STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:3443–3457
Published online 12 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2380

A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints

Roger M. Harbord[1,*,†], Matthias Egger[1,2] and Jonathan A. C. Sterne[1]

[1]*MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol, U.K.*
[2]*Department of Social and Preventive Medicine, University of Berne, Switzerland*

SUMMARY

Publication bias and related bias in meta-analysis is often examined by visually checking for asymmetry in funnel plots of treatment effect against its standard error. Formal statistical tests of funnel plot asymmetry have been proposed, but when applied to binary outcome data these can give false-positive rates that are higher than the nominal level in some situations (large treatment effects, or few events per trial, or all trials of similar sizes). We develop a modified linear regression test for funnel plot asymmetry based on the efficient score and its variance, Fisher's information. The performance of this test is compared to the other proposed tests in simulation analyses based on the characteristics of published controlled trials. When there is little or no between-trial heterogeneity, this modified test has a false-positive rate close to the nominal level while maintaining similar power to the original linear regression test ('Egger' test). When the degree of between-trial heterogeneity is large, none of the tests that have been proposed has uniformly good properties. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: meta-analysis; systematic reviews; publication bias; funnel plot; small-study effects; binary data

1. INTRODUCTION

While systematic reviews and meta-analyses have the potential to produce precise estimates of treatment effects that reflect all of the relevant literature, they are not immune to bias. Publication bias—the association of publication probability with the statistical significance of study results—is well documented as a problem in the medical research literature [1–5]. Funnel plots (scatter plots of treatment effects estimated from individual studies against study size or standard error) have been proposed as a visual tool to detect publication bias [6]. Under plausible assumptions about the nature of publication bias, a funnel plot will often be visibly asymmetric, as small imprecise studies that find small or unfavourable treatment

*Correspondence to: R. M. Harbord, Department of Social Medicine, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, U.K.
†E-mail: roger.harbord@bristol.ac.uk

Received 17 January 2005
Accepted 5 July 2005

Copyright © 2005 John Wiley & Sons, Ltd.

effects are missing. However, publication bias is not the only explanation for such funnel plot asymmetry. Smaller studies may also be biased due to poor methodological quality in design, execution or analysis [7, 8]. On the other hand, studies may be smaller because they are targeted at high-risk groups or groups where the treatment is particularly likely to be beneficial, resulting in larger true treatment effects and genuine heterogeneity. Funnel plots are therefore best viewed as a method for checking for 'small-study effects' (a trend for smaller studies to show larger treatment effects) in general rather than publication bias in particular [9, 10].

Visual inspection of a funnel plot has been advocated as a routine step in any meta-analysis [11]. However, there is an inevitable degree of subjectivity in the interpretation of funnel plots, which has led to the development of several statistical tests for funnel plot asymmetry. The most widely used such tests are based on the association between treatment effect and its standard error, using methods based on either rank correlation [12] or linear regression [7]. However, there has been considerable debate regarding the properties of such tests [9, 13–15]: in particular their type I error (false-positive) rate when applied to 2×2 tables of a binary outcome in two groups, the most common form of results in medical applications. An alternative test has been proposed [14] based on a weighted linear regression of treatment effect on total sample size. In simulations [14] this test was found to have a lower false-positive rate than the linear regression test proposed by Egger *et al.* [7], though at the expense of lower power (higher type II error rate). Of the previous studies of the properties of these tests [9, 14, 15], only one [15] included heterogeneity in the true treatment effect in the simulations: this examined the type I error but not the power.

The aims of this paper are to propose a new test for small-study effects in meta-analysis and to compare the properties of this test and two existing tests [7, 14] in scenarios typical of controlled trials, including varying degrees of heterogeneity in effect size. Section 2 introduces notation and the example which we will use to illustrate the different tests. Section 3 describes the regression tests proposed by Egger *et al.* [7] and by Macaskill *et al.* [14], while Section 4 introduces our proposed modification of the Egger test. We present three sets of simulation studies evaluating the properties of the three tests in Section 5. We briefly outline in Section 6 some possible extensions of the modified test to measures of treatment effect other than the odds ratio, which is the focus in the rest of the paper. In Section 7, we discuss the implications of the simulation analyses and make some recommendations based on them.

2. PRELIMINARIES

We shall be concerned with meta-analysis of 2×2 tables, where each study contains a treatment group and a control group and the outcome is binary. We shall use the notation shown in Table I for a single 2×2 table, using the letter d to denote those who experience the event of interest and h for those who do not, with subscripts 0 and 1 to indicate the control and intervention groups. We shall concentrate on the log odds ratio θ as the measure of treatment effect, estimated by $\hat{\theta} = \log(d_1 h_0 / d_0 h_1)$. The usual estimate of the variance of the log odds ratio is the Woolf formula [16] $\mathrm{Var}(\hat{\theta}) = 1/d_0 + 1/h_0 + 1/d_1 + 1/h_1$, the square-root of which gives the estimated standard error $\mathrm{SE}(\hat{\theta})$.

