STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:641–654

A comparison of methods to detect publication bias in meta-analysis

Petra Macaskill[1,*,†], Stephen D. Walter[2] and Les Irwig[1]

[1]*Department of Public Health and Community Medicine, University of Sydney, NSW, Australia 2006*
[2]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5*

SUMMARY

Meta-analyses are subject to bias for many of reasons, including publication bias. Asymmetry in a funnel plot of study size against treatment effect is often used to identify such bias. We compare the performance of three simple methods of testing for bias: the rank correlation method; a simple linear regression of the standardized estimate of treatment effect on the precision of the estimate; and a regression of the treatment effect on sample size. The tests are applied to simulated meta-analyses in the presence and absence of publication bias. Both one-sided and two-sided censoring of studies based on statistical significance was used. The results indicate that none of the tests performs consistently well. Test performance varied with the magnitude of the true treatment effect, distribution of study size and whether a one- or two-tailed significance test was employed. Overall, the power of the tests was low when the number of studies per meta-analysis was close to that often observed in practice. Tests that showed the highest power also had type I error rates higher than the nominal level. Based on the empirical type I error rates, a regression of treatment effect on sample size, weighted by the inverse of the variance of the logit of the pooled proportion (using the marginal total) is the preferred method.

1. INTRODUCTION

Checking for evidence of publication bias should be undertaken routinely at a preliminary stage of a meta-analysis [1]. This bias arises when the published studies identified for inclusion in the meta-analysis do not represent all studies on the topic of interest [2]. For example, studies with statistically significant effects and positive treatment outcomes are more likely to be published [3, 4], resulting in a biased estimate of the effect of treatment in the meta-analysis. Both the decision to submit a study for publication and the probability that a journal will accept it for publication are associated with the study results [5]. In practice, the studies which are less likely to appear in the published literature tend to be the less conclusive ones (because of smaller sample sizes or less statistical precision) or those

*Correspondence to: Petra Macaskill, Department of Public Health and Community Medicine, University of Sydney, NSW, Australia 2006
†E-mail: petram@health.usyd.edu.au

Received 1998
Accepted April 2000

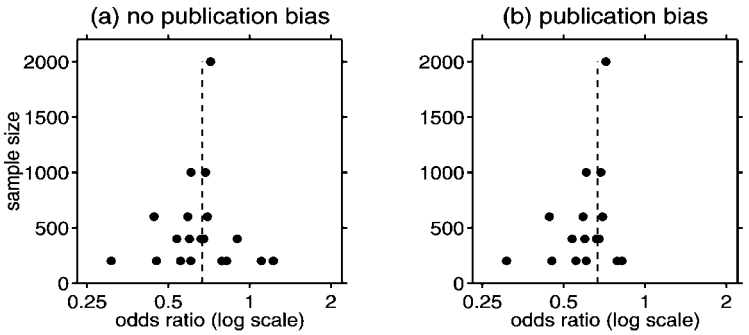

Figure 1. Examples of funnel plots based on simulated meta-analyses when there is (*a*) no publication bias and (*b*) publication bias present. The dotted line indicates the true treatment effect.

where the treatment effect is small. Although searching for relevant unpublished studies is important and may sometimes alleviate publication bias, identifying such studies may be difficult. Hence we need methods to detect publication bias, based on the data in the available studies.

Creating a funnel plot of sample size against estimated treatment effect (for example, the relative risk of death in treated versus control subjects) is the most commonly used method for detecting publication bias [1, 6]. In the absence of publication bias, the plot is expected to take on its characteristic funnel shape, with the amount of scatter about the true effect (a vertical line of symmetry) decreasing with increasing sample size (see Figure 1(*a*)). Typically, there will be relatively few large studies and relatively more small ones.

The presence of publication bias will change the shape of this plot. For instance, if the true effect is not zero, asymmetry in the plot will occur because some small studies with an effect near zero will be missed in the meta-analysis, but studies of similar size which show a large estimated effect will be included (Figure 1(*b*)). This will result in an association between sample size and treatment effect among the included studies. Other factors such as heterogeneity in treatment effect between low and high risk groups can also lead to asymmetry in the funnel plot [7].

Two simple methods have been proposed for detecting publication bias. First, Begg's rank correlation method [1, 8] uses Kendall's tau to evaluate the association between the (standardized) effect and variance of the treatment effect. Second, Egger *et al.* [9] fit a simple linear regression to the Galbraith plot [10] of the standardized treatment effect against its precision (the inverse of its standard error), but unlike the Galbraith approach the line is not forced through the origin. It has been shown that Begg's method has low power with continuous (normally distributed) data, particularly when the number of studies is small [8]. We are not aware of any such evaluation of the Egger approach, but it is known to be intrinsically biased because: (i) the independent variable is subject to sampling variability; (ii) the standardized treatment effect is correlated with its estimated precision; and (iii) for binary data, the independent regression variable is a biased estimate of the true precision, with larger bias for smaller sample sizes [11–13]. Note that the term bias is used in two ways in this paper. We will refer explicitly to publication bias as described above to distinguish it from bias in the statistical methods used to detect it.

