Supporting material for JL Peters, AJ Sutton, DR Jones, KR Abrams, L Rushton (2006) Comparison of two methods to detect publication bias in meta-analysis *JAMA* 295; 676-680

Taken from Technical Report 05-01:
Performance of tests and adjustments for publication bias in the presence of heterogeneity Jaime L Peters, Alex J Sutton, David R Jones, Keith R Abrams, Lesley Rushton. Department of Health Sciences, University of Leicester, UK

Full paper version of Technical Report 05-01 available on request to **medstats@le.ac.uk**

Simulations

Parameters

Both fixed and random effects models have been used to simulate the meta-analyses. The fixed effects model is given by

$$y_i = \theta + \varepsilon_i , \qquad (1)$$

where θ is the true underlying effect, lnOR.

The random effects model is

$$y_i = \theta_i + \varepsilon_i , \ \theta_i \sim N(\mu, \tau^2) \qquad (2)$$

where θ_i is the true effect in study i, μ is the true underlying effect, lnOR, and τ^2 is the between-study variance.

For the random effects meta-analysis, the between-study variance is defined to be 20%, 150% and 500% of the average within-study variance for studies from the corresponding simulation. This compares with specification of I^2, describing the percentage of total variation across studies that is due to between-study heterogeneity rather than chance (Higgins and Thompson, 2002). Here, 20%, 150% and 500% of the within-study variation corresponds to an I^2 of 16.7%, 60% and 83.3%, respectively.

Characteristics of the simulated meta-analyses were determined from a systematic review of meta-analyses of animal toxicology studies (Peters et al, 2004), but findings can be applied generally. The characteristics took the following values:

- The number of primary studies in a meta-analysis was 6, 16, 30 or 90 studies.

- In each meta-analysis, the probability of an event (death) in the control group was sampled from a uniform distribution (0.3, 0.7).
- The number of control subjects within each primary study is taken from the distribution N(5, 0.3).
- For simplicity the ratio of exposed to control subjects is one.
- The underlying ORs were 1, 1.2, 1.5, 3 or 5.

Publication bias was induced in two ways:

1. The first was based on the assumption that studies are censored as a result of the *one-sided* p-value associated with the effect estimate of interest (as in Begg and Mazumdar (1994) and Macaskill et al (2001)). For example, studies showing the chemical exposure to be protective compared to the control could be more likely to be censored than studies showing a significant harmful effect of the chemical exposure, and vice versa.
2. The second is based on the assumption that the size of the effect estimate influences whether a study is censored or no. Studies with the most extreme estimates of effect are censored (as in Duval and Tweedie (2002)).

Publication bias based on p-values

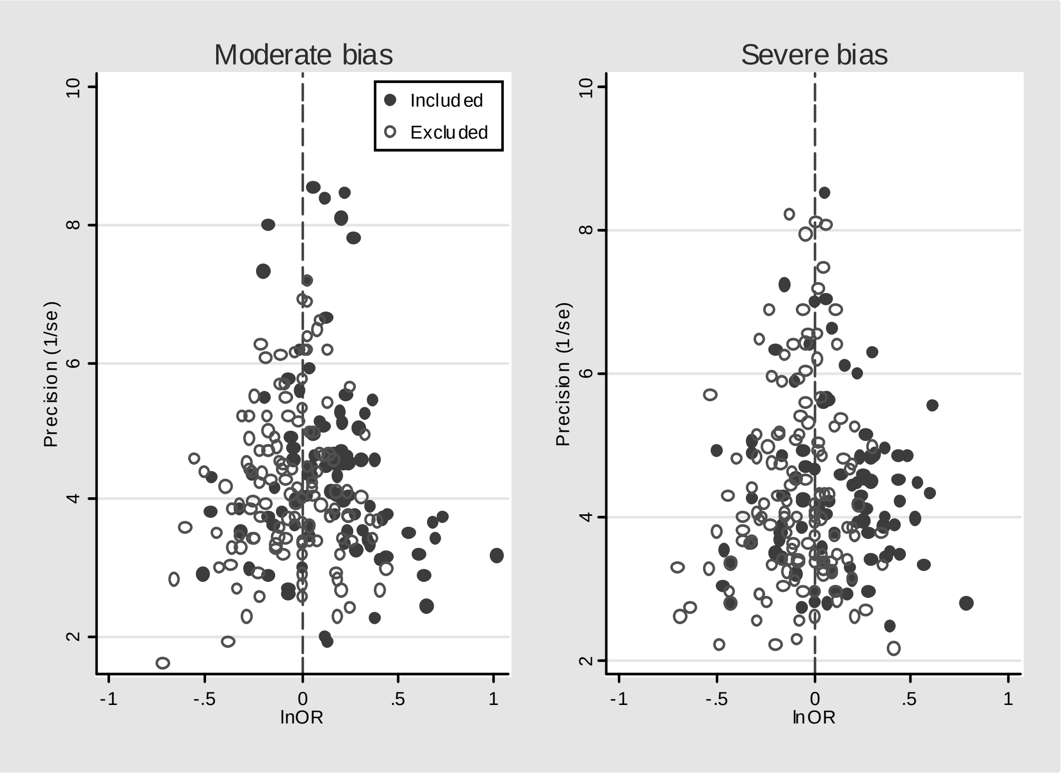
Two levels of publication bias were simulated based on that specified in Hedges & Vevea (1996) (Table 1).

Table 1 *Specification of publication bias severity based on one-sided significance*

Severity of publication bias	p-value	Probability for selection
Moderate	<0.05	1
	0.05 – 0.2	0.75
	0.2 – 0.5	0.5
	>0.5	0.25
Severe	<0.05	1
	0.05 – 0.2	0.75
	> 0.2	0.25

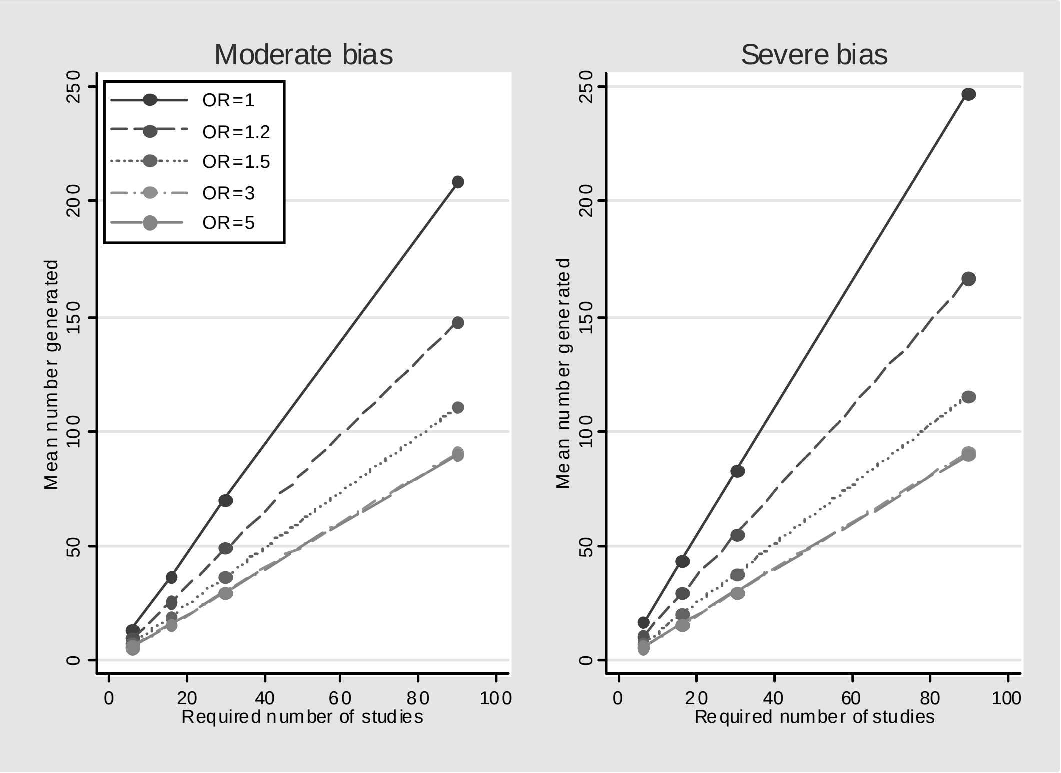
All simulations were repeated until the desired number of studies (6, 16, 30 or 90) was obtained. Figure 1 shows the included and excluded studies from the two levels of severity of selection bias in a particular scenario (underlying OR = 1 and the number of included studies = 90).

