

# Reporting bias

## Jonathan Sterne

Cochrane Collaboration training meeting  
Cambridge, 28 July 2008

Chapter editors: Jonathan Sterne, Matthias Egger and David Moher on behalf of the Bias Methods Group

Contributing authors: James Carpenter, Matthias Egger, Roger Harbord, Julian Higgins, David Jones, David Moher, Jonathan Sterne, Alex Sutton, Jennifer Tetzlaff.

## Reporting biases

- arise when the dissemination of research findings is influenced by the nature and direction of results. Statistically significant, 'positive' results that indicate that an intervention works are more likely to be published, more likely to be published rapidly, more likely to be published in English, more likely to be published more than once, more likely to be published in high impact journals and, related to the last point, more likely to be cited by others.

## Outline of chapter

1. Introduction
2. Types of reporting bias and the supporting evidence
3. Avoiding reporting biases
4. Detecting reporting biases

## Outline of chapter

1. Introduction
2. Types of reporting bias and the supporting evidence
3. Avoiding reporting biases
4. **Detecting reporting biases**
  - Funnel plots
  - Different reasons for funnel plot asymmetry
  - Tests for funnel plot asymmetry
  - Sensitivity analyses
  - Summary

## Acknowledgements

- James Carpenter, LSHTM
- Matthias Egger, ISPM, Berne
- Julian Higgins, MRC BSU, Cambridge
- Roger Harbord, University of Bristol
- David Jones, University of Leicester
- Alex Sutton, University of Leciester

## Declaration of interest....

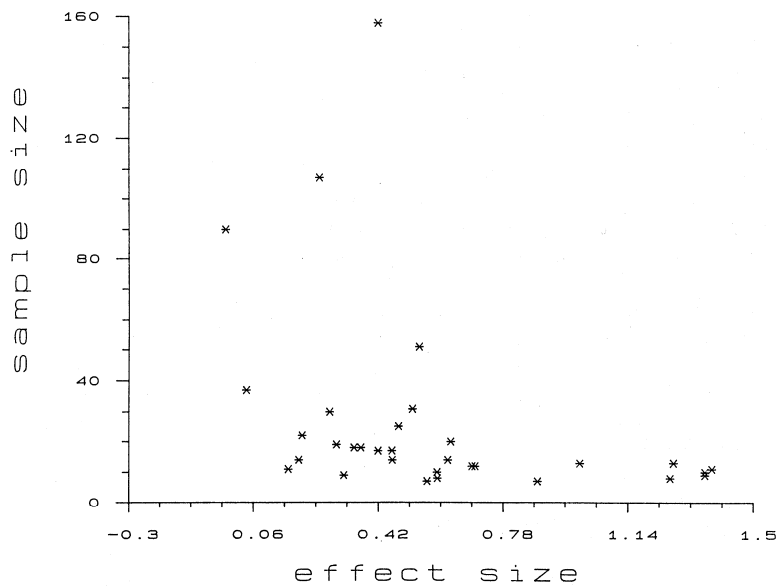
- James Carpenter, LSHTM
  - Matthias Egger, ISPM, Berne
  - Julian Higgins, MRC BSU, Cambridge
  - Roger Harbord, University of Bristol
  - David Jones, University of Leicester
  - Alex Sutton, University of Leciester
  - Jonathan Sterne, University of Bristol
- With the exception of JH (who is much more interested in promoting the  $I^2$  statistic) all of the above are co-authors on papers proposing tests for funnel plot asymmetry