Given these concerns about the Egger method, we have conducted a simulation study to investigate its performance and that of Begg's rank correlation method. We focus on meta-analyses involving binary data using scenarios that reflect the number of studies commonly included in practice. Notation and underlying assumptions are given in Section 2. The Egger and Begg methods are described in Section 3 and an alternative approach (funnel plot regression) intended to overcome some of the problems with the Egger approach is outlined. The methods and scenarios used for the simulation study are described in Section 4 and the simulation results are given in Section 5. A discussion of the results and the relative performance of the different methods is given in Section 6.

2. NOTATION

Our investigations consider the situation where each study comprises a treated and a control group, the outcome is assumed to be binary, and the measure of effect is the log odds ratio. Without loss of generality, the true underlying log-odds ratio is assumed to be $\leqslant 0$ (odds ratio $\leqslant 1$) to reflect either no effect or a protective treatment.

For each of k studies of size n_i $(i=1,\ldots,k)$, let a_i and b_i represent the observed number who experience the outcome of interest in the treated and control groups, respectively. The corresponding numbers not developing the outcome are denoted by c_i and d_i. Using the observed frequencies, the estimated treatment effect can be expressed as $t_i=\ln[(a_i d_i)/(b_i c_i)]$, and the corresponding sampling variance of the observed treatment effect (v_i) is computed as $1/a_i+1/b_i+1/c_i+1/d_i$. Assuming that the underlying effect of treatment is the same across studies (fixed effect), a meta-analysis of the k studies gives some estimate $\hat{\theta}$ of the true effect θ.

A 'pooled' estimate of the proportion with the outcome in study i is given by $p_i^*=(a_i+b_i)/n_i$, with variance $1/[p_i^*(1-p_i^*)n_i]$ (or $1/(a_i+b_i)+1/(c_i+d_i)$). The inverse of this variance is used as a weight in our alternative funnel plot regression method (see Section 3).

3. METHODS FOR DETECTING PUBLICATION BIAS

3.1. Begg's rank correlation method

Begg's method (BV) uses Kendall's tau to test for correlation between the standardized treatment effect t_i^* and the variance of the treatment effect (v_i), where $t_i^*=(t_i-\bar{t})/\sqrt{v_i^*}$, $\bar{t}=(\sum t_j/v_j)/\sum(1/v_j)$, and $v_i^*=v_i-(\sum 1/v_j)^{-1}$ [8]. Alternatively, the test can be based on the correlation between t_i^* and the sample size for each study (n_i) (BN) [1]. Treatment effects are standardized to obtain a set of estimates that can be assumed to be independent and identically distributed under the null hypothesis of no publication bias [1, 8].

Each study is assigned ranks for both t_i^* and v_i (or n_i). The correlation between the two sets of ranks can be computed by ordering the (t_i^*, v_i) pairs according to the value of t_i^* and then comparing the ranks v_i and $v_j (i \neq j)$ for all $k(k-1)/2$ possible pairs of studies. If P represents the number of pairs ranked in the same order by t_i^* and v_i, and Q the number in which the order is reversed, a normalized test statistic is computed as $z=(P-Q)/[k(k-1)(2k+5)/18]^{0.5}$

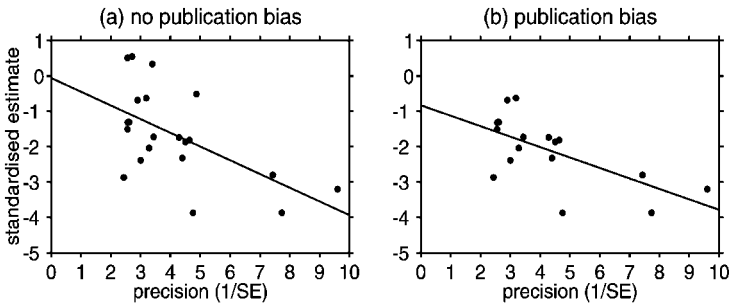

Figure 2. Examples of the Egger regression method using a simulated meta-analysis when there is (*a*) no publication bias and (*b*) publication bias present. The fitted (unweighted) regression line is shown on each graph.

if there are no ties. A more complex formula is used when there are ties present. (See Begg [1] for a more detailed description and a worked example.)

3.2. Egger regression method

A regression model is fitted using the standardized estimate of the treatment effect ($z_i = t_i/\sqrt{v_i}$) as the dependent variable and its precision ($1/\sqrt{v_i}$) as the independent variable [9]. Either an unweighted model (EU), or a model weighted by the inverse of the variance for each study (EW) can be used.

This approach assumes that when there is no publication bias, the intercept will have an expected value of zero and the slope will be an unbiased estimate of the true (underlying) effect (Figure 2(*a*)). If, on the other hand, smaller studies show effects that differ on average from larger studies, the fitted line will not pass through the origin (Figure 2(*b*)). Hence, the size of the intercept is taken as the basis of a test for publication bias.