Figure 1 Funnel plot of studies from simulation of ‘moderate’ and ‘severe’ publication bias; included and excluded studies are indicated



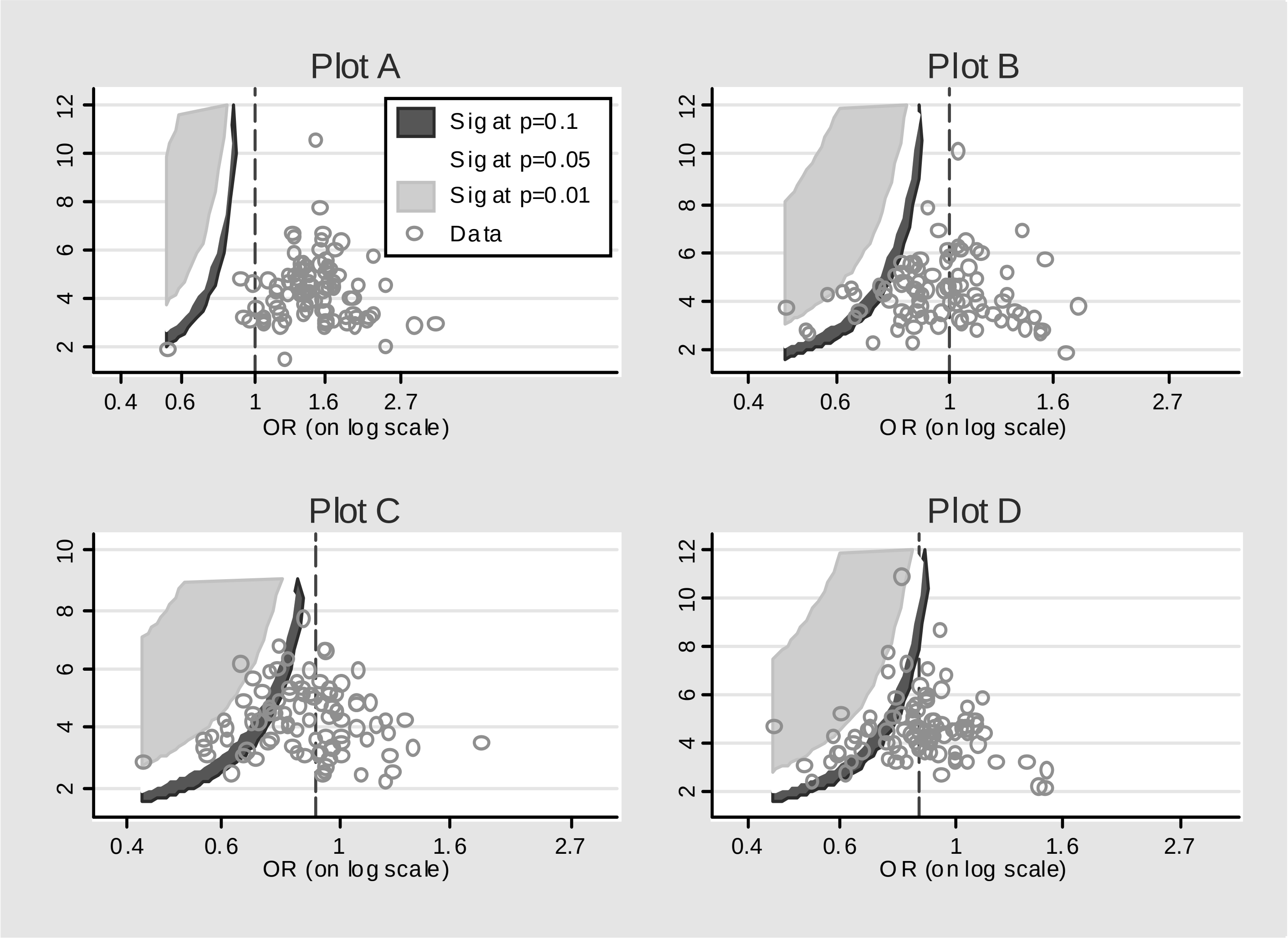
The following plot (Figure 2) shows the mean number of studies that were generated so that each simulated meta-analysis had the required number of included studies.

Figure 2 Mean number of studies generated to obtain the number of studies required under 'moderate' and 'severe' bias



As one would expect, for severe bias, the average number of studies generated to obtain the required number of included studies is greater than for moderate bias, since fewer studies are likely to be censored under moderate bias. Of particular interest is that the number of studies generated reduces as the underlying OR gets further from 1. This is because when the underlying OR is far from the null, most studies are likely to have a statistically significant effect estimate and so not have the potential to be censored. This is clearly seen in the examples in Figure 3. When the underlying OR is relatively large (plot A), few, if any of the estimates will be non-significant, and so will be less likely to be censored, thus little publication bias is induced. In the remaining plots in Figure 3 (B-D), as the underlying OR decreases, it can be seen that the number of studies having the potential to be censored increases. Therefore, inducing publication bias on the basis of significance leads to a somewhat distorted picture. Studies from a meta-analysis with a large underlying OR are less likely to be subject to censoring than studies from a meta-analysis where the underlying OR is close to the null (i.e. OR=1).

Figure 3 *Examples of censoring by level of statistical significance of the effect estimate*



This phenomenon leads to consideration of an alternative method for inducing funnel plot asymmetry.

Publication bias based on size of effect

Again, two levels of bias were induced to represent ‘moderate’ and ‘severe’ publication bias. Either the *14%* or *40%* most extreme studies showing a negative effect of the exposure (i.e. OR < 1) were censored such that the final number of studies in a meta-analysis was still 6, 16, 30 or 90. So, for 40% censoring:

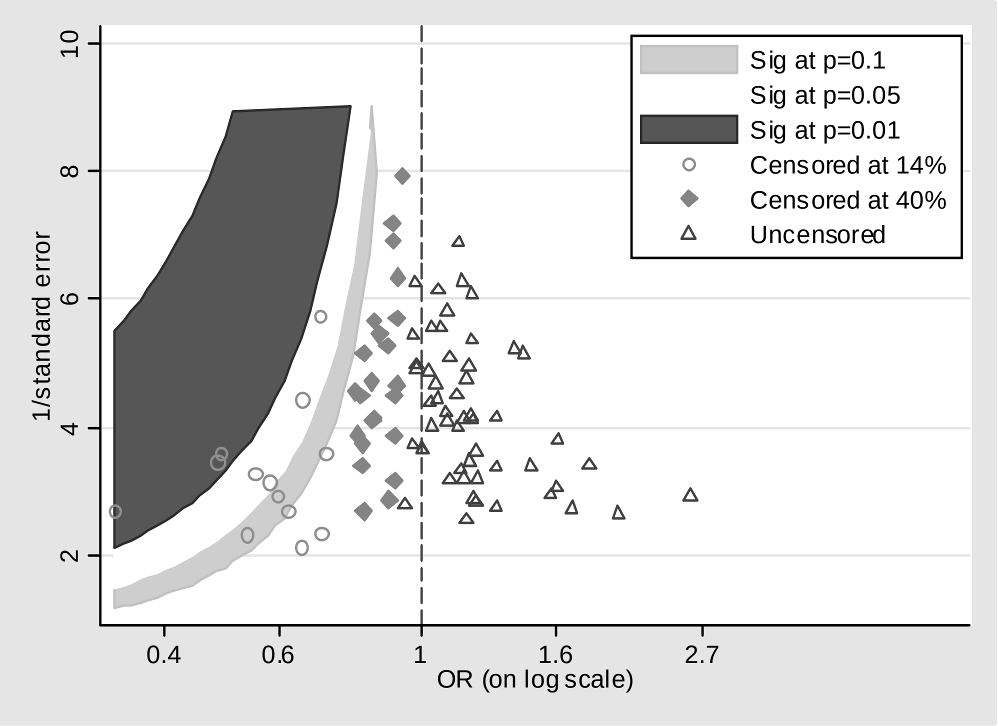
- Where the final number of studies is 6, 10 studies have been generated and the 4 studies (= 40% of the original 10 studies) giving the most extreme negative estimates (i.e. the chemical exposure is protective) have been censored.
- Where the final number of studies is 90, 150 studies have been generated and the most extreme negative 60 (= 40% of 150) studies have been censored.

This second method of inducing publication bias (by censoring x% of studies giving the most extreme negative results) is much more intuitive when looking at a funnel plot. More importantly, the number of studies censored does not depend on the size of the underlying OR as it does when publication bias is induced on the basis of p-values.

Relationship between publication bias based on p-values and effect size

How one induces publication bias for these simulations may impact on how well the different models are seen to work. The relationship between the p-values of an estimate and its size is not immediately obvious. Figure 4 may help in understanding the connection.

Figure 4 *Inducing publication bias by p-value and by size of effect*



In this example the underlying OR = 1. The circle and diamond symbols represent studies that would be censored using the 14% and 40% cut-off points, respectively, for the most extreme effect sizes. The shaded areas show the level of (one-sided) significance of the effect estimate, illustrating the studies that are more likely to be censored (i.e. those found in the shaded areas).

