How to estimate treatment effects from reports of clinical trials. II: Dichotomous outcomes

Robert D Herbert
The University of Sydney

This paper, the second in a series of two, discusses how readers of clinical trials can extract simple estimates of treatment effect size from trial reports when trial outcomes are measured on a dichotomous scale. A method is given to quantify the degree of uncertainty of these estimates. Estimates of treatment effect size can be adjusted on the basis of baseline risk to determine the probability that treatment will help a particular patient. The probability that the treatment will be helpful should be weighed against the costs and risks of the treatment. **[Herbert RD (2000): How to estimate treatment effects from reports of clinical trials. II: Dichotomous outcomes. *Australian Journal of Physiotherapy* 46: 309-313]**

Key words: Confidence Intervals; Decision-Making; Evidence-Based Medicine

Introduction

The preceding paper in this series (Herbert 2000) considered how reports of clinical trials can be used to obtain unbiased estimates of the size of a treatment's effects. That paper discussed how readers of clinical trials can extract simple estimates of treatment effect size when trial outcomes are measured on a continuous scale. The approach incorporates clinical intuition and patient preferences into clinical decision-making. It was suggested that, when making decisions about therapy for individual patients, optimal decision-making involves modifying effect size estimates on the basis of patient characteristics and comparing such estimates with the "smallest clinically worthwhile effect". In this paper, the same process is applied to clinical trials in which outcomes are measured on a dichotomous scale.

Dichotomous outcomes The examples in the preceding paper were of clinical trials in which outcomes were measured as continuous variables. Continuous variables are those that can take on any of an infinite number of values between their upper and lower extremes. Oedema, self-reported pain on a VAS scale and lung function tests are all examples of continuous variables.

Other outcomes are measured as "dichotomous" variables. Dichotomous outcomes are discrete events - things that either happen or do not - such as death, injury, or "satisfied with treatment". We quantify these outcomes of therapy in terms of the proportion of subjects who experienced the event of interest, usually within some specified period of time. This tells us about the risk of the event for individuals from that population. A good example is provided by a recent trial of the effects of prophylactic chest physiotherapy on respiratory complications following major abdominal surgery (Olsen et al 1997). In this study, the event of interest was the development of a respiratory complication. Fifty-two of 192 subjects in the control group experienced respiratory complications within six days of surgery, so the risk of respiratory complications for these subjects was 0.27 (or 27%).

In clinical trials with dichotomous outcomes, we are interested in whether treatment reduces the risk of the event of interest. Thus we need to determine if the risk differs between treatment and control groups. The magnitude of the risk reduction, which tells us about the degree of effectiveness of the treatment, can be expressed in a number of different ways (Guyatt et al 1994, Sackett et al 1998). Three common measures are the *absolute risk reduction, number needed to treat* and *relative risk reduction*.

Absolute risk reduction The absolute risk reduction is simply the difference in risk between treatment and control groups. In the trial by Olsen et al, a relatively small proportion of subjects in the treatment group (10/172 = 0.06 or 6%) experienced respiratory complications, so the risk of respiratory

complications for subjects in the group was relatively small compared with the 27% risk in the control group. The absolute reduction in risk is 0.27-0.06 = 0.21, or 21%. This means that treated subjects were at a 21% lower risk than control group subjects of experiencing respiratory complications in the six days following surgery. Big absolute risk reductions indicate treatment is very effective. Negative absolute risk reductions indicate that risk is greater in the treatment group than in the control group and that the treatment is harmful. (An exception to this rule is when the event is a positive event, such as return to work, rather than a negative event).

Number needed to treat Understandably, many people have difficulty in appreciating the magnitude of absolute risk reductions. A consequence is that it is often difficult to specify the smallest clinically worthwhile effect in terms of absolute risk reduction (especially when the absolute risk reduction is small). How big is a 21% reduction in absolute risk? Is a 21% absolute risk reduction clinically worthwhile? A second measure of risk reduction, the number needed to treat, makes the magnitude of an absolute risk reduction more explicit. The number needed to treat (NNT) is obtained by taking the inverse of the absolute risk reduction. In our example, the absolute risk reduction is 0.21, so the NNT is 1/0.21, or ~ 5. This is the number of people who would need to be treated, on average, to prevent the event of interest once. In our example, one respiratory complication is prevented for every five people given the treatment. For the other four out of every five patients, the treatment made no difference; some would not have developed a respiratory complication anyhow, and the others developed a respiratory complication despite treatment. A small NNT (such as five) is better than a large NNT (such as 100) because it indicates that a relatively small number of patients need to be treated before the treatment makes a difference to one of them.