As noted earlier, this method violates the usual assumptions of simple linear regression. Measurement error in the independent variable is present because the standard error is estimated from the observed data, and is therefore subject to sampling error. This results in a biased estimate of the regression slope. The extent and direction of the bias depends on the variances and covariance of the true value of the independent variable and the measurement error. Attenuation of the regression slope, with a corresponding shift in the intercept, will occur when there is no correlation between the measurement error and the true value of the independent variable [12]. If we assume that the treatment being evaluated is protective ($\theta < 0$), we would expect attenuation of the slope to result in a negative intercept even in the absence of publication bias.

3.3. Alternative method (Funnel plot regression)

To overcome the problems with the Egger approach, we consider fitting a regression directly to the data using the treatment effect (t_i) as the dependent variable, and study size (n_i) as the independent variable. The observations are weighted by the inverse variance of the estimate to allow for possible heteroscedasticity (FIV).

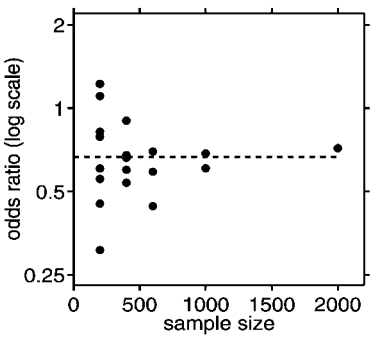

Figure 3. Example of the funnel plot regression method for a simulated meta-analysis when there is no publication bias. The dotted line indicates the true treatment effect.

When there is no publication bias, the regression slope has an expected value of zero (Figure 3). A non-zero slope would suggest an association between treatment effect and sample size, possibly due to publication bias. If we were now to introduce publication bias by, for example, removing the studies corresponding to the three highest odds ratios, the slope of the regression line would be positive and the intercept would have negative bias. Unfortunately, as with the Egger method, the weights are based on the observed data and are subject to random variability; hence there may be bias in the slope because the variance is a function of the estimated log-odds ratio. If, for a given study, the observed odds ratio happens to be closer to 1 than the true underlying value, then the weight given to that observation will tend to be higher. Thus a negative slope may occur even though there is no publication bias. This problem led us to consider an alternative weighted regression (FPV) with the weights being the reciprocal of the pooled variance for each study, that is, the variance of the pooled estimate resulting from combining the data for the two groups. This form of weighting should reduce the correlation between the weight and the dependent variable.

4. SIMULATIONS

4.1. Scenarios

The scenarios used in our simulations were chosen to reflect the number of studies and their sample sizes commonly included in meta-analyses in practice; these parameters were based on characteristics observed in 70 meta-analyses published in seven leading medical journals between 1990 and 1995 [14]. Each simulated meta-analysis comprised a total of 21 studies. Two configurations of study size were used: (i) configuration A, 11 studies of 100 per treatment group, 6 of 200/group and 4 of 300/group (7000 total subjects) (ii) configuration B, 10 of 100/group, 5 of 200/group, 3 of 300 per group, 2 of 500/group, 1 of 1000/group (9800 subjects). The underlying treatment effects (θ) considered were log-odds ratios of $0.0, -0.405, -0.693$ and -1.386, corresponding to odds ratios of $1, 2/3, 1/2$ and $1/4$. Each scenario was repeated with and without publication bias being applied.

The underlying outcome proportion in the control group was taken to have a uniform distribution between 0.1 and 0.5. Once this proportion had been determined by a random

Table I. Mean intercept (Egger regressions) and slope (funnel plot regressions) of simulated meta-analyses at each chosen treatment effect (true log-odds ratio) for configurations A and B under the scenario of no publication bias.

True log-odds ratio	Configuration	Egger method (intercept)		Funnel plot regression (slope)	
		EU	EW	FIV	FPV
0.0	A	-4.3×10^{-3}	2.2×10^{-3}	-4.9×10^{-7}	-4.5×10^{-7}
	B	3.6×10^{-3}	1.2×10^{-2}	-1.1×10^{-6}	-1.0×10^{-6}
−0.405	A	$-1.2 \times 10^{-1*}$	$-6.5 \times 10^{-2*}$	$-1.1 \times 10^{-5*}$	7.5×10^{-6}
	B	$-3.8 \times 10^{-2*}$	1.4×10^{-2}	$-3.5 \times 10^{-6*}$	-1.1×10^{-6}
−0.693	A	$-1.6 \times 10^{-1*}$	$-6.4 \times 10^{-2*}$	$-4.1 \times 10^{-5*}$	-4.2×10^{-6}
	B	$-4.0 \times 10^{-2*}$	$5.6 \times 10^{-2*}$	$-1.0 \times 10^{-5*}$	-1.0×10^{-6}
−1.386	A	$-3.7 \times 10^{-1*}$	$-1.4 \times 10^{-1*}$	$-1.0 \times 10^{-4*}$	3.8×10^{-6}
	B	$-1.1 \times 10^{-1*}$	$1.1 \times 10^{-1*}$	$-2.6 \times 10^{-5*}$	1.5×10^{-6}
Max SE	A	8.8×10^{-3}	9.6×10^{-3}	4.1×10^{-6}	4.3×10^{-6}
	B	5.9×10^{-3}	7.5×10^{-3}	9.4×10^{-7}	9.6×10^{-7}