In summary, the simulations take the following form:

- Fixed effects meta-analysis where there is no publication bias
  - » 5 different sizes of OR (ORs of 1, 1.2, 1.5, 3, 5) and 4 sizes of meta-analysis (6, 16, 30, 90 studies) = **20 situations**
- Random effects meta-analysis where there is no publication bias
  - » 5 different magnitudes of OR, 4 sizes of meta-analysis and 3 levels of between-study heterogeneity (20%, 150%, 500% of the within-study variance) = **60 situations**
- Fixed effects meta-analysis where there is publication bias

   - » 5 different magnitudes of OR, 4 sizes of meta-analysis and four types of induced publication bias (by p-value: moderate and severe; by effect size: 14% and 40% censored) = **80 situations**

- Random effects meta-analysis where there is publication bias
   - » 5 different magnitudes of OR, 4 sizes of meta-analysis, four types of induced publication bias and three levels of between-study heterogeneity = **240 situations**

The results are based on 1000 repetitions of each of the above 400 situations. All analyses were carried out in Stata 8.2 (StataCorp 2004). Performance of the rank correlation test (Begg and Mazumdar 1994), Egger's regression test (Egger et al 1997), alternative regression tests are assessed. The alternative regression models are now specified, in addition to Egger's regression model.

Egger's fixed effects regression on the standard error

Egger's regression test is given by $\frac{y_i}{se_i} = \beta + \frac{\alpha}{se_i} + \varepsilon_i$. This is equivalent to

$y_i = \alpha + \beta.se_i + \varepsilon_i.se_i$ weighted by $\frac{1}{se_i^2}$ **(Model 1)**

where y_i is the lnOR from study i and se_i is the standard error of y_i (Egger et al 1997).

To avoid effects of the correlation between the lnOR and its standard error, the inverse of the total sample size is used as the dependent variable in an alternative Egger regression model (Model 2).

Egger's fixed effects regression on the inverse of sample size

$y_i.size_i = \alpha + \beta.size_i + \varepsilon_i$ **(Model 2)**

where $size_i$ is the total sample size of study i.

In Egger's regression model (Model 1), the error term is multiplicative: $\varepsilon_i.se_i$. This feature is not consistent with usual simple regression models. A more intuitive regression model would have error as an additive component. Because of this the following two models (Models 3 and 4a) are considered.

Linear fixed effects regression on standard error (weighted by $\frac{1}{se_i^2}$)

$$y_i = \alpha + \beta . se_i + \varepsilon_i \qquad \textbf{(Model 3)}$$

Linear fixed effects regression on sample size (weighted by $\frac{1}{se_i^2}$)

$$y_i = \alpha + \beta . size_i + \varepsilon_i \qquad \textbf{(Model 4a)}$$

The performance of Model 4a was assessed in Macaskill et al 2001[12]. In their paper they also specified this model with a different weighting structure where *A*, *B*, *C* and *D* are the values in the usual 2x2 table for calculation of the OR in each primary study. We also investigate the performance of this model (Model 4b).

Linear fixed effects regression on sample size (weighted by $\left(\frac{1}{A+B}+\frac{1}{C+D}\right)^{-1}$)

$$y_i = \alpha + \beta . size_i + \varepsilon_i \qquad \textbf{(Model 4b)}$$

To complement Model 4b, a model where the inverse of the total sample size is the dependent variable, is also assessed (Model 4c).

Linear fixed effects regression on inverse of sample size (weighted by $\left(\frac{1}{A+B}+\frac{1}{C+D}\right)^{-1}$)

$$y_i = \alpha + \frac{\beta}{size_i} + \varepsilon_i \qquad \textbf{(Model 4c)}$$

Finally, two random effects linear models are considered (Model 5 and Model 6). These random effects models are considered since between-study heterogeneity is induced in some simulations and it is therefore of interest to assess how well a random effects model for publication bias will fair in those simulations. They are the random effect versions of Model 3 and Model 4a.

Linear random effects regression on standard error (weighted by $\frac{1}{se_i^2}$)

$$y_i = \alpha + \beta . se_i + \mu_i + \varepsilon_i \qquad \textbf{(Model 5)}$$

Linear random effects regression on sample size (weighted by $\frac{1}{se_i^2}$)

$$y_i = \alpha + \beta . size_i + \mu_i + \varepsilon_i \qquad \textbf{(Model 6)}$$

Table 2 summarises these eight models.

Table 2 *Summary of the regression models assessed in these simulations*

Model based on…	Egger's fixed effects model	Linear fixed effects model	Linear random effects model
Some transformation of standard error	Model 1	Model 3	Model 5
Some transformation of sample size	Model 2	Model 4 *	Model 6

* three different weightings for each study are implemented (Models 4a, 4b and 4c)

All of these regression models test an association between the (standardized) effect size from each study in the meta-analysis, y_i, and some measure of its (standardized) precision or sample size. Evidence of an association may suggest that the meta-analysis is subject to publication bias if the smaller, less powerful studies have larger effect sizes than the more precise studies.

Analyses

The *type I error rates* (false positive rates: percentage of simulations where publication bias is incorrectly indicated at $p < 0.1$) for the rank correlation test, Egger's regression test (Model 1) and the alternative regression tests (Models 2, 3, 4a, 4b, 4c, 5 and 6) are investigated in a number of scenarios.

In the presence of publication bias, the *power* (true positive rate: percentage of simulations where publication bias is correctly indicated at $p < 0.1$) of these tests is explored. An ideal test would have type I error rates of 10% and good power to detect publication bias when it is present, regardless of the size of the true effect, the number of primary studies in the meta-analysis and the amount of between-study heterogeneity.

The two-tailed p-value for the coefficient of interest in each regression model is calculated in two ways: from the usual t-test and from a permutation test (or randomization test). The permutation test has recently been proposed by Higgins and Thompson (2004) as an alternative in meta-regression to temper high type I error rates. This approach does not rely on an assumed distribution when calculating statistical significance. Applying the permutation test in these regression models for publication bias is a novel approach.

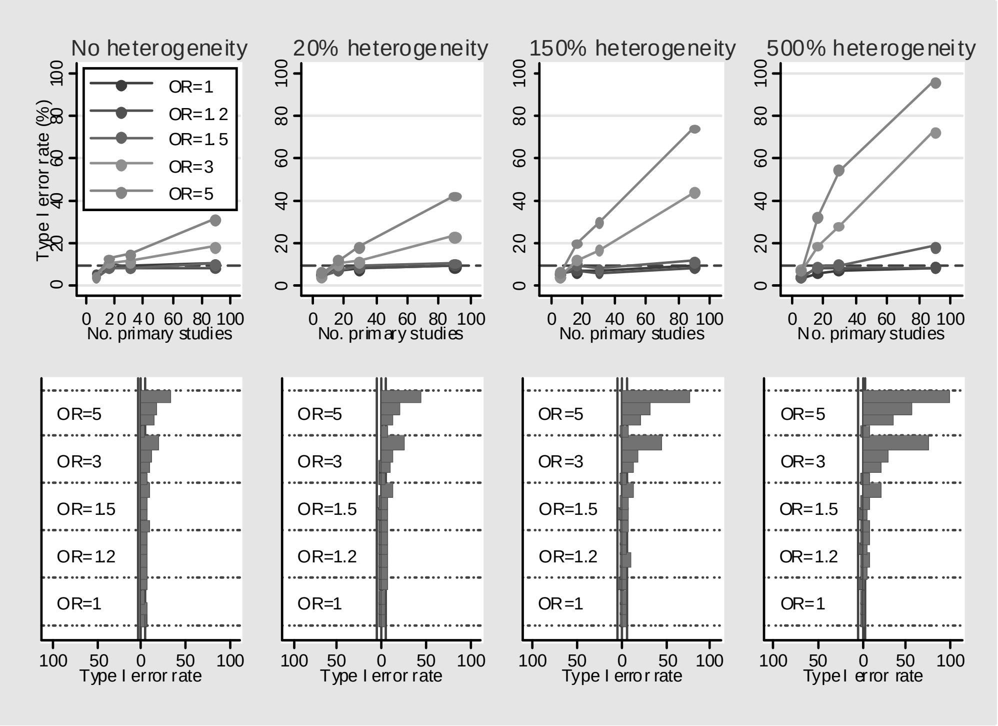
When looking at the performance of these methods one must bear in mind that when publication bias is induced on the basis of the p-value associated with the effect estimate from a study, a meta-analysis with a large true OR is unlikely to contain many studies with a non-significant estimate. These studies are unlikely to be censored and so little publication bias will be induced. Hence, for meta-analyses where the true OR ≥ 3, and publication bias has been induced by p-value, any effect of publication bias will be trivial. This must be taken into account when interpreting the performance of these methods to detect, and adjust for, possible publication bias. These results are given in the next section.

Results

The rank correlation test

Our results show that when the number of primary studies in the meta-analysis is small (n ≤ 16) and the true OR ≤ 1.5, the type I error rates for the rank correlation test are lower than the expected 10% level, regardless of the amount of between-study heterogeneity in the meta-analysis (top row of Figure 5). When the true OR is large (≥ 3), the type I error rates exceed the expected 10% level. As the amount of between-study heterogeneity increases, the type I error rates increase for the large ORs.

