How to estimate treatment effects from reports of clinical trials. I: Continuous outcomes

Robert D Herbert
The University of Sydney

Properly conducted randomised trials can aid clinical decision-making by providing unbiased estimates of the average size of treatment effects. This paper, the first 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 continuous scale. When making decisions about therapy for individual patients, these estimates can be modified on the basis of patient characteristics. Modified estimates of treatment effect size can be used to determine if the effect of treatment is likely to be large enough to be "clinically worthwhile". This approach optimises clinical decision-making by combining unbiased estimates of the size of treatment effect from clinical trials with clinical intuition and patient preferences. **[Herbert RD (2000): How to estimate treatment effects from reports of clinical trials. I: Continuous outcomes. *Australian Journal of Physiotherapy* 46: 229-235]**

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

Introduction

Randomised controlled trials and systematic reviews of randomised controlled trials potentially provide unbiased and precise estimates of the effects of therapy. For this reason they can be used, when available, as arbiters of which interventions are effective and which are not.

Unfortunately not all clinical trials are valid, and the implications of valid trials are not always apparent in trial reports. Consequently, if readers of clinical trials are not to be misled, they must critically appraise trial reports to determine if the findings of the trials are likely to be valid and to ascertain what the trials mean. The process of determining the *validity* of a trial has been described in many publications (eg Guyatt et al 1993, Sackett et al 1998), and involves deciding if the trial satisfies key methodological criteria such as proper randomisation, adequate blinding and sufficient follow-up. The purpose of the present paper is to consider some issues related to the *interpretation* of valid clinical trials. Specifically, it will consider how readers can determine from a trial report if the effects of a treatment are likely to be large enough to be worthwhile. These issues have also been discussed elsewhere (Guyatt et al 1994, McAlister et al 2000, Sackett et al 1998) although they have received little attention in the physiotherapy literature.

Why do we need to know about the magnitude of a treatment's effects?

In controlled clinical trials, attention is often focused on the "*p* value" of the difference between groups. The *p* value is used to determine if the difference between groups is likely to represent a real treatment effect or could have occurred simply by chance ie "*p*" is the probability of the observed difference between groups occurring by chance alone. A small probability (conventionally, $p < 5$ per cent) means that it is unlikely that the difference would have occurred by chance alone, so it constitutes evidence of a treatment effect. Higher probabilities (conventionally, probabilities ≥ 5 per cent) indicate that the effect could have occurred by chance alone. High *p* values are properly interpreted as a lack of evidence of a treatment effect.

A consequence of this tortuous logic is to distract readers from the most important piece of information that a trial can provide, that is, information about the magnitude of the treatment's effects. If clinical trials are to influence clinical practice, they must determine more than simply whether the treatment has an effect.

They must, in addition, ascertain how big the treatment effect is. Good clinical trials provide unbiased estimates of the size of a treatment's effects. Such estimates can be used to determine if a treatment has a big enough effect to be clinically worthwhile.

What is a clinically worthwhile treatment effect? That depends on the costs and risks of treatment. Costs most obviously include monetary costs (to the patient, health provider or state), but they also include the inconvenience, discomfort and side-effects of the intervention. If a treatment is to be clinically worthwhile, its positive effects must exceed its costs, so it does more good than harm. Clinical trials often provide information about the size of treatment effects, but they rarely provide information about the costs of treatment. Thus the evaluation of whether a treatment provides a clinically worthwhile effect usually requires balancing objective information about benefical treatment effects (provided by clinical trials) against subjective impressions of the costs and risks of treatment (necessarily generated by therapists and patients).

What can trials tell us about the effects of treatment?

All treatments have variable effects. Many will have beneficial effects on some patients and have no effect or be harmful to others. Thus, strictly speaking, we cannot talk of "the effect" of a treatment. What useful information can a clinical trial provide if it cannot tell us how all patients (or any individual patient) will respond to treatment? Clinical trials can provide an estimate of the average effects of treatment. Fortunately, knowing about the *average* effects of treatment is usually the same as knowing about the *most probable* effects of treatment - usually they are, in fact, the same thing. Thus, while clinical trials cannot tell us what the effect of a treatment will be for a particular patient, they can tell us what the most likely effect will be. The same limitation applies to all sources of information about treatment effects - this is not a limitation unique to clinical trials.

