Department of Health Sciences

M.Sc. Module: Systematic Reviews

Meta-analysis: heterogeneity and publication bias

Martin Bland

Professor of Health Statistics

University of York

http://martinbland.co.uk/msc/

Heterogeneity

- Galbraith plots
- Meta-regression
- Random effects models

Publication bias

- Funnel plots
- Begg and Eggar tests
- Trim and fill
- Selection modelling
- Meta-regression

Heterogeneity

Studies differ in terms of

- Patients
- Interventions
- Outcome definitions
- Design

⇒ **Clinical heterogeneity**

- Variation in true treatment or risk factor effects in magnitude or direction

⇒ **Statistical heterogeneity**

Heterogeneity

- Statistical heterogeneity may be caused by
  - clinical differences between studies
  - methodological differences between studies
  - unknown study characteristics
- Even if studies are clinically homogeneous there may be statistical heterogeneity

Heterogeneity

How to identify statistical heterogeneity

Test the null hypothesis that the studies all have the same effect in the population.

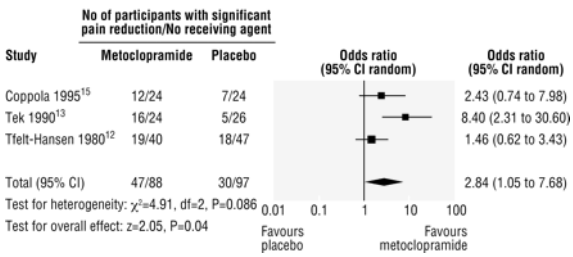
The test looks at the differences between observed effects for the studies and the pooled effect estimate.

Square, divide by variance, sum.

This gives a chi-squared test with degrees of freedom = number of studies – 1.

Expected chi-squared if null hypothesis true = degrees of freedom.

Heterogeneity



Test for heterogeneity: $\chi^2 = 4.91$, df = 2, P=0.086

Heterogeneity

Heterogeneity not significant

- No statistical evidence for difference between studies
- But, test for heterogeneity has low power - the number of studies is usually low - and may fail to detect heterogeneity as statistically significant when it exists.
- This cannot be interpreted as evidence of homogeneity.
- To compensate for the low power of the test a higher significance level is sometimes taken, $P < 0.1$ for statistical significance.

Heterogeneity

Significant heterogeneity

- differences between studies exist
- it may be invalid to pool the results and generate a single summary result
- describe variation
- investigate sources of heterogeneity
- account for heterogeneity

Dealing with heterogeneity

- Do not pool — narrative review.
- Ignore heterogeneity and use fixed effect model:
  - confidence interval too narrow,
  - difficult to interpret pooled estimate,
  - may be biased.
- Explore heterogeneity, can we explain it and remove it?
- Allow for heterogeneity and use random effects model.

Measuring heterogeneity

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, 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.

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

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

Measuring heterogeneity

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

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 percentage of the chi-squared statistic which is not explained by the variation within the studies.

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

It represents the percentage of the total variation which is due to variation between studies.

Measuring heterogeneity

Interpreting I^2

Higgins *et al.* (2003) suggest:

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

These are arbitrary, except for 0.

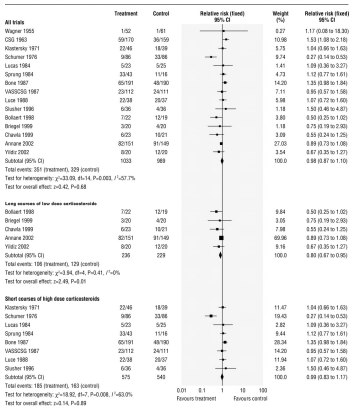
I^2 can never reach 100% and values above 90% are very rare.

Investigating sources of heterogeneity

- Subgroup analysis:
  - subsets of studies,
  - subsets of patients,
  - subsets should be pre-specified to avoid bias.
- Relate size of effect to characteristics of the studies, e.g.:
  - average age,
  - proportion of females,
  - intended dose of drug,
  - baseline risk.
- 'Meta-regression' can be used.

