PERSONAL VIEW

Balancing statistical and clinical significance in evaluating treatment effects

W-C Leung

Epidemiology and Public Health, Newcastle General Hospital, Newcastle-upon-Tyne, UK

Correspondence to:
Dr W–C Leung, Health Policy and Practice,
Elizabeth Fry Building,
University of East Anglia,
Norwich NR4 7TJ, UK
Wai_chingleung@hotmail.com

Submitted 24 May 2000
Accepted 29 August 2000

To decide whether a new treatment should be used, statistical significance of its effectiveness over current treatment alone is insufficient. Measures of the size of the treatment effects (that is, clinical significance) are also necessary.[1]

Statistical significance measures how likely that any apparent differences in outcome between treatment and control groups are real and not due to chance. p Values and confidence intervals (CI) are the most commonly used measures of statistical significance. The p values give the probability that any particular outcome would have arisen by chance with the assumption that the new and the control treatments are equally effective as the null hypothesis. CI estimate the range within which the real results would fall if the trial is conducted many times. Hence, 95% CI of the difference in treatment outcomes between the two groups would indicate the range which the differences between the two treatments would fall on 95% of the occasions, if the trial is carried out many times.[2]

Clinical significance measures how large the differences in treatment effects are in clinical practice. Different measures have been devised. Relative risk is independent of the prevalence of the disease and can be applied to populations with different prevalence of the disease. Relative risk is the ratio of the risks in the treatment group to the event rate in the control group. However, patients may not consider this measure relevant to them as it does not specify the size of the absolute risk. The measures absolute risk reduction (ARR) and numbers needed to treat (NNT) vary with the prevalence of the disease. ARR is simply the difference in the absolute risks between the treatment group and the control group. NNT is the number of patients needed to treat to prevent one adverse event, and is numerically equal to 1/ARR. NNT has been highlighted as a meaningful measure of clinical significance.[3] The level of treatment effect regarded as clinically significant also depends on the severity of the disease and any potential side effects of the treatment.

A common measure of combined statistical and clinical significance is to state a measure of clinical significance (for example, relative risk, ARR, or NNT) with its 95% CI. For example, taking into account the prevalence of the condition, an ARR of 0.2 (95% CI 0.1 to 0.4) may indicate that the expected ARR is 0.2, and the real ARR is likely to lie between 0.1 and 0.4 on 95% of the occasions if the trial is carried out many times. The fact that the lower confidence limit is greater than zero means that the treatment is significantly more effective than control at $p<0.05$.

However, this method has its drawbacks for the clinicians who wish to apply the results in their clinical practice. First, it only indicates the statistical significance with the null hypothesis that the treatments are equally effective. However, as there may be direct and indirect costs and risks of the new treatment, the levels of treatment effects regarded as clinically worthwhile to introduce are likely to differ among clinicians and settings. If one clinician considers that introduction of the new treatment is worthwhile only if the actual risk is reduced by 15%, say, it is important to know how likely the clinical trial observations would have arisen by chance with the null hypothesis that the new treatment has an ARR of less than 0.15 and not with the null hypothesis that the two treatments are equally effective. Secondly, the 95% CI for ARR tend to be interpreted either as statistically significant (if the CI does not include zero) or not significant (if the CI includes zero). In clinical practice, such dichotomy may not be useful, and the size of the treatment effects have to be balanced with the statistical significance.[4] For example, compare the ARR of 0.1 (95% CI 0.05 to 0.2) with an ARR of 0.5 (95% CI –0.01 to 0.8). The former is regarded as statistically significant while the latter is not. However, clinical significance is likely to be higher in the latter than the former. The absolute risk reduction with the confidence limits does not guide the clinicians how to balance these two factors.

Bayesian statistical methods are advocated as alternatives to avoid these problems.[4–6] However, subjective synthesis of all available information to determine the prior distribution and complex computation required has rendered this method often impractical.[4]

I propose an alternative that a plot of p value-ARR or p value-NNT will be useful to the clinicians who practise evidence based medicine.

p Value-ARR or p value-NNT plots

The null hypothesis is that the ARR for the new treatment is less than x. p Values are calculated for a range of values of x. These p values are plotted on the y axis and the ARR on the x axis. Hence, for a range of values of ARR, we have the corresponding probability that the clinical trial observations would have arisen by chance if the real ARR were less than the given values.

