Department of Health Sciences

M.Sc. Module: Systematic Reviews

Meta-analysis: dealing with heterogeneity

Dealing with heterogeneity

We have already discussed the meaning and detection of heterogeneity in the previous lecture, 'Meta-analysis: methods for quantitative data synthesis'. In this lecture we look at how to deal with it when we have it. There are a number of possibilities.

First, we could decide not to pool the study estimates at all. Instead, we would carry out a narrative review. We do not get any numerical estimate.

Second, we could ignore the heterogeneity and analyse the data as described in the previous lecture. We would use what is called a **fixed effect model**, assuming that the underlying effects are the same for all studies. As we shall see, this can result in a confidence interval which is too narrow, a pooled estimate which is difficult to interpret, and which may be biased.

Third, we could explore the heterogeneity and try to explain it and remove it. We may be able to find a variable or variables which explains this heterogeneity and so give our meta-analysis estimate depending on this variable.

Fourth, we could allow for the heterogeneity in our analysis and produce a much wider confidence interval, using what is called a random effects model.

We shall look at all these options below.

Measuring heterogeneity

First, we ask how much heterogeneity is there? The chi-squared test provides a test of significance for heterogeneity, but it does not measure it. An index of heterogeneity can be defined as I^2 (Higgins and Thompson 2002), where

$$I^2 = 100 \times \frac{X^2 - df}{X^2}$$

and X^2 is the chi-squared heterogeneity statistic with *df* degrees of freedom. If I^2 is negative we set it to zero.

The value which we expect chi-squared to have if there is no heterogeneity is equal to its degrees of freedom. Hence I^2 is the percentage of the chi-squared statistic which is not explained by the variation within the studies. It represents the percentage of the total variation which is due to variation between studies.

I^2 without the 100 is essentially an intraclass correlation coefficient.

For interpreting I^2, Higgins *et al.* (2003) suggest:

- $I^2 = 0\% \rightarrow$ no heterogeneity,
- $I^2 = 25\% \rightarrow$ low heterogeneity,
- $I^2 = 50\% \rightarrow$ moderate heterogeneity,
- $I^2 = 75\% \rightarrow$ high heterogeneity.

These are arbitrary, except for 0%. I^2 can never reach 100% and values above 90% are very rare.