Copyright © 2005 John Wiley & Sons, Ltd.

Table I. Notation for a single 2×2 table.

	Outcome		
	Experienced event D (Disease)	Did not experience event H (Healthy)	Total
Group 1 (intervention)	d_1	h_1	n_1
Group 0 (control)	d_0	h_0	n_0
Total	d	h	n

2.1. Zero cells and continuity corrections

When any of the four cells of the 2×2 table is zero, neither $\hat{\theta}$ nor $\mathrm{Var}(\hat{\theta})$ is defined with the above formulae. A number of continuity corrections, which have differing effects on the small-sample bias of the estimates, have been used to deal with this problem [17]. We shall follow Macaskill *et al.* [14] in always adding $\frac{1}{2}$ to all four cell counts before calculating $\hat{\theta}$ and $\mathrm{SE}(\hat{\theta})$ using the above formulae, in order to aid comparability with their results. We do not believe that the choice of continuity correction is likely to materially alter the properties of tests of small-study effects in realistic situations. The test introduced in Section 4 below does not require such a continuity correction.

2.2. Example: meta-analysis of nicotine gum for smoking cessation

We shall use an example taken from a systematic review of randomized trials of nicotine replacement therapies in smoking cessation [18], restricting to the 51 trials that used chewing gum as the method of delivery. There is a small degree of heterogeneity between the trials (I^2 statistic [19, 20] $= 19$ per cent), although this does not reach conventional levels of statistical significance (heterogeneity statistic $Q = 62.0$ (50 df), $p = 0.119$). The random-effects summary odds ratio for abstinence was 1.68 (95 per cent CI 1.51, 1.87) comparing active with placebo gum, close to the fixed-effect estimate (OR = 1.60, 95 per cent CI 1.50, 1.80). Figure 1 shows a funnel plot of these trials following the conventions and recommendations of Sterne and Egger [21], centred at the fixed-effect summary odds ratio. Visual inspection of the funnel plot suggests some asymmetry.

3. EXISTING REGRESSION TESTS FOR SMALL-STUDY EFFECTS

3.1. Treatment effect and standard error (*Egger*)

Egger *et al.* [7] suggested a linear regression method based on Galbraith's radial plot [22, 23] in which the standard normal deviate $\hat{\theta}/\mathrm{SE}(\hat{\theta})$ is plotted against the precision $1/\mathrm{SE}(\hat{\theta})$. This is the test denoted by 'EU' in Reference [14]. (The alternative 'weighted' version of the test also suggested by Egger *et al.* [7], denoted by 'EW' in Reference [14], is seldom used and lacks a theoretical justification [24].) The slope of a linear regression through the origin is equal to the summary estimate of a fixed-effect meta-analysis. Egger *et al.* proposed that if the regression line is not constrained to pass through the origin, the intercept may be taken as a

Copyright © 2005 John Wiley & Sons, Ltd.

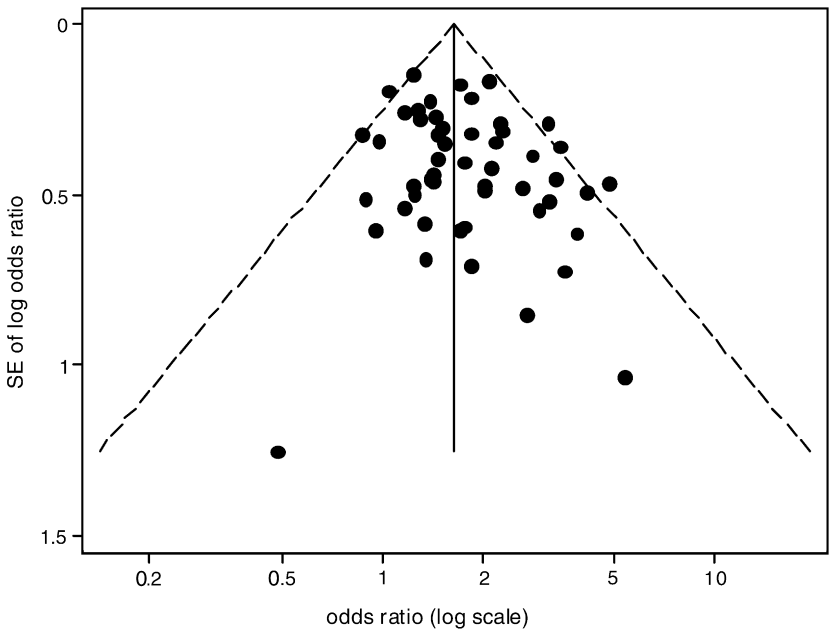

Figure 1. Funnel plot (with pseudo 95 per cent confidence limits) for meta-analysis of nicotine replacement gum *versus* placebo.

measure of the magnitude of small-study effects, with a two-sided t-test of the null hypothesis of zero intercept giving a formal test for small-study effects. This is identical to a test of non-zero slope in a weighted regression of $\hat{\theta}$ on $\mathrm{SE}(\hat{\theta})$ with weights $1/\mathrm{Var}(\hat{\theta})$ [9].

