When is a further clinical trial justified?

Manuela L Ferreira,[1] Robert D Herbert,[2] Michael J Crowther,[3] Arianne Verhagen,[4] Alex J Sutton[3]

[1]The George Institute for Global Health, University of Sydney, PO Box M201, Missenden Road, NSW 2050, Australia

[2]Neuroscience Research Australia, Sydney, Australia

[3]Department of Health Sciences, University of Leicester, UK

[4]Department of General Practice, Erasmus Medical Centre University, Rotterdam, Netherlands

Correspondence to: M L Ferreira mferreira@georgeinstitute.org.au

Accepted: 21 August 2012

High quality randomised trials provide unbiased estimates of the effects of health interventions, but the findings of a single trial are rarely conclusive. Usually data from several similar trials must be pooled together to draw firm conclusions about the effectiveness of an intervention. This article considers how to determine whether data from existing trials are conclusive and, if not, whether a further trial is justified.

Even on their own the best randomised trials do not usually provide convincing evidence of the effectiveness (or lack of effectiveness) of an intervention. Consensus about effectiveness is usually only achieved when a meta-analysis of several high quality trials shows a statistically significant effect. Until such a consensus is achieved researchers may claim that a new trial is justified. We will refer to this as the "conventional approach" to justification of a new trial.

Two problems, and solutions

There are two problems with the conventional approach. The first concerns the interpretation of existing evidence: interpretation of data from clinical trials or meta-analyses is reduced to a decision about whether the intervention is "effective" or "ineffective."[1] The simplistic classification of interventions as effective or ineffective fails to make the important distinction between interventions that have trivial effects and those that have worthwhile effects.

A health intervention produces "worthwhile" effects when it does more good than harm. Here we must use the term "harm" to include all of the negative aspects of interventions: from a patient's perspective, these could include risks of adverse events, pain or discomfort, cost, or inconvenience. The role of clinical trials is to provide unbiased estimates of the beneficial effects of health interventions so that it can be ascertained whether those effects outweigh risks, costs, and inconvenience.

Several methods have been developed to determine what beneficial effect would outweigh the risks, costs, and inconvenience of an intervention. In a recent review we argued such methods should be based on patients' (usually not researchers' or clinicians') perceptions, intervention specific, and expressed in terms of differences between outcomes with and without intervention.[2] One method potentially satisfies all three criteria. The "benefit-harm trade-off method" involves presenting patients with hypothetical scenarios about the effects of an intervention and identifying the smallest hypothetical effect for which the patient would choose to receive the intervention.[3] We refer to this as the "smallest worthwhile effect" of intervention to differentiate it from similar constructs such as the "minimum clinically important difference" derived with alternative methods.

Figure 1 illustrates how information about the smallest worthwhile effect of an intervention can be used to interpret an estimate of the effect of intervention from a meta-analysis (or, for that matter, from an individual clinical trial). The figure shows six hypothetical pooled estimates of the effect of a health intervention (A-F) obtained from six hypothetical meta-analyses of randomised trials. The magnitude of the pooled estimate of effect is indicated by the location of the diamonds along the horizontal axis. For meta-analyses in which the effect is quantified as a mean difference in outcomes of treated and control groups (for example, weighted mean difference or standardised mean difference) the line of no effect would have a value of 0. For meta-analyses in which the effect is quantified as a ratio of outcomes of treated and control groups (for example, relative risk, odds ratio, hazard ratio, or incidence rate ratio) the line of no effect would have a value of 1. For each possible outcome, the statistical significance of a test of the null hypothesis (that is, no effect of intervention) is shown, along with an interpretation of the outcome based on consideration of the estimated size of the effect.

SUMMARY POINTS

A further clinical trial may be justified when the existing evidence from high quality clinical trials does not clearly indicate whether an intervention is or is not worth while

Benefit-harm trade-off studies can determine what constitutes a "worthwhile" effect

When a meta-analysis of existing trials does not provide clear findings about whether an intervention has worthwhile effects, and a further trial is being considered, extended funnel plots can be used to explore the potential impact of a new trial on the updated meta-analysis

