Department of Health Sciences

M.Sc. in Evidence Based Practice, M.Sc. in Health Services Research

Meta-analysis: publication bias

Publication bias

One of the great problems with systematic review is that not all studies carried out are published. Those which are published may be different from those which are not. Research with statistically significant results is more likely to be submitted and published than work with null or non-significant results. Research with statistically significant results is likely to be published more prominently than work with null or non-significant results, for example in English, in higher impact journals. To make things worse, well designed and conducted research is less likely to produce statistically significant results than badly designed and conducted research. Combining only published studies may lead to an over-optimistic conclusion.

In this lecture we shall look at how to identify publication bias from the available data and what, if anything, we can do about it.

Graphical methods for identifying publication bias

The main graphical method for identifying publication bias is the use of funnel plots. A **funnel plot** is a plot of effect size against sample size or some other indicator of the precision of the estimate.

To illustrate funnel plots I have used simulated data, where we know that there is no publication bias. I generated 50 simulated studies of varying sample size where the true effect size was = 0.5.

Figure 1 shows a funnel plot of effect against sample size. If no bias is present this should be symmetrical about the true population effect size and get narrower as the sample size increases. This is said to be shaped like a funnel, hence the name. It is shaped like a funnel in the same way the Normal distribution curve is shaped like a bell, i.e. not really. 95% of studies should lie within the two limit lines. Usually funnel plots do not show these because they depend on the population and in particular on the true effect size, which we do not usually know. As this is a simulation, they can be included.

Figure 2 shows a funnel plot of effect against standard error. The boundaries are now straight lines. This is probably the easiest version to use, but there are plenty of others.

Figure 3 shows a funnel plot of effect against 1/standard error. The boundaries become curves again, but the curvature is so dramatic as in Figure 1.

Figure 4 shows a funnel plot of effect against meta-analysis weight. This is almost identical to Figure 1.

Figure 5 shows a funnel plot turned the other way round, plotting meta-analysis weight against effect size. It is just Figure 4 rotated through 90^o.

Figure 1. Funnel plot: effect against sample size for 50 simulated studies of varying sample size where there is no publication bias

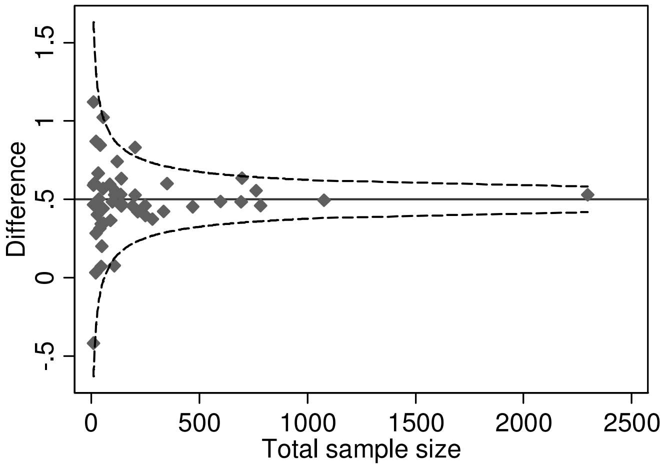


Figure 2. Funnel plot: effect against standard error for 50 simulated studies of varying sample size where there is no publication bias

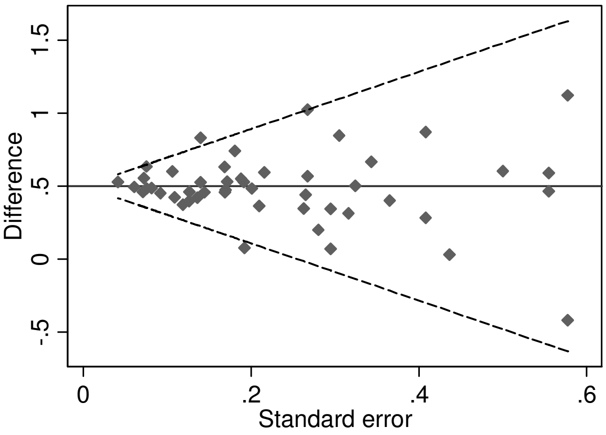


Figure 3. Funnel plot: effect against 1/standard error for 50 simulated studies of varying sample size where there is no publication bias

