BJA

"I would have everie man write what he knowes and no more."—Montaigne

BRITISH JOURNAL OF ANAESTHESIA

Volume 89, Number 6, December 2002

Editorial I

Confidence in statistical analysis

In this era of evidence-based medicine, randomized controlled trials are crucial sources for making clinical decisions. But what should we expect from a randomized controlled trial? First of all, we would like to know if there is a difference between treatments—for example, if a new drug is better than an old one. As we cannot obtain data from the entire population, we set up a randomized controlled trial on a sample of subjects and apply each treatment to a different group. We will almost certainly find some difference between the groups, because even if the same treatment had been applied to both groups we would be very likely to have some difference because of the variability between subjects in their response to the treatment. So, how do we know whether the observed difference is a reflection of a systematic difference in the population as a whole or just a reflection of patient variability? This is usually answered by using a hypothesis test. The way of proceeding is to set up a null hypothesis—that there is no difference between groups. We can calculate a probability (P value) that the observed difference arose only by chance from the variability between patients. For example, if P=0.04, this means that there is a 4% probability that the observed difference is due to chance, and therefore a 96% probability that the alternative hypothesis (there is a difference in the population) is true. The difference observed in any trial is usually regarded as statistically significant when P<0.05 (or sometimes <0.01).

Once we have learnt that there is a significant difference between groups, we would like to know how large the difference is. The difference can be estimated (point estimate) by simply calculating the difference between the mean responses in the two groups. Suppose in one study of 20 patients the mean consumption of morphine during the postoperative period was 30 mg in the control group, in which a conventional analgesic was given during anaesthesia, and 20 mg in the study group, in which a new drug was used. The difference between groups is calculated to be 10 mg.

Next, we recognize that, because we have studied only a sample of subjects, the true difference in the population will probably be a bit bigger or smaller than our point estimate. Therefore we would like to know how much bigger or smaller the true difference might probably be. Hypothesis tests do not quantify this. Most probably as a result, clinicians have tended to ignore this important question of the magnitude of the difference between groups (how much better is the new drug than the old one?) and simply consider whether 'there is a difference' or 'there is no difference'. In fact, we can calculate a range of values that will, with high probability (commonly 95%, less often 99%), contain the true difference in the population. This range of values is called the confidence interval.

The main purpose of the confidence interval is to indicate the (im)precision of the point estimates of population values.[1] More exactly, in a statistical sense, the confidence interval means that if a long series of identical studies were carried out on different samples from the same population and a 95% confidence interval for the difference between groups calculated in each study, then, in the long run, 95% of these confidence intervals would include the population difference. In the study mentioned above, the 95% confidence interval for the difference in the postoperative morphine consumption between groups is calculated to be 2–18 mg. This indicates that the true difference should be, with 95% certainty, somewhere between these values. In other words, the true difference may be as low as 2 mg or as high as 18 mg.

The variability of the subjects studied and the sample size affect the width of the confidence interval: the greater the variability of subjects studied, or the smaller the sample size, the wider the interval. Suppose a study similar to the above was done, but in 200 patients, and that this study again showed the difference in the postoperative morphine consumption between groups to be 10 mg (with similar variability between patients). If the 95% confidence interval for the difference were 2–18 mg for the study of 20 patients, calculation would show it to be 8–12 mg for the study of 200 patients. Therefore, the precision of a point estimate is indicated by the width of the confidence interval: the narrower the interval the more precise the point estimate of the population difference.

There is a close link between the confidence interval and the hypothesis test. When the 95% confidence interval does not contain 0, there is a significant difference (P<0.05), whereas when the interval contains 0 there is no significant difference between groups. Thus, the confidence interval