The NNT is very useful because it makes it relatively easy to specify a smallest clinically worthwhile effect. With the NNT, we can more easily weigh up the benefits of preventing the event in one subject against the costs and risks of giving the therapy (particularly the costs and risks to the subjects on whom the treatment had no effect). In our example, most would agree that an NNT of 10 would be worthwhile, because preventing one respiratory complication is a very desirable thing, and the risks and costs of this simple intervention are minimal, so little is lost from ineffectively treating nine out of every 10 patients. Most would agree, however, that an NNT of 100 would be too large to make the intervention worthwhile. There may be little risk associated with this intervention but it probably is too great a cost to ineffectively treat 99 people to make the prevention of one respiratory complication worthwhile. What, then, is the largest NNT for prophylactic chest physiotherapy that we would accept as being clinically worthwhile (ie what is the smallest clinically worthwhile effect)? When the author polled some experienced cardiopulmonary therapists they indicated that they would not be prepared to instigate this therapy if they had to treat more than about 20 patients to prevent one respiratory complication. That is, they nominated an NNT of 20 as the smallest clinically worthwhile effect. This corresponds to an absolute risk reduction of 5%. It would be interesting to survey patients facing major abdominal surgery to determine what they considered to be the smallest clinically worthwhile effect. The size of the treatment effect demonstrated in the trial by Olsen et al (NNT = 5) is much greater than most therapists would consider to be minimally clinically worthwhile (NNT ~ 20; remember that a small NNT indicates large treatment effects).

Clearly there is no one value for the NNT that can be deemed to be the smallest clinically worthwhile effect. The size of the smallest clinically worthwhile effect will depend on the seriousness of the event and the costs and risks of treatment. Thus the smallest clinically worthwhile effect for a three-month exercise program may be as little as two or three if the event being prevented is infrequent giving way of the knee, whereas the smallest clinically worthwhile effect for the use of incentive spirometry in the immediate post-operative period after chest surgery may be an NNT of 100 if the event being prevented is death. When therapy is ongoing, the NNT, like the absolute risk reduction, should be related to the period of therapy. An NNT of 10 for a three-month course of therapy aimed at reducing respiratory complications in children with cystic fibrosis is similar in the size of its effect to another therapy which has an NNT of five for a six-month course of therapy.

Relative risk reduction A more common but less immediately helpful way of expressing the reduction in risk is as a proportion of the risk of untreated

patients. This is termed the relative risk reduction. The relative risk reduction is obtained by dividing the absolute risk reduction by the risk in the control group. Thus the relative risk reduction produced by prophylactic chest physiotherapy is 0.21/0.27, which is 0.78 or 78%. In other words, prophylactic chest physiotherapy reduced the risk of repiratory complications by 78% *of the risk of untreated patients*. It can be seen that the relative risk reduction (78%) is much larger than the absolute risk reduction (21%). In fact, the relative risk reduction is always larger than the absolute risk reduction, because it is obtained by dividing the absolute risk reduction by a number which is always less than 1. Which, then, is the best measure of the magnitude of a treatment's effects? Should we use the absolute risk reduction, its inverse (the NNT), or the relative risk reduction?

The relative risk reduction has some properties that make it useful for comparing the findings of different studies, but it can be deceptive when used for clinical decision-making. For example, Lauritzen et al (1993) showed that the provision of hip protector pads to residents of nursing homes produced large relative reductions in risk of hip fracture (relative risk reduction of 56%). This might sound like the treatment has a big effect, and it may be tempting to conclude, on the basis of this statistic, that the hip protectors are clinically worthwhile. However, the incidence of hip fractures in a nursing home population is about 5% per year (Lauritzen et al 1993), so the absolute reduction of hip fracture risk with hip protectors in this population is 56% of 5%, or less than 3%. By converting this to an NNT, it can be seen that 36 people would need to wear hip protectors for one year to prevent one fracture. When the risk reduction is expressed as an absolute risk reduction or, better still as an NNT, the effects appear small, and may be too small to be clinically worthwhile. This example illustrates that it is probably better to make decisions about the magnitude of treatment effects in terms of absolute risk reductions or numbers needed to treat than relative risk reductions.