These procedures can be used to determine if a further clinical trial is justified
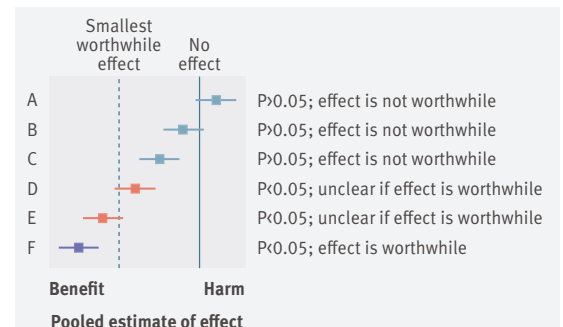

Fig 1 | Interpretation of six pooled estimates of effect from six hypothetical meta-analyses of randomised trials. Error bars are 95% confidence intervals
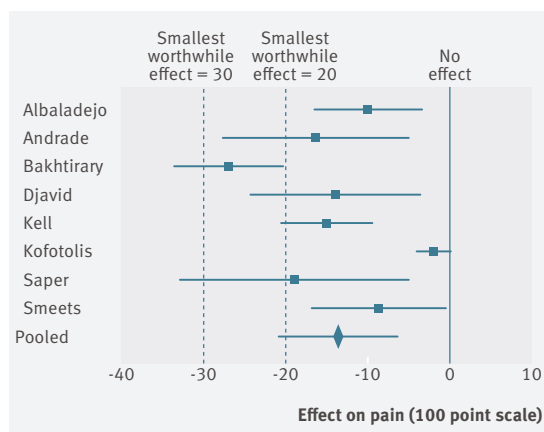
Fig 2 | Forest plot showing estimates of effect of exercise on chronic low back pain. Summary data (mean effect, expressed as a difference between treated and control groups) from eight randomised trials (squares) with pooled estimate from random effects meta-analysis (diamond). Values less than 0 are indicative of a beneficial effect (reduction in pain). Error bars are 95% confidence intervals. Two values for the smallest worthwhile effect are shown, corresponding, respectively, to the 50th and 80th centiles of estimates of the smallest worthwhile effect obtained in a benefit harm trade-off study

In each of the six scenarios one of three conclusions may be drawn: the effect is worth while (outcome F in fig 1), it is unclear if the effect is worth while (outcomes D and E), or the effect is not worth while (outcomes A, B, and C). A further trial is only justified when it is unclear if the effect is worth while. Note that the statistical significance of a test of the null hypothesis is not relevant when determining if a further trial is justified. Further trials should only be done in scenarios D and E.

A second problem with the conventional approach to determining whether a further trial is justified is that the decision is often made without consideration of how the findings of the new trial might contribute to existing evidence.[4] Usually it will be an updated meta-analysis of all trial data, not just the data from the new trial, that provides the most precise estimate of effect of intervention. A further trial is only justified when addition of data from the new trial to a meta-analysis of existing trials could convert uncertainty about whether the effects of intervention are worthwhile (outcomes D and E in fig 1) to the conclusion that the intervention is worth while (outcome F) or is not worthwhile (outcomes A-C).

It is widely believed that provided a clinical trial is large enough it will confer a high degree of certainty about the effects of intervention. (Here we put aside important issues about risk of bias and quality of interventions.) A surprising new idea is that sometimes no trial of any size, when added to an existing random effects meta-analysis, will be able to provide a high degree of certainty about the effects of intervention.[4] This is because when the findings of the new trial are sufficiently extreme to have a noticeable effect on the pooled estimate of effect, they may increase between trial heterogeneity to the extent that the data from the new trial increase, rather than decrease, the width of the confidence interval about the pooled estimate of effect (see supplementary file).

Two of the authors have developed methods for evaluating whether it is plausible that a new trial could convert existing uncertainty about the effects of intervention into clear evidence for or against the existence of a worthwhile effect.[4 5] These methods include "extended funnel plots"—graphical augmentations of the funnel plots traditionally used to investigate small study bias in meta-analysis.[6 7] The shaded contours of an extended funnel plot show how the conclusions of an updated meta-analysis can be influenced by the findings and size of the new trial: particular combinations of trial findings and trial sizes may result in the conclusion that the effect of intervention is clearly worth while, or clearly not worth while, or remains uncertain. Extended funnel plots can be constructed using the extfunnel macro in Stata.[6]

In the following section we briefly illustrate the use of benefit-harm trade-off studies and extended funnel plots to determine whether a further trial is justified. The example of exercise for chronic low back pain is used. We focus on the short term effects of exercise on pain, measured on a 100 point scale.