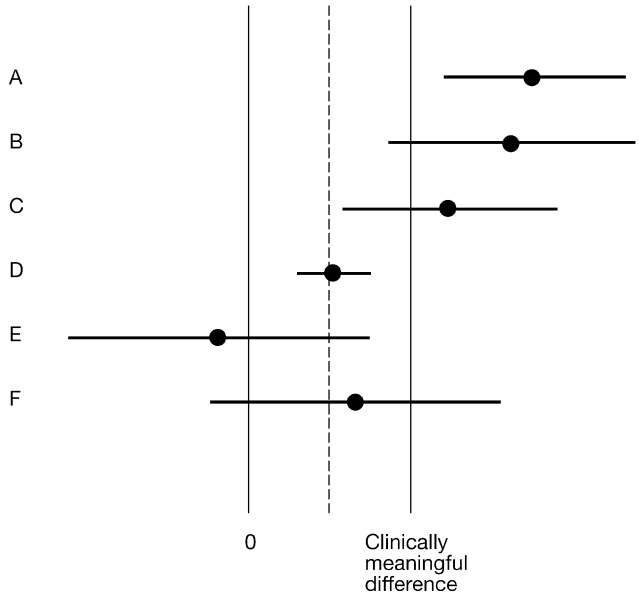

Fig 1 Examples of confidence intervals in relation to zero difference and a proposed clinically meaningful difference. The dashed line indicates half the proposed clinically meaningful difference (see text for details).

allows readers to assess whether or not the difference between groups is statistically significant.

Over the last few decades, statistical analysis has increasingly been used in medical studies. Nevertheless, it is a pity that hypothesis tests have been used predominantly and the usefulness of the confidence interval—another major statistical analysis—has not been well appreciated. In addition, the hypothesis test has often been misused and misinterpreted. Even now, it often appears that the main purpose of studies is to obtain $P<0.05$: P values less than 0.05 are regarded as meaningful and publishable in medical journals, whereas values equal to or greater than 0.05 are considered meaningless and not publishable. However, suppose two new drugs (A and B) have recently been developed and one study showed that drug A was significantly better than the old drug ($P=0.048$), whereas another study showed that there was no significant difference between drug B and the old drug ($P=0.052$). Should we then consider using drug A, but never drug B? The answer should be no. Similarly, both $P=0.053$ and $P=0.12$ are judged as non-significant, but the meaning of these can be quite different. Such dichotomy (significant and non-significant, yes and no, black and white) can produce publication bias and can lead to an unsound clinical judgement.

What we should not forget is that a significant difference between groups does not necessarily mean that there is a clinically meaningful difference. Conversely, a non-significant difference between groups does not necessarily mean that there is no clinically meaningful difference. Why? It is because, in hypothesis testing, a difference which is too small to be clinically meaningful can be statistically significant when a large number of subjects has been studied, whereas a clinically meaningful difference may be non-significant if the number of subjects studied was too small.

Then how can we decide the number of subjects that is appropriate for testing the hypothesis? To do this, it is mandatory to decide, before the start of the study, the minimum true difference between groups that we can regard as clinically meaningful. The number of subjects required is calculated on the basis of the proposed clinically meaningful difference and the anticipated variability of the data. This calculation is called power analysis. Without this calculation, reliable conclusions cannot be drawn from the P values. In particular, when the P value turns out to be $\geqslant 0.05$ (or $\geqslant 0.01$), the hypothesis test indicates a non-significant difference, even though there may in fact be a meaningful population difference (false negative, type II or beta error).[2]

Even if the required number of subjects is calculated correctly, there are still two dangers. First, a statistically significant result may not be genuinely clinically meaningful. For instance, suppose a group of researchers studying a new but expensive antihypertensive drug considered that a decrease in arterial pressure by 10 mm Hg (compared with the placebo group) would be clinically meaningful. Other clinicians might consider that the reduction needed to be 20 mm Hg to be clinically meaningful. The researcher would study a large number of subjects and may achieve a significant difference, even if the drug in fact reduces the blood pressure by only 10 mm Hg. Therefore, this study may not be useful, since clinicians would not use such an expensive drug with a weak effect in daily practice. Secondly, a statistically non-significant result may conceal a meaningful difference. Suppose another group of researchers, studying the same drug, decided that only a decrease of 30 mm Hg was clinically meaningful. They would study a much smaller number of patients so that, if the population difference was only a little larger than the 20 mm Hg that other clinicians considered meaningful, they might report a non-significant result, thereby perhaps inappropriately discouraging clinicians from using the new drug.

