What does ‘clinically important’ really mean?

Manuela L Ferreira¹ and Robert D Herbert²

¹Back Pain Research Group, Faculty of Health Sciences; ²The George Institute for International Health
The University of Sydney, Australia

Trials and meta-analyses provide estimates of the effects of intervention. Interpreting these estimates, however, is not always as straightforward as it appears. Interpretation of the estimates of effects of intervention usually involves consideration of two independent issues (Herbert 2000a, 2000b). First, it is important to consider the clinical importance of the effect. Clinically important interventions are those whose effects are large enough to make the associated costs, inconveniences, and harms worthwhile. Second, it is necessary to consider the precision of the estimate. While statistical methods required to describe precision are now well developed, methods for determining clinical importance are still very much in their infancy.

Different methodologies have been proposed and developed in the attempt to determine the clinical importance of intervention. These approaches are usually based either on the observed change of outcomes in a sample, or on an external criterion which determines whether patients have deteriorated or improved with intervention (Crosby et al 2003, de Vet et al 2007). These are known as the distribution-based approach and the anchor-based approach, respectively.

With a distribution-based approach, clinical importance is evaluated by comparing the size of the effect of intervention to some measure of variability, such as the between-person variability in outcomes or the variability associated with repeated measures of the outcome. For example, some researchers calculate the standardised effect size, which represents the difference in outcome between the intervention and control groups in units of the combined standard deviation of the two groups. By convention, standardised effects sizes of 0.2 are considered small, effects of 0.5 are considered medium, and effects of 1.0 are considered large (Cohen 1988). Some researchers claim, explicitly or implicitly, that if standardised effects are small then the difference between groups is clinically unimportant, whereas large effects are clinically important. For instance, if the difference between manipulation and placebo for neck pain is 0.12 (where 0.12 = mean post-intervention pain intensity level in the manipulation group, minus mean post-intervention pain intensity level in the placebo group, divided by the combined standard deviation), it could be argued that the effect of manipulation on neck pain intensity levels is not clinically relevant.

Alternatively, some researchers report the standard error of the measurement (an index that quantifies the amount of error that is typically associated with measurement of the outcome) or the minimum detectable change (the smallest difference between two successive measures that can be expected to be greater than measurement error), and claim that an intervention is clinically important if its effects on a certain outcome are clearly greater than could be attributed to measurement error. For example, if the minimum detectable change for the 6-min Walk Test in ambulatory children with cerebral palsy is 62 m (Thompson et al 2008), then any improvement on that scale that is short of 62 m would be regarded as not being clinically important.

All of these indices have potential uses (eg, standardised effects can be used for meta-analysis, and the standard error of the mean and the minimum detectable change are useful clinimetric indices). However, in our opinion they should not be used to evaluate clinical importance. Reasons for this are discussed further in this Editorial.

An alternative approach when measuring the effects of intervention is to use anchor-based methods. In this approach an anchor (such as the global rating of perceived effect) is used to associate descriptors with particular outcomes (such as change in pain). For example, Farrar and colleagues reported that patients whose pain reduced by more than about 30% typically considered themselves ‘much improved’ (Farrar et al 2001). Hence the researchers concluded that reductions in pain of greater than 30% represented clinically important improvements. We will argue that, while anchor-based methods can help researchers associate descriptors with outcomes, they do not assist researchers to decide if an effect of intervention is clinically important.

Recently, anchor-based methods have been combined with distribution-based methods (Crosby et al 2003, De Vet et al 2007). In this approach an external anchor is used to define three levels of outcome: a) importantly improved, b) not importantly changed, and c) importantly deteriorated. Outcomes of patients who have just been treated are defined according to the criteria, and a Receiving Operating Characteristic (ROC) curve is constructed (De Vet et al 2007). Such methods share the limitations of both anchor-based and distribution-based methods.