Max SE = maximum standard error of parameter estimates.
*Significantly different from 0 ($|z| > 1.96$).
Configuration A (21 studies: 11×100/grp, 6×200/grp, 4×300/grp).
Configuration B (21 studies: 10×100/grp, 5×200/grp, 3×300/grp, 2×500/grp, 1×1000/grp).

number generator, the corresponding proportion in the treated group was calculated from the assumed value of the true odds ratio. The simulated results for each study were then generated using a binomial random number generator. All three methods of analysis were applied to each realized set of studies. The process was repeated 10000 times to provide distributions of the parameter estimates and their statistical significance. A two-sided 10 per cent level of significance was chosen to conform with that used by Egger *et al.* [9] and also with the recommendation by Begg [1, 8] to use a more liberal significance level than the usual 5 per cent two-sided level.

The empirical type I error rate (test size) for each test was ascertained for the scenario of no publication bias. The empirical power (1-type II error rate) was determined for situations where publication bias was present. The size and power of the tests for publication bias are analogous to (1-specificity) and sensitivity of the methods. Based on 10000 replications, the maximum standard error for test size was 0.4 per cent and for power was 0.5 per cent. The mean and standard error (SE) of the distribution of the estimate (intercept for Egger regressions, slope for funnel plot regressions) was computed when no publication bias was present. The z-score ($z = \text{mean}/\text{SE}$) was used to test whether the mean differed significantly from zero. The maximum standard error across treatment effects is given in Table I for each method and configuration of study size.

Additional simulations were conducted to examine how the power of the tests varied when (i) the number of studies per meta-analysis was doubled or trebled; and (ii) fixing the number studies per meta-analysis at 21, the sample size for each study was increased by a factor of 2 and 3. The extra simulations were restricted to one-sided censoring of studies (see below)

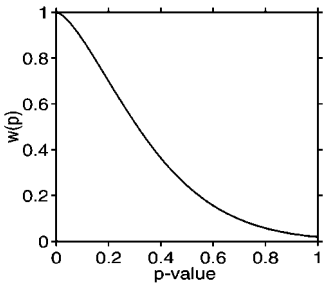

Figure 4. The weight function for selecting studies for inclusion in the meta-analysis as a function of the p-value.

and a true treatment effect of 0 or -0.405, based on the findings in the main analyses. All simulations and analyses were undertaken using SAS [15].

4.2. Study selection

One- and two-sided censoring mechanisms used by Begg and Mazumdar [8] were adopted to introduce publication bias to the data. Briefly, the probability of selection of a study for inclusion in a meta-analysis is given by the weight function $w_i(p_i) = \exp\{-\beta p_i^{\alpha}\}$, where the function is evaluated at $p_i = \Phi(t_i/\sqrt{v_i})$ for one-sided censoring (assuming a protective treatment effect, that is, $t_i < 0$), or $p_i = 2\Phi(-|t_i|/\sqrt{v_i})$ for two-sided censoring. The values of α and β were set to 1.5 and 4, respectively, to correspond to strong selection bias [8] as shown in Figure 4. The probability of selection is high for studies with low p-values, but decreases rapidly as the p-value increases. After computing w_i for each study, the decision to select a study for inclusion was based on a simulated biased coin toss. The number of selected studies included in each meta-analysis was fixed at 21, and the specified distribution of sample size was maintained by repeating the above procedure for each chosen sample size until the required number of studies were selected. This was to ensure that the final number of studies in the simulated meta-analyses agreed with the typical number observed in the literature and that the required distribution of study size (configuration A or B) was obtained.

4.3. Zero cells

We added 0.5 to all cells for all studies; this reduces small sample bias in the estimated log-odds ratio for a given study [16] and improves its asymptotic standard error estimate, [13]. No adjustment was made to cell frequencies in the calculation of the Mantel–Haenszel odds ratio.

5. RESULTS

5.1. Publication bias absent

Table I presents results for the case where there is no publication bias for both configurations A and B. Here, if the method were unbiased, we would expect the mean intercept to be zero for the Egger method, and the mean slope to be zero for the funnel plot regression.