A sensible way to use estimates of average treatment effects provided by clinical trials is to take them as a best first guess or prior expectation of what the size of the treatment effect is likely to be. This can then be modified up or down depending on the characteristics of the particular patients to whom the therapy is to be applied. For example, a recent trial by Dean and Shepherd (1998) showed that two weeks of task-specific motor training after stroke increased maximum seated reaching distance by about eight centimetres. Subjects in the study had had strokes more than one year previously but did did not have dementia or receptive aphasia. We might anticipate bigger effects than those reported by Dean and Shepherd if training was conducted sooner after stroke, but less effect when training demented or very aphasic patients. This approach combines the objectivity of clinical trials (which provide unbiased estimates of average effects of therapy) with the richness of clinical acumen (which may be able to distinguish between probable good and poor responders to therapy). Of course, care must be taken when using clinical intuition to modify estimates of treatment effects sizes provided by clinical trials. A conservative approach would be to ensure that the estimate of treatment effect size is modified downwards as often as it is modified upwards, although it may be reasonable to depart from this approach if the patients in the trial differ markedly, on average, from the clinical population being treated. Particular caution ought to be applied when a clinical trial provides evidence of no effect of therapy.

Weighing a treatment's effects against its costs: is this effect clinically worthwhile?

The easiest way to make decisions about whether a treatment has a clinically worthwhile effect is to first nominate the smallest treatment effect that is clinically worthwhile. This is a subjective decision that involves consideration of patients' perceptions of both the benefits and costs of treatment. Most therapists routinely consider the smallest clinically worthwhile effect when deciding whether or not to administer a particular treatment. Sometimes decisions about what is the smallest clinically worthwhile effect are explicitly negotiated with the patient.

To illustrate this process, we will consider if the application of a pneumatic compression pump produces clinically worthwhile reductions in post-mastectomy lymphoedema. We might begin by nominating the smallest reduction in lymphoedema that would make the costs of the compression therapy worthwhile. Most therapists, and perhaps even most patients, would agree that a short course of daily

compression therapy would be clinically worthwhile if it produced a sustained 75 per cent reduction in oedema. Most would also agree that a 15 per cent decrease was not clinically worthwhile. Somewhere in between these values lies the smallest clinically worthwhile effect. This value is best arrived at by discussion with the particular patients for whom the therapy is intended. Let us assume for the moment that a particular patient (or typical patients) considers that the smallest reduction in oedema that would make therapy worthwhile is around 40 per cent.

Does compression therapy produce reductions in lymphoedema of this magnitude? Perhaps the best answer to this question comes from a randomised trial by Dini et al (1998) that compared two weeks (10 days) of daily intermittent pneumatic compression to advice only. We will use the findings of this trial to estimate what the effect of compression therapy is likely to be.

Estimating the size of a treatment's effects

The best estimate of the treatment's effect is simply the difference in the means (or, in some trials, the medians) of the treatment and control groups. In the trial by Dini et al (1988), oedema was measured by measuring arm circumference at seven locations, summing the measures, and then taking the difference of the summed circumference of affected and unaffected arms (positive numbers indicate that the affected arm had a larger circumference than the unaffected arm). After the two-week experimental period, the oedema was 14.1cm (SD 5.6) in the control group and 14.2cm (SD 6.0) in the treatment group. Thus the best estimate of the treatment effect is that it increases oedema by 0.1cm (as 14.1cm - 14.2cm = -0.1cm). As the level of oedema averaged about 15.5cm prior to the experimental period, this corresponds to an increase of oedema of less than 1 per cent (100 x 0.1/15.5). Clearly this treatment effect is smaller than the smallest clinically worthwhile effect (which we had decided might be about 40 per cent). In fact, the treatment effect is slightly in the wrong direction, as the treated group had very slightly more oedema than controls. We can anticipate that, when pressure therapy is applied to this population in the manner described by Dini et al, there will be little effect of therapy. Our best guess is that the effect of therapy will be, on average (ie, most probably), too small to be clinically worthwhile.