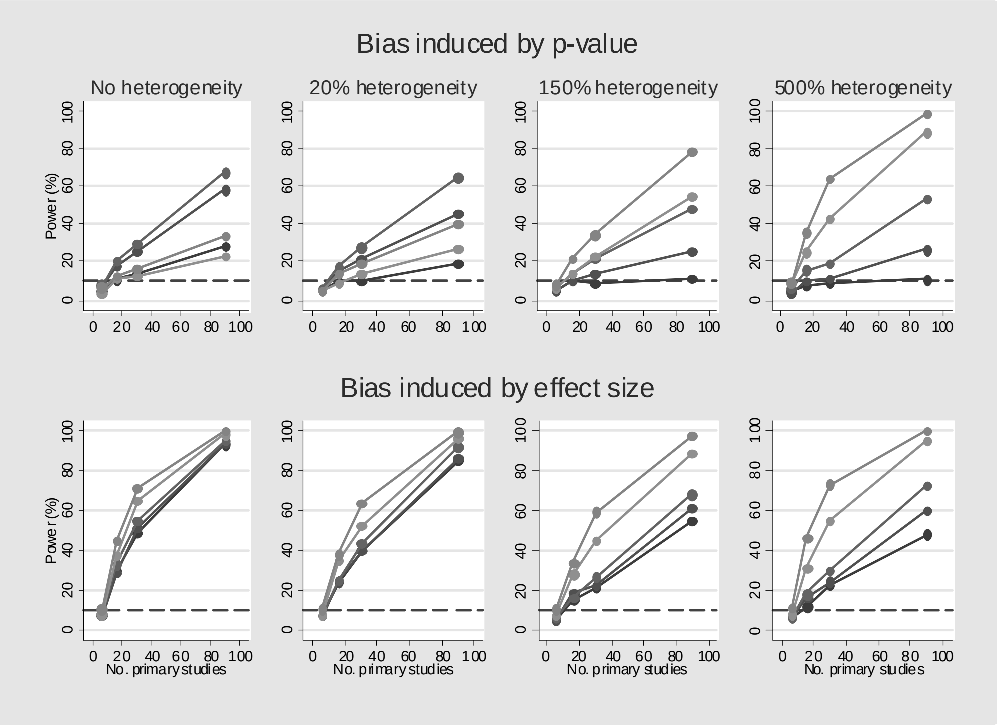
***Figure 5** Type I error rates for the rank correlation test*



From the bottom row of Figure 5 one can see that there is great imbalance in the tail probability areas, with the test suggesting a greater percentage of significant positive scores from the correlation, than significant negative scores.

The power of the rank correlation test must be interpreted in light of these type I error rates. As such, one can be confident about the power of this test when the true OR ≤ 1.5, since the type I error rates are as expected. However, this is not the case for the larger ORs (≥ 3), where power cannot be distinguished from high type I error rates. Figure 6 illustrates the power of the rank correlation test under different scenarios. Results in the top row correspond to severe publication bias induced by p-value. On the bottom row, severe publication bias is induced based on the size of effect.

Figure 6 *Power of the rank correlation test to detect 'severe' bias*



The rank correlation test appears more powerful when severe publication bias is based on effect size (bottom row), rather than p-value (top row). Except for when the true OR is large, power tends to decrease with increasing amounts of between-study heterogeneity, and it is difficult to distinguish power (Figure 6) from the type I error rates (Figure 5). Although power of the rank correlation test is much lower when there is 'moderate' publication bias, the same general trend seen for 'severe' bias is observed (results not shown).

The regression model tests

Results based on the p-values from the t-test are considered first, comparisons with the permutation test are made later. Egger's test (Model 1) has type I errors which exceed the expected 10% level particularly when the true OR is large and as between-study heterogeneity increases (Figure 7). There is an imbalance in the tail probability areas for Egger's test. This imbalance gets more noticeable as the underlying OR and between-study heterogeneity increase, as seen in the bottom row of Figure 7.

Figure 7 *Type I error rates for Egger's regression test*

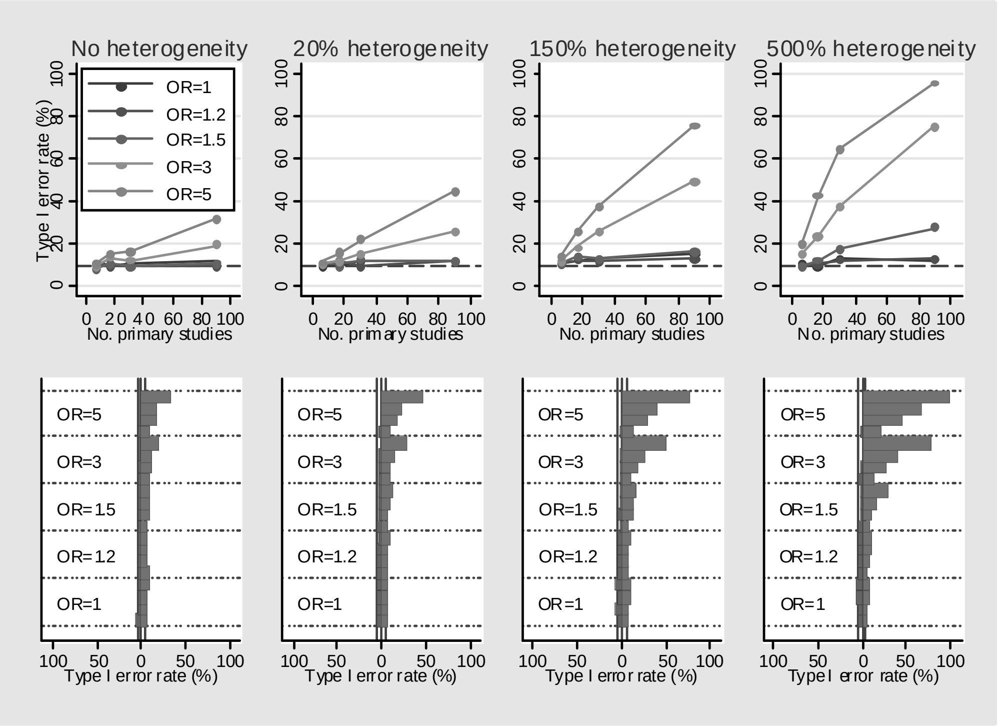
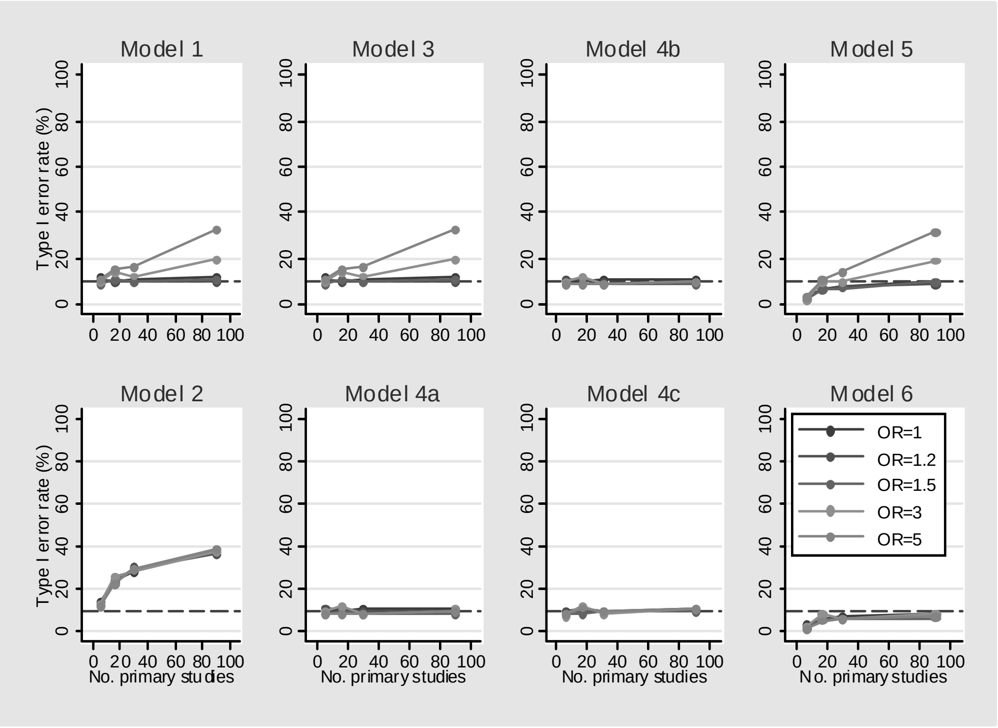