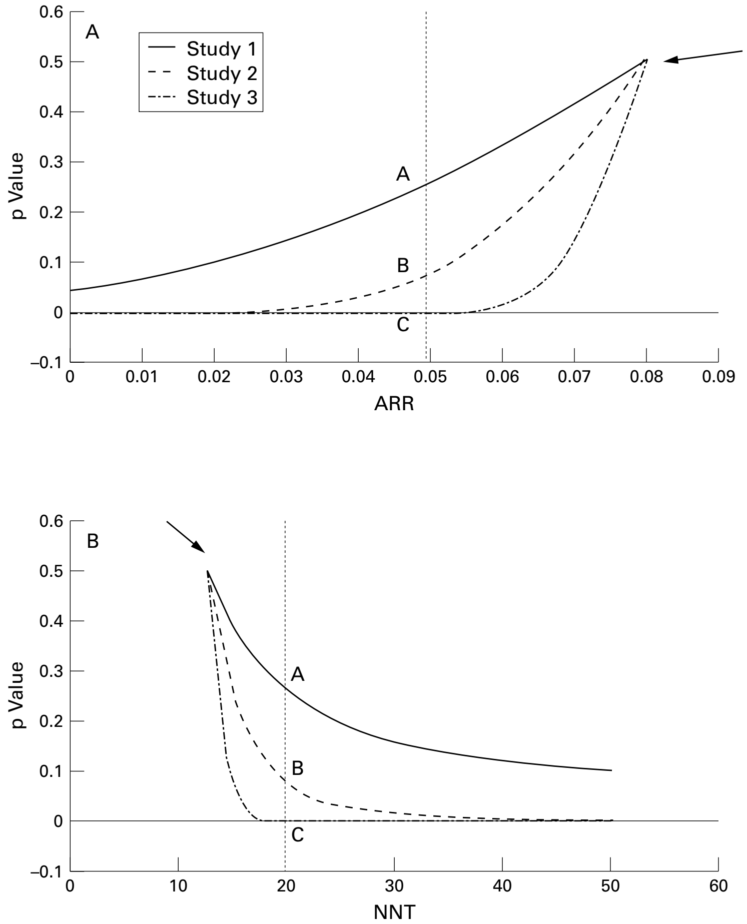

Figure 1 Plots for (A) p value-ARR and (B) p value-NNT.

The method of computation is shown in appendix 1 (see p 203).

EXAMPLE 1—STUDIES WITH SAME TREATMENT EFFECTS BUT DIFFERENT SAMPLE SIZES

I will illustrate this plot by comparing the results of three fictitious studies with the same expected relative risk, ARR, and NNT. However, the sample sizes are different and the levels of statistical significance are also different. In the first study, p>0.05 and is not statistically significant, while in the second and third studies, p<0.05 and is statistically significant.

Study 1

Treatment group: 50 subjects; 49 survived, one died.

Control group: 50 subjects; 45 survived, five died.

ARR = 0.08 (95% CI −0.012 to 0.172)

NNT = 12.5 (95% CI [5.8 to infinity] and [−infinity to −84.9])

Study 2

Treatment group: 250 subjects; 245 survived, five died.

Control group: 250 subjects; 225 survived, 25 died.

ARR = 0.08 (95% CI 0.039 to 0.121)

NNT = 12.5 (95% CI 8.3 to 25.7)

Study 3

Treatment group: 1250 subjects; 1225 survived, 25 died.

Control group: 1250 subjects; 1125 survived, 125 died.

ARR = 0.08 (95% CI 0.062 to 0.100)

NNT = 12.5 (95% CI 10.2 to 16.2)

The calculations for ARR and NNT are detailed in appendix 2. The p value-ARR plots are shown in fig 1A. As the expected value for ARR is 0.08, the p values with the null hypothesis that the real absolute risk reduction is less than 0.8 is 0.5 for each of the three studies (see arrow in the figure). That is, the actual ARR values are as equal likely to be over 0.08 as under 0.08. The p values with the null hypothesis that the treatment and controls are equally effective (that is, ARR = 0) are less than 0.05 for studies 2 and 3 (see intercept of the curves with the vertical axis). This is consistent with the 95% CI not including zero. However, fig 1A also shows that in study 1, the corresponding p value is over 0.05 (see intercept with the vertical axis). This is consistent with the 95% CI being inclusive of zero.