The smallest worthwhile effect of exercise for chronic back pain

Using the benefit-harm trade-off method we obtained estimates of the smallest worthwhile effect of exercise for chronic low back pain. We ranked the smallest worthwhile effects reported by 95 participants and calculated the 50th and 80th centiles—that is, the effects considered large enough to be worth while by 50% and 80% of participants. On the 100 point pain scale the 50th and 80th centiles corresponded to treatment effects of 20 and 30 points, respectively.

Existing evidence of the effect of exercise on chronic back pain

We used optimised search strategies to update to August 2011 a recent meta-analysis of the effects of exercise on chronic low back pain.[8] Pain data were rescaled to a common 0-100 scale. Data from eight trials[9-15] were pooled in a random effects meta-analysis (fig 2). These analyses showed that compared with no treatment, exercise reduces pain on average by 14 points (95% confidence interval 6 to 21 points) on a 100 point scale. The effect is clearly significant (P of test of no effect <0.001), but the more interesting issue concerns whether the effect is large enough to be worth while.

Does exercise produce worthwhile reductions in chronic back pain?

Use of the 80th centile of the smallest worthwhile effect (a 30 point reduction in pain) might be justified if it was believed that the interpretation of clinical trials should be based on an effect of intervention considered to be worth while by most people in the population of interest. As fig 2 shows, if the 80th centile is used, one would conclude that exercise is clearly not worth while for treatment of chronic back pain (because the confidence interval for the effect of exercise only includes effects <30). In that case there would be no need to do a further trial.

Alternatively, use of the 50th centile of the smallest worthwhile effect (a 20 point reduction in pain) might be justified if it was believed that the interpretation of clinical trials should be based on effects considered to be worth while by typical people in the population of interest. As fig 2 shows, if the 50th centile is used, it is not clear whether exercise is worth while for treatment of chronic back pain because the confidence interval for the effect of exercise includes values above and below 20.

The 50th centile (the median) of the smallest worthwhile effect may be more justifiable as it may be particularly important to determine if the expected effect of intervention is of interest to a typical person in the population of interest. Consequently we concluded it is not clear if exercise has worthwhile effects on chronic back pain and we considered whether a new trial could resolve this uncertainty.

What influence would the findings of a new trial have?

Figure 3A is an extended funnel plot showing the potential conclusions arising from a new trial of exercise for chronic back pain. In fig 3B the plot has been modified so that the sample size, rather than the standard error, is shown on the vertical axis. In both panels the horizontal axis represents the size of the effect of intervention that could be observed in a new trial. The horizontal axis has been drawn so that its extremes are the limits of the 95% prediction interval.[16] (The prediction interval is the range of effects that is predicted could occur in a new trial, given the effects observed in the eight existing trials.) The vertical axis reflects the size of the trial. In fig 3A the size of the new trial is expressed in terms of its standard error. In fig 3B the size of the trial is expressed as the number of participants in the trial. (Here it is assumed the standard deviation is known—it was assigned a value of 15—and that there will be two equally

sized groups.) Data from the existing meta-analysis (fig 2) have been superimposed on the extended funnel plots (same data in both panels). The mean effects from each of the eight randomised trials are shown as circles, the solid vertical line is the pooled estimate of the mean effect of exercise for chronic back pain, and the dashed line is the smallest worthwhile effect. A new trial could produce a finding that is located anywhere on the plot—the location in the horizontal direction depends on the estimate of effect provided by the new trial and the location in the vertical direction depends on the size of the new trial. Each point on the plot is coloured red or blue to indicate the conclusion that would be drawn from the updated meta-analysis if the new trial had that particular combination of effect size (horizontal axis) and trial size (vertical axis). The colour codes are the same as in fig 1. Thus these plots show that if the findings of a new trial of exercise for chronic back pain were to be added to the existing meta-analysis, the updated conclusion of the meta-analysis could be that the effect of exercise is too small to be worth while (blue regions) or that it is uncertain if the effect is worth while (red regions). There are no purple regions on the plot, indicating that it is unlikely a new trial could show that exercise has a worthwhile effect.