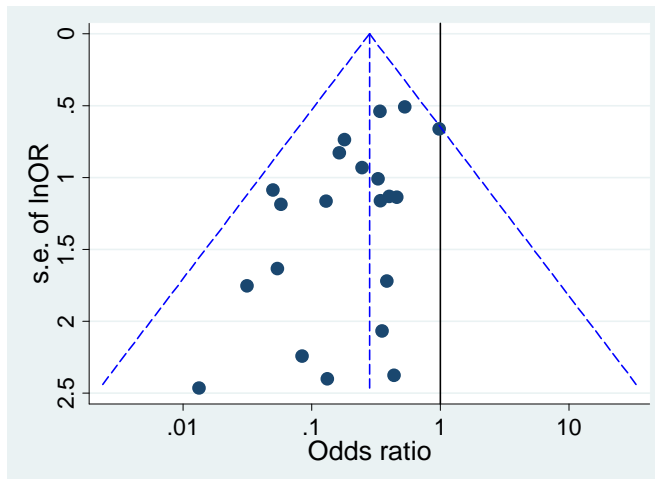
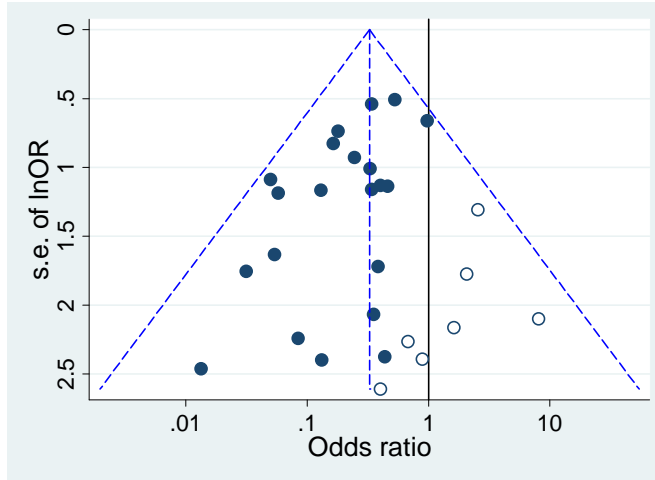
## It all started here....

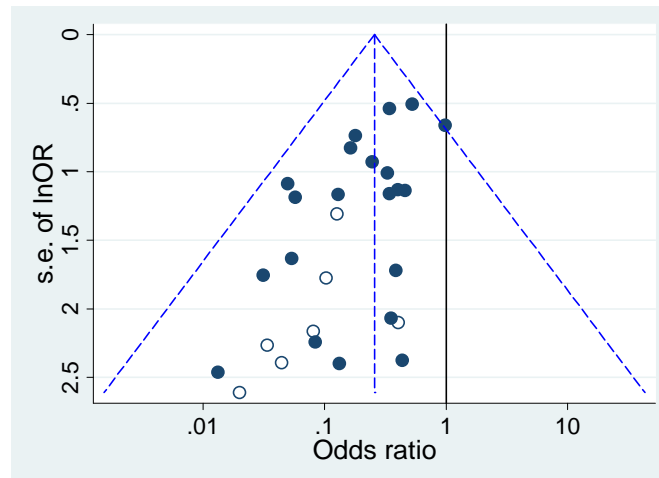
- If all studies come from a single underlying population, this graph should look like a funnel, with the effect sizes homing in on the true underlying value as  $n$  increases. [If there is publication bias] there should be a bite out of the funnel.”

Light RJ, Pillemer DB. Summing up. The science of reviewing research. *Harvard University Press*, 1984.

## Funnel plot from Begg and Berlin (*JRSS A* 1988)





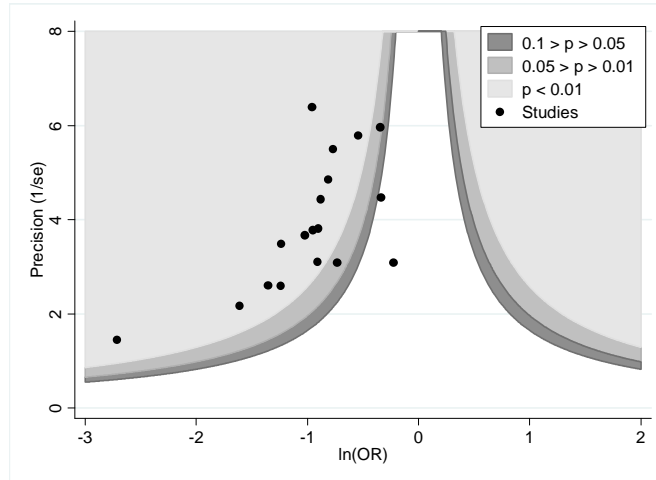


## Different reasons for funnel plot asymmetry

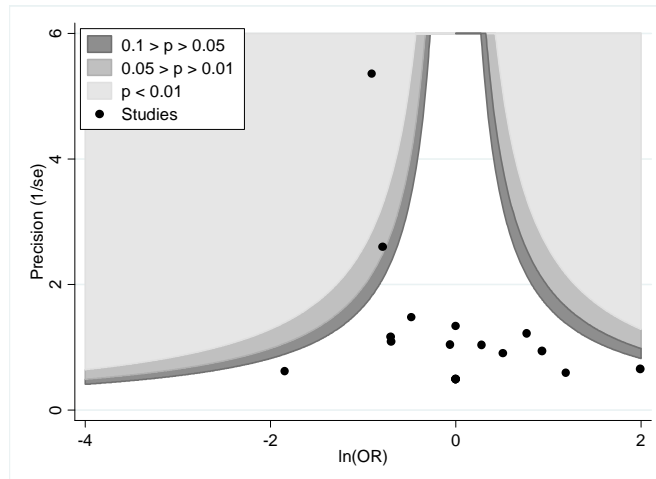
(Adapted from Egger et al. *BMJ* 1997)