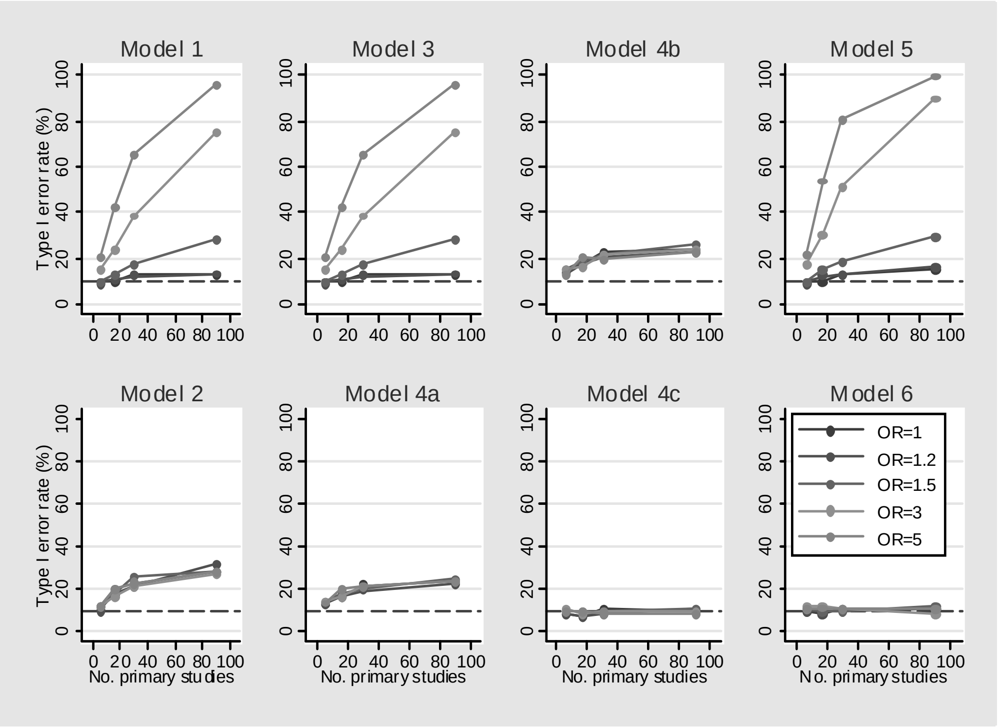
Figure 8 *Type I error rates for all the regression tests when there is no induced between-study heterogeneity*



The type I error rates for Model 2 (the alternative Egger regression model based on sample size) are also higher than expected, though unlike with Model 1, it is the case regardless of the size of the underlying OR.

Of the 6 remaining alternative regression models (Models 3, 4a, 4b, 4c, 5 and 6), the type I error rates are as expected in many of the different scenarios. Both of the linear regression models specifying standard error as the independent variable, Model 3 (fixed effects) and Model 5 (random effects), have slightly elevated type I error rates when the true OR is large (≥ 3) (Figure 8). These type I error rates increase with increasing amounts of between-study heterogeneity present in the meta-analysis (see Figure 9 for when between-study heterogeneity is 500% of within-study heterogeneity).

***Figure 9** Type I error rates for the regression tests when induced between-study heterogeneity is 500% of within-study heterogeneity*



Model 6 has slightly lower type I error rates that one would expect (Figure 8), especially when the number of primary studies in the meta-analysis is small. But when a great deal of between-study heterogeneity exists, this model has the expected type I error rate of 10% (Figure 9). Models 4a, and 4b appear to perform well (i.e. the expected type I error rate of 10% is attained) regardless of the size of the true OR or the number of primary studies in the meta-analysis (Figure 8). However, as the amount of between-study heterogeneity in the meta-analysis increases, the type I error rates also increase (Figure 9).

On the other hand, Model 4c appears to have type I error rates of 10% regardless of the size of the true OR, the number of primary studies and the amount of between-study heterogeneity in the meta-analysis (Figure 9). These error rates appear balanced in the tail probability areas (results not shown) unlike Egger's test (Figure 7). Model 4c has relatively good power to detect severe publication bias when there is no between-study heterogeneity compared to the other models, regardless of how publication bias is induced (Figures 10 and 11).

***Figure 10** Power of all models to detect publication bias induced by effect size when there is no between-study heterogeneity*

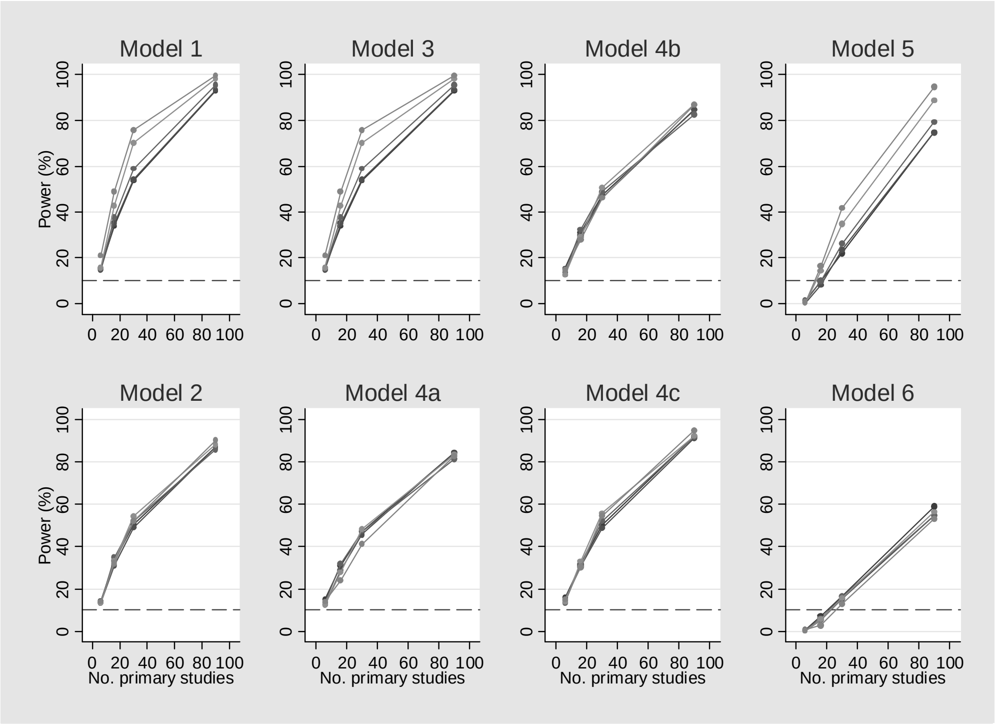
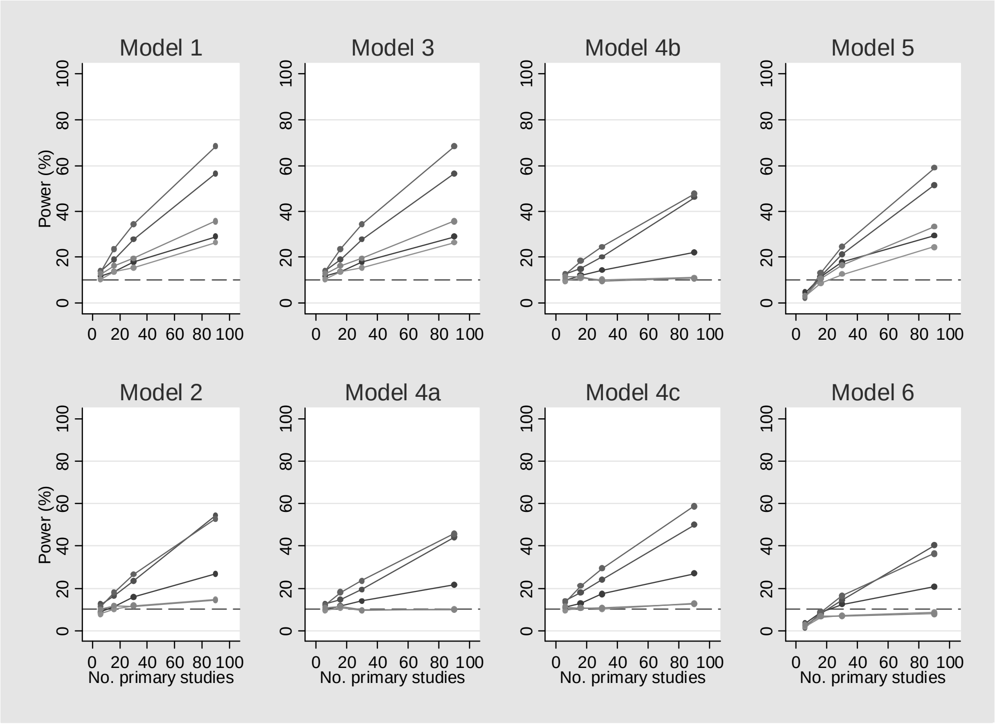


Figure 11 *Power of all models to detect publication bias induced by p-value when there is no between-study heterogeneity)*



When there is considerable between-study heterogeneity all models, except Models 1, 3 and 5 have very low power to detect publication bias whether it is induced by effect size or p-value (Figures 12 and 13).