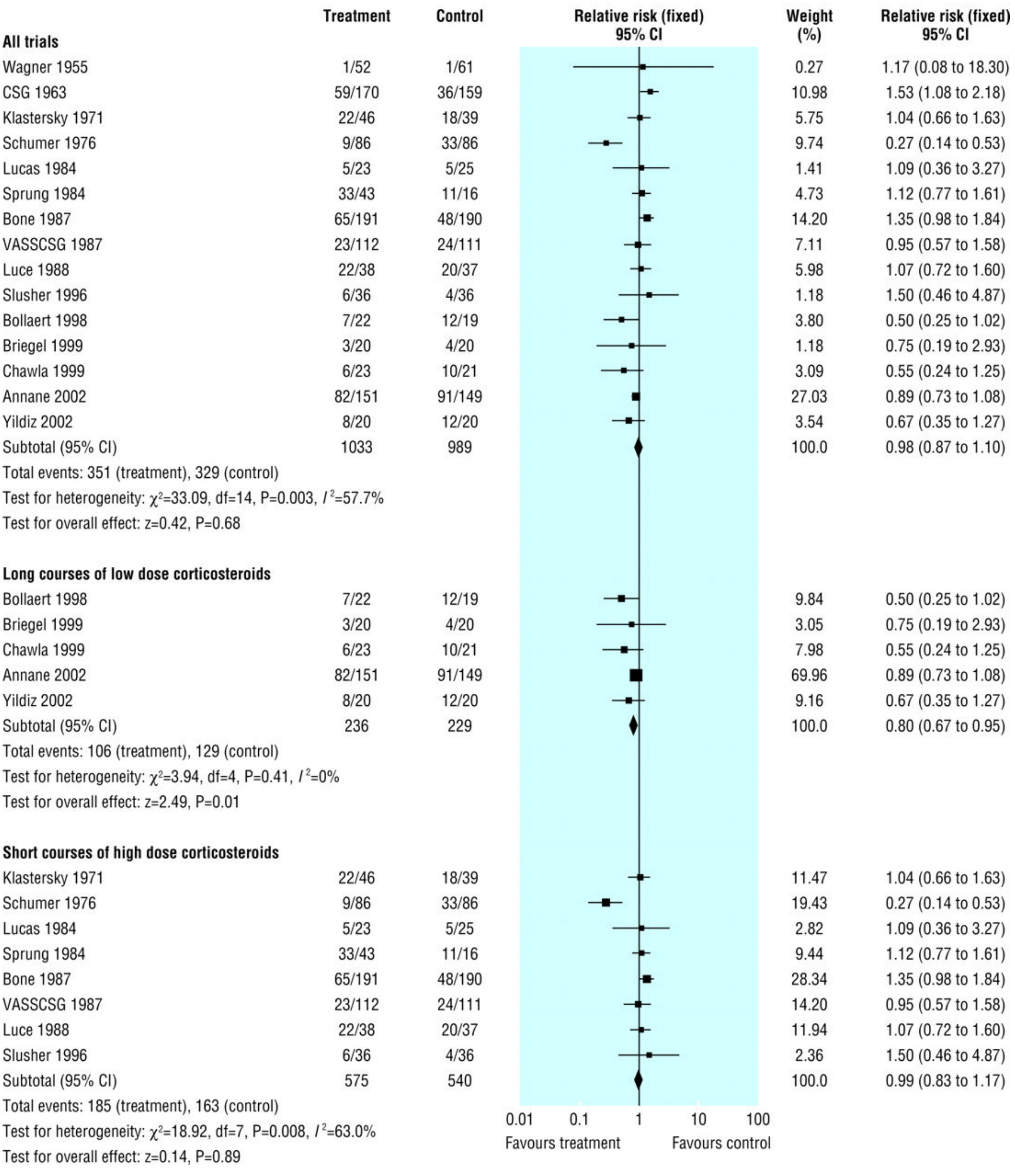
Investigating sources of heterogeneity

Heterogeneity comes about because the effects in the populations which the studies represent are not the same. We can look for possible explanations of this in variations in study characteristics. For example, there may be subsets of studies within which there is little heterogeneity. These may be defined by different subsets of subjects, such as hospital and community subjects, or variations in treatments, such as different antibiotics being compared to none, or different study designs, cohort or case-control studies, or cross-over or parallel groups trials. These subsets should be pre-specified, if possible, so as to avoid bias.

Figure 1 shows a meta-analysis for trials of corticosteroids for the treatment of severe sepsis and septic shock. The authors found moderate and highly significant heterogeneity, I^2 = 57.7%, P = 0.003. They split the trials according to type of treatment and found that long courses of low dose corticosteroids produced no evidence of heterogeneity (I^2 = 0%, P = 0.4) and good evidence for an effect on outcome, whereas another group of trials with short courses of high dose corticosteroids produced evidence of heterogeneity (I^2 = 63.0%, P = 0.008) and no consistent evidence of any effect on outcome, despite one trial reporting a substantial effect.

We may try to relate the size of the effect to characteristics of the studies and their subjects, such as average age, proportion of females, intended dose of drug, or baseline risk. For example, Figure 2 shows the percentage reduction in risk of ischaemic heart disease (and 95% confidence intervals) associated with 0.6 mmol/l serum cholesterol reduction in 10 prospective studies of men (Thompson 1994). The heterogeneity is obvious in the forest plot as many of the confidence intervals do not overlap. It is highly significant, X^2 = 127, d.f. = 9, P<0.001. Although the I^2 statistic had not been invented at the time, it is easy to calculate, giving I^2 = 92.9%. These studies can be broken down into 26 sub-studies with fairly narrow age ranges and the percentage reduction in risk of ischaemic heart disease plotted against mean age at experiencing a coronary event (Figure 3). This shows a clear relationship, the effect of cholesterol reduction being much greater at younger ages. We can carry out a regression analysis, fitting a relationship between % reduction in mortality and age. We do this weighted for the precision of the estimate, as for the ordinary weighted average. Such a regression analysis is called **meta-regression**. (This term upsets boring pedants even more than does 'meta-analysis'. 'Meta-analytic regression' would be better, but it is too late!) If we then look at the differences between the observed effect for each study and the effect predicted by the regression, rather than the weighted average, we can test the heterogeneity after adjustment for age. There was still moderate heterogeneity, X^2 = 45, d.f. = 23, P = 0.005, I^2 = 48.8%. The heterogeneity can be seen clearly in the scatter plot, as there are confidence intervals for studies with very similar ages which do not overlap. This must be due to other differences between these groups. The situation has clearly improved, however. The conclusion of the analysis was that a decrease in cholesterol concentration of 0.6 mmol/l was associated with a decrease in risk of ischaemic heart disease of 54% at age 40, 39% at age 50, 27% at age 60, 20% at age 70, and 19% at age 80.