Investigating sources of heterogeneity

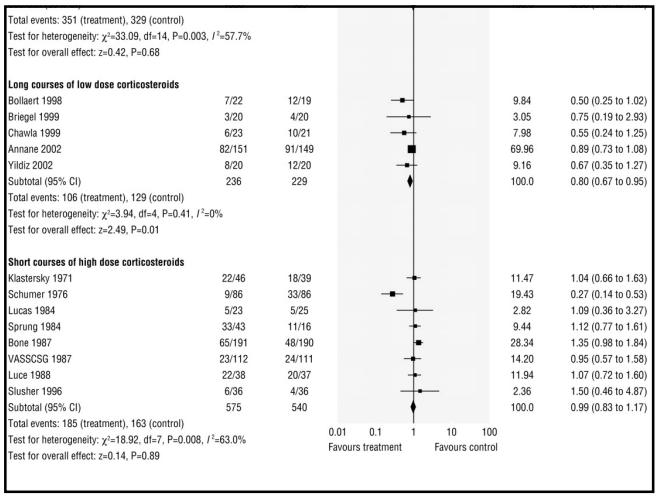
Corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004)



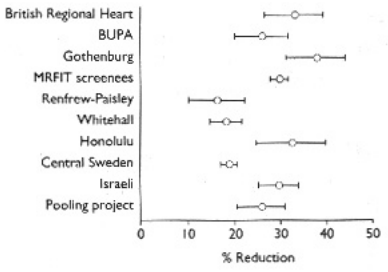
Annane D, Bellissant E, Bollaert PE, Briegel J, Keh D, Kupfer Y. (2004) Corticosteroids for severe sepsis and septic shock: a systematic review and meta-analysis. *BMJ* **329**, 480.

Investigating sources of heterogeneity

	Treatment	Control	Relative risk (fixed) 95% CI	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					



Investigating sources of heterogeneity

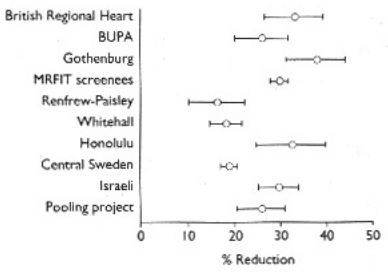


Heterogeneity X^2 = 127, df=9, P<0.001

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 SG. Systematic review: why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994; **309**: 1351-1355.

Investigating sources of heterogeneity



Heterogeneity X^2 = 127, df=9, P<0.001

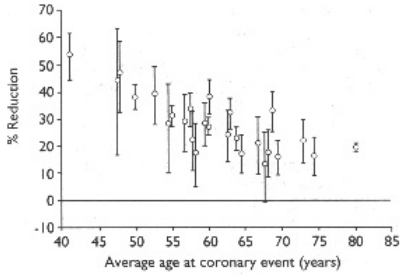
Studies varied in:

- age of men,
- cholesterol reduction achieved.

Split into sub-studies with more uniform age groups.

Investigating sources of heterogeneity

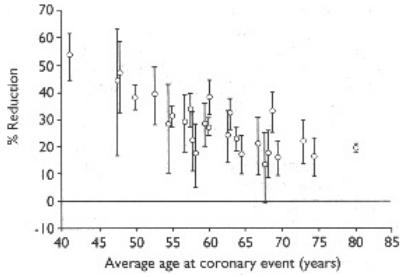
Split into 26 sub-studies with more uniform age groups.



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.

Investigating sources of heterogeneity

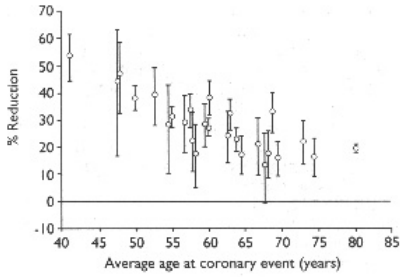
Split into 26 sub-studies with more uniform age groups.