***Figure 12** Power of all models to detect publication bias induced by effect size when there is 500% between-study heterogeneity*

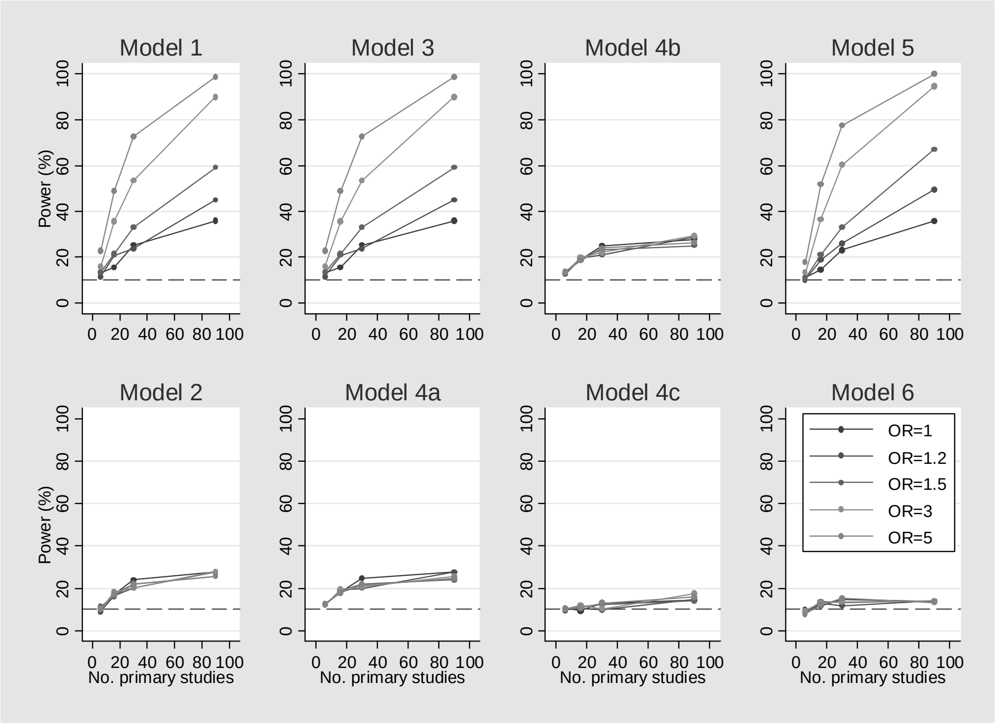
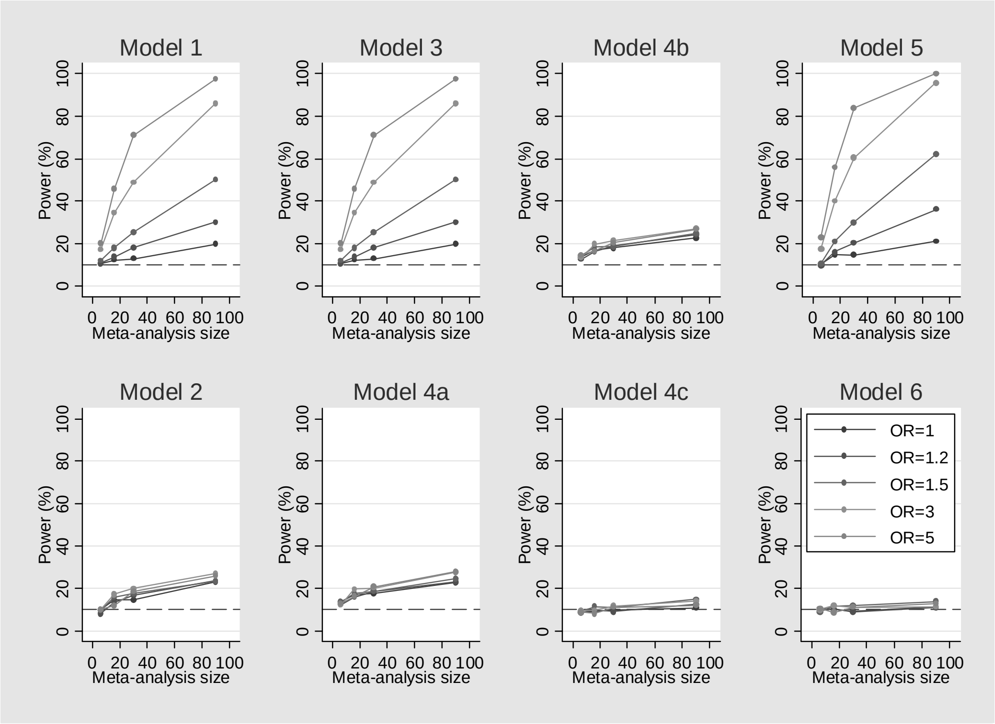


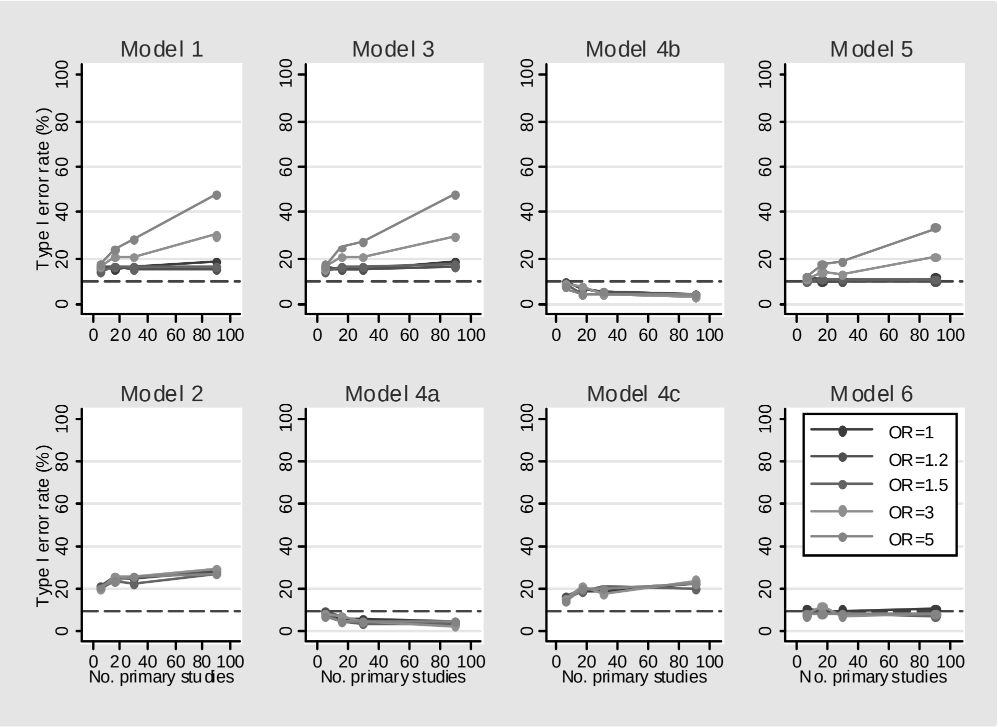
***Figure 13** Power of all models to detect publication bias induced by p-value when there is 500% between-study heterogeneity*



However, it is difficult to distinguish the power of Models 1, 3 and 5 from the type I error rates when between-study heterogeneity is large, as the patterns seen in Figures 12 and 13 reflect those seen when there is no induced publication bias (Figure 9). Clearly there is a trade-off between the type I error rates and power of these tests when assessing their performance. In these situations power cannot be interpreted due to the high type I error rates. Therefore, although Model 4c may not be powerful when a great deal of between-study heterogeneity exists, the power is distinguishable from the type I error rates. These results suggest that Model 4c is superior to all other regression models assessed here, because of the appropriate type I error rates and reasonable power to detect publication bias. Power to detect 'moderate' publication bias is lower than that to detect 'severe' publication bias for all models. However, the same general trend in power is seen for 'moderate' publication bias as the level of between-study heterogeneity increases, as it is for severe publication bias (results not shown). These results are based on p-values from the usual t-test. Results based on p-values from the permutation test are now described.