In the example that has just been used, outcomes were measured in terms of the amount of oedema *at the end of the experimental period*. Some trials will, instead, report the *change* in outcome variables over the treatment period. In such trials, the measure of the size of the treatment's effect is still the difference of the means (this time of the difference of the mean *change*) in treatment and control groups.

Estimating uncertainty

Even when clinical trials are well designed and conducted, their findings are associated with uncertainty. This is because the difference between group means observed in the study is only an estimate of the true effect of treatment derived from the sample of 80 subjects employed in the study by Dini et al. The outcomes in this sample, as in any sample, approximate but do not exactly equal the average outcomes in the populations which the sample represents. Thus the size of the treatment effect reported in the study approximates but does not equal the true size of the treatment effect. Rational interpretation of the clinical trial requires consideration of how good an approximation the study provides. That is, to properly interpret a study's findings, it is necessary to know how much uncertainty is associated with its results.

The degree of uncertainty associated with the size of a treatment effect can be described with a confidence interval (Gardiner and Altman 1989, Sim and Read 1999). Most often the 95 per cent confidence interval is used. Roughly speaking, this is the range of treatment effects within which we can be 95 per cent certain that the true average treatment effect actually lies. (Note that the confidence interval describes the degree of uncertainty about the average effect on the population, not the degree of uncertainty of the effect on individuals.) The 95 per cent confidence interval for the difference between means in the trial by Dini et al extends from -2.9cm to 2.7cm (methods used to calculate confidence intervals are presented below) or, if changes in oedema are expressed as a percentage of initial oedema, from -19 per cent to 17 per cent. This indicates that we can be confident that the true average effect of pressure therapy lies somewhere between an increase in oedema of 19 per cent and a reduction in oedema of 17 per cent. All of the values encompassed by the 95 per cent confidence interval are smaller than the smallest clinically worthwhile effect. Thus we can conclude

that not only is the best estimate of the magnitude of the treatment effect less than the smallest clinically worthwhile effect (-1 per cent < 40 per cent), but also that no value of the treatment effect that is plausibly consistent with the findings of this study exceeds the smallest clinically worthwhile effect. These data strongly suggest that pressure therapy, at least as administered by Dini et al, does not produce clinically worthwhile reductions in lymphoedema.

Some readers will find confidence intervals easier to interpret if they sketch the confidence intervals on a "tree" graph, as in Figure 1. The tree graph consists of a line along which varying treatment effects lie. The middle of the line represents no effect (difference between group means of 0). The right end of the line represents a very good treatment effect (treatment group mean minus control group mean is a large positive number) and the left end represents a very harmful treatment (treatment group mean minus control group mean is a large negative number). For any trial, we can draw three variables on this graph (Figure 2): the smallest clinically worthwhile effect (in our example this is 40 per cent), the best estimate of the treatment effect (the difference between group means from Dini et al's randomised controlled trial, or -1 per cent), and the 95 per cent confidence interval about that estimate (-19 per cent to 17 per cent). The region to the right of the smallest clinically worthwhile effect is the domain of clinically worthwhile treatment effects. The graph for the Dini trial (Figure 2B) clearly shows that there is not a clinically worthwhile effect, because neither the best estimate of the treatment effect, nor any point encompassed by the 95 per cent confidence interval lie in the region of a clinically worthwhile effect.

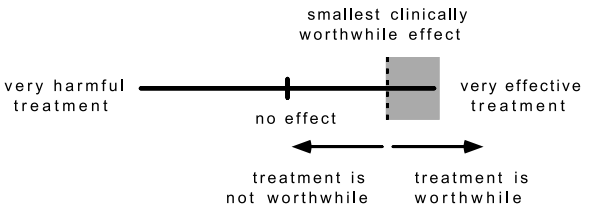

Figure 1. "Tree plots" of effect size. The tree plot consists of a horizontal line representing treatment effect. At the extremes are very harmful and very effective treatments. The smallest clinically worthwhile effect is represented as a vertical broken line. The region to the right of this line (shaded) represents clinically worthwhile effects.