The main limitation of both distribution-based methods and anchor-based methods is that neither attempts to evaluate whether patients feel that the effect is large enough to make the costs, inconvenience, and harms associated with intervention worthwhile. Distribution-based methods determine whether the effects of intervention are large compared to between-patient variability or between-measurement variability, but not whether effects are large enough to be worthwhile. Anchor-based methods determine whether patients feel they experienced small improvements or large improvements, but not whether they feel that the effects of intervention were large enough to be worthwhile. In anchor-based studies, that decision is made by the researchers. For example, in the study by Farrar and colleagues (2001) it was the researchers who decided that outcomes must be ‘much improved’ for the intervention to be clinically important. Decisions about whether effects of treatment are large enough to make the costs, inconvenience or harm worthwhile are best made by patients, not by clinicians or researchers. In future research, recipients of care should be involved in this decision.
A second important issue that is not dealt with by either distribution-based or anchor-based approaches is that these approaches evaluate properties of outcome measures, not of interventions. Thus Farrar and colleagues (2001) used anchor-based methods to claim that a 30% reduction in pain was clinically important, but they made no distinction between interventions as diverse as an information booklet and major surgery. In reality, provision of an information booklet is inexpensive and associated with little inconvenience or harm, so provision of an information booklet may be considered to have clinically important effects even if the effects are small. On the other hand, major surgery is associated with obvious costs and inconveniences so even quite large effects may be of little or no importance. The clinical importance of an intervention cannot be evaluated without reference to the costs and inconveniences of that intervention. This means that any evaluation of clinical importance must be intervention-specific, not only outcome-specific.

One further limitation of most methods used to evaluate the clinical significance of intervention is that such estimates are almost always based on within-group changes (ie, changes in outcome from baseline) rather than between-group changes (ie, the difference in outcome between the intervention and control groups). Once again, using the anchor-based estimates of Farrar and colleagues as an example, the 30% reduction in pain considered to be clinically important is calculated from patients’ improvements from baseline and thus represents a within-group change rather than a between-group change (Farrar et al 2001). However, this estimate is used to interpret between-group differences reported in clinical trials. Within-group changes may be due to the intervention, but they may also be due to natural recovery, statistical regression, and placebo (Herbert et al 2005). Consequently, we cannot use within-group changes to tell us about the clinical importance of the effects of intervention. Between-group differences in outcomes can provide unbiased estimates of effects of intervention that are not distorted by natural recovery or statistical regression or placebo effects. So the proper focus of the analysis should be on between-group differences (ie, on the difference in outcome of people who do and do not receive the intervention).

In recent years, Barrett and colleagues (Barrett et al 2005a, Barrett et al 2005b) have developed a new method, the ‘benefit-harm trade-off method’, for assessing clinically important effects of intervention. This method involves presenting patients with estimates of the benefits and harms associated with the intervention. Patients are asked to comment on whether they would choose the intervention. Then, holding the costs, inconveniences, and harms constant, the patient is asked to imagine that the benefit of the intervention is larger (or smaller). This is repeated until it is possible to identify, with sufficient precision, the threshold benefit for which the patient would choose to have the intervention. The threshold is called the ‘sufficiently important difference’. The method has been used successfully to estimate sufficiently important differences of four interventions for the common cold (Barret et al 2005b).

Importantly, the benefit-harm trade-off method overcomes all of the main shortcomings of distribution-based and anchor-based evaluations of clinical importance: patients (not researchers or clinicians) estimate directly how large the effects of intervention must be to make the intervention worthwhile, the estimate is intervention-specific, and (if conducted properly) the method focuses on the effects of intervention (between-group differences) rather than on change over time (within-group differences).

Estimates of the sufficiently important differences obtained with the benefit-harm trade-off method can be compared with estimates of the expected effects of intervention obtained from randomised trials. This information could be used at the individual patient level: a patient’s sufficiently important difference could be compared with expected effects of intervention and, if the expected effect exceeded that estimate, the patient could be offered the intervention. Perhaps more practically, the same approach could be applied at population level: the intervention could be offered to patients if studies on representative samples suggested patients’ sufficiently important differences typically do not exceed the expected effect of intervention. Estimates of sufficiently important differences obtained from representative samples could also inform the design of clinical trials because the sample size of trials should be sufficient to detect effects that are considered to be worthwhile by most patients.

The benefit-harm trade-off method is clearly superior to distribution and anchor-based methods for determining whether the effects of intervention are clinically important. It should be the method of choice to inform decisions about whether to offer an intervention.

References