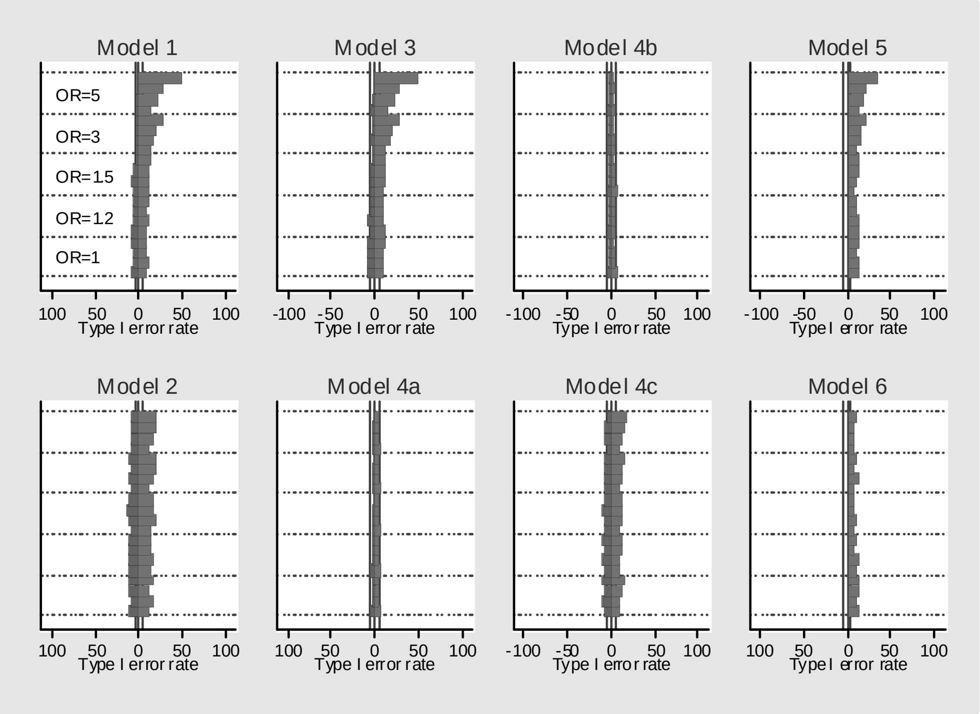
In the absence of induced publication bias, Model 6 (random effects model on sample size) looks to be the most appropriate model since it is the only model providing the expected type I error rates regardless of the underlying OR or the number of primary studies in the meta-analysis for no between-study heterogeneity (Figure 14).

Figure 14 *Type I error rates for all models based on p-values from the permutation test when there is no between-study heterogeneity*



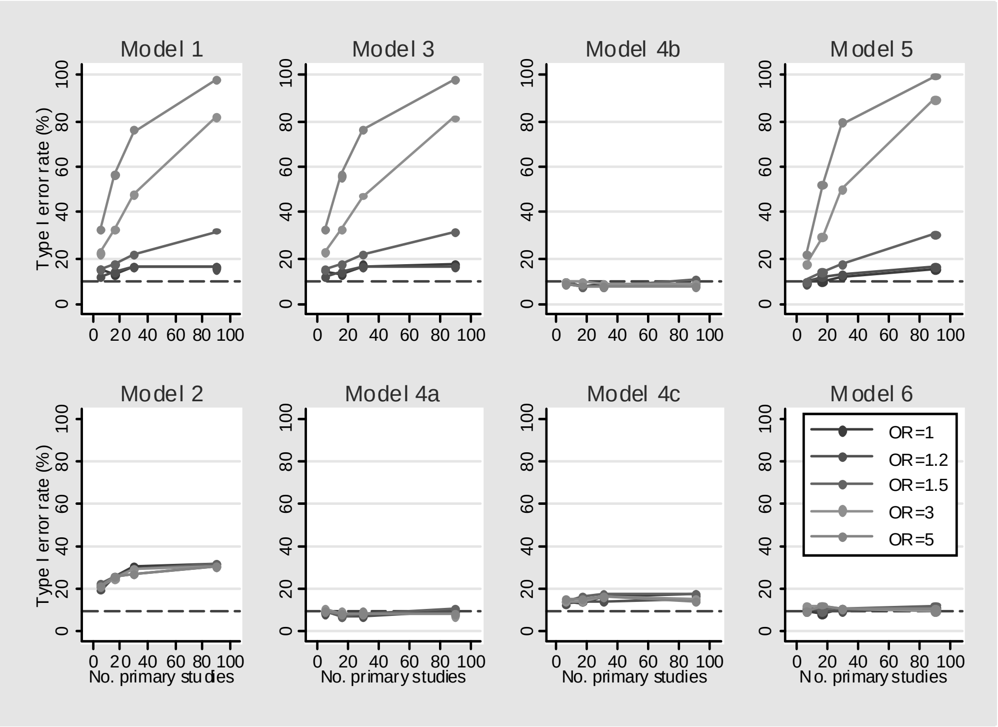
However, there is substantial imbalance in the tail probability areas for this model (Figure 15).

Figure 15 *Tail probability areas for all models using the permutation test*



Models 4a and 4b have lower than expected type I error rates, but the tail probability areas look quite balanced. As the amount of between-study heterogeneity increases, models 4a and 4b demonstrate the expected type I error rates (see Figure 16).

***Figure 16** Type I error rates for all models using the permutation test when between-study heterogeneity is 500% of the within-study heterogeneity*



All other models have inappropriately high type I error rates, increasing as the amount of between-study heterogeneity increases (see Figures 14 and 16 for comparison), although the type I error rates for Model 4c are only moderately above that expected.

When p-values are based on the permutation test, of the three models performing reasonably well in terms of type I error rates (Models 4a, 4b and 6), Model 6 has the highest power to detect publication bias, regardless of how it is induced (Figures 17 and 18).

Figure 17 *Power for all models using the permutation test based on effect size when there is no between-study heterogeneity*

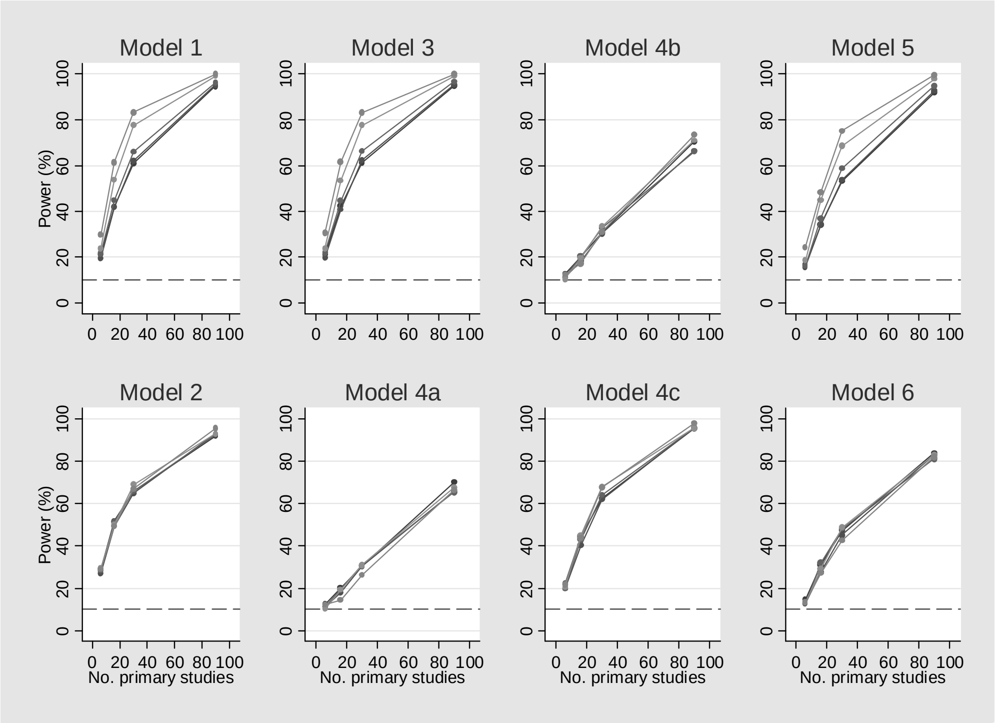
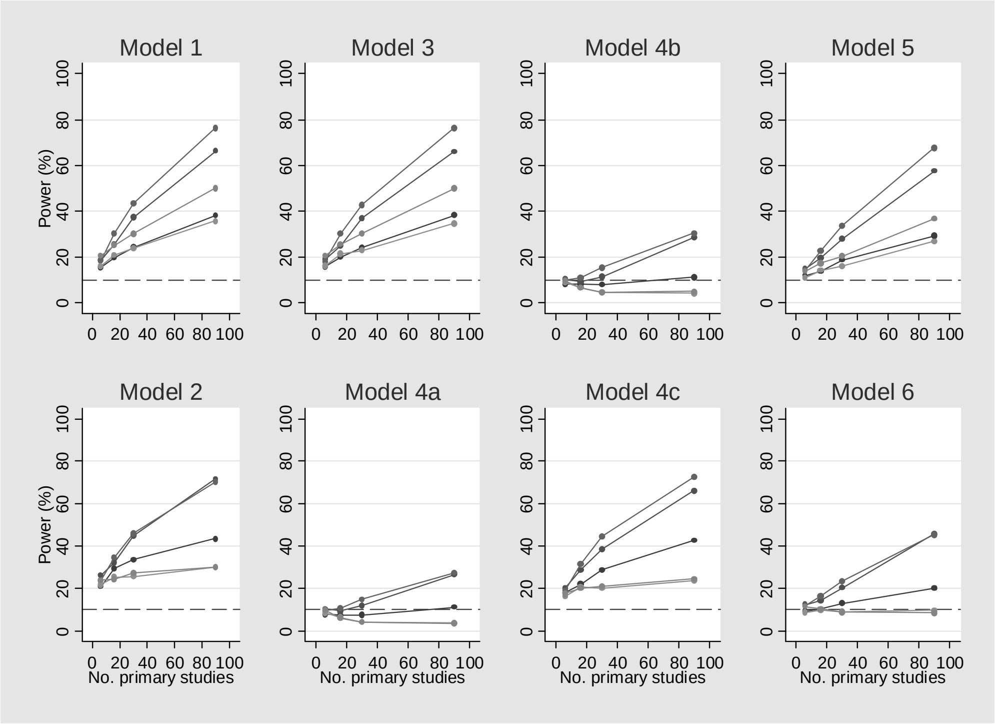