Living with uncertainty

In the example just used, the treatment effect was clearly not large enough to be clinically worthwhile. Sometimes, of course, the effect of treatment is found to be clearly clinically worthwhile. Often, however, the results will be less clear. Ambiguity arises when the confidence interval spans the smallest clinically worthwhile effect, because then it is plausible both that the treatment does and does not have a clinically worthwhile effect (part of the confidence interval is less than the smallest clinically worthwhile effect and part of the confidence interval is greater than the smallest clinically worthwhile effect; either result could be the true one). For example, Sand et al (1995) showed that 15 weeks of pelvic floor electrical stimulation for women with genuine stress incontinence produced large reductions in urine leakage (average of 32mL or 70 per cent reduction) compared with sham stimulation. This result is shown on a tree plot in Figure 2B. The mean difference suggests a large and worthwhile treatment effect, but the 95 per cent confidence interval spanned from a 7 per cent to a 100 per cent reduction. There is, therefore, a high degree of uncertainty about how big the effect actually is and, because the lower end of the confidence interval includes trivially small reductions in urine loss, it is not certain, on the basis of this trial alone, that the therapy is worthwhile.

This situation, when the confidence interval spans the smallest worthwhile effect, arises commonly for two reasons. First, the designers of most clinical trials use sample sizes that are sufficient only to rule out a treatment effect of zero if there truly is a clinically worthwhile effect, but such samples may be too small to prevent their confidence intervals spanning the smallest clinically worthwhile effect. Second, many treatments have modest effects (their true effects are close to the smallest clinically worthwhile effect), so their confidence intervals must be very narrow if they are not to span the smallest clinically worthwhile effect. Consequently, few studies provide unambiguous evidence of a treatment effect or lack of treatment effect.

There are two ways to respond to the uncertainty that is often provided by single trials. First, we can accept uncertainty and proceed on the basis of the best available evidence. In this approach, clinical decisions are based on the difference between group

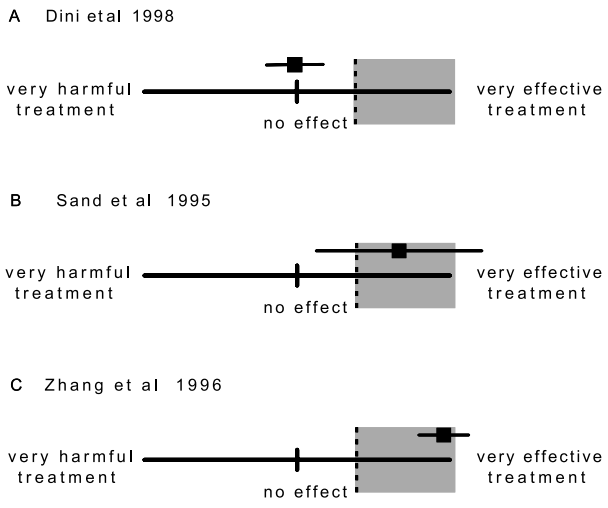

Figure 2. (A) Data from Dini et al (1998) on reduction of oedema. The smallest clinically worthwhile effect has been nominated as 40 per cent. The best estimate of the size of the treatment effect (-1 per cent) has been illustrated as a small square, and the 95 per cent confidence interval about this estimate (-19 to 17 per cent) is shown as a horizontal line. The effect is clearly smaller than the smallest clinically worthwhile effect. (B) Data from Sand et al (1995) on reduction in urine leakage. The smallest clinically worthwhile effect has been nominated as 40 per cent. The best estimate of the size of the treatment effect (70 per cent) and the 95 per cent confidence interval about this estimate (7 to 100 per cent) are shown. The best estimate of the treatment effect is that it is clinically worthwhile, but this conclusion is subject to a high degree of uncertainty. (C) Data from Zhang et al (1996) on reduction in labour time. The smallest clinically worthwhile effect has been nominated as 1 hour. The best estimate of the size of the treatment effect (2.8 hours) and the 95 per cent confidence interval about this estimate (2.2 to 3.4 hours) are shown. This treatment clearly has a clinically worthwhile effect.

means. When the difference exceeds the smallest clinically worthwhile effect, the treatment is thought to be worthwhile and when the difference between group means is less than the smallest clinically worthwhile effect, the treatment is thought to be insufficiently effective. With this approach, the role of confidence intervals is to provide an indicator of the degree of self-doubt that should be applied, but they do not otherwise affect clinical decisions. An alternative is to seek certainty by determining if the findings of individual studies are replicated in other, similar studies. This is one of the main reasons why systematic reviews of randomised controlled trials have become popular (Chalmers and Altman 1995).