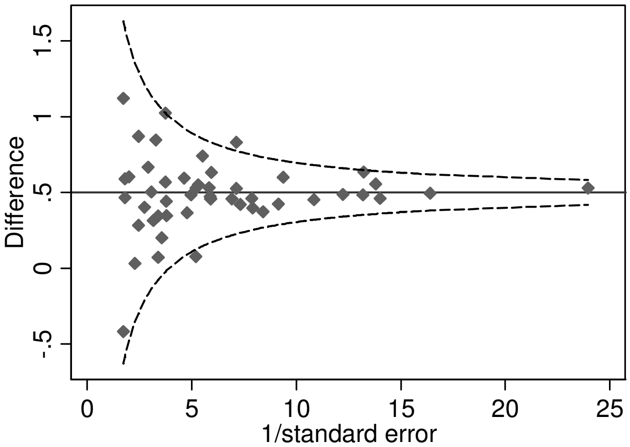


Figure 4. Funnel plot: effect against meta-analysis weight for 50 simulated studies of varying sample size where there is no publication bias

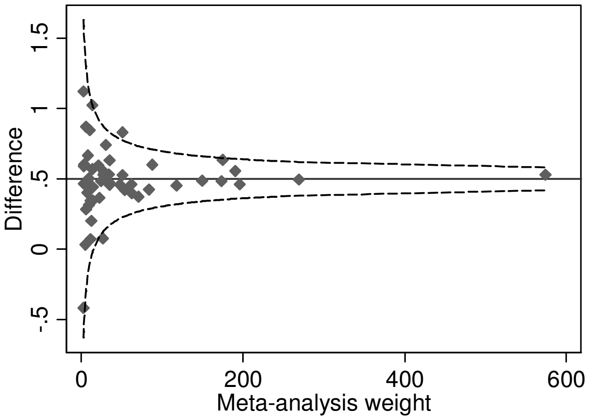


Figure 5. Funnel plot: meta-analysis weight against effect size for 50 simulated studies of varying sample size where there is no publication bias

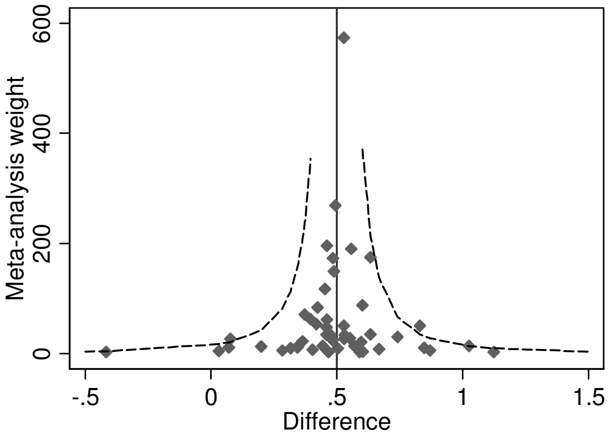
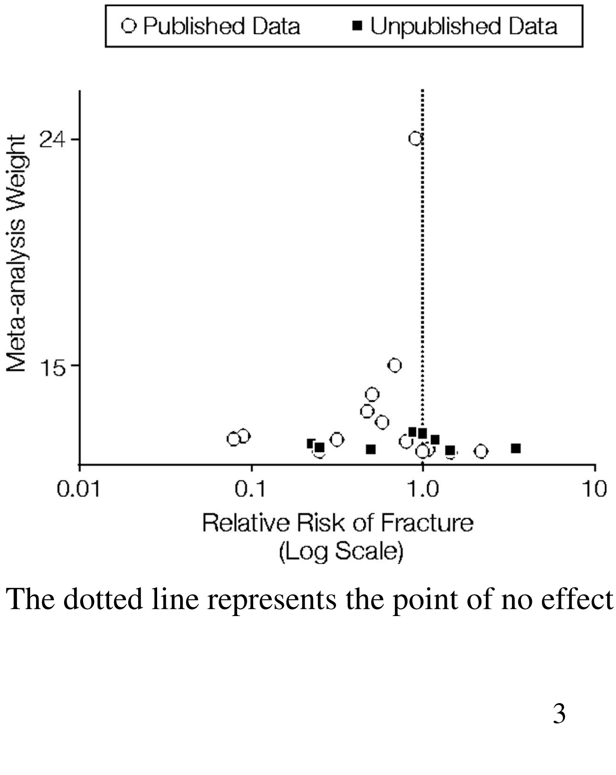


Figure 6. Funnel plot for trials of hormone replacement therapy for the prevention of nonvertebral fractures (Torgerson and Bell-Syer 2001)



The dotted line represents the point of no effect.