Although the extended funnel plot indicates there is no real possibility a new trial could show that exercise produces worthwhile reductions in chronic back pain, some clinical trialists might none the less consider carrying out a further trial to determine if exercise does not have worthwhile effects. (It would be useful to know if this were true as exercise is currently used widely for the management of chronic back pain.) However, sample size calculations[4] (supplementary file) suggest that this may be futile: with a sample size of 500 there is only a 55% probability of the updated meta-analysis concluding that exercise does not produce worthwhile effects and a 45% chance of concluding ongoing uncertainty. Note that conventional power calculations[17] would conclude very differently: they would suggest that a new trial with 500 participants would have a power of nearly 100% to detect the smallest worthwhile effect.

In summary, while existing trials clearly show statistically significant effects of exercise for chronic back pain, they neither confirm nor rule out the possibility that exercise has worthwhile effects; and it is likely that a new trial, even a large trial, would not resolve uncertainty about whether the effects are large enough to be worth while. The latter conclusion is based on consideration of the contribution of just one new trial. We acknowledge that several trials (perhaps even several small trials) may yield certainty where one large trial does not.[4] Alternatively, it may be possible to use individual patient data meta-analyses or metaregression methods to explain some of the between trial heterogeneity. An additional consideration may be the cost of a new trial. A complete justification of a new trial would weigh the cost of the trial against the value of the information it generates.[18]

The methods we have described can be based on either fixed effect or random effects meta-analyses. We recommend routine use of random effects meta-analysis. When heterogeneity is present, however, random effects models limit the impact of any individual study.[4] A consequence is
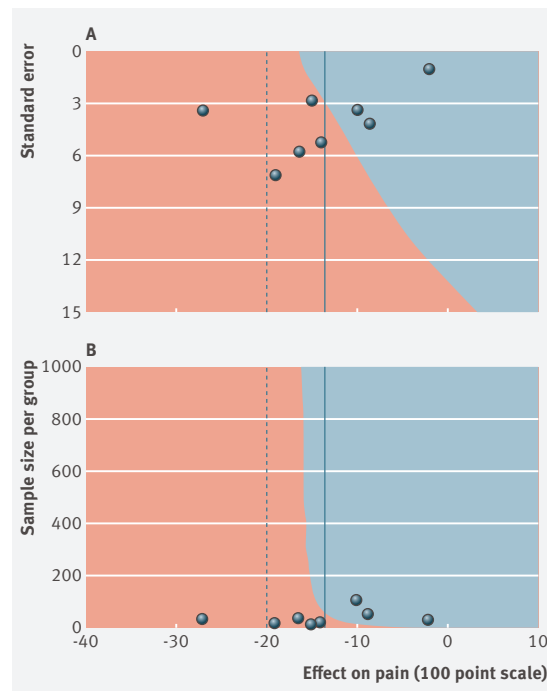

Fig 3 | Extended funnel plots showing the predicted consequences of a further trial on the updated conclusions of a meta-analysis of the effects of exercise for back pain

that in some circumstances, as in the example used here, new studies can have little impact on existing conclusions.[4]

Conclusions

This article has described the use of benefit-harm studies to define the smallest worthwhile effect of intervention, and the use of extended funnel plots to explore the potential influence of a new trial on the findings of an updated meta-analysis. Clinical trialists should consider using these procedures when deciding whether to carry out a further clinical trial. Peer reviewers and granting bodies could seek evidence from these sorts of analyses when assessing requests for funding to conduct a further trial.[4] [19]

