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

<table>
<thead>
<tr>
<th>Study</th>
<th>No of participants with significant pain reduction</th>
<th>Odds ratio (95% CI range)</th>
<th>Odds ratio (95% CI range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doublet 1980</td>
<td>12/24</td>
<td>2.43 (0.74 to 7.86)</td>
<td></td>
</tr>
<tr>
<td>Tru 1990</td>
<td>16/24</td>
<td>5.40 (2.31 to 12.05)</td>
<td></td>
</tr>
<tr>
<td>Thiit Finnian 1990</td>
<td>19/49</td>
<td>1.46 (0.48 to 4.43)</td>
<td></td>
</tr>
<tr>
<td>Total (95% CI)</td>
<td>47/88</td>
<td>2.84 (1.05 to 7.68)</td>
<td></td>
</tr>
</tbody>
</table>

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

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

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


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

- 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.

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

Investigating sources of heterogeneity

Heterogeneity $\chi^2 = 127$, 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

Investigating sources of heterogeneity

Heterogeneity $\chi^2 = 127$, df=9, P<0.001

Studies varied in:
- age of men,
- cholesterol reduction achieved.

Split into sub-
states 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

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

[Diagram of Galbraith plot with data points indicating long, low dose trials]
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?

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.)


\[ X^2 = 59.4, \text{25df, P<0.001} \]

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

We assume that the effect is the same in all studies.

We use only the sampling variation within the studies.

Random effects model

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 the sampling variation within the studies and the sampling variation between studies.
### Fixed and random effects models

<table>
<thead>
<tr>
<th>Fixed effects model</th>
<th>Random effects model</th>
</tr>
</thead>
<tbody>
<tr>
<td>If the effect is the same in all studies, it is more powerful and easier.</td>
<td>Less powerful because P values are larger and confidence intervals are wider.</td>
</tr>
<tr>
<td>No assumption about representativeness.</td>
<td>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.</td>
</tr>
</tbody>
</table>

#### Random effects model

Variance of effect in study = standard error squared plus inter-trial variance, $\tau^2$ (tau squared).

Weight = $1/\text{variance}$

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

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

#### Fixed effects model

Variance of effect in study = standard error squared.

Weight = $1/\text{variance}$

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

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 exists we get:

- possibly a different pooled estimate with a different interpretation,
- a wider confidence interval,
- a larger P-value.
Fixed and random effects models

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

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study</td>
<td>n</td>
</tr>
<tr>
<td>1</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
</tr>
<tr>
<td>4</td>
<td>147</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
</tr>
</tbody>
</table>

Heterogeneity: chi-squared = 20.97 (d.f. = 6), P = 0.002
I² = 71.4%

Fixed and random effects models

Fixed effects model:

<table>
<thead>
<tr>
<th>Study</th>
<th>Weighted Mean diff. (95% CI)</th>
<th>% Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.70 (-2.45,5.85)</td>
<td>6.1</td>
</tr>
<tr>
<td>2</td>
<td>-7.20 (-16.25,1.85)</td>
<td>1.3</td>
</tr>
<tr>
<td>3</td>
<td>-9.70 (-15.14,-4.26)</td>
<td>3.6</td>
</tr>
<tr>
<td>4</td>
<td>3.00 (-1.16,7.16)</td>
<td>6.1</td>
</tr>
<tr>
<td>5</td>
<td>1.47 (0.33,2.61)</td>
<td>81.3</td>
</tr>
<tr>
<td>6</td>
<td>-4.58 (-15.19,6.03)</td>
<td>0.9</td>
</tr>
<tr>
<td>7</td>
<td>4.30 (9.26,17.86)</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Overall (95% CI) | -1.02 (-0.01,2.05) | P = 0.05

Favours treatment | Favours control

Fixed and random effects models

Random effects model:

<table>
<thead>
<tr>
<th>Study</th>
<th>Weighted Mean diff. (95% CI)</th>
<th>% Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.70 (-2.45,5.85)</td>
<td>18.7</td>
</tr>
<tr>
<td>2</td>
<td>-7.20 (-16.25,1.85)</td>
<td>9.4</td>
</tr>
<tr>
<td>3</td>
<td>-9.70 (-15.14,-4.26)</td>
<td>18.6</td>
</tr>
<tr>
<td>4</td>
<td>3.00 (-1.16,7.16)</td>
<td>24.7</td>
</tr>
<tr>
<td>5</td>
<td>1.47 (0.33,2.61)</td>
<td>7.6</td>
</tr>
<tr>
<td>6</td>
<td>-4.58 (-15.19,6.03)</td>
<td>5.3</td>
</tr>
<tr>
<td>7</td>
<td>4.30 (9.26,17.86)</td>
<td></td>
</tr>
</tbody>
</table>

Overall (95% CI) | -1.08 (-4.58,2.41) | P = 0.5

Favours treatment | Favours control
**Fixed and random effects models**

<table>
<thead>
<tr>
<th>Fixed effects model</th>
<th>Random effects model</th>
</tr>
</thead>
<tbody>
<tr>
<td>When heterogeneity does not exist:</td>
<td>When heterogeneity does not exist:</td>
</tr>
<tr>
<td>a pooled estimate which is correct,</td>
<td>a pooled estimate which is correct,</td>
</tr>
<tr>
<td>a confidence interval which is correct,</td>
<td>a confidence interval which is too wide,</td>
</tr>
<tr>
<td>a P-value which is correct.</td>
<td>a P-value which is too large.</td>
</tr>
</tbody>
</table>

**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.

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

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

Significance tests

‘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?’


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.

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.)

\[ \text{Tau b} = 0.09, \text{P} = 0.7. \]

---

**Problem:**

Power very low at small numbers of trials.

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

---

**Egger’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

Egger’s test:

Based on the Galbraith plot.

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

![Diagram](image)

Regres 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

Egger’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

Egger’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

**Egger's test:**

Based on the Galbraith plot.

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

Unweighted:

\[
\frac{D}{SE} = -1.14 + 0.39 \times \frac{1}{SE}
\]

Intercept P = 0.22.

Weighted:

\[
\frac{D}{SE} = -2.01 + 0.67 \times \frac{1}{SE}
\]

Intercept P = 0.17.

**Is this test biased?**

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

The regression intercept is a biased estimate.

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

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.

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.

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).

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).

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.’