Figure 6 shows a funnel plot for some real data, from a meta-analysis of hormone replacement therapy and the prevention of nonvertebral fractures. This is in the same style as Figure 5.

If only significant studies are published, part of the funnel will be sparse or empty. In Figure 6, the unpublished studies which have been identified are small, low weight studies and appear to have mean relative risk closer to 1.0 that the published studies of similar size.

We can simulate publication bias by dropping the studies which do not produce significant results. Figure 7 shows the funnel plot of Figure 2 with the studies which were not significant shown as open symbols. In this version of the funnel plot, small studies are on the right, because they have large standard errors.

If studies where the difference is not significant are not published, we won't see them. The funnel plot will not look like Figure 2. We won't have the guide lines in a real study, either, because they depend on the true effect size. It would look like Figure 8. The asymmetry about any possible horizontal line is fairly obvious here.

Significance tests for publication bias

There have been at least two attempts to produce significance tests to identify publication bias:

- 'Begg's test' (Begg and Mazumdar 1994)
- 'Egger's test' (Egger *et al.*, 1997)

Both ask: 'Is the study estimate related to the size of the study?'

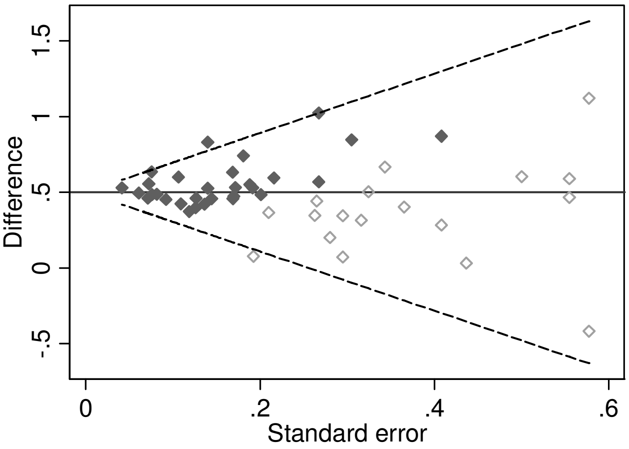
Begg's test starts with a funnel plot. We shall illustrate it using the systematic review of trials of corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004). Figure 9 shows a funnel plot. Is the study estimate (log odds ratio in this example) related to the size of the study? We look for a correlation between the log odds ratio and the meta-analysis weight.

There is a problem: the variance of the effect estimate is not the same for all points. Begg's solution is to divide each estimate by its standard error. To be more precise, Begg subtracts the pooled estimate first then divides by SE of the deviation, which makes it a bit more complicated. The effect is to give us estimates which have the same variance. The division by standard error makes it very similar to the Galbraith plot, described in 'Meta-analysis: dealing with heterogeneity', which is used in the Egger test, described below.

Now we can find a correlation between the adjusted effect size and the meta-analysis weight. This is equivalent to looking whether there is a correlation with sample size. We could use any suitable variable on the X axis (SE, 1/SE, etc.). Begg uses Kendall's rank correlation coefficient, rather than the ordinary product moment correlation, because Normal distributions are unlikely here. For Figure 10, tau b = 0.09, P = 0.7, so there is no evidence of publication bias using this test.

A problem identified with this test is that the power is very low if there are small numbers of studies. Begg and Mazumdar (1994) say that their test is 'fairly powerful with 75 studies, moderate power with 25 studies'. Even 25 studies is a pretty large meta-analysis.

Figure 7. Funnel plot: effect against standard error for 50 simulated studies of varying sample size where there is no publication bias, with significant and not significant effects identified



Open diamonds are studies where the difference is not significant.