1   Silva-Aycaguer LC, Suarez-Gil P, Fernandez-Somoano A. The null hypothesis significance test in health sciences research (1995-2006): statistical analysis and interpretation. *BMC Med Res Methodol* 2010;10:44.
2   Ferreira ML, Herbert RD, Ferreira PH, Latimer J, Ostelo RW, Nascimento DP, et al. A critical review of methods used to determine the smallest worthwhile effect of interventions for low back pain. *J Clin Epidemiol* 2012;65:253-61.
3   Barrett B, Brown R, Mundt M, Dye L, Alt J, Safdar N, et al. Using benefit harm tradeoffs to estimate sufficiently important difference: the case of the common cold. *Med Decis Making* 2005;25:47-55.
4   Sutton AJ, Cooper NJ, Jones DR, Lambert PC, Thompson JR, Abrams KR. Evidence-based sample size calculations based upon updated meta-analysis. *Stat Med* 2007;26:2479-500.
5   Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Med Res Methodol* 2009;9:29.
6   Langan D, Higgins JP, Gregory W, Sutton AJ. Graphical augmentations to the funnel plot assess the impact of additional evidence on a meta-analysis. *J Clin Epidemiol* 2012;65:511-9.
7   Crowther M, Langan D, Sutton A. Graphical augmentations to the funnel plot to assess the impact of a new study on an existing meta-analysis. *Stata J* 2012;(In press).
8   Ferreira ML, Smeets RJ, Kamper SJ, Ferreira PH, Machado LA. Can we explain heterogeneity among randomized clinical trials of exercise for chronic back pain? A meta-regression analysis of randomized controlled trials. *Phys Ther* 2010;90:1383-403.
9   Albaladejo C, Kovacs FM, Royuela A, del Pino R, Zamora J. The efficacy of a short education program and a short physiotherapy program for treating low back pain in primary care: a cluster randomized trial. *Spine* 2010;35:483-96.
10   Bakhtiary AH, Safavi-Farokhi Z, Aminian-Far A. Influence of vibration on delayed onset of muscle soreness following eccentric exercise. *Br J Sports Med* 2007;41:145-8.
11   Djavid GE, Mehrdad R, Ghasemi M, Hasan-Zadeh H, Sotoodeh-Manesh A, Pouryaghoub G. In chronic low back pain, low level laser therapy combined with exercise is more beneficial than exercise alone in the long term: a randomised trial. *Aust J Physiother* 2007;53:155-60.
12   Kell RT, Asmundson GJ. A comparison of two forms of periodized exercise rehabilitation programs in the management of chronic nonspecific low-back pain. *J Strength Cond Res* 2009;23:513-23.
13   Kofotolis N, Kellis E. Effects of two 4-week proprioceptive neuromuscular facilitation programs on muscle endurance, flexibility, and functional performance in women with chronic low back pain. *Phys Ther* 2006;86:1001-12.
14   Saper RB, Sherman KJ, Cullum-Dugan D, Davis RB, Phillips RS, Culpepper L. Yoga for chronic low back pain in a predominantly minority population: a pilot randomized controlled trial. *Altern Ther Health Med* 2009;15:18-27.
15   Smeets RJ, Vlaeyen JW, Hidding A, Kester AD, van der Heijden GJ, Knottnerus JA. Chronic low back pain: physical training, graded activity with problem solving training, or both? The one-year post-treatment results of a randomized controlled trial. *Pain* 2008;134:263-76.
16   Riley RD, Higgins JP, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ* 2011;342:d549.
17   Julious S. Sample sizes for clinical trials with Normal data. *Stat Med* 2004;23:1921-86.
18   Willan AR, Pinto EM. The value of information and optimal clinical trial design. *Stat Med* 2005;24:1791-806.
19   Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A Stat Soc* 2009;172:137-59.

ANSWERS TO ENDGAMES, p 50   For long answers go to the Education channel on bmj.com

PICTURE QUIZ

Hip pain in an adolescent after injury while playing football

1   A simple muscular strain is the most likely diagnosis given the history. Hip dislocation, slipped capital femoral epiphysis, pelvic and femoral fractures, and infection are all potentially serious causes of pelvic pain in adolescents. Neoplastic and rheumatological disorders also merit consideration.

2   The anterioposterior radiograph of the pelvis shows a displaced avulsion fracture of the left anterior superior iliac spine.

3   Avulsion fractures are a consequence of secondary ossification centres being the weakest point in the pelvis in adolescents. Contraction of the sartorius muscle and its subsequent avulsion from its proximal insertion were facilitated by the patient accidentally striking the firm earth, rather than the ball.

4   The patient initially requires ice, regular analgesia including non-steroidal anti-inflammatory drugs, and rest (non-weight bearing for about four weeks). Surgery is rarely indicated.

ANATOMY QUIZ

Axial computed tomography of a male pelvis in the portal venous phase post-contrast

A: Right femoral artery

B: Right femoral vein

C: Right sciatic nerve

D: Right gluteus maximus muscle

E: Left rectus abdominis muscle

F: Urinary bladder

G: Left seminal vesicle

H: Rectum

I: Left sartorius muscle

STATISTICAL QUESTION

What is a P value?

Statement *c* best describes the P value.