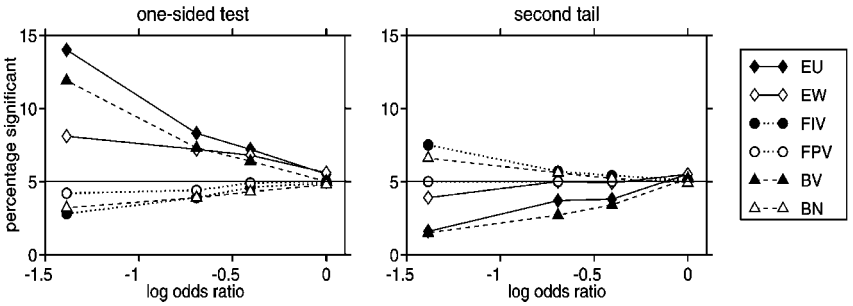

Figure 5. Percentage of simulated meta-analyses under configuration A (21 studies: 11×100/grp, 6×200/grp, 4×300/grp) which incorrectly show significant publication bias. The graph on the left shows results for a one-sided test (5 per cent level) and the right graph shows percentages for the other tail. Combining the tails gives two-sided test results (10 per cent level). A horizontal line is shown at the 5 per cent nominal significance level on each graph.

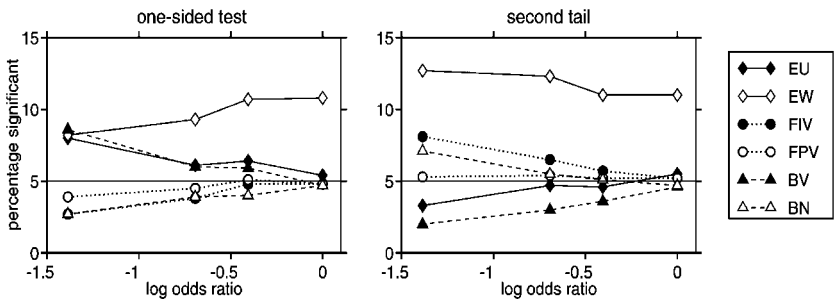

Figure 6. Percentage of simulated meta-analyses under Configuration B (21 studies: 10×100/grp, 5×200/grp, 3×300/grp, 2×500/grp, 1×1000/grp) which incorrectly show significant publication bias. The graph on the left shows results for a one-sided test (5 per cent level) and the right graph shows percentages for the other tail. Combining the tails gives 2-sided test results (10 per cent level). A horizontal line is shown at the 5 per cent nominal significance level on each graph.

In fact, the mean intercepts for the unweighted Egger method are mostly negative. For the EW model the direction of the bias depends on the distribution of study size. In both the EU and EW analyses, the bias increases as the treatment effect increases. Similarly, the bias in the slope for FIV tends to increase as the treatment effect increases. None of the methods shows significant bias when there is no underlying effect of treatment. Only the FPV method shows no significant bias across all treatment effects.

Figures 5 and 6 show the proportion of meta-analyses identified as showing significant publication bias by the various methods. Using a two-sided test of significance at the 10 per cent level, we would expect about 5 per cent of these estimates to be statistically significant by chance in each direction. A one-sided test of significance is appropriate if, for instance, we expect that the effect of treatment is protective ($\theta < 0$) and that publication bias arises from the omission of studies which show little or no treatment effect. Hence, we would use the left tail of the distribution for the BV and Egger (EW and EU) methods and the right tail of the distribution for the funnel plot (FIV and FPV) and BN methods.

In the one-tailed test results (graph of the left in each figure), all of the methods except EW give close to the nominal 5 per cent level for low to moderate effects when the meta-analysis includes a large, relatively precise study (Figure 6, configuration B). When there is less variation in the size of the studies (Figure 5, configuration A), the EU and the BV methods become increasingly liberal when the effect is further from the null, in contrast to FIV and BN methods which become increasingly conservative under both configurations. The FPV method gives the best overall results for a one-sided test in both configurations.

The graph on the right in each figure shows the percentage of meta-analyses that are incorrectly declared to have publication bias in the second tail. Comparing the graphs for the two tails, there is marked asymmetry between them that tends to increase as the treatment effect increases. If we add the area in the two tails, the EU, BV, BN and FIV methods give percentages close to the nominal two-sided significance level (10 per cent) for configuration B, but the EU method becomes increasingly liberal as the treatment effect increases for configuration A. The BV method behaves similarly to the EU regression. The EW, FPV and BN methods show a less marked imbalance in the tails. However, when there is one or more large studies (for example, configuration B), the area in both tails for the EW regression is about double the nominal level. In this situation the large studies will have a high leverage in determining the slope, resulting in increased variability in the intercept.

5.2. Publication bias present

Tables II and III provide comparisons of the various methods under one-sided and two-sided censoring of studies, respectively. It is important to note that the percentage of studies censored decreases with increasing treatment effect, and hence the problem of publication bias decreases. This is evident in Table II for one-sided censoring, where the overall proportion of meta-analyses showing significant publication bias at the two-sided 10 per cent level decreases as the treatment effect increases.

The overall power for detecting publication bias is low even when almost two-thirds of generated studies are censored. The highest power is shown by the EU method. However, we must interpret this in the context of our previous results, which show that this method is too liberal when publication bias is absent. The EW and BV methods give similar results.