Figure 8. Funnel plot: effect against standard error for 50 simulated studies of varying sample size where there is no publication bias, significant effects only

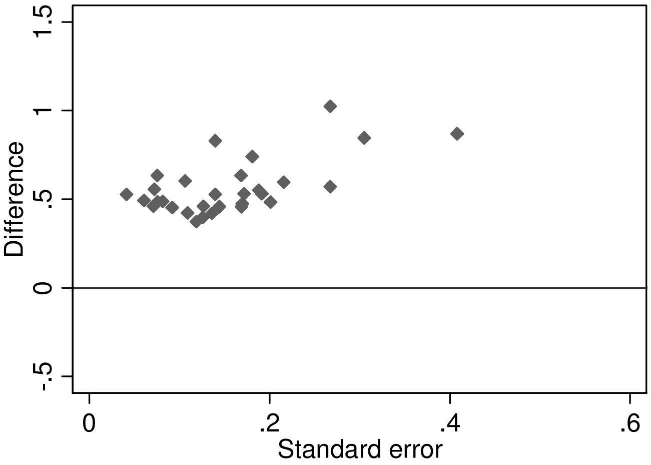
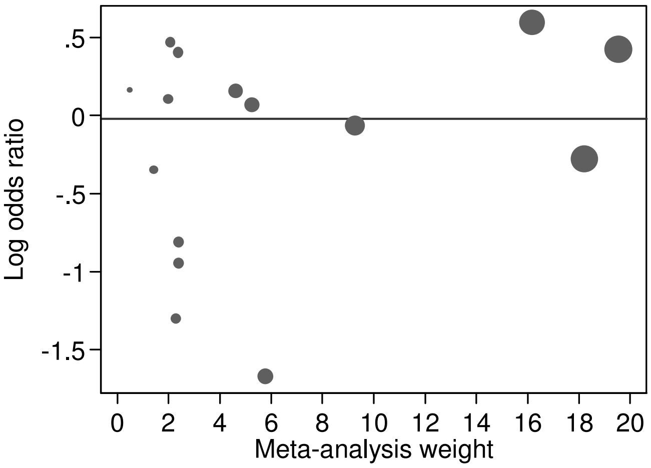
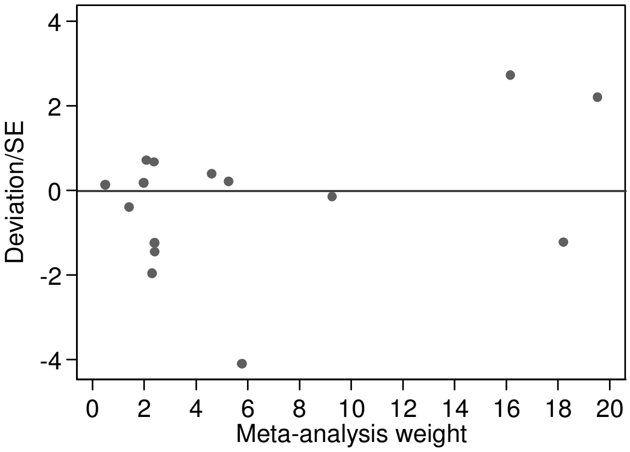


Figure 9. Funnel plot for trials of corticosteroids for severe sepsis and septic shock



Area of circles is proportional to meta-analysis weight, 1/variance of the estimate

Figure 10. Funnel plot for trials of corticosteroids for severe sepsis and septic shock after subtracting the pooled estimate from the observed effect sizes and dividing by the standard error



The alternative, Egger's test, is based on the Galbraith plot. This is a plot of difference over standard error against one over standard error. Figure 11 shows the Galbraith plot for the corticosteroid trials. Egger's suggestion is that we carry out a regression rather than a correlation. He suggests that we calculate the regression of study difference (log odds ratio in this case) over standard error on 1/standard error. We have seen that this line theoretically should go through the point (zero, zero), the origin. Does it? Clearly not in Figure 12, but it is very unlikely to do so because of random variation. Is it further from zero than we would expect by chance? Egger suggests that we test the null hypothesis that the intercept is equal to zero in the population.

For Figure 12, the regression line is D/SE = –1.14 + 0.39×1/SE. The intercept is –1.14, which has standard error = 0.88, P = 0.22, 95% CI = –3.05 to 0.77. Any regression program will print this out for you. So, again we have no evidence for publication bias.