Figure 1. Corticosteroids for severe sepsis and septic shock: effect on all cause mortality (Annane *et al.*, 2004)



	Treatment	Control	Weight (%)	Relative risk (fixed) 95% CI
All trials				
Wagner 1955	1/52	1/61	0.27	1.17 (0.08 to 18.30)
CSG 1963	59/170	36/159	10.98	1.53 (1.08 to 2.18)
Klastersky 1971	22/46	18/39	5.75	1.04 (0.66 to 1.63)
Schumer 1976	9/86	33/86	9.74	0.27 (0.14 to 0.53)
Lucas 1984	5/23	5/25	1.41	1.09 (0.36 to 3.27)
Sprung 1984	33/43	11/16	4.73	1.12 (0.77 to 1.61)
Bone 1987	65/191	48/190	14.20	1.35 (0.98 to 1.84)
VASSCSG 1987	23/112	24/111	7.11	0.95 (0.57 to 1.58)
Luce 1988	22/38	20/37	5.98	1.07 (0.72 to 1.60)
Slusher 1996	6/36	4/36	1.18	1.50 (0.46 to 4.87)
Bollaert 1998	7/22	12/19	3.80	0.50 (0.25 to 1.02)
Briegel 1999	3/20	4/20	1.18	0.75 (0.19 to 2.93)
Chawla 1999	6/23	10/21	3.09	0.55 (0.24 to 1.25)
Annane 2002	82/151	91/149	27.03	0.89 (0.73 to 1.08)
Yildiz 2002	8/20	12/20	3.54	0.67 (0.35 to 1.27)
Subtotal (95% CI)	1033	989	100.0	0.98 (0.87 to 1.10)

Total events: 351 (treatment), 329 (control)

Test for heterogeneity: χ^2=33.09, df=14, P=0.003, I^2=57.7%

Test for overall effect: z=0.42, P=0.68

	Treatment	Control	Weight (%)	Relative risk (fixed) 95% CI
Long courses of low dose corticosteroids				
Bollaert 1998	7/22	12/19	9.84	0.50 (0.25 to 1.02)
Briegel 1999	3/20	4/20	3.05	0.75 (0.19 to 2.93)
Chawla 1999	6/23	10/21	7.98	0.55 (0.24 to 1.25)
Annane 2002	82/151	91/149	69.96	0.89 (0.73 to 1.08)
Yildiz 2002	8/20	12/20	9.16	0.67 (0.35 to 1.27)
Subtotal (95% CI)	236	229	100.0	0.80 (0.67 to 0.95)

Total events: 106 (treatment), 129 (control)

Test for heterogeneity: χ^2=3.94, df=4, P=0.41, I^2=0%

Test for overall effect: z=2.49, P=0.01

	Treatment	Control	Weight (%)	Relative risk (fixed) 95% CI
Short courses of high dose corticosteroids				
Klastersky 1971	22/46	18/39	11.47	1.04 (0.66 to 1.63)
Schumer 1976	9/86	33/86	19.43	0.27 (0.14 to 0.53)
Lucas 1984	5/23	5/25	2.82	1.09 (0.36 to 3.27)
Sprung 1984	33/43	11/16	9.44	1.12 (0.77 to 1.61)
Bone 1987	65/191	48/190	28.34	1.35 (0.98 to 1.84)
VASSCSG 1987	23/112	24/111	14.20	0.95 (0.57 to 1.58)
Luce 1988	22/38	20/37	11.94	1.07 (0.72 to 1.60)
Slusher 1996	6/36	4/36	2.36	1.50 (0.46 to 4.87)
Subtotal (95% CI)	575	540	100.0	0.99 (0.83 to 1.17)

Total events: 185 (treatment), 163 (control)

Test for heterogeneity: χ^2=18.92, df=7, P=0.008, I^2=63.0%

Test for overall effect: z=0.14, P=0.89

Figure 2. Percentage reduction in risk of ischaemic heart disease (and 95% confidence intervals) associated with 0.6 mmol/l serum cholesterol reduction in 10 prospective studies of men (Thompson 1994)

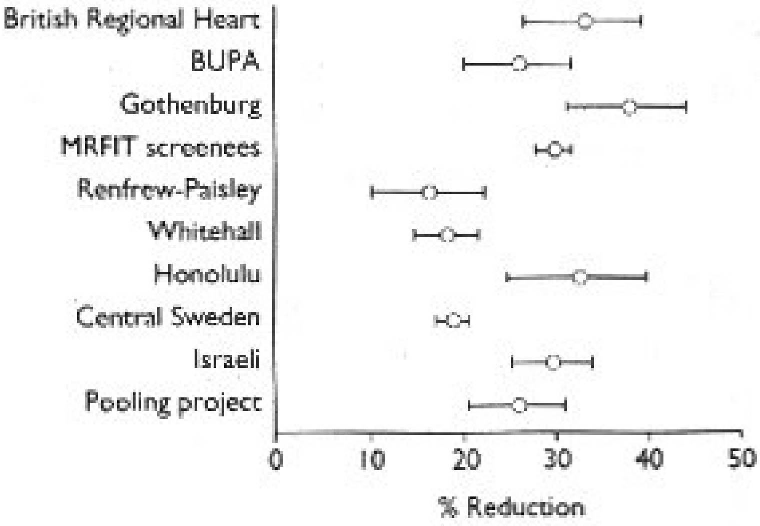


Figure 3. Percentage reduction in risk of ischaemic heart disease (and 95% confidence intervals) associated with 0.6 mmol/l serum cholesterol reduction, according to age at experiencing a coronary event (Thompson 1994)