The importance of baseline risk In general, even the best treatments (those with large relative risk reductions) will produce only small absolute risk reductions when the risk of the event in untreated subjects (the "baseline risk") is low. Perhaps this is intuitively obvious - if few people are likely to experience the event, it is not possible to prevent it very often. There are two very practical implications. First, even the best treatments are unlikely to produce clinically worthwhile effects if the event that is to be prevented is unlikely. The converse of this is that a treatment is more likely to be clinically worthwhile when it reduces risk of a high risk event. (For a particularly clear discussion of this issue see Glasziou and Irwig 1995.) Second, as the magnitude of the treatment effect is likely to depend very much on the risk that untreated subjects are exposed to, care is needed when applying the results of a clinical trial to a particular patient if the risk to patients in the trial differs markedly from the risk in the patient for whom the treatment is being considered. If the risk in control subjects in the trial is much higher than in the patient in question, the size of the treatment effect will tend to be overestimated (that is, the absolute risk reduction calculated from trial data will be too high, and the NNT will be too low).

There is a simple work-around that makes it possible to apply the results of a clinical trial to patients with higher or lower levels of risk. The approach described here is based on the method used by Straus and Sackett (1999; see also McAlister et al 2000). The absolute risk reduction or NNT is calculated as described above directly from the results of the trial, but is then adjusted by a factor, *f*, which describes how much more at risk subjects are, than the untreated (control) subjects in the trial. An *f* of greater than 1 is used when the patients to whom the result is to be applied are at a greater risk than control subjects in the trial, and an *f* of less than 1 is used when patients to whom the result is to be applied are at a lower risk than untreated subjects in the trial. The absolute risk reduction is adjusted by multiplying by *f* and the NNT is adjusted by dividing by *f*.

The following example illustrates how this approach might be used. A therapist treating a morbidly obese patient undergoing major abdominal surgery might estimate that the patient was at twice the risk of respiratory complications as subjects in the trial by Olsen et al (1997). To obtain a reasonable estimate of the effects of therapy (that is, to take into account the greater baseline risk in this subject than in subjects in the trial) the NNT (which we previously calculated as 5) could be divided by 2. This gives an NNT of 2.5 (which rounds to 3) for morbidly obese subjects. Thus we can anticipate an even larger effect of prophylactic physiotherapy amongst high-risk patients. This approach can be used to adjust estimates of the likely

effects of treatment for any individual patient up or down on the basis of therapists' perceptions of that patient's risk.

Calculating confidence intervals for treatment effects on dichotomous outcomes As with trials that measure continuous outcomes, many trials with dichotomous outcomes do not report confidence intervals about the relative risk reduction, absolute risk reduction or NNT. Most do, however, supply sufficient data to calculate the confidence interval. A rough 95% confidence interval for the absolute risk reduction can be obtained simply from the average sample size (n) of the experimental and control groups:

$$95\%\ CI = difference\ in\ risk \pm 1/\sqrt{n}$$

(for the derivation of this equation, see the Appendix). This approximation works quite well (it gives an answer that is quite close to that provided by more complex equations) when the average risk of the events of interest in treated and control groups is greater than ~ 10% and less than ~ 90%.

To illustrate the calculation of confidence intervals for dichotomous data, recall that in the study by Olsen et al (1997) the risk to control subjects was 0.27, the risk to experimental subjects was 0.06, and the average size of each group was 182, so:

$$95\%\ CI = 0.27 - 0.06 \pm 1/\sqrt{182}$$

$$95\%\ CI = 0.21 \pm 0.07$$

Thus the best estimate of the absolute risk reduction is 21% and its 95% confidence interval extends from 14% to 28%. Inverting the 95% confidence interval of the absolute risk reduction gives a 95% confidence interval for the NNT that extends from 4 to 7.

This result has been illustrated on a tree graph of the absolute risk reduction in Figure 1. The logic of this tree graph is exactly the same as that used for the tree graph of a continuous variable which was presented in the preceding paper (Herbert 2000) except that, by convention, a positive effect (a risk *reduction*) is to the left. Again we plot the smallest clinically worthwhile effect (absolute risk reduction of 5%, corresponding to an NNT of 20), the treatment effect size (absolute risk reduction of 21%) and its confidence interval (14% to 28%) on the graph. In this example, the estimated absolute risk reduction and its confidence interval are clearly greater than the smallest clinically worthwhile effect, so we can confidently conclude that this treatment is clinically worthwhile. For morbidly obese patients (for whom we could multiply the absolute risk reduction by an f of 2 to take into account their greater untreated risk), the therapy is even more worthwhile.