Conclusion:
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.

Investigating sources of heterogeneity

Split into 26 sub-studies with more uniform age groups.

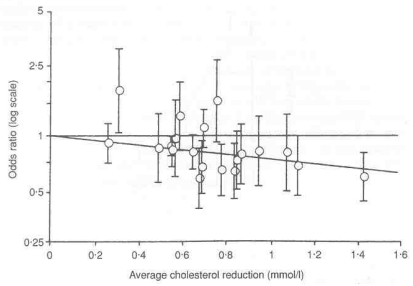


Before adjustment for age: $X^2 = 127$, df=9, P<0.001.

After adjustment for age: $X^2 = 45$, df=23, P=0.005.

A considerable improvement, but still some heterogeneity present.

Investigating sources of heterogeneity



Line fitted by meta-regression.

Thompson SG. Systematic review: why sources of heterogeneity in meta-analysis should be investigated. *BMJ* 1994; **309**: 1351-1355.

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

Investigating sources of heterogeneity

Galbraith plot

Alternative graphical representation to forest plot.

Horizontal axis: 1/standard error.

Horizontal axis will be zero if standard error is infinite, a study of zero size.

Vertical axis: effect/standard error.

This is the test statistic for the individual study.

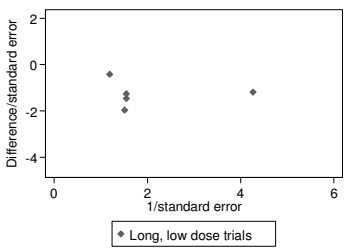
For 95% of studies, we expect this to be within 2 units of the true effect.

Investigating sources of heterogeneity

Galbraith plot

Corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004), trials of treatments with low doses and long duration.

Galbraith plot for log OR

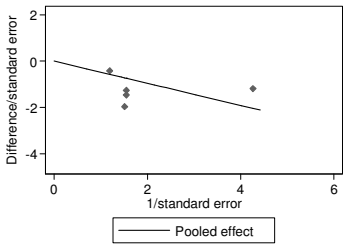


Investigating sources of heterogeneity

Galbraith plot

Corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004), trials of treatments with low doses and long duration.

Galbraith plot for log OR



Plot effect/se against 1/se.

We can add a line representing the pooled effect.

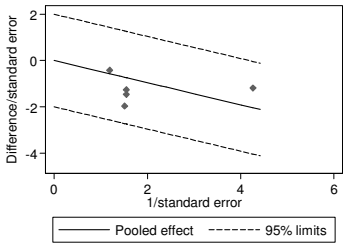
Slope = pooled effect

Investigating sources of heterogeneity

Galbraith plot

We expect 95% of points to be between these limits if there is no heterogeneity.

This is true for low dose, long duration trials.



Plot effect/se against 1/se.

We can add a line representing the pooled effect.

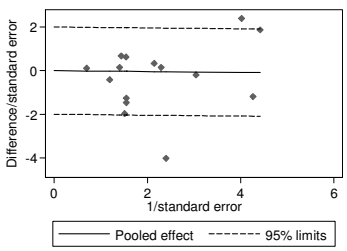
95% limits will be 2 units above and below this line.

Investigating sources of heterogeneity

Galbraith plot

Corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004), all trials.

The pooled effect is smaller so the line is less steep.



We have two points outside the 95% limits and one on the line.

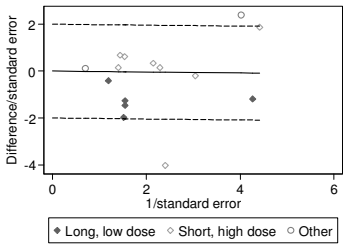
We can investigate them to see how these trials differ from the others.

Investigating sources of heterogeneity

Galbraith plot

Corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004), all trials.

These trials are all of high dose or short duration treatments.