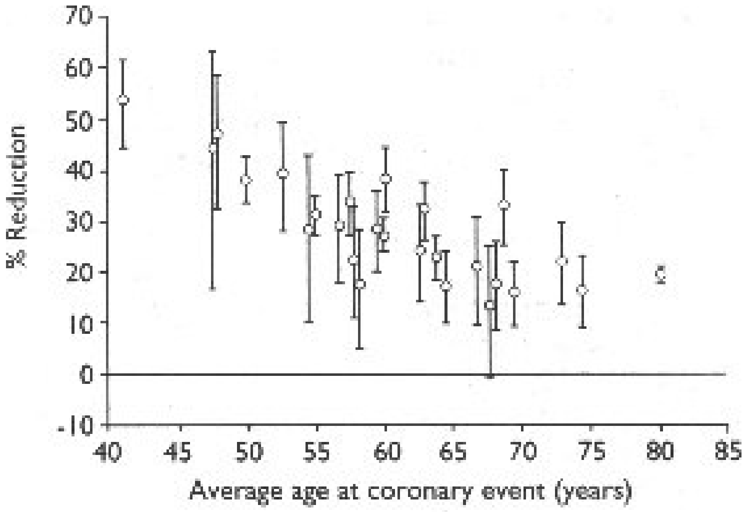
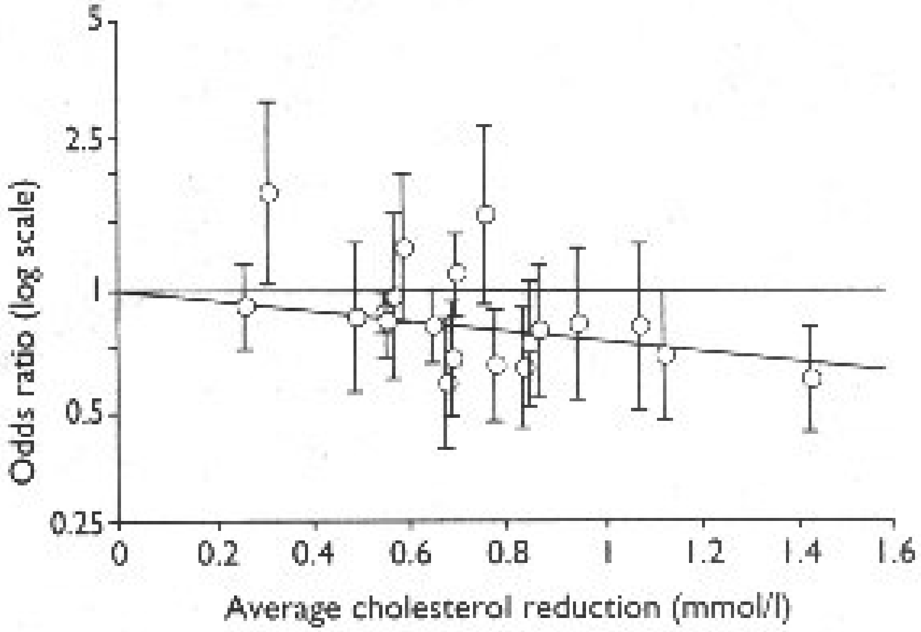


Figure 4. Odds ratios of ischaemic heart disease (and 95% confidence intervals) according to the average extent of serum cholesterol reduction achieved in each of 28 trials (Thompson 1994)



Overall summary of results is indicated by sloping line. Results of the nine smallest trials have been combined.

Figure 5. Galbraith plot for log odds ratio of death for corticosteroids in patients with severe sepsis and septic shock (Annane *et al.*, 2004), trials of treatments with low doses and long duration.

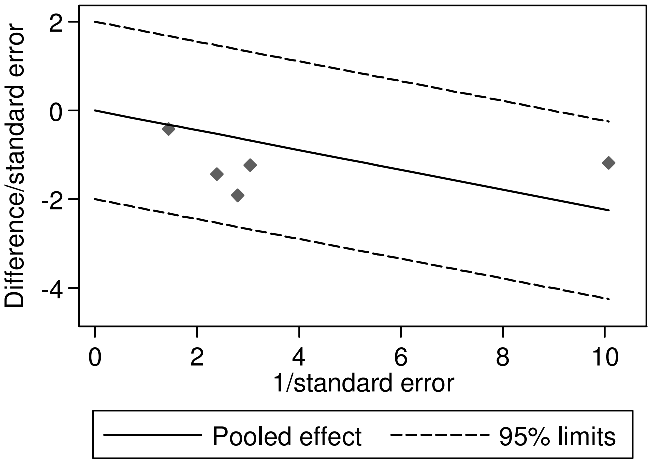


Figure 4 shows another example, explaining why some studies of cholesterol lowering interventions produced no reduction in the risk of ischaemic heart disease and others did. Those trials where the intervention produced little reduction in serum cholesterol produced no discernable effect, whereas interventions which were successful in lowering cholesterol were also successful in reducing heart disease risk.