For two-sided censoring (Table III), none of the methods can detect the large proportion of studies that are censored when there is no treatment effect. This is because the censoring tends to occur in the centre of the funnel plot, and hence its symmetry is maintained. The EW method is too liberal in the situation where there is a large study, as occurred when no studies were censored.

As the treatment effect increases (that is, is further from 0), the results with two-sided censoring closely resemble those with one-sided censoring. The proportion of studies that are censored then decreases and the percentage of meta-analyses identified as having publication bias decreases. The overall power of the tests is low, with the EU method giving the highest power, but only slightly better than BV.

5.3. Effect of increasing the number of studies per meta-analysis

These additional simulations are restricted to treatment effects of 0 and -0.405 as they correspond to the situations where publication bias is most evident under one-sided censoring. The percentage of censored studies is within 0.1 per cent of those given in Table II.

Table II. Power to detect publication bias of the Egger regressions, funnel plot regressions and rank correlation method using a one-sided test of significance ($P<0.05$) and a two-sided test ($P<0.1$) (sum of one-sided and second tail) at each chosen treatment effect for configurations A and B under the scenario of one-sided censoring of studies.

True log-odds ratio (θ)	Configuration	Mean $\hat{\theta}_{MH}-\theta$	Per cent censored	Egger method Power %				Funnel plot regression Power %				Rank correlation Power %			
				EU		EW		FIV		FPV		BV		BN	
				one-sided	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail
0.0															
	A	−0.20	64.5	**37.0**	0.1	**31.8**	0.3	**25.6**	0.4	**26.3**	0.4	**33.4**	0.2	**25.0**	0.4
	B	−0.16	64.5	**58.2**	0.1	**56.8**	0.3	**40.2**	0.1	**40.8**	0.1	**43.3**	0.0	**39.4**	0.1
−0.405															
	A	−0.066	21.5	**29.8**	0.3	**23.9**	0.8	**16.4**	1.0	**17.5**	0.9	**26.1**	0.4	**16.8**	0.8
	B	−0.044	19.5	**31.8**	0.4	**29.7**	3.2	**18.3**	1.6	**19.5**	1.5	**26.4**	0.2	**19.3**	0.4
−0.693															
	A	−0.025	7.7	**18.6**	1.1	**13.6**	2.4	**8.3**	3.3	**9.2**	2.9	**16.3**	0.8	**8.1**	2.7
	B	−0.016	7.0	**15.0**	1.3	**15.4**	7.7	**7.8**	4.0	**8.6**	3.4	**14.0**	0.7	**8.3**	2.0
−1.386															
	A	−0.0034	0.8	**15.4**	1.3	**9.6**	3.3	**3.4**	7.0	**4.7**	4.7	**13.5**	1.2	**3.5**	6.3
	B	−0.0026	0.7	**9.2**	2.5	**9.2**	12.0	**3.1**	7.9	**4.7**	5.1	**9.7**	1.3	**2.8**	5.8

Configuration A (21 studies: 11×100/grp, 6×200/grp, 4×300/grp).
Configuration B (21 studies: 10×100/grp, 5×200/grp, 3×300/grp, 2×500/grp, 1×1000/grp).

Table III. Power to detect publication bias of the Egger regressions, funnel plot regressions and rank correlation method using a one-sided test of significance ($P < 0.05$) and a two-sided test ($P < 0.1$) (sum of one-sided and second tail) at each chosen treatment effect for configurations A and B under the scenario of two-sided censoring of studies.

True log-odds ratio (θ)	Configuration	Mean $\hat{\theta}_{MH} - \theta$	Per cent censored	Egger method Power %				Funnel plot regression Power %				Rank correlation Power %			
				EU		EW		FIV		FPV		BV		BN	
				one-sided	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail	**one-sided**	2nd tail
0.0															
	A	−0.00061	64.6	**5.2**	5.5	**5.4**	6.0	**4.9**	5.0	**4.9**	5.0	**4.8**	5.0	**4.7**	4.8
	B	−0.000067	64.5	**5.5**	4.9	**11.5**	10.7	**4.3**	4.0	**4.3**	4.0	**5.1**	4.5	**4.8**	4.7
0.405															
	A	−0.099	34.5	**39.5**	0.4	**32.7**	0.5	**23.8**	0.5	**24.8**	0.5	**39.3**	0.2	**26.4**	0.5
	B	−0.065	32.1	**47.8**	0.0	**42.5**	1.5	**28.7**	0.5	**29.7**	0.4	**43.8**	0.1	**34.6**	0.1
−0.693															
	A	−0.042	14.2	**29.9**	0.4	**21.1**	1.0	**13.2**	1.5	**14.6**	1.3	**26.1**	0.2	**13.1**	1.2
	B	−0.028	12.9	**24.0**	0.7	**20.2**	5.8	**11.3**	2.9	**12.2**	2.4	**22.8**	0.4	**13.5**	0.8
−1.386															
	A	−0.0053	1.7	**18.1**	1.1	**10.0**	3.5	**3.7**	6.2	**5.0**	4.4	**15.2**	0.9	**3.8**	5.2
	B	−0.0042	1.5	**10.3**	2.2	**9.8**	12.4	**3.9**	8.0	**5.2**	5.2	**10.1**	1.1	**3.4**	5.5

Configuration A (21 studies: 11×100/grp, 6×200/grp, 4×300/grp).
Configuration B (21 studies: 10×100/grp, 5×200/grp, 3×300/grp, 2×500/grp, 1×1000/grp).