Therefore, the implications of a hypothesis test depend on the investigators' opinion of the clinically meaningful difference and thus the test is, in a sense, subjective. This is why it is important that the authors should state unambiguously the null hypothesis and a proposed clinically meaningful difference, to allow readers to assess to what extent the significant or non-significant difference supports or refutes the existence of a difference which they would regard as clinically meaningful. In this sense, the hypothesis test should not, in principle, be used for unplanned comparisons (such as an unexpected difference) where there is no clear null hypothesis.

The confidence interval solves the problems associated with hypothesis tests in several ways. First, the confidence interval facilitates the distinction between a significant

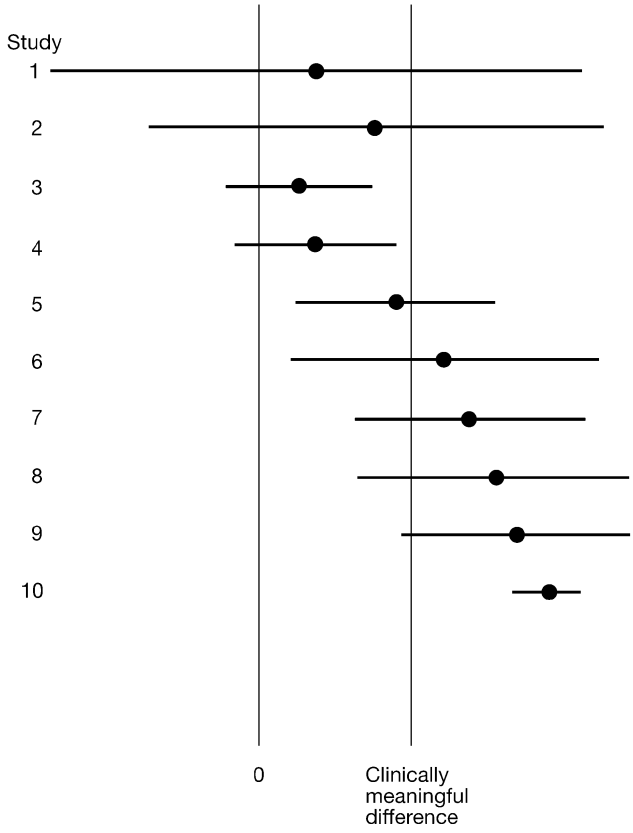

Fig 2 Examples of point estimates with their confidence intervals in relation to zero difference and a proposed clinically meaningful difference.

difference and a clinically meaningful difference. There are several possible confidence intervals in relation to a clinically meaningful difference and zero difference between groups (Fig. 1). In Figure 1, hypothesis tests would show significant differences for examples A, B, C and D. In contrast, the confidence intervals indicate that only in example A is the difference almost certainly clinically meaningful. For example B, the difference is likely to be clinically meaningful, but it may not be. For example C, the difference could be clinically meaningful but is doubtful. Lastly, for example D, the confidence interval would indicate that the difference is almost certainly not clinically meaningful.

Secondly, because the confidence interval does not require a particular hypothesis, the results will not be directly affected by the clinically meaningful difference proposed by the researchers. Therefore, by using their own values for the smallest clinically meaningful difference, readers can interpret whether or not the study indicates a difference which they would regard as clinically meaningful. For example, if readers consider that the minimum clinically meaningful difference is half the proposed value in Figure 1 (dotted line), examples A–C can be regarded as showing almost certainly a clinically meaningful difference (because the lower limit is now greater than the smaller proposed value), whereas example D may or may not be clinically meaningful.

Thirdly, the confidence interval may differentiate between no clinically meaningful difference and a beta error. In examples E and F, the hypothesis tests would indicate no significant difference between groups. The confidence interval indicates that in example E there would be no clinically meaningful difference between groups, as the upper limit of the interval is less than the value for the clinically meaningful difference. For example F, the confidence interval cannot differentiate between zero difference and a clinically meaningful difference between groups. This is most likely to occur when the number of subjects studied is too small. Nothing can be inferred from this study, which would give a false impression of the absence of an effect (beta error).