The Galbraith plot

The Galbraith plot is an alternative to a forest plot as a graphical representation of the study data. On the horizontal axis we plot 1/standard error of the study effect estimate. The horizontal axis will be zero if standard error is infinite, a study of zero size. This cannot happen, so there should never be a point actually at zero. On the vertical axis we plot the study effect estimate divided by its standard error. This is the test statistic for the individual study. For 95% of studies, we expect this to be within 2 units of the true or population effect, because we expect 95% of studies to have effect estimates within two standard errors of the population effect. If the horizontal axis variable were zero, the standard error would be infinite and so the vertical axis, which is effect divided by standard error, would be zero also.

For example, Figure 5 shows a Galbraith plot for the log odds ratio of death for corticosteroids in patients with severe sepsis and septic shock for the group of trials of treatments with low doses and long duration. These were the trials in which there was no significant heterogeneity. We can add a line representing the pooled effect. This is a straight line going through the point (zero, zero), i.e. of the form

effect/se = (pooled effect) × 1/se

I.e. the slope of the line is equal to the pooled effect. The 95% limits will be 2 units above and below this line and we can add these to the plot, too. We expect 95% of points to be between these limits if there is no heterogeneity. This is true for the low dose, long duration trials.

Figure 6 shows the Galbraith plot for all the corticosteroid trials, where there was significant heterogeneity. The pooled effect is smaller so the line is less steep. We have two points outside the 95% limits and one on the line. This is what we would expect given the presence of significant heterogeneity. We can investigate them to see how these trials differ from the others.

We could reanalyse taking dosage and duration separately, as shown in Figure 7. The trials which stand out as producing heterogeneity are clearly seen to be not the trials of low dose long course treatments.

Is a Galbraith plot preferable to a forest plot? Thompson (1994) wrote “Conventional meta-analysis diagrams . . . are not very useful for investigating heterogeneity. A better diagram for this purpose was proposed by Galbraith . . .”. Is this really true? Figure 8 shows the Galbraith plot for the corticosteroid trials with the trials identified and the forest plot with a vertical line drawn through the pooled estimate. Trials outside the Galbraith limits will be trials where the 95% confidence interval does not contain the pooled estimate. We can spot them from the forest plot.

Figure 6. Galbraith plot for log odds ratio of death for corticosteroids in patients with severe sepsis and septic shock (Annane *et al.*, 2004), all trials.

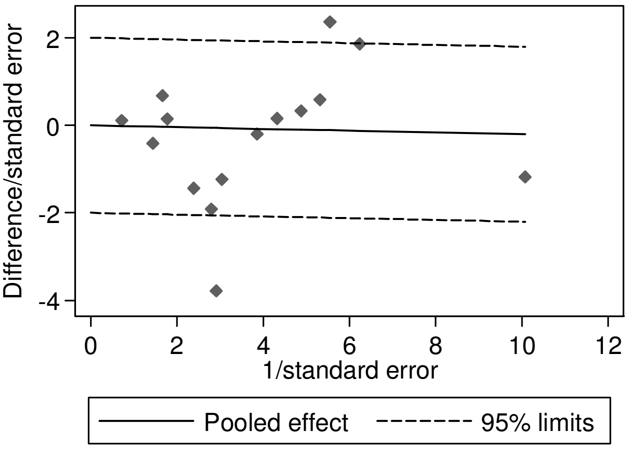


Figure 7. Galbraith plot for log odds ratio of death for corticosteroids in patients with severe sepsis and septic shock (Annane *et al.*, 2004), all trials.

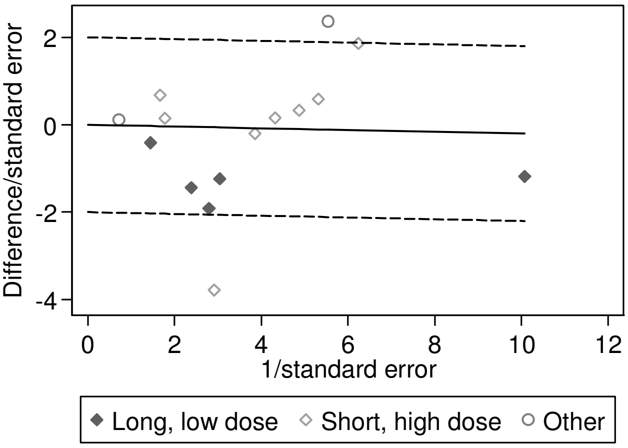


Figure 8. Galbraith and forest plots for corticosteroids in patients with severe sepsis and septic shock, all trials, with trials identified and a vertical line through the pooled estimate.