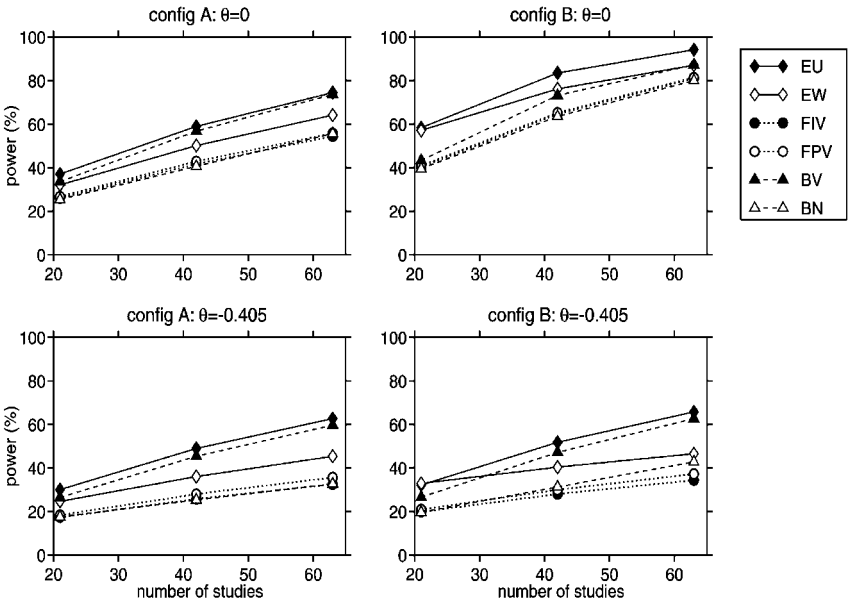

Figure 7. Increase in power of the tests as the number of studies in each simulated meta-analysis is increased under configurations A and B.

The increase in power of the tests as the number of studies is increased to 42 and 63 is shown in Figure 7. The BV and EU methods again show the highest power. When there is no treatment effect and 63 studies, the power of the different methods under configuration B varies between 81 per cent and 94 per cent and from 54 per cent to 75 per cent for configuration A. When the treatment effect is -0.405, the percentage of censored studies decreases and the power is lower (B: 34–66 per cent; A: 32–63 per cent).

5.4. Effect of increasing the sample size per study

With the number of studies fixed at 21, increasing the sample size per study resulted in a moderate increase in power to detect publication bias when there is no underlying treatment effect. For example, doubling the sample size under scenario B increases the power of the EU method from 58 per cent to 70 per cent and that of BV from 43 per cent to 57 per cent. The corresponding increases under scenario A are 37 per cent to 50 per cent and 33 per cent to 47 per cent. Trebling the sample size results in only a very small additional increase in power (<1 per cent) because the power is limited by the number of studies due to the small within-study variation.

For $\theta = -0.405$, the power decreases as the sample size increases because the percentage of studies that are censored decreases from about 20 per cent to 10 per cent when the sample size is doubled and to 6 per cent when it is trebled. For example, under configuration A the power for the EU method decreases from 30 per cent to 19 per cent to 13 per cent as the sample size increases. The corresponding figures for configuration B were 32 per cent, 18 per cent and 11 per cent.

6. DISCUSSION

Our results indicate that the performance of the available methods of detecting publication bias vary with the magnitude of the treatment effect, the distribution of study size and whether a one- or two-tailed test is used.

Before considering the question of power, it is informative to focus on the type I error rates, that is, the percentage of meta-analyses showing significant publication bias when none exists. The FPV method performs well overall with both a one-sided and two-sided test. The EW method is too likely to falsely identify significant publication bias when a two-sided test is used, particularly when there is a wide range of study size. Whilst the overall percentage of meta-analyses showing significant publication bias is consistently closer to the nominal 10 per cent two-sided level for the other methods, the imbalance in the tails increases as the treatment effect increases, particularly for the EU and BV methods. Using a one-tailed test, these two methods become increasingly likely to incorrectly detect publication bias for larger effects, whereas the FIV and BN methods become increasingly conservative.

When the number of studies is set to 21, the methods do not display a high level of power for detecting publication bias when it is present, even under the strong censoring applied in the simulations. The results are consistent with Begg's estimates of power (using BV) for continuous data [8].