1. Selection biases
  - Publication bias
    - Delayed publication bias
    - Location biases (Language bias, Citation bias, Multiple publication bias)
  - Selective outcome reporting
2. Poor methodological quality leading to spuriously inflated effects in smaller studies
  - Poor methodological design
  - Inadequate analysis
  - Fraud
3. True heterogeneity
  - Size of effect differs according to study size
4. Artefactual
5. Chance

## Contour-enhanced funnel plots



## Contour-enhanced funnel plots



## Tests for funnel plot asymmetry

- A test for funnel plot asymmetry (small study effects) formally examines whether the association between estimated intervention effects and a measure of study size (such as the standard error of the intervention effect) is greater than might be expected to occur by chance
- For outcomes measured on a continuous (numerical) scale this is reasonably straightforward...

### Original regression test (“Egger test”)

Egger, Davey Smith, Schneider, Minder. *BMJ* 1997; 315:629-34

$\theta$  : treatment effect

- Regress  $\theta$  on  $SE(\theta)$  with weights  $1/Var(\theta)$
- $t$ -test of slope = 0

*equivalently:*

- Regress  $\theta / SE(\theta)$  on  $1/SE(\theta)$  without weights
- $t$ -test of intercept = 0

**By far the most widely used and cited approach**

## Unfortunately....

- When outcomes are dichotomous, and intervention effects are expressed as odds ratios, there are statistical problems with this approach, because the standard error of the log odds ratio is mathematically linked to the size of the odds ratio
- This can cause funnel plots plotted using log odds ratios (or odds ratios on a log scale) to appear asymmetric and can mean that P values from the test of Egger et al. are too small
- These problems are especially prone to occur when the intervention has a large effect, there is substantial between-study heterogeneity, there are few events per study, or when all studies are of similar sizes

## Developing recommendations

- We tried to base recommendations both on theory and on findings from published simulation studies
- We also identified areas where absence of evidence meant that definitive recommendations were not possible

## A plethora of alternative (published) tests

Reference	Basis of test
Begg 1994	Rank correlation between standardised treatment effect and its standard error
Egger 1997	Linear regression of effect estimate against its s.e., weighted by inverse variance of effect estimate
Tang 2000	Linear regression of treatment effect estimate on $1/\sqrt{N}$ with weights $N$
Macaskill 2001	Linear regression of treatment effect estimate on $N$ , with weights $dh/N$
Deeks 2005	Linear regression of log OR on $1/\sqrt{ESS}$ with weights $ESS$ , where effective sample size $ESS = n_0 n_1 / N$
Harbord 2006	Modified version of the Egger test, based on the “score” ( $O-E$ ) and “score variance” ( $V$ ) of the log OR
Peters 2006	Linear regression of treatment effect estimate on $1/N$ , with weights $dh/N$ .
Schwarzer 2006	Rank correlation test, using mean and variance of the non-central hypergeometric distribution
Rücker 2007	Test based on arcsine transformation of observed risks, with explicit modelling of between-study heterogeneity

## Recommendations on testing for funnel plot asymmetry (f.p.a.)

- **For all types of outcome**
  - tests for f.p.a. should be used only when there are  $\geq 10$  studies in the meta-analysis. When there are fewer studies the power is too low to distinguish chance from real asymmetry
  - Tests for f.p.a. should not be used if all studies are of similar sizes. However we are not aware of simulation evidence providing specific guidance on when study sizes should be considered ‘too similar’
  - Results of tests for f.p.a. should be interpreted in the light of visual inspection of the funnel plot
  - When there is evidence of small-study effects, publication bias should be considered as one of a number of possible explanations. Although funnel plots may alert review authors to a problem which needs considering, they do not provide a solution to this problem.