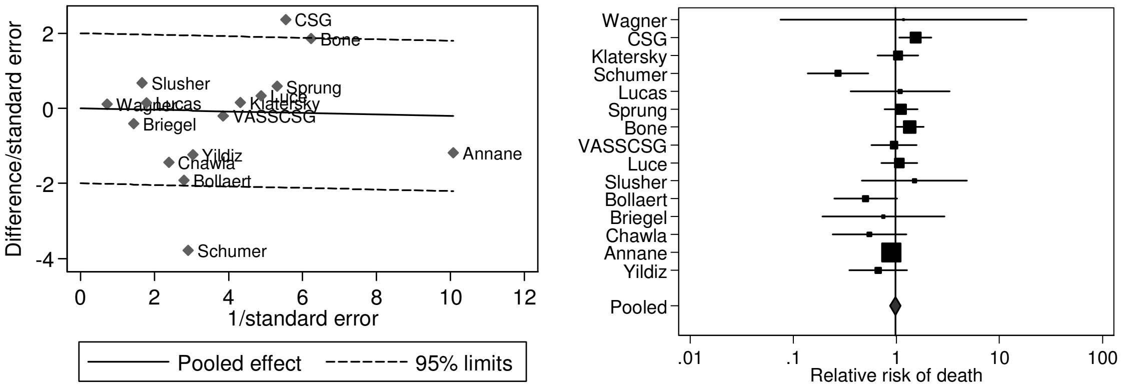
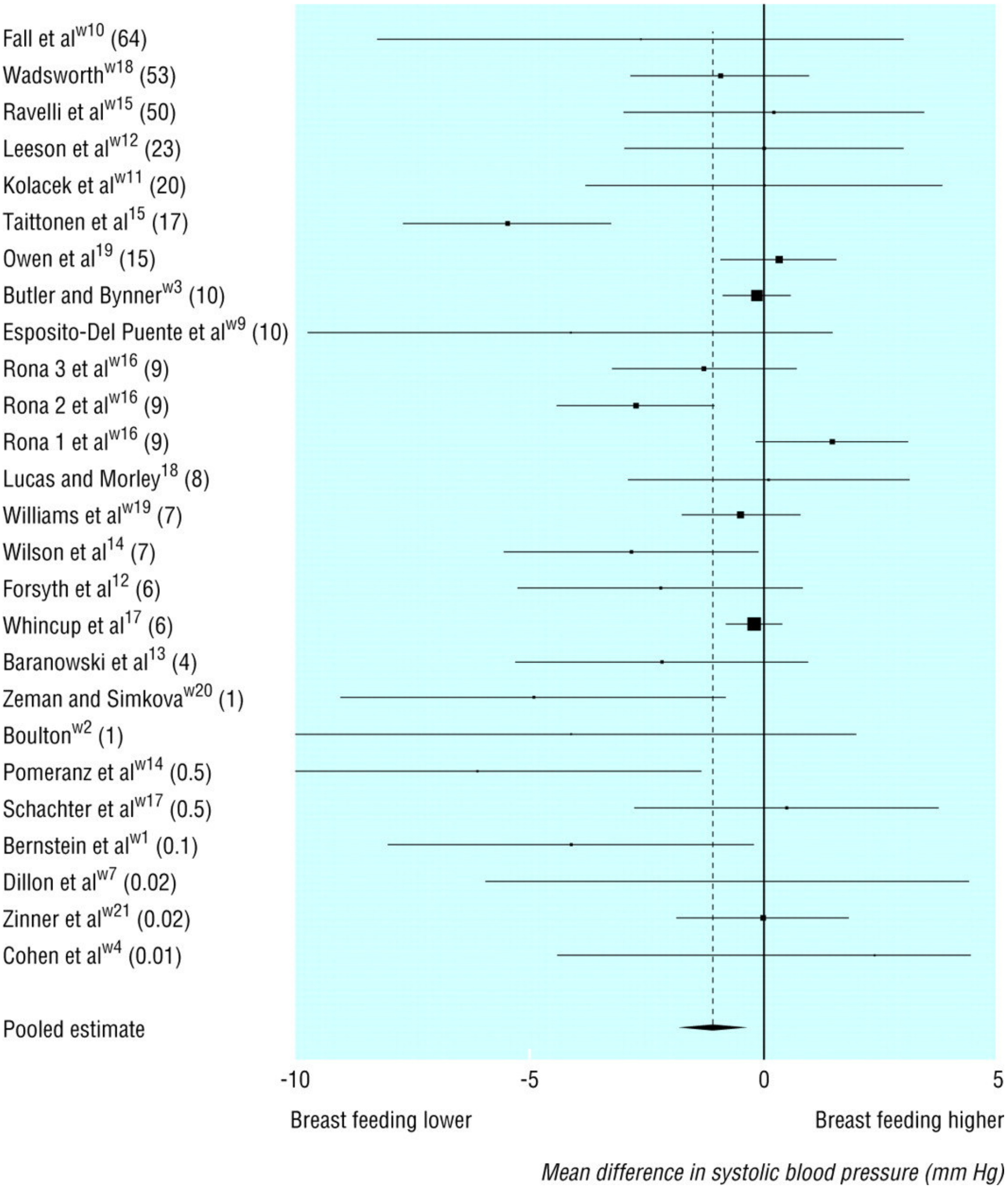


Figure 9. : Effect of breast feeding in infancy on blood pressure in later life (Owen *et al.*, 2003)



(In parenthesis: age in years at which blood pressure measured. 0.5 represents 6 months.)