The decrease in power with increasing treatment effect is not surprising given that the probability of a study being censored also decreases as the treatment effect becomes large. Larger sample sizes also reduce the power when the true effect is non-zero, because fewer studies are censored. The relationship between the extent of publication bias and the true effect was also noted by Dear and Begg [17]. Increasing the number of studies to 63 results in high power when a large proportion of studies are censored (no treatment effect) and there is a wider range of study size. However, the power is still relatively poor for our smallest, non-zero effect.

When we compare the methods at a given treatment effect, the Egger (EU and EW) and Begg (BV) methods appear to have higher power than the funnel plot (FIV and FPV) or Begg (BN) methods. However, we must interpret this comparison in the light of the differences in the type I error rates noted earlier, which indicate that there is a trade-off in the sensitivity and specificity of the tests.

The methods discussed in this paper test for the presence of publication bias. Alternative, more complex analytic methods for assessing and also adjusting the treatment effect for publication bias have been developed, but do not appear to be widely used [1, 17–20]. A more recent, potentially simpler approach, which assumes that selection of studies is based on the size and direction of the treatment effect, uses an iterative non-parametric method to provide a corrected estimate of treatment based on the symmetry property of the funnel plot [21]. These methods may provide a useful adjunct to the simple tests considered in this paper as they provide a means of assessing how sensitive the results of a meta-analysis are to differing assumptions of the underlying causes of publication bias.

When considering possible approaches for detecting and adjusting for publication bias, it is important to recognize that missing studies (publication bias) are not necessarily the only cause of asymmetry in the funnel plot or source of bias in a meta-analysis. Underlying heterogeneity arising from factors such as temporal changes in treatment effect or differences in the effect of treatment across risk groups can also lead to asymmetry [7, 9].

We have evaluated several methods for detecting publication bias under conditions that reflect situations that often arise in practice, using the log-odds ratio as the measure of effect. None of the methods display consistently good performance in detecting publication bias under one- or two-sided censoring. However, based on our results for the type I error rates, the funnel plot regression weighted by the inverse of the pooled variance (FPV) is the preferred approach. Despite their low power when the number of studies is consistent with that often observed in practice, the tests do provide a more objective means of identifying the presence of publication bias than a highly subjective visual inspection of the funnel plot. The tests are most useful for large meta-analyses and when there is a wide range of study size.

ACKNOWLEDGEMENTS

The authors would like to thank Geoffrey Berry and Norma Terrin for their helpful comments on an earlier draft, Joe Lau for providing us with information on the distribution of study size, Paul Glasziou for insights which helped to motivate this work, and the reviewers for their constructive comments.

REFERENCES

1. Begg CB. Publication bias. In *The Handbook of Research Synthesis*, Cooper H, Hedges LV (eds). Russell Sage Foundation: New York, 1994; Chapter 25.
2. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. *British Medical Journal* 1994; **309**(6964):1286–1291.
3. Begg CB, Berlin JA. Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society*, *Series A* 1988; **151**(3):419–463.
4. Dickersin K. How important is publication bias? A synthesis of available data. *AIDS Education and Prevention* 1997; **9**(1 Suppl):15–21.
5. Dickersin K. The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association* 1990: **263**(10):1385–1389.
6. Light RJ, Pillemer DB. *Summing Up: The Science of Reviewing Research.* Harvard University Press: Cambridge MA, 1984.
7. Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *Journal of the American Medical Association* 1998; **279**(14):1089–1093.
8. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; **50**(4):1088–1101.
9. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; **315**(7109):629–634.
10. Galbraith RF. A note on the graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine* 1988; **7**(8):889–894.
11. Irwig L, Macaskill P, Berry G, Glasziou P. Bias in meta-analysis detected by a simple, graphical test. Graphical test is itself biased (Letter). *British Medical Journal* 1998; **316**(7129):470.
12. Draper NR, Smith H. *Applied Regression Analysis*, 2nd edn. Wiley: New York, 1981.
13. Agresti A. *Categorical Data Analysis*. Wiley: New York, 1990.
14. Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics in Medicine* 1998; **17**(17):1923–1942.
15. SAS Institute Inc. *SAS Language*: *Reference*, *Version* 6, First Edition. SAS Institute Inc: Cary, NC, 1990.
16. Walter SD, Cook RJ. A comparison of several point estimators of the odds ratio in a single 2×2 contingency table. *Biometrics* 1991; **47**(3):795–811.
17. Dear KB, Begg CB. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* 1992; **7**(2):237–245.
18. Greenhouse JB, Iyengar S. Sensitivity analysis and diagnostics. In *The Handbook of Research Synthesis*, Cooper H, Hedges LV (eds). Russell Sage Foundation: New York, 1994; Chapter 24.
19. Hedges LV. Modeling publication selection effects in meta-analysis. *Statistical Science* 1992; **7**(2):246–255.
20. Givens GH, Smith DD, Tweedie RL. Publication bias in meta-analysis: a Bayesian data augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science* 1997; **12**(4):221–250.
21. Duval S, Tweedie RL. Trim and Fill: a simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics* 2000; **56**(2):455–463.