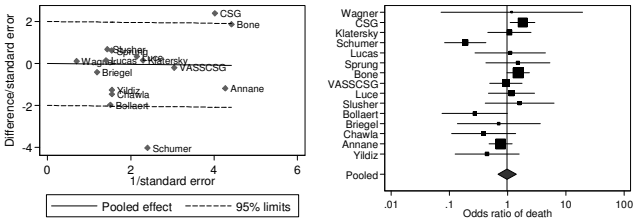
We could reanalyse taking dosage and duration separately.

Investigating sources of heterogeneity

Galbraith plot or forest plot?

"Conventional meta-analysis diagrams . . . are not very useful for investigating heterogeneity. A better diagram for this purpose was proposed by Galbraith . . ." (Thompson, 1994).

Is this really true?

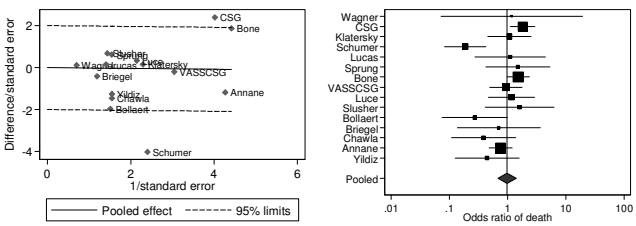


Investigating sources of heterogeneity

Galbraith plot or forest plot?

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.



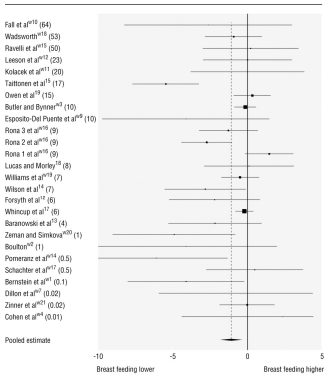
Investigating sources of heterogeneity

Cannot always explain heterogeneity

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

(In parenthesis: age at which blood pressure measured.)

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. *BMJ* **327**, 1189-1195.



Investigating sources of heterogeneity

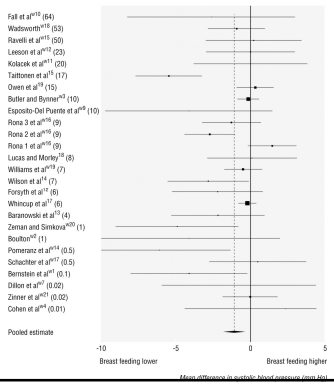
Cannot always explain heterogeneity

X^2=59.4, 25df, P<0.001

Three age groups: P=0.6.

Born before or after 1980: P=0.8.

Have to accept it and take it into account by using a random effects model.



Fixed and random effects models

Fixed effects model	**Random effects model**
We assume that the effect is the same in all studies.	We assume that the effect is not the same in all studies. The studies are a sample of possible studies where the 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.

Fixed and random effects models

Fixed effects model	Random effects model
If the effect is the same in all studies, it is more powerful and easier.	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 and random effects models

Fixed effects model	Random effects model
Variance of effect in study = standard error squared.	Variance of effect in study = standard error squared plus inter-trial variance, τ^2 (tau squared).
Weight = 1/variance $= 1/SE^2$	Weight = 1/variance. $= \frac{1}{SE^2 + \text{inter-trial variance}}$
	Inter-trial variance has degrees of freedom given by number of studies minus one.
	Typically small.

Fixed and random effects models

Fixed effects model	Random effects model
When heterogeneity exists we get:	When heterogeneity exists we get:
a pooled estimate which may give too much weight to large studies,	possibly a different pooled estimate with a different interpretation,
a confidence interval which is too narrow,	a wider confidence interval,
a P-value which is too small.	a larger P-value.