Random effects models

We cannot always explain heterogeneity. For example, (Owen *et al.*, 2003) carried out a review of the effect of breast feeding in infancy on blood pressure in later life (Figure 9). Although there is clear heterogeneity, the authors were unable to explain it. The obvious candidate explanatory variable, the age at which the blood pressure was measured, was unable to explain the heterogeneity. Under these circumstances, we have to accept the existence of the heterogeneity and say that the greater uncertainty which this adds to our estimate should be reflected in the method of estimation and calculation of the confidence interval. We do this using a **random effects model**, where we regard each study as estimating a different effect. The study effects for all the studies which could be done form a population, of which the studies actually carried out are a sample. The mean of this population will be our best measure of the overall effect. We estimate both the variability between subjects within the studies and the variation between studies. The usual method of analysis assumes that all the studies are estimating the same effect and only random variation between research subjects leads the observed study effects to vary. We call this the **fixed effects model**. We assume that the effect to be estimated is the same in all studies and use only the sampling variation within the studies.

The two models can be compared as follows:

Fixed effects model	**Random effects model**
We assume that the effect to be estimated is the same in all studies.	We assume that the effect is the not same in all studies. The studies are a sample of possible of studies where the treatment effect varies.
We use only the sampling variation within the studies.	We use the sampling variation within the studies and the sampling variation between studies.
If the effect is the same in all studies, a fixed effects model is more powerful and easier.	If the effect is the same in all studies, less powerful because P values are larger and confidence intervals are wider.
No assumption about representativeness	The studies are a sample from a population of possible of studies where the effect varies. They must be a representative or random sample. Very strong assumption.

Fixed effects model

Variance of effect in study = standard error squared.

Weight = 1/variance

$$= 1/\mathrm{SE}^2$$

When heterogeneity exists we get:

- a pooled estimate which may give too much weight to large studies,
- a confidence interval which is too narrow,
- a P-value which is too small.

When heterogeneity does not exist:

- a pooled estimate which is correct,
- a confidence interval which is correct,
- a P-value which is correct.

Random effects model

Variance of effect in study = standard error squared plus inter-study variance

Weight = 1/variance.

$$= \frac{1}{\mathrm{SE}^2 + \text{inter-study variance}}$$

Inter-study variance has degrees of freedom given by number of studies minus one. Typically small.

When heterogeneity exists we get:

- possibly a different pooled estimate with a different interpretation,
- a wider confidence interval,
- a larger P-value.

When heterogeneity does not exist:

- a pooled estimate which is correct,
- a confidence interval which is too wide,
- a P-value which is too large.

Using a random effects model affects not only the confidence interval but also the estimate itself. Figure 10 show two Cochrane meta-analyses of the same trials, using fixed and random effects models. The data are shown in Table 1. There is a lot of heterogeneity (I^2 = 74%, P = 0.002) and the random effects method would be preferred. The fixed effect model shows a marginally significant effect in favour of control, whereas the random effects model produces a non-significant effect in favour of active treatment. The fixed effect model would be misleading here.

On the other hand, if there is no heterogeneity, although the estimate will be the same the confidence interval will be wider and the random effects method may be misleading.

Table 1. Raw data for a meta-analysis of oral rehydration in cholera, reduced osmolarity versus standard, duration of diarrhea

	Intervention			Control		
Study	n_1	$mean_1$	s_1	n_2	$mean_2$	s_2
1.	82	44.4	13.3	78	42.7	13.5
2.	34	49.9	18.7	29	57.1	17.9
3.	33	37.2	9.9	30	46.9	11.9
4.	147	46.0	18.2	153	43.0	18.6
5.	19	21.44	1.32	16	19.97	1.99
6.	19	33.89	16.4	20	38.47	17.4
7.	26	82.9	27.5	32	78.6	24.5

Heterogeneity: chi-squared = 20.97 (d.f. = 6), P = 0.002

$I^2 = 71.4\%$

Fixed or random effects?

There is no universally accepted method for choosing whether to use a random effects or a fixed effects model. I think that this would be a reasonable approach.

1. Irrespective of the numerical data, decide whether the assumption of a fixed effects model is plausible. Could the studies all be estimating the same effect? This depends on whether there is clinical heterogeneity. If not, consider a random effects model.