In systematic reviews, the results of individual trials can be combined statistically in a meta-analysis, effectively providing a single result from many studies. The combined result is derived from a relatively large sample size, so it usually provides a more precise estimate of treatment effect size (its confidence intervals are relatively narrow), and it is more likely to provide unambiguous information about the size of the treatment effect (narrow confidence intervals are less likely to span the smallest clinically worthwhile effect). One example is the systematic review and meta-analysis by Zhang et al (1996) which showed that primiparous women who received professional support during labour had shorter labours than women who received no support (mean pooled difference 2.8 hours, 95 per cent CI 2.2 to 3.4). This effect (illustrated on a tree plot in Figure 2C) is large, and the confidence intervals are sufficiently narrow that even the most pessimistic interpretation of the data (support reduces labour time by over two hours) suggests a clinically worthwhile effect.

Calculating confidence intervals for differences between means

When confidence intervals about differences between group means are not explicitly supplied in reports of clinical trials, it is usually an easy matter to calculate these from the data reported in trials.

The confidence intervals for the difference between the means for two groups can be calculated from the difference between the two means (*difference*), their standard deviations, and the group sizes. An approximate 95 per cent confidence interval is given by first obtaining the average of the two standard deviations (*SD*) and the average of the group sizes (*n*). Then the 95 per cent confidence interval (95 per cent CI) is calculated from:

$$95 \text{ per cent CI} = \text{difference} \pm 3 \times SD / \sqrt{n}$$

(a derivation of this approximation is given in Appendix 1). In other words, the confidence interval spans an interval from $3SD/\sqrt{n}$ below the difference in group means to $3SD/\sqrt{n}$ above the difference in group means. This equation is an approximation to the more complex equation that should be used when triallists analyse their data, but it is an adequate approximation for readers of clinical trials to use for clinical

decision-making. It has the advantage that it is simple enough to be routinely calculated whenever a clinical trial does not report the confidence interval for the difference between group means.

In the trial by Dini et al on 80 subjects (average group size of 40) the authors reported mean measures of oedema for both treatment and control groups (14.2cm and 14.1cm respectively), and the SDs about those means (6.0cm and 5.6cm respectively; the average of these two SDs is 5.8cm), but they did not report the 95 per cent confidence interval for the difference between means. The 95 per cent confidence interval can be calculated from this data and is:

$$95\%\ CI \approx (14.1 - 14.2) \pm 3 \times 5.8 / \sqrt{40}$$
$$\approx -0.1 \pm 2.8$$
$$\approx -2.9 \text{ to } +2.7\text{cm}$$

Often papers will report standard errors (SEs) rather than SDs. In that case, the approximation is even simpler:

$$95\%\ CI \approx difference \pm 3 \times SE$$

Many trials have more than two groups (as there may be more than one treatment group, or more than one control). The reader must then decide which between-groups comparison is (or are) of most interest, and then the 95 per cent confidence intervals for differences between these groups can be calculated in the same way as above. Similarly, most trials report several, and sometimes many, outcomes. It is tedious to calculate 95 per cent confidence intervals for all outcomes, and the best approach is usually to decide which few outcomes are of greatest interest, and then calculate 95 per cent confidence interval for those outcomes only.

Sometimes a degree of detective work is required to find the SDs or SEs of the group means. If the SD or SE are not explicitly given they may sometimes be obtained from the error bars in figures. In other trial reports, there may be inadequate reporting of trial outcomes and it will not be possible to calculate 95 per cent confidence intervals. Such trials are difficult to interpret. Some trials report medians and interquartile ranges (or other measures of central tendency and dispersion) instead of means and SDs and it usually will not be possible to calculate confidence intervals for these trials.