Fixed and random effects models

Example: oral rehydration in cholera, reduced osmolarity versus standard, duration of diarrhea

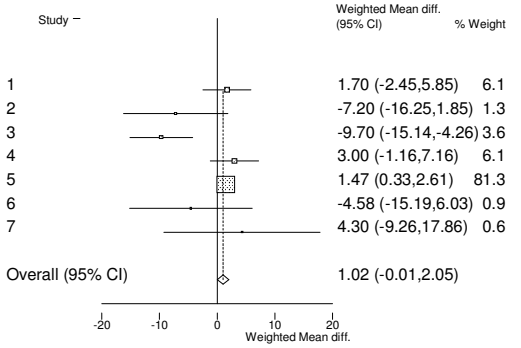
	Intervention			Control		
Study	n_1	$mean_1$	s_1	n_2	$mean_2$	s_2
1.	82	44.4	13.3	78	42.7	13.5
2.	34	49.9	18.7	29	57.1	17.9
3.	33	37.2	9.9	30	46.9	11.9
4.	147	46.0	18.2	153	43.0	18.6
5.	19	21.44	1.32	16	19.97	1.99
6.	19	33.89	16.4	20	38.47	17.4
7.	26	82.9	27.5	32	78.6	24.5

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

I^2 = 71.4%

Fixed and random effects models

Fixed effects model:

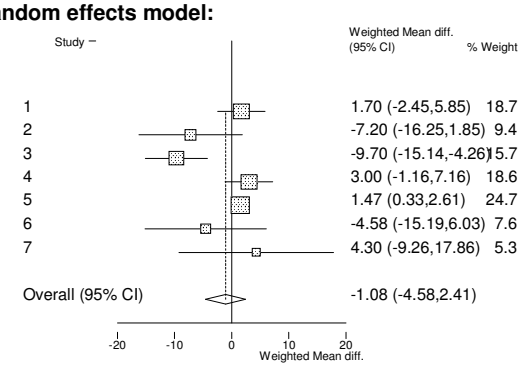


Favours treatment Favours control

P = 0.05

Fixed and random effects models

Random effects model:



Favours treatment Favours control

P = 0.5

Fixed and random effects models

Fixed effects model	Random effects model
When heterogeneity does not exists:	When heterogeneity does not exist:
a pooled estimate which is correct,	a pooled estimate which is correct,
a confidence interval which is correct,	a confidence interval which is too wide,
a P-value which is correct.	a P-value which is too large.

Fixed or random effects?

No universally accepted method for choosing.

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? If not, consider a random effects model.
2. If fixed effects assumption is plausible, are the data compatible?

   Graphical methods: forest plot, Galbraith plot.

   Analytical methods: heterogeneity test, I^2 statistic.

   If assumption looks compatible with the data, use fixed effects, otherwise consider random effects.

Fixed or random effects?

3. If we consider a random effects model, do studies represent a population where the average effect is interesting? Do we want to pool them?

   If yes: use a random effects model.

   If no: do a narrative review.

Publication bias

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 — in English, in higher impact journals.

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.

Identifying publication bias

Funnel plots

A plot of effect size against sample size.

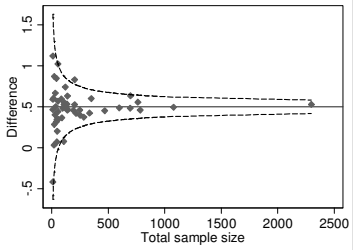
No bias is present → shaped like a funnel.

50 simulated studies with true effect = 0.5.

Funnel plot: effect against sample size.

95% of studies should lie within the lines.

Usually do not show these because they depend on population.



Identifying publication bias

Funnel plots

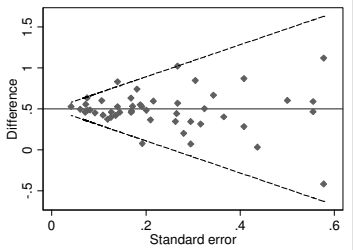
A plot of effect size against sample size.

No bias is present → shaped like a funnel.

50 simulated studies with true effect = 0.5.

Funnel plot: effect against standard error.

Boundaries are now straight lines.