## Recommendations on testing for funnel plot asymmetry

- For **continuous outcomes with intervention effects measured as mean differences**
  - The test proposed by Egger et al (1997) may be used to test for f.p.a. There is currently no reason to prefer any of the more recently proposed tests in this situation, although their relative advantages and disadvantages have not been formally examined
  - General considerations suggest that the power will be greater than for dichotomous outcomes, but that use of the method with substantially fewer than 10 studies would be unwise

## Recommendations on testing for funnel plot asymmetry

- For **dichotomous outcomes with intervention effects measured as odds ratios:**
  - The tests of Harbord et al. and Peters et al. avoid the mathematical association between the log odds ratio and its standard error (and hence false-positive test results) that occurs for the Egger test when there is a substantial intervention effect, while retaining power compared with alternative tests. However, false-positive results may still occur in the presence of substantial between-study heterogeneity
  - The test proposed by Rücker et al. (2007) avoids false-positive results both when there is a substantial intervention effect and in the presence of substantial between-study heterogeneity. As a rule of thumb, when  $\tau^2 > 0.1$ , only the version of the arcsine test including random-effects has been shown to work reasonably well. However it is slightly conservative in the absence of heterogeneity, and its interpretation is less familiar because it is based on an arcsine transformation

## Recommendations on testing for funnel plot asymmetry

- For dichotomous outcomes with intervention effects measured as odds ratios:
  - When the  $\tau^2 < 0.1$ , one of the Harbord, Peters or Rucker tests can be used. (Test performance generally deteriorates as  $\tau^2$  increases)
  - As far as possible, review authors should specify their testing strategy in advance (noting that test choice may be dependent on the degree of heterogeneity observed). They should apply only one test, appropriate to the context of the particular meta-analysis, from the above-recommended list and report only the result from their chosen test. Application of two or more tests is undesirable since the most extreme (largest or smallest) P value from a set of tests does not have a well-characterized interpretation
  - **None of these tests is implemented in RevMan, and consultation with a statistician is recommended for their implementation**

## Recommendations on testing for funnel plot asymmetry

- For dichotomous outcomes with intervention effects measured as risk ratios or risk differences, and continuous outcomes with intervention effects measured as standardized mean differences :
  - Potential problems in funnel plots have been less extensively studied for these effect measures than for odds ratios, and firm guidance is not yet available
  - Funnel plots using risk differences should seldom be of interest

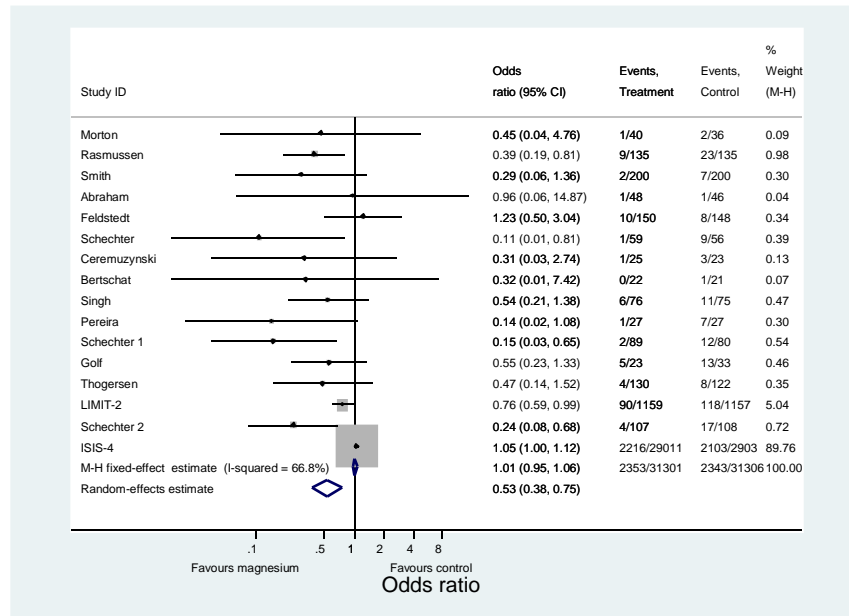
## **Tests for which there is insufficient evidence to recommend use**

- The test proposed by Begg and Mazumdar (1994) has the same statistical problems as but lower power than the Egger test and is therefore not recommended
- The test proposed by Tang and Liu (2000) has not been evaluated in simulation studies, while the test proposed by Macaskill et al. (2001) has lower power than more recently proposed alternatives.
- The test proposed by Deeks et al. (2005) is aimed at meta-analyses of diagnostic test accuracy studies: it is likely to have lower power than more recently proposed alternatives.
- The test proposed by Schwarzer et al. (2006) avoids the mathematical association between the log odds ratio and its standard error, but has low power relative to other tests

## **Sensitivity analyses**

- Comparing fixed and random-effects estimates
- Trim and Fill
- Fail-safe N
- Other selection models
- Sensitivity analyses based on selection models