Figure 18 *Power for all models using the permutation test based on p-values when there is no between-study heterogeneity*



When there is considerable between-study heterogeneity, we see again (Figures 19 and 20) that most models have very little power to detect publication bias, while for the remaining models (Models 1, 3 and 5), power and type I error rates are difficult to disentangle (compare Figure 16 with Figures 19 and 20).

Figure 19 *Power for all models using the permutation test based on effect size when there is 500% between-study heterogeneity*

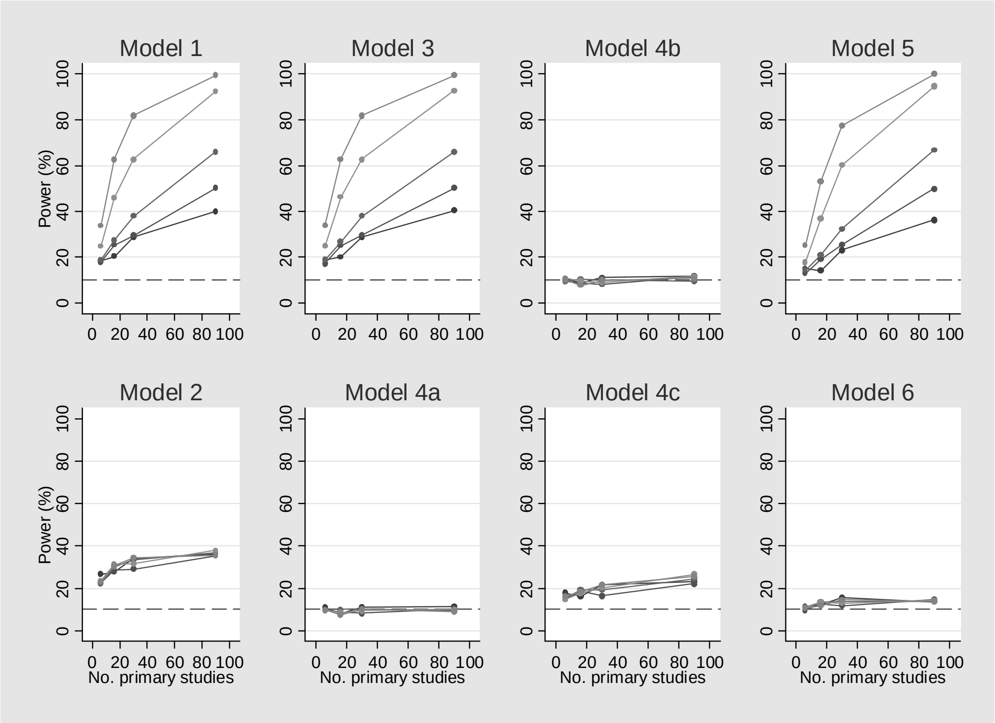
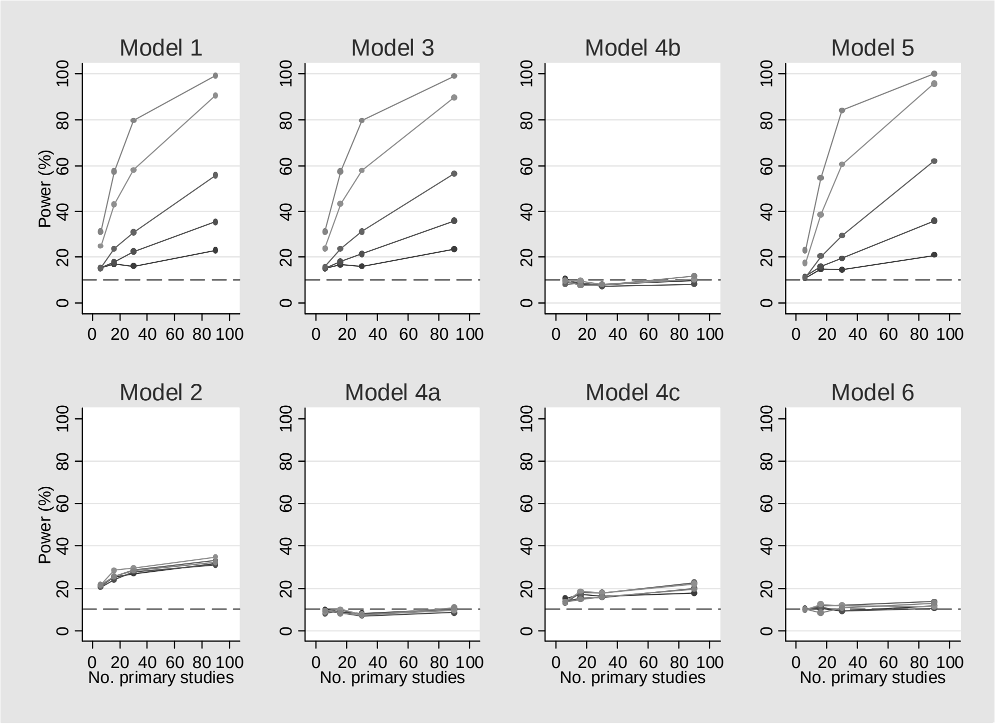


Figure 20 *Power for all models using the permutation test based on p-value when there is 500% between-study heterogeneity*



Based on estimates of type I error rates and power, initial findings suggest that use of models based on the permutation test (in order to address inflated type I errors) are not necessarily superior to those based on the usual t-test.

Implementing the alternative regression model in Stata

Firstly the weighting given to each study in the regression, must be calculated. This is obtained by

```
gen prec_lnodds = 1/(1/(rt + rc) + 1/(rtn + rcn))
```

where `rt` is the number of subjects in the treatment group for which an event was observed, `rtn` is the number of treatment group subjects for which an event was not observed, `rc` is the number of subjects in the control group for which an event was observed and `rcn` is the number of control group subjects for which an event was not observed[3]. The regression test is implemented using

```
regress ln_ES inv_SS [aweight=prec_lnodds]
```

where `ln_ES` is the natural log of the OR and `inv_SS` is the inverse of the total sample size. The following is the output given by Stata:

```
(sum of wgt is   6.8567e+03)

      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  1,    88) =    0.21
       Model |  .012798488     1  .012798488           Prob > F      =  0.6459
    Residual |  5.29906085    88  .060216601           R-squared     =  0.0024
-------------+------------------------------           Adj R-squared = -0.0089
       Total |  5.31185934    89  .059683813           Root MSE      =  .24539

------------------------------------------------------------------------------
        ln_ES |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      inv_SS |   6.867071   14.89534     0.46   0.646    -22.73428    36.46842
       _cons |   1.636909   .0494751    33.09   0.000     1.538588    1.735231
------------------------------------------------------------------------------
```

In this example, the coefficient for the inverse sample size variable, `inv_SS`, is not significant (p=0.65), suggesting there is little evidence to reject the null hypothesis (of no association between lnOR and inverse sample size), providing no evidence for the presence of publication bias.

References

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

Duval S, Tweedie R (2000) Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 56:455-463.

Egger M, Davey Smith G, Schneider M, Minder C (1997) Bias in meta-analysis detected by a simple, graphical test. BMJ 315:629-634.

Hedges LV, Vevea JL (1996) Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. Journal of Educational and Behavioral Statistics 21(4):299-332.

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

Higgins JPT, Thompson SG (2004) Controlling the risk of spurious findings from meta-regression. Statistics in Medicine 23:1663-1682.

Macaskill P, Walter SD, Irwig L (2001) A comparison of methods to detect publication bias in meta-analysis. Statistics in Medicine 20:641-654.

Peters JL, Sutton AJ, Jones DR, Rushton L, Abrams KA (2004). A review of the use of systematic review and meta-analysis methods to evaluate animal toxicology studies. Technical Report 04-02. Department of Health Sciences, University of Leicester.

StataCorp (2004) Stata Statistical Software. Release 8.2. College Station, TX: Stata Corporation.