Identifying publication bias

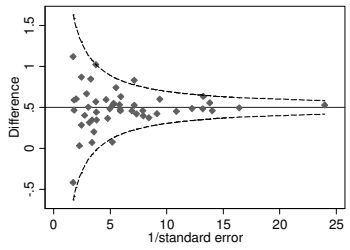
Funnel plots

A plot of effect size against sample size.

No bias is present → shaped like a funnel.

50 simulated studies with true effect = 0.5.

Funnel plot: effect against 1/standard error.



Identifying publication bias

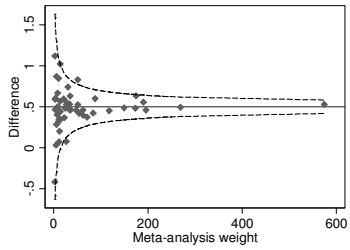
Funnel plots

A plot of effect size against sample size.

No bias is present → shaped like a funnel.

50 simulated studies with true effect = 0.5.

Funnel plot: effect against meta-analysis weight.



Identifying publication bias

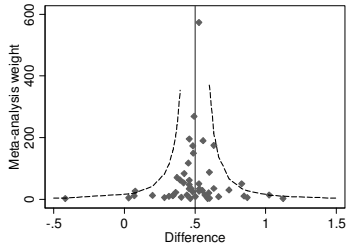
Funnel plots

Sometimes plot of sample size (etc.) against effect size.

Turned round through 90 degrees.

50 simulated studies with true effect = 0.5.

Funnel plot: meta-analysis weight against effect size.



Identifying publication bias

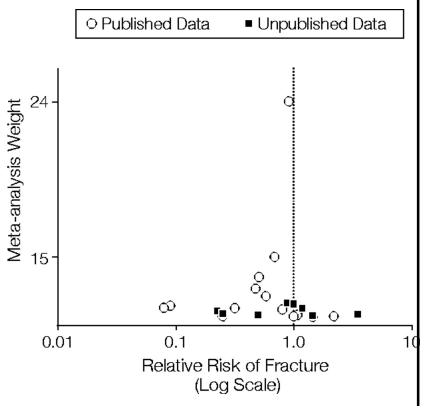
Funnel plots

A real one:

Hormone replacement therapy and prevention of nonvertebral fractures

The dotted line represents the point of no effect.

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



Identifying publication bias

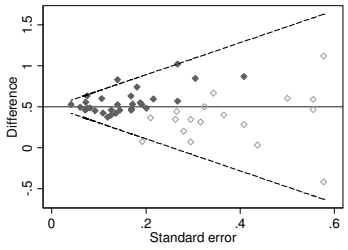
Funnel plots

If only significant studies are published, part of the funnel will be sparse or empty.

50 simulated studies with true effect = 0.5.

Funnel plot: effect against standard error.

Open diamonds are studies where the difference is not significant.



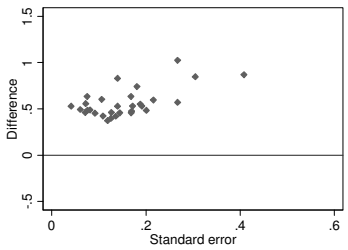
Identifying publication bias

Funnel plots

If only significant studies are published, part of the funnel will be sparse or empty.

If studies where the difference is not significant are not published, we won't see them.

We won't have the guide lines, either.



Identifying publication bias

Significance tests

'Begg's test' (Begg and Mazumdar 1994)

'Eggar's test' (Egger *et al.*, 1997)

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

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

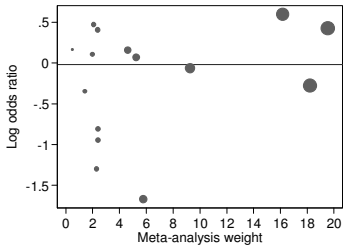
Egger M, Smith GD, Schneider M, Minder C. (1997) Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629-634.

Identifying publication bias

Begg's test

Starts with the funnel plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004), all trials.



Is the study estimate (log odds ratio in this example) related to the size of the study?

Correlation between log odds ratio and weight?