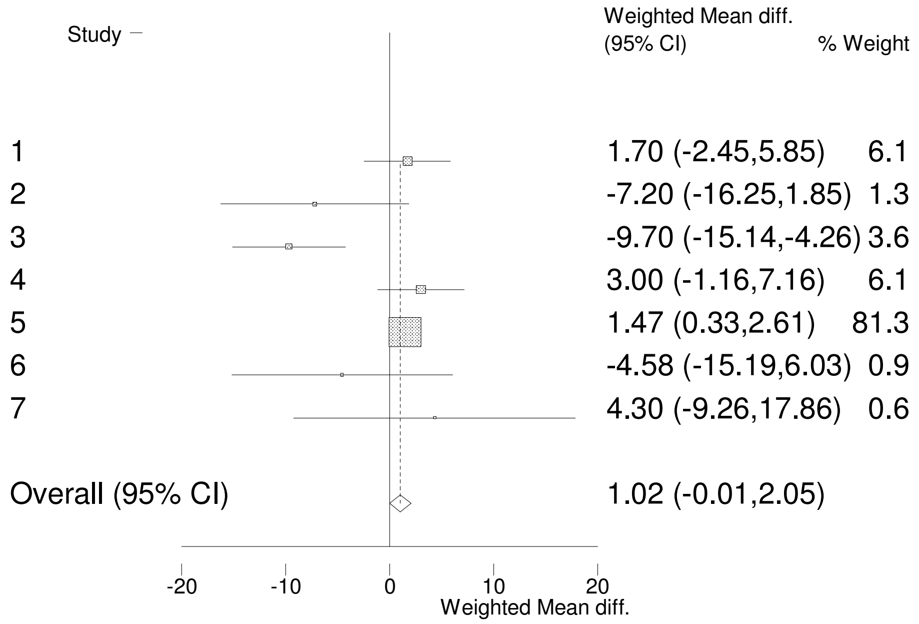
2. If a fixed effects assumption is plausible, are the data compatible with it? We can do this using both graphical methods, such as forest or Galbraith plots, and analytical methods, such as a heterogeneity test and I^2 statistic. If the assumption looks compatible with the data, use a fixed effects model, otherwise consider a random effects model.

3. If we consider a random effects model, we ask whether these studies represent a population where the average effect is interesting. Do we want to pool them at all? If yes, we can use a random effects model to do this. If they do not, we can do a narrative review.

Martin Bland
19 February 2009

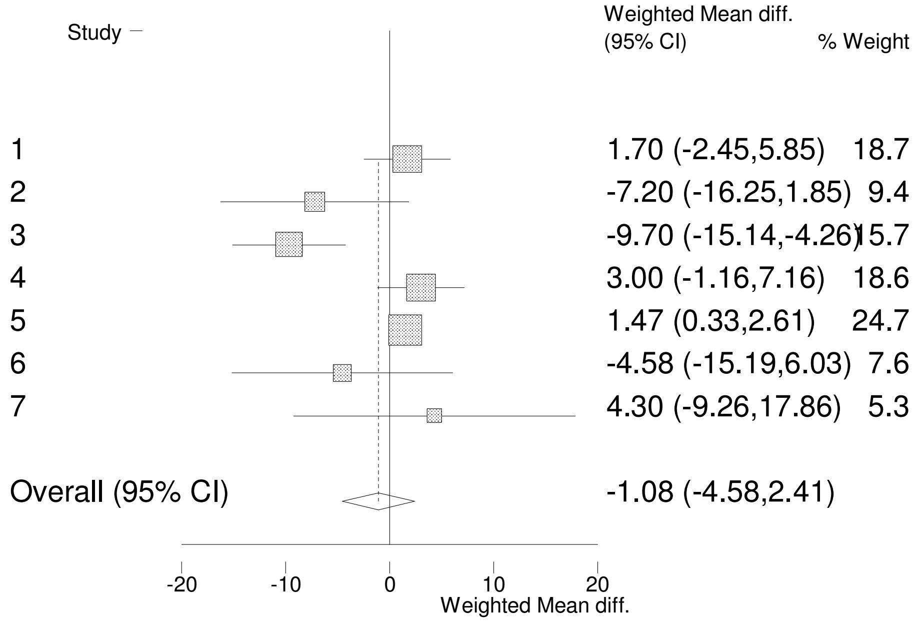
Figure 10. Meta-analyses with fixed effects model analysis and random effects model analysis, showing the effect on the estimate

Fixed effect:



Favours treatment Favours control P = 0.05

Random effect:



Favours treatment Favours control P = 0.5

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.

Higgins JPT, Thompson SG. (2002) Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21**, 1539-1558.

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. (2003) Measuring inconsistency in meta-analyses. *British Medical Journal* **327**, 557-560.

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.

Thompson SG. (1994) Systematic review: why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal*, **309**, 1351-1355.