Lastly, the confidence interval is potentially useful in examining results from several similar studies. Suppose 10 studies (of various numbers of patients) have been carried out comparing the efficacy of a new drug with that of an old one (Fig. 2). When we examine the results of hypothesis tests, we will find it difficult to draw a conclusion about whether the new drug is clinically more useful than the old one, as six studies (studies 5–10) showed significant differences whereas the other four (studies 1–4) showed non-significance. We may hope to reach a conclusion by repeating similar studies, but this would often aggravate the inconsistencies between studies because, as the number of studies increases, the risk of false-positive and false-negative results also increases. The confidence interval may shed a quite different light. From Figure 2, we can deduce that studies 1 and 2 are useless, since their confidence intervals contain both zero difference and the clinically meaningful difference. Only two of the remaining eight studies (studies 3 and 4) contain zero difference, and the negative segments of the confidence intervals of these two studies are much smaller than the positive segments. In eight of the 10 studies, the confidence interval contains (or exceeds) the value for a clinically meaningful difference, and in five studies the point estimates are greater than the clinically meaningful difference. By seeing these, we can conclude that the new drug is very likely to be genuinely more effective and is likely to produce a clinically meaningful difference in the population at large. In addition, if a new trial of a large number of subjects is done (such as study 10 in Fig. 2), the confidence interval will be narrower than others so that more weight can be given to such a study than to others. Therefore, by examining the degree of difference in each study, and the confidence intervals, we can summarize all the studies. Thus, increasing the number of studies would help us to reach a more reliable conclusion. The most effective way of analysing randomized controlled trials using this concept is by meta-analysis.[3]

Taking these considerations together, the following recommendations can be made. First, authors should state

unambiguously a null hypothesis and their proposed value for the clinically meaningful difference (as well as the estimated variability between patients). The excessive use of hypothesis testing should be avoided not only because multiple testing increases the risk of false-positives (alpha error), but also because it is usually unreasonable to produce multiple hypotheses from a single study. Secondly, authors should calculate, before the start of the trial, the number of subjects required (and show the calculation in the Method section), to reduce the risk of a beta error to an acceptable level, such as 5 or 10% (corresponding to a power of 0.95 and 0.9 respectively). Thirdly, the authors should state the point estimate and its confidence interval. Finally, regardless of the presence or absence of a significant difference between groups, efforts should be made to clarify the likelihood of a clinically meaningful difference.

Gardner and Altman[4] have formulated a policy for the *British Medical Journal* that confidence intervals, if appropriate to the type of study, should be used for major findings in both the main text of a paper and its abstract. The *British Journal of Anaesthesia* states in its Guide to Contributors that 'Confidence intervals provide a more informative way to deal with a significance test than a simple *P* value'.[5] Hypothesis tests answer these questions only partially, and the *P* value in itself does not tell us if there is a clinically meaningful difference between groups. Therefore, authors submitting manuscripts to this journal are strongly encouraged to show the confidence interval, at least for the main findings (and perhaps even more so for any supplementary findings), to permit the drawing of a reliable clinical inference.

T. Asai
Department of Anaesthesiology
Kansai Medical University
Osaka 570-8507
Japan
E-mail: asait@takii.kmu.ac.jp

Acknowledgement. I thank Professor Emeritus W. W. Mapleson for his helpful comments on the manuscript.

References

1 Gardner MJ, Altman DG. Estimating with confidence. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines.* London: BMJ Books, 2000; 3–5

2 Yentis SM. The struggle for power in anaesthetic studies. *Anaesthesia* 1996; **51**: 413–4

3 Altman DG. Clinical trials and meta-analyses. In: Altman DG, Machin D, Bryant TN, Gardner MJ, eds. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: BMJ Books, 2000; 120–138

4 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986; **292**: 746–50

5 *British Journal of Anaesthesia*. Extended guide to contributors. *Br J Anaesth* 2000; **84**: 131–7