This regression ignored the different variances of the observations. Should we weight the observations before carrying out the regression. Egger *et al.* (1997) say:

'In some situations (for example, if there are several small trials but only one larger study) power is gained by weighting the analysis by the inverse of the variance of the effect estimate.

'We performed both weighted and unweighted analyses and used the output from the analysis yielding the intercept with the larger deviation from zero.'

The weighted regression gives D/SE = –2.01 + 0.67×1/SE, the intercept is not significantly different from zero, P = 0.17.

Is this test biased? I think so for two reasons. First, doing both regressions and choosing the more significant is multiple testing. Second, the regression intercept is a biased estimate. When there is error in the X variable, this stretches out the horizontal scale and makes the slope less steep than it would be if there were no error. This has the effect of making the intercept move away from the value it would have if there were no error in X. If it should be zero, the error in X prevents it from being zero. This is called the regression dilution effect. We are therefore testing a null hypothesis which we should not expect to be true. I think that Egger's test is suspect and would not use it.

For an example, in their meta-analysis of the effect of breast feeding in infancy on blood pressure in later life (Owen *et al.*, 2003) the authors published a funnel plot (Figure 14). There is considerable heterogeneity, shown by the points outside the lines, but nothing to suggest that there are missing studies. Owen *et al.* (2003) reported that 'The Egger test was significant (P = 0.033) for publication bias but not the Begg test (P = 0.186)'. The Begg test is non-parametric, which reduces its power, but even so the discrepancy is large. I think the Egger test is not to be trusted.

Figure 11. Galbraith plot for trials of corticosteroids for severe sepsis and septic shock

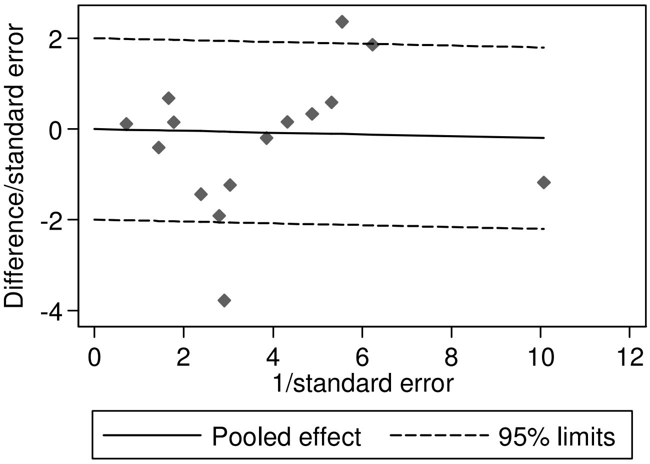


Figure 12. Regression of difference/standard error on 1/standard error

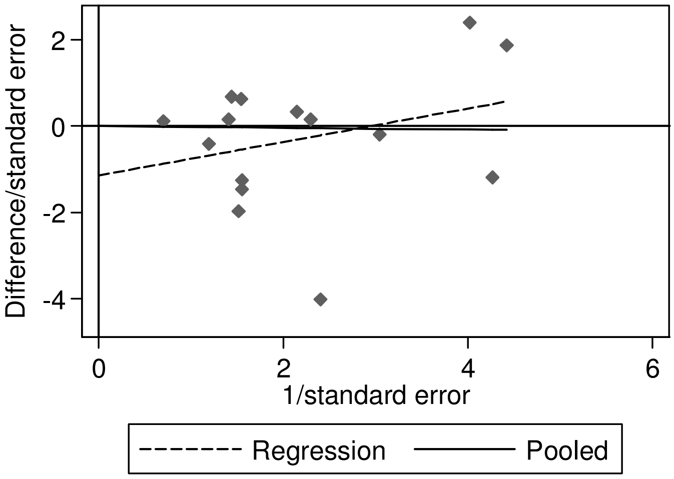


Figure 13. Weighted and unweighted regression lines of difference/standard error on 1/standard error