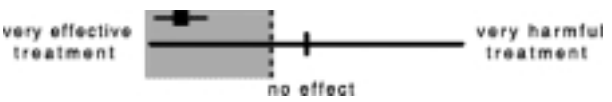

Figure 1. A "tree plot" of the size of the treatment effect reported by Olsen et al (1997). The tree plot consists of a horizontal line representing treatment effect. At the extremes are very harmful and very effective treatments. The smallest clinically worthwhile effect is represented as a vertical dashed line. The region to the left of this line represents clinically worthwhile effects. In this example, on absolute reduction of risk of respiratory complications following abdominal surgery, the smallest clinically worthwhile effect has been nominated as 5%. The best estimate of the size of the treatment effect (21%) has been illustrated as a small square, and the 95% confidence interval about this estimate (14 to 28%) is shown as a horizontal line. The treatment effect is clearly greater than the smallest clinically worthwhile effect.

This paper and the preceding paper in this two-part series have described a process which can be used to incorporate information from clinical trials in clinical decision-making. Good clinical trials provide reasonably unbiased estimates of the size of treatment effects. These estimates can be modified on the basis of patient characteristics and then compared with the smallest clinically worthwhile effect to determine if the therapy is likely to do more good than harm.

Author Robert D Herbert, School of Physiotherapy, The University of Sydney, PO Box 170, Lidcombe NSW 1825. E-mail: r.herbert@cchs.usyd.edu.au (for correspondence).

References

Gardner MJ and Altman DG (1989): Statistics with Confidence - Confidence Intervals and Statistical Guidelines. London: BMJ, pp. 20-33.

Glasziou PP and Irwig LM. An evidence based approach to individualising treatment. *British Medical Journal* 311: 1356-1359

Guyatt GH, Sackett DL and Cook DJ (1994): User's guide to the medical literature: II. How to use and article about therapy or prevention: B. What were the results and will they help me in caring for my patients? *Journal of the American Medical Association* 271: 59-63.

Herbert RD (2000): How to estimate treatment effects from reports of clinical trials. 1: Continuous outcomes. *Australian Journal of Physiotherapy*. 46: 229-235.

Newcombe RG (1999): Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 17: 873-890

Newcombe RG (1998): Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 17: 2635-2650.

Olsen MF, Hahn I, Nordgren S, Lonroth H and Lundholm K (1997): Randomized controlled trial of prophylactic chest physiotherapy in major abdominal surgery. *British Journal of Surgery* 84: 1535-1538.

Piantadosi S (1997): Clinical Trials: A Methodological Perspective. New York: Wiley

Sackett DL, Richardson WS, Rosenberg W and Haynes RB (1998): Evidence-Based Medicine. How to Practice and Teach EBM. Edinburgh: Churchill Livingstone, pp. 91-96.

Straus SE and Sackett DL (1999): Applying evidence to the individual patient. *Annals of Oncology* 10: 29-32.

Appendix.
Approximate 95% confidence intervals for absolute risk reduction.

The usual equation for the confidence interval about the absolute risk reduction (ie the difference between two proportions) is:

$$95\%\ CI = difference\ in\ risk \pm z_{(1-\alpha/2)} \times \sqrt{\frac{R_c(1-R_c)}{n_c} + \frac{R_t(1-R_t)}{n_t}}$$

where $z_{(1-a/2)}$ is the appropriate value from a z-distribution, R_c is the risk to subjects in the control group, R_t is the risk to subjects in the treatment group, and n_c and n_t are the number of subjects in control and treatment groups respectively (Gardner and Altman, 1989). Under some conditions, particularly when R_c and R_t are near 0 or 100% and n is small, this equation is not very accurate (Newcombe 1998, 1999), but for most clinical purposes it is sufficient. In randomised clinical trials the group sizes are usually similar (i.e. $n_t \approx n_c$), so we can replace n_t and n_c with n_{av}, the average size of the two groups. Also, $z_{(1-a/2)} \approx 2$, so this expression simplifies to

$$95\%\ CI = difference\ in\ risk \pm 2 \times \sqrt{\frac{R_c(1-R_c) + R_t(1-R_t)}{n}}$$

x Piantadosi (1997) has pointed out that $R[1-R]$ varies relatively little with R, at least over the range of $0.1 < R < 0.9$. To simplify the equation, we can assign a value of 1 to the term $2 \times \sqrt{(R_c[1-R_c] + R_t[1-R_t])}$, and the equation reduces to:

$$95\%\ CI = difference\ in\ risk \pm 1/\sqrt{n}$$

Analysis of the errors associated with this approximation is complex, but the approximation appears to work reasonably well. Even with small sample sizes (eg $n = 30$), the relative error (error as a proportion of the width of the "true" confidence interval) is typically less than 20%, which is probably sufficient for clinical decision making.