## Comparing fixed and random-effects estimates



## Comparing fixed and random-effects estimates

When review authors are concerned about small-study effects and there is evidence of between-study heterogeneity ( $I^2 > 0$ ), then compare the fixed- and random-effects estimates of the treatment effect. If the estimates are similar then small study effects have little effect on the treatment effect estimate. If the random-effects estimate is more beneficial, then consider whether it is reasonable to conclude that the treatment was more effective in the smaller studies. If the larger studies are those conducted with more methodological rigour, or in circumstances typical of the use of the intervention in practice, consider reporting meta-analyses restricted to the larger, more rigorous studies. Formal evaluation of such strategies in simulation studies would be desirable. Formal statistical comparisons of the fixed and random-effects estimates are not possible. It is still possible for small study effects to bias the results of a meta-analysis in which there is no evidence of heterogeneity.

## Trim and Fill

There is no guarantee that the adjusted treatment effect matches what would have been observed in the absence of publication bias, since we cannot know the true mechanism for publication bias. Equally importantly, the trim and fill method does not take into account reasons for funnel plot asymmetry other than publication bias. Therefore, “corrected” treatment effect estimates from this method should be interpreted with great caution. The method is known to perform poorly in the presence of substantial between-study heterogeneity. Additionally, estimation and inferences are based on a dataset containing imputed treatment effect estimates. Such estimates, it can be argued, inappropriately contribute information that reduces the uncertainty in the summary intervention effect.

## Fail-safe N

- This and related methods are not recommended

## **Other selection models**

- The complexity of the statistical methods, and the large number of studies needed, probably explain why selection models have not been widely used in practice

## **Sensitivity analyses based on selection models**

- Copas developed a model in which the probability that a study is included in a meta-analysis depends on its standard error. Because it is not possible to estimate all model parameters precisely, he advocates sensitivity analyses in which the value of the estimated treatment effect is computed under a range of assumptions about the severity of the selection bias (Copas 1999). Rather than a single estimate treatment effect “corrected” for publication bias, the reader can see how the estimated effect (and confidence interval) varies as the assumed amount of selection bias increases. Application of the method to epidemiological studies of environmental tobacco smoke and lung cancer suggests that publication bias may explain some of the association observed in meta-analyses of these studies (Copas 2000).

## **Testing for excess of studies with significant results**

- Ioannidis and Trikalinos (2007) proposes a simple test that compares the number of studies that have formally statistically significant results with the number of statistically significant results expected under different assumptions about the magnitude of the effect size. Unlike either the regression tests or contour funnel plots, the test does not make any assumption about small-study effects. An excess of significant results can reflect either suppression of whole studies or related selective/manipulative analysis and reporting practices that would cause similar excess. The test has limited power when there are few studies and when there are few studies with significant results. Because it has not been rigorously evaluated through simulation, we currently do not recommend the test as an alternative to those described earlier.

## **Next steps**

- Convert to stand-alone consensus statement on testing for publication bias/small study effects as soon as possible

## Summary (1)

- Although there is clear evidence that publication and other reporting biases lead to over-optimistic estimates of intervention effects, overcoming, detecting and correcting for publication bias is problematic.
- Comprehensive searches are important, but are not sufficient to prevent some substantial potential biases.
- Publication bias should be seen as one of a number of possible causes of 'small study effects' – a tendency for estimates of the intervention effect to be more beneficial in smaller studies.
- Funnel plots allow review authors to make a visual assessment of whether small study effects may be present.
- For continuous (numerical) outcomes with intervention effects measured as mean differences, funnel plots and statistical tests for funnel plot asymmetry are valid.

## Summary (2)

- For continuous outcomes with intervention effects measured as mean differences, funnel plots and tests for f.p.a. are valid
- For dichotomous outcomes with intervention effects expressed as ORs, the s.e. of the log OR is mathematically linked to the OR, even in the absence of small study effects
  - This can cause funnel plots to appear asymmetric and can mean that P values from the Egger test are too small. For other effect measures, firm guidance is not yet offered
- Three tests for small study effects are recommended for use in Cochrane reviews, if there are  $\geq 10$  studies. However, none is implemented in RevMan and statistical support is usually required. Only one test has been shown to work when the between-study heterogeneity variance exceeds 0.1

## Summary (3)

- Results from tests for funnel plot asymmetry should be interpreted cautiously. When there is evidence of small-study effects, publication bias should be considered as only one of a number of possible explanations. In these circumstances, review authors should attempt to understand the source of the small study effects, and consider their implications in sensitivity analyses