The procedures described above for calculating the confidence interval of the difference between two means will tend to produce overly conservative confidence intervals (confidence intervals that are too broad) in some circumstances. In particular, this procedure will tend to produce confidence intervals that are too broad when the study is a cross-over study, a study in which subjects are matched prior to randomisation, or a study in which statistical procedures are used to partition out explainable sources of variance (such as ANCOVA). Less often, if the sample size is small and the group sizes are very unequal, the confidence interval may be too narrow. In such studies it is highly desirable that the authors report confidence intervals for the differences between groups. Unless authors provide confidence intervals for differences between groups it usually will not be possible for the reader to obtain more accurate estimates of the 95 per cent confidence intervals.

The next paper in this two-part series will describe how to determine the size of a treatment's effects on dichotomous outcomes.

Author Rob Herbert, School of Physiotherapy, The University of Sydney, Post Office Box 170, Lidcombe, New South Wales 1825. Email: r.herbert@cchs.usyd.edu.au (for correspondence).

References

Chalmers I and Altman D (1995): Systematic Reviews. London: British Medical Journal.

Dean CM and Shepherd RB (1997): Task-related training improves performance of seated reaching tasks after stroke. A randomized controlled trial. *Stroke* 28:722-728.

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

Guyatt GH, Sackett DL and Cook DJ (1993): User's guide to the medical literature: II. How to use an article about therapy or prevention: A. Are the results of the study valid? *Journal of the American Medical Association* 270: 2598-2601.

Guyatt GH, Sackett DL and Cook DJ (1994): User's guide to the medical literature: II. How to use an 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.

McAlister FA, Straus SE, Guyatt GH, Haynes RB (2000): User's guide to the medical literature. XX. Integrating research evidence with the care of the individual patient. *Journal of the American Medical Association* 283: 2829-2836.

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.

Sand PK, Richardson DA, Staskin DR, Swift SE, Appell RA, Whitmore KE and Ostergard DR (1995): Pelvic floor electrical stimulation in the treatment of genuine stress incontinence: a multicenter, placebo-controlled trial. *American Journal of Obstetrics and Gynecology* 173: 72-79.

Sim J and Reid N (1999): Statistical inference by confidence intervals: issues of interpretation and utilization. *Physical Therapy* 79: 186-195.

Zhang J, Bernasko JW, Leybovich E, Fahs M and Hatch MC (1996): Continuous labor support from labor attendant for primiparous women: a meta-analysis. *Obstetrics and Gynecology* 88: 739-744.

Appendix.
Approximate 95 per cent confidence intervals for the difference between two means

The usual equation for the confidence interval about the difference between two means is:

$$CI = difference \pm t_{(1-a/2)} \times \sqrt{\frac{(n_t - 1)\, SD_t^2 + (n_c - 1)\, SD_c^2}{n_t + n_c - 2}} \times \sqrt{\frac{1}{n_t} + \frac{1}{n_c}}$$

where *difference* is the difference between group means, $t_{(1-a/2)}$ is the appropriate value from a *t*-distribution, *n* is the number of subjects and *SD* is the standard deviation in a group, and the subscripts *t* and *c* mean "of the treatment group" and "of the control group" respectively (Gardner and Altman 1989). In randomised clinical trials the group sizes are usually similar (i.e. $n_t \approx n_c$) and it is usually assumed that the variances are equal (so that $SD_t = SD_c$). When $n_t = n_c = n$, and $SD_t = SD_c = SD$, these expressions simplify to:

$$CI = difference \pm t_{(1-a/2)} \times \sqrt{2} \times SD / \sqrt{n}$$

As $t_{(1-a/2)}$ for the 95 per cent CI ≈ 2, and $\sqrt{2} \approx 1.5$, this simplifies further to:

$$95 \text{ per cent } CI \approx difference \pm 3 \times SD / \sqrt{n}$$

The adequacy of this approximation was tested by comparing the width of the confidence intervals produced with "exact" and approximate equations using group sizes of between 10 and 100 subjects and effect sizes (expressed here as the inverse of the SD of the difference between group means) of 0.2 to 0.8. The approximate equation tended to produce confidence intervals that were too wide, but the magnitude of this error was always less than 8 per cent. This was so even when n_t and n_c or SD_t and SD_c differed by 20 per cent. The mean absolute error was 5 per cent. These small and conservative errors are probably acceptable for clinical decision-making.