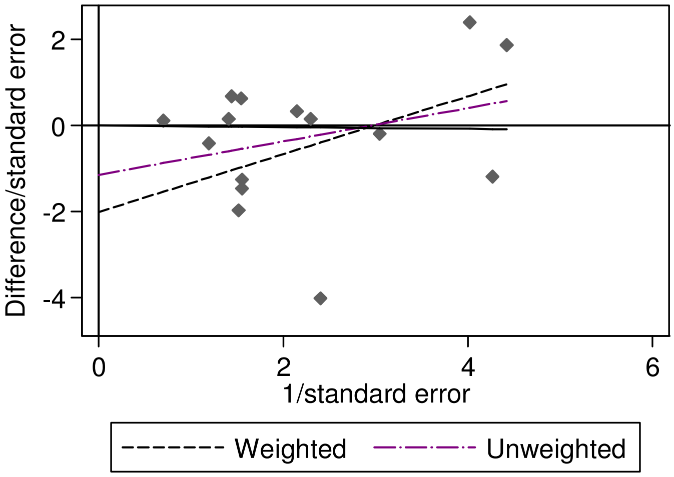
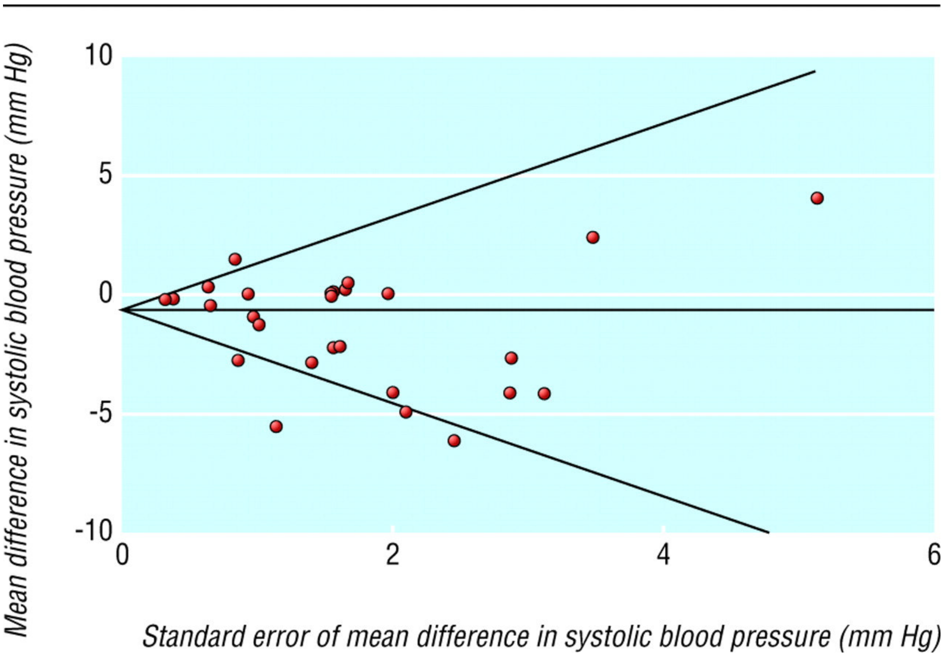


Figure 14. Begg's funnel plot (pseudo 95% confidence limits) showing mean difference in systolic blood pressure by standard error of mean difference (Owen *et al.*, 2003)



Dealing with publication bias

Can we correct for the effects of publication bias using the data we have? Several approaches can be tried, including

- trim and fill,
- selection models,
- meta-regression.

In the trim and fill method, there are two stages. First we trim the data, meaning that we eliminate studies, starting with the least powerful, until we have symmetry in the funnel plot. From the remaining studies we get a new pooled estimate. Second, we fill in the holes we think we have identified: for the studies eliminated, we reflect them in the pooled estimate line and put in new studies.

