoint is to quote the “protective efficacy” (by analogy with vaccine trials). In this study the protective efficacy is 89%, with the 95% confidence interval ranging from 67 to 97%.

If the drug did not work, we would expect a protective efficacy of 0%. The confidence interval in this case indicates that we can reasonably claim that the true protective efficacy of the drug is somewhere between 67% and 97% – much higher than the “no-effect” value of 0%. Had the confidence interval contained 0%, we could not have been sure that the true effect was different from 0%. The confidence interval alone therefore convincingly demonstrates that oseltamivir is superior to placebo in this indication. This is illustrated in Figure 1.

Equivalence: Two devices for administering albuterol
In some clinical trials, however, it is not expected that a new treatment will be superior to an existing standard. It may be realistic only to expect that a new treatment is in some sense “equivalent” to an existing established one, and it may be the objective of a clinical trial to provide adequate evidence of such equivalence. Standard statistical methods have been developed to support such objectives, and these too involve the use of confidence intervals.

The standard method requires that a definition of equivalence should be stated in advance for the two treatments that are to be compared. This means setting numerical limits for the allowable difference between two treatments, such that any difference within these limits would be accepted by the clinical community as indicating that the two treatments produce essentially the same clinical effect.

Consider, for example, the study recently reported by Ploin et al\(^3\). These authors set out to establish that a metered dose-inhaler device was clinically equivalent to a nebulizer in treatment of children with recurrent wheezing. The primary endpoint in the study was a measure of lung function known as the Pulmonary Index (PI).

Before the trial started, it was specified that if the difference between the two groups in respect of change in PI was within ±1.5 units, then the two treatments would be considered clinically equivalent. In fact, in the final data analysis, the 90% confidence interval for the difference between the two groups in respect of PI change was ±1.0 units. This is illustrated in Figure 2. Because the 90% confidence interval limits lie entirely within the pre-specified bounds, it is unlikely that the two devices really do differ by more than 1.5 units; hence, clinical equivalence has been adequately demonstrated.

Note that when establishing equivalence, 90% confidence intervals are usually used, whereas in demonstrations of superiority, a 95% interval is more common. There are some technical reasons for this, which are beyond the scope of this summary; it should be noted, though, that this convention is arbitrary, is not universal, and is the subject of some debate amongst medical statisticians.

It is interesting to note in passing that clinical equivalence would still be concluded using this type of approach in some situations in which a
treatment difference was statistically significant. Consider the hypothetical situation (B) in Figure 2. Here, the 95% confidence interval for the difference does not include the “no-effect” value 0, and therefore there would be a conventional statistically significant difference. However, when the likely size of the treatment difference in this hypothetical example is examined, it can be seen that the confidence interval for the difference still lies wholly within the pre-defined ±1.5 unit limits, and therefore the two devices would still be considered clinically equivalent. This is a simple demonstration of the often-stated fact that statistical significance and clinical significance are not necessarily the same thing.

Non-inferiority: Valganciclovir and ganciclovir in CMV retinitis

By a simple extension of the notion of equivalence, it is also possible to set up a trial with the intention of demonstrating that a new treatment is, at least, not worse than an existing standard. Such trials are often referred to as “non-inferiority” trials.

In such cases, the objective is to show that the lowest estimate of the size of the treatment effect (i.e. the lower bound of the confidence interval) in a particular study is higher than some pre-determined value. This pre-determined value is chosen such that any value lower would commonly be held to be an inferior result, and the choice should be made before the trial starts (or at least before data are unblinded).

As an example, consider the situation in Figure 3, representing a recent study of the use of the antiviral drug valganciclovir in treatment of CMV retinitis in HIV positive patients (FDA Advisory Committee 4):

In this study, the standard control treatment is ganciclovir. The new drug, valganciclovir, was not expected to perform better than the existing standard, but it was hoped to demonstrate that the new drug was not clinically inferior.

Clinicians determined in advance that, as long as the proportion of patients who showed progression of CMV retinitis on valganciclovir was not more than 25 percentage points worse than on ganciclovir, then valganciclovir could be regarded as no worse than (i.e. not inferior to) ganciclovir in clinical practice in this indication.

The actual results of the study were that, on both the ganciclovir and valganciclovir arms, 10% of patients showed progression of CMV retinitis. The 90% confidence interval for the difference between these two proportions is ±10 percentage points, and the lower limit of this interval (-10 percentage points) is not less than the pre-defined lower bound of -25 percentage points. Hence it was concluded that valganciclovir is not inferior; this conclusion was accepted by the FDA Antiviral Drugs Advisory Committee.

Summary

Confidence intervals can be used to provide evidence that a new treatment is superior to an existing standard, or equivalent to an existing standard, or, at least, not inferior to an existing standard. A simple summary of the principles is given in Table 1. These methods have been established within medical statistics for a considerable time, but only in recent years, as the use of statistical principles in clinical research has come to be considered essential, have these formal approaches become more widely used in practice, and they are now commonly encountered in the literature.

The main difference in design terms between the different types of trial described above is in the required sample sizes to give adequate power in each situation. Generally, an equivalence or a non-inferiority trial will require more patients than a superiority study: an extra requirement of about 10% is a good rule of thumb.

“It is important not to conduct a superiority trial and then... try to claim equivalence or non-inferiority”

It is important not to conduct a superiority trial and then, when results become available and it is found that superiority is not demonstrated, try to claim equivalence or non-inferiority. Apart from the fact that it is likely that not enough patients will have been recruited for such a demonstration to be made, it will also be the case that such a post-hoc claim would be regarded sceptically by reviewers and especially by regulatory authorities, because the equivalence or non-inferiority criteria were not set out in advance.
At the design stage, it is important to be realistic about what is expected of a new treatment in a clinical trial, and to state clearly at the outset what the study is intended to demonstrate. An appropriate sample size calculation can then be performed, and sufficient patients recruited to give assurance that application of the appropriate statistical method will be likely to give the desired result.

Finally, it should be stressed that there are a number of other important technical issues to consider when designing equivalence and non-inferiority trials: issues such as blinding, trial competence, and which sets of patients to analyse. Discussions with a statistician at an early stage of trial planning are advised.

Les Huson is a Consultant Medical Statistician with GNB Limited, Chislehurst.

References
Figure 1: Superiority of Oseltamivir compared with placebo in prevention of influenza

Protective efficacy of oseltamivir
Figure 2: Equivalence of two inhaler devices in effect on pulmonary index

Difference between devices in respect of change in Pulmonary Index
Difference between percentage of patients progressing (Ganciclovir - Valganciclovir)

Figure 3:
Non-inferiority of valganciclovir compared with ganciclovir in treatment of CMV retinitis
Figure 4: Superiority, equivalence and non-inferiority in terms of confidence intervals

- **Superiority**
  - Confidence interval for treatment difference does not include zero (or other "null" value if appropriate, e.g., 1 for relative risk).

- **Equivalence**
  - Upper and lower bounds of confidence interval for treatment difference are entirely within pre-specified bounds defining clinical equivalence.

- **Non-inferiority**
  - Lower bound of the confidence interval for the treatment difference is greater than a pre-specified minimum value.