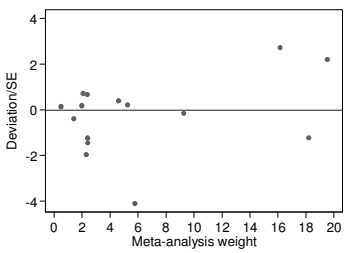
Problem: variance is not the same for all points.

Identifying publication bias

Begg's test

Starts with the funnel plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al.*, 2004), all trials.



Problem: variance is not the same for all points.

Solution: divide each estimate by standard error.

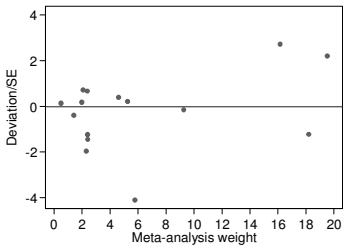
Begg subtracts pooled estimate first then divides by SE of the deviation.

Identifying publication bias

Begg's test

Starts with the funnel plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al*., 2004), all trials.



Now find Kendall's rank correlation between deviation/SE and weight.

Could use any suitable variable on x axis (SE, 1/SE, etc.)

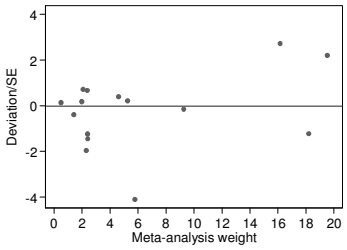
Tau b = 0.09, P = 0.7.

Identifying publication bias

Begg's test

Starts with the funnel plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al*., 2004), all trials.



Problem:

Power very low at small numbers of trials.

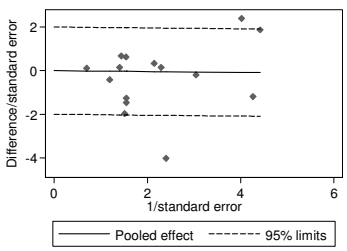
'Fairly powerful with 75 studies, moderate power with 25 studies'. (Begg and Mazumdar 1994).

Identifying publication bias

Eggar's test:

Based on the Galbraith plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al*., 2004), all trials, log odds ratio.



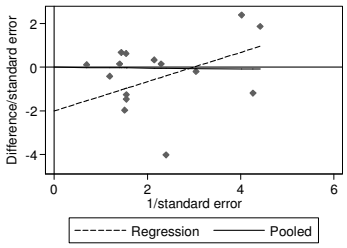
Regress study difference (log odds ratio) over standard error on 1/standard error.

Identifying publication bias

Eggar's test:

Based on the Galbraith plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al*., 2004), all trials, log odds ratio.



Regress study difference (log odds ratio) over standard error on 1/standard error.

Does the line go through the origin?

Test intercept against zero.

Identifying publication bias

Eggar's test:

Should we weight the observations?

'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.

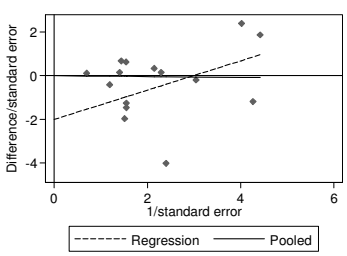
'We performed both weighted and unweighted analyses and used the output from the analysis yielding the intercept with the larger deviation from zero.' (Egger *et al*., 1997).

Identifying publication bias

Eggar's test:

Based on the Galbraith plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al*., 2004), all trials, log odds ratio.



Unweighted:

D/SE = –1.14 + 0.39×1/SE

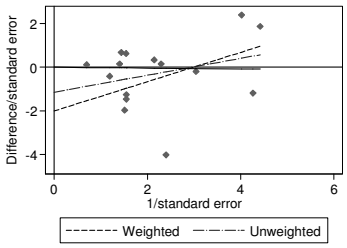
Intercept = –1.14, se = 0.88, P = 0.22, 95% CI = –3.05 to 0.77.

Identifying publication bias

Eggar's test:

Based on the Galbraith plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al*., 2004), all trials, log odds ratio.