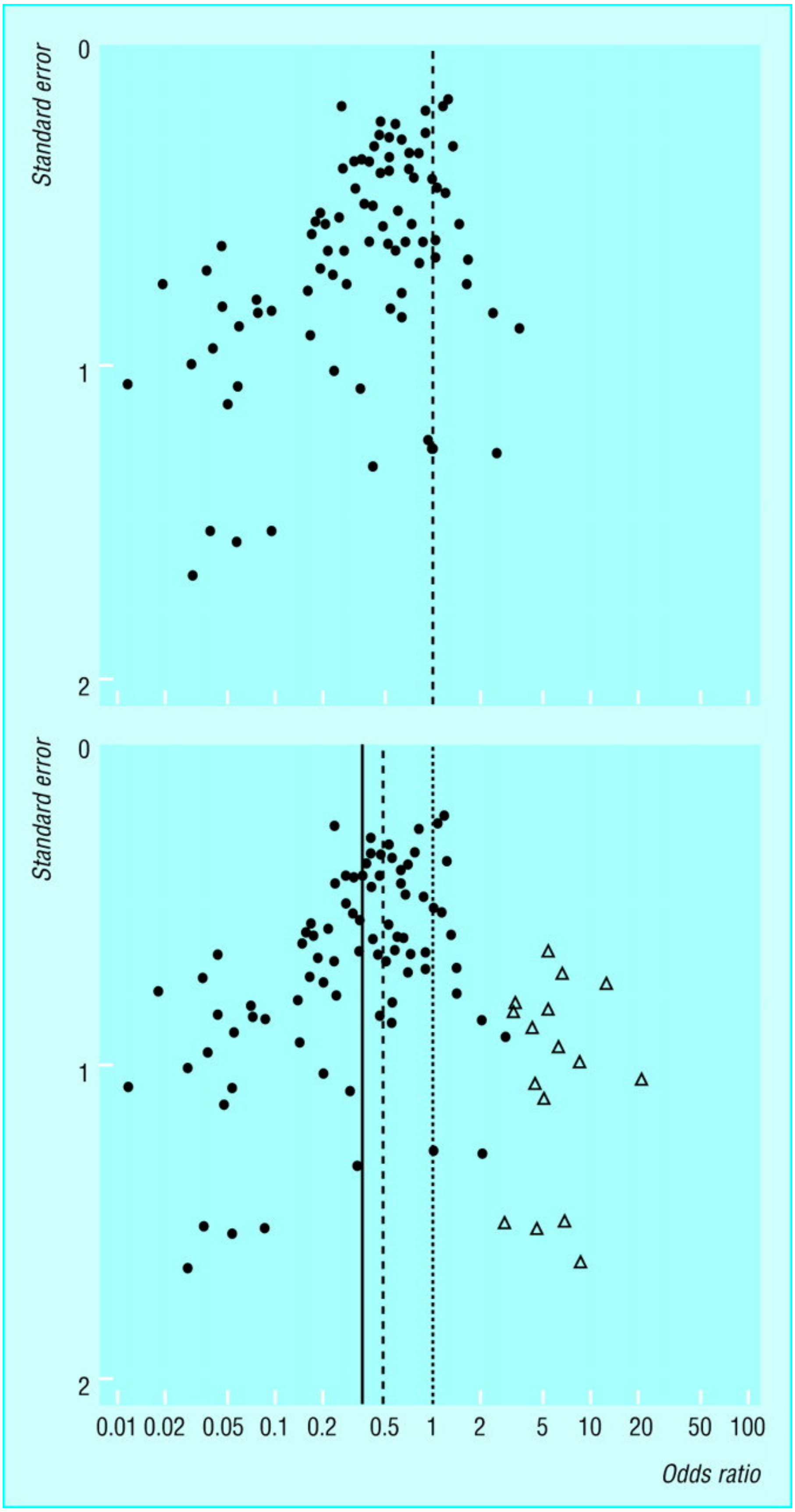
For example, Sterne *et al.* (2001) carried out a trim and fill analysis of 89 trials comparing homeopathic medicine with placebo. Figure 15 shows their analysis. The upper panel is the funnel plot, a vertically oriented plot showing standard error against odds ratio (on a logarithmic scale). I have no idea why they chose to plot the vertical axis with ascending numbers going downwards rather than upwards in the usual way. There is clear asymmetry about any vertical line we cared to draw, suggesting considerable publication bias. The dotted line is the no effect line, where the odds ratio is 1.0.