Suppose a clinician considers the treatment worthwhile only if the ARR is more than 0.05. Using traditional 95% CI for ARR, it will be concluded that the 95% CI for ARR includes 0.05 for both studies 1 and 2. In study 2, the 95% CI for ARR range from 0.04 to 0.121. ARR of 0.05 is close to the lower CI, but it is unclear how likely that the real ARR exceeds 0.05. Should the clinician proceed with the treatment with the results in study 2?

However, it can be seen from fig 1A that the probabilities that the study observations would have arisen by chance with the null hypothesis that the real ARR is less than 0.05 are 0.26, 0.08, and 0.000 for the first to third studies respectively (see intersections between the three curves and the vertical dotted line, labelled A, B, and C respectively). While p = 0.05 is the level usually used as the threshold in statistical tests used to decide whether a treatment is more effective than placebo, this threshold may not be applicable for testing probabilities that the ARR is less than a predetermined clinical significance. In fact, most clinicians would use a much higher threshold—perhaps up to as much as 0.2 (one in five probability of results having arisen by chance). For example, in study 2, the level p = 0.08 means that the chance that the observations have arisen by chance if ARR <0.05 is only 0.08. Hence, it is much more likely than not that the required level of clinical significance is achieved. This is a good argument to use the treatment. The p value-ARR plot gives a p value for every clinical significance level appropriate for the particular clinicians.

Similarly, p values with their 95% CI can be calculated and plotted for different NNT values, using the relationship NNT = 1/ARR. This is shown in fig 1B. The equivalence of ARR <0.05 is NNT >20. Figure 1B can be interpreted in a similar way giving similar conclusions to those from fig 1A.

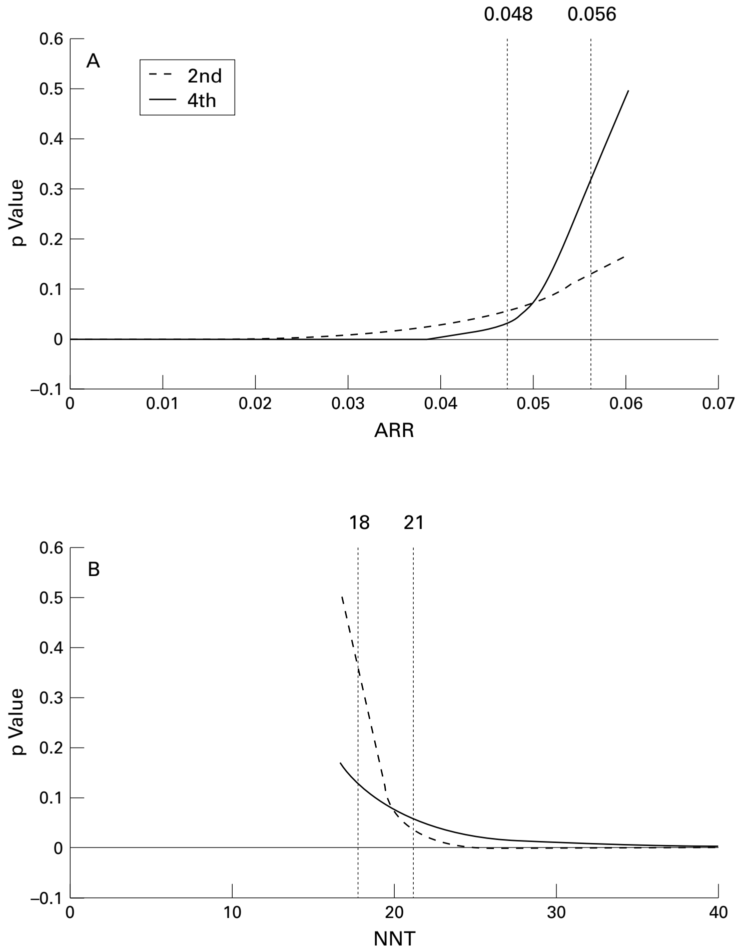

Figure 2 Plots for (A) p value-ARR and (B) p value-NNT.