Unweighted:

D/SE = –1.14 + 0.39×1/SE

Intercept P = 0.22.

Weighted:

D/SE = –2.01 + 0.67×1/SE

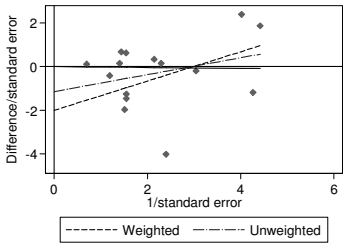
Intercept P = 0.17.

Identifying publication bias

Eggar's test:

Based on the Galbraith plot.

Corticosteroids for severe sepsis and septic shock (Annane *et al*., 2004), all trials, log odds ratio.



Is this test biased?

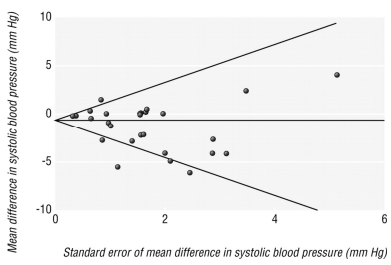
Doing both regressions and choosing the more significant is multiple testing.

The regression intercept is a biased estimate.

Identifying publication bias

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

Begg's funnel plot (pseudo 95% confidence limits) showing mean difference in systolic blood pressure by standard error of mean difference.



'The Egger test was significant (P = 0.033) for publication bias but not the Begg test (P = 0.186).'

Dealing with publication bias

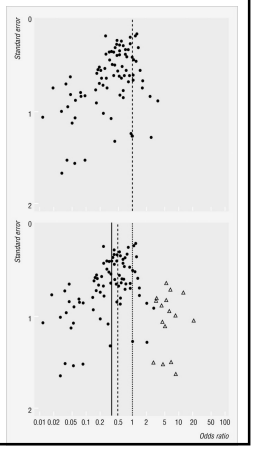
- Trim and fill
- Selection models
- Meta-regression

Dealing with publication bias

Trim and fill

Trim: we eliminate studies, starting with the least powerful, until we have symmetry. Get a new pooled estimate.

Fill: for the studies eliminated, we reflect them in the pooled estimate line and put in new studies.



Dealing with publication bias

Trim and fill

Example: 89 trials comparing homeopathic medicine with placebo.

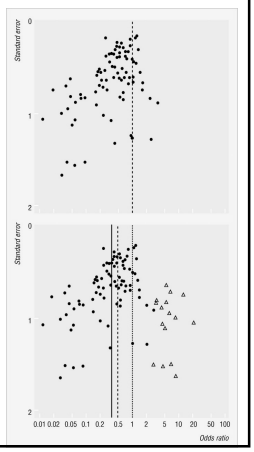
Dotted line: no effect.

Solid line: pre trim and fill estimate.

Open triangles are filled trials.

Broken line: post trim and fill estimate.

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.



Dealing with publication bias

Trim and fill

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) .

Dealing with publication bias

Selection models

Model the selection process that determines which results are published.

Based on the assumption that the study's P value affects its probability of publication.

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.

Not widely used.

Dealing with publication bias

Meta-regression

Use study characteristics, e.g. Jadad score, sample size, to predict outcome.

Example, 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)

Dealing with publication bias

A note of caution

- These methods require large numbers of studies. They are not powerful in most meta-analyses.
- 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!)
- Publication bias may not result from significance or sample size. Researchers or sponsors may not like the result. Most healthcare researchers are amateurs with other demands on their attention (e.g. their patients).

Dealing with publication bias

A note of caution

Better to think of these methods as a way of exploring possibilities than to produce definitive answers.

Example: homeopathy versus placebo (Sterne *et al*., 2001)

Regression of trial effect on **asymmetry coefficient**, **language English/other**, allocation concealment, **blinding**, handling of withdrawals, indexed by Medline (**bold** were significant).

Dealing with publication bias

Example: homeopathy versus placebo (Sterne *et al*., 2001)

'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.'