The lower panel shows the trim and fill procedure. The solid vertical line is the pre trim and fill pooled estimate. Starting at the bottom of the graph and moving upwards, there is clear asymmetry, as there are trials with odds ratios well to the left of the this line with no counterparts equally far to the right. We start dropping them out and recalculating the pooled estimate. We go on doing this until we have symmetry, achieved in Figure 15 when the standard error gets down to about 0.6. The final pooled estimate is shown by the broken line. We then put back the eliminated trials and we try to put in the missing trials whose existence we have deduced. The open triangles in Figure 15 are filled trials. For each eliminated trial, we invent another trial of the same size an equal distance from the post trim and fill pooled estimate, but on the other side.

Simulation studies have found that the trim and fill method detects ‘missing’ studies in a substantial proportion of meta-analyses in the absence of bias. Application of trim and fill could mean adding and adjusting for non-existent studies in response to funnel plot asymmetry arising from nothing more than random variation (Sterne *et al.*, 2001). Clearly, you need a lot of studies to attempt anything like this.

Do people really do this? I could find only five examples in a literature search.

Figure 15. Trim and fill analysis of 89 trials comparing homeopathic medicine with placebo (Sterne *et al.*, 2001).



A second approach which has been proposed is the use of selection models. We try to model the selection process that determines which results are published. The method is based on the assumption that the study's P value affects its probability of publication. However, many factors may affect the probability of publication of a given set of results, and it is difficult, if not impossible, to model these adequately. This approach is not widely used. I could find no examples at all in a quick search.

A third approach is to use study characteristics, e.g. Jadad score, sample size, to predict outcome. Regression methods are used here and would usually be described as meta-regression. An example occurred in the Owen *et al.* (2003) study of breast feeding and blood pressure:

‘The estimate of effect size decreased with increasing study size: –2.05 mm Hg in the 13 studies with fewer than 300 participants, –1.13 mm Hg in the seven studies (nine observations) with 300 to 1000 participants, and –0.16 mm Hg in the four studies with more than 1000 participants (test for trend between groups P = 0.046). However, a test for trend with study size treated as a continuous variable, was not significant (P = 0.209).’ (Owen *et al.*, 2003)

This approach is much more widely used. A note of caution should be sounded. These methods require large numbers of studies. They are not powerful in most meta-analyses. Furthermore, a relationship between trial outcome and sample size may not result from publication bias. Small trials may differ in nature, e.g. have more intensive treatment or treatment by more committed clinicians (i.e. more committed to the technique, not to their work!) Furthermore, publication bias may not result from significance or sample size. Researchers or sponsors may simply not like the result. Most healthcare researchers are amateurs with other demands on their attention (e.g. their patients). It is easy for them not to publish their work. It is better to think of these methods as a way of exploring possibilities than to produce definitive answers.

For a final example, we return to the comparison of homeopathy versus placebo (Sterne *et al.*, 2001). These authors carried out regression of the trial effect on **asymmetry coefficient**, **language English/other**, allocation concealment, **blinding**, handling of withdrawals, whether journal indexed by Medline (**bold** were significant).

Sterne *et al.* (2001) report that

‘The largest trials of homoeopathy (those with the smallest standard error) that were also double blind and had adequate concealment of randomisation show no effect.’

‘The evidence is thus compatible with the hypothesis that the clinical effects of homoeopathy are completely due to placebo and that the effects observed . . . are explained by a combination of publication bias and inadequate methodological quality of trials.’

‘We emphasise, however, that these results cannot prove that the apparent benefits of homoeopathy are due to bias.’

Martin Bland

6 March 2006

References

Annane D, Bellissant E, Bollaert PE, Briegel J, Keh D, Kupfer Y. (2004) Corticosteroids for severe sepsis and septic shock: a systematic review and meta-analysis. *British Medical Journal*, **329**, 480.

Begg CB, Mazumdar M. (1994) Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50**, 1088-1101.

Egger M, Smith GD, Schneider M, Minder C. (1997) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629-634.

Owen C, Whincup PH, Gilg JA, Cook DG. (2003) Effect of breast feeding in infancy on blood pressure in later life: systematic review and meta-analysis. *British Medical Journal*, **327**, 1189-1195.

Sterne JAC, Egger M, Smith GD. (2001) Systematic reviews in health care - Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal* **323**, 101-105.

Torgerson DJ, Bell-Syer SEM. (2001) Hormone replacement therapy and prevention of nonvertebral fractures. A meta-analysis of randomized trials. *JAMA* **285**, 2891-2897.