EXAMPLE 2—STUDIES WITH DIFFERENT SIZES OF TREATMENT EFFECT AND SAMPLE SIZES

The advantages of the p value-ARR or p value-NNT plots over CI may be better shown with two studies: one with a lower statistical but a higher clinical significance (study 2 below), and one with a higher statistical but a lower clinical significance (study 4 below).

Study 2

Treatment group: 250 patients; 245 survived, five died.

Control group: 250 patients; 225 survived, 25 died.

ARR = 0.08 (95% CI 0.04 to 0.121)

NNT = 12.5 (95% CI 8.3 to 25.7)

Study 4

Treatment group: 2500 patients; 2400 survived, 100 died

Control group: 2500 patients; 2250 survived, 250 died.

ARR = 0.06 (95% CI 0.05 to 0.075)

NNT = 16.7 (95% CI 13.5 to 21.8)

First, suppose the clinician decides that a clinical significance of ARR >0.056 (that is, NNT <18) is required, and that different treatments are studied in studies 2 and 4. For both studies, the NNT value of 18 is within the 95% CI. It is not clear which, if any, of the treatments should the clinician use. What if the clinician decides that the clinical significance level is ARR >0.048 (that is, NNT <21)? Again, this NNT value is within the CI of both studies.

Figure 2A shows the p value-ARR plot and 2B shows the p value-NNT plot of studies 2 and 4. It is clearly seen from both fig 2A and 2B that at ARR >0.056 (that is, NNT <18), the p value of study 4 much exceeds study 2, while the reverse is the case at the level of ARR >0.048 (that is, NNT <21). On the basis of this plot, the clinician might choose the treatment in study 4 if the required clinical significance is NNT <21, but choose the treatment in study 2 if the required clinical significance is NNT <18.

Conclusion

The p value-ARR plot allows the clinicians to make a more informed decision than the traditional measure of ARR with CI. It allows clinicians to set different thresholds which they regard as clinically significant and evaluate the statistical significance that the treatment exceeds such thresholds. Traditional measures give an indication of the range within which the ARR is likely to lie, but does not provide information on the likelihood for each value within this range.

1 Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988;**318**:1728–33.
2 Greenhalgh T. How to read a paper: statistics for the non-statistician. II: “Significant” relations and their pitfalls. *BMJ* 1997;**315**:422–5.
3 Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ* 1995;**310**: 452–4.
4 Burton PR, Gurrin LC, Campbell MJ. Clinical significance not statistical significance: a simple Bayesian alternative to p values. *J Epidemiol Community Health* 1998;**52**:318–23.
5 Bland JM, Altman DG. Bayesians and frequentists. *BMJ* 1998;**317**:1151–60.
6 Lilford RJ. The statistical basis of public policy: a paradigm shift is overdue. *BMJ* 1996;**313**:603–7.

Appendix 1

Let r_1 and r_2 deaths observed among n_1 and n_2 patients in the control and treatment groups respectively. Hence, $p_1 = r_1 / n_1$ and $p_2 = r_2 / n_2$.

The null hypothesis is that the real absolute risk reduction (ARR) is less than x, where x is a variable. The deviation of the observed ARR value from the expected value is:

$p_1 - p_2 - x$

The standard error (SE) for this term is:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Hence, the term:

$$\frac{p_1 - p_2 - x}{SE}$$

should be normally distributed.

The p value is looked up in a normal distribution table. The procedure is repeated for different values of ARR (x).

Appendix 2

Study 1: In the treatment group, one out of 50 subjects died. Hence, risk of death = 1/50 = 0.02. In the control group, five out of 50 subjects died. Hence, risk of death = 5/50 = 0.1. Hence, absolute risk reduction (ARR) = 0.1−0.02 = 0.08. Number needed to treat (NNT) = 1/ARR = 1/0.08 = 12.5 The ARR and NNT for the other studies can be similarly calculated as follows:

Study	*Risk in experimental group*	*Risk in control group*	*ARR*	*NNT*
1	1/50 = 0.02	5/50 = 0.1	0.1−0.02 = 0.08	1/0.08 = 12.5
2	5/250 = 0.02	25/250 = 0.1	0.1−0.02 = 0.08	1/0.08 = 12.5
3	25/1250 = 0.02	125/1250 = 0.1	0.1−0.02 = 0.08	1/0.08 = 12.5
4	100/2500 = 0.04	250/2500 = 0.1	0.1−0.04 = 0.06	1/0.06 = 16.7