The theoretical basis of any test of small-study effects that is based on the association between the treatment effect and its standard error can be questioned when applied to measures of association for 2×2 tables such as the log odds ratio [13, 14]. The principal objection is that the estimate of the effect size and the estimate of its variance are correlated. An additional issue is that linear regression ignores the sampling variability in the covariate. In practice these factors do not have a large impact on the performance of the linear regression test when treatment effects are modest, the number of trials is reasonable and there is clear variation in study sizes, with at least one study of medium or large size [9]. Such situations may form the large majority of meta-analyses. However, a test that can reliably be used in a wider range of situations is clearly desirable.

Figure 2 shows a Galbraith plot with fitted regression line for the nicotine gum example introduced above. The estimated intercept is 0.705 with standard error 0.357, giving a p-value of 0.054. This may be interpreted as an increase of 0.705 in the log odds ratio for each unit increase in the standard error of the log odds ratio.

3.2. *Effect and sample size* (*Macaskill*)

Macaskill *et al.* [14] proposed a test for non-zero slope in a weighted regression of the treatment effect $\hat{\theta}$ on sample size n. In order to reduce the correlation between the weight

Copyright © 2005 John Wiley & Sons, Ltd.

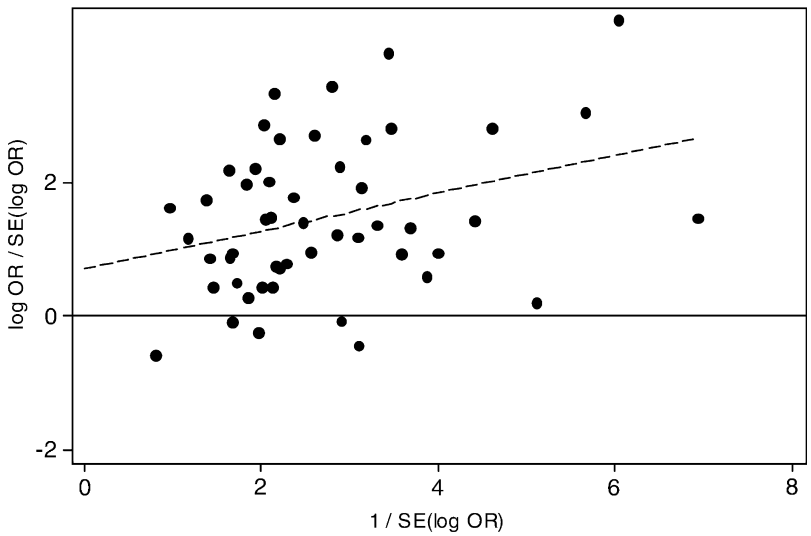

Figure 2. Galbraith plot with fitted linear regression line (dashed line) for meta-analysis of nicotine replacement gum *versus* placebo in smoking cessation.

and the effect estimate, they suggested weighting each study by the reciprocal of the variance of the log-odds of the event rate, with the event rate estimated by pooling both trial groups using the marginal totals, which assumes the null hypothesis is true. (This method is denoted by 'FPV' in their paper; they also assessed inverse variance weighting (FIV) but found it gives a greater type I error rate and recommended FPV.) In the notation of Table I, this weight is dh/n. (Tang and Liu [25] suggested a similar regression of treatment effect $\hat{\theta}$ on sample size n, but weighting simply by sample size.)

The dashed line in Figure 3 illustrates the result of applying this to the nicotine gum data set. The estimated slope is 1.40×10^{-4} with standard error 0.88×10^{-4}, giving a p-value of 0.118 from a two-sided t-test. This may be interpreted as an increase of 0.14 in the log odds ratio when the sample size increases by 1000 subjects.

4. A MODIFIED REGRESSION TEST

We suggest a modification to the Egger test based on the component statistics of the score test, namely the efficient score Z and the score variance (Fisher's information) V. Z is the first derivative and V is minus the second derivative of the log-likelihood with respect to θ evaluated at $\theta = 0$ [26, 27]. (In the presence of nuisance parameters, it is necessary to use profile or conditional likelihood functions—see Reference [27] for details).

We propose using the intercept in a regression of $Z/\sqrt{V}$ against $\sqrt{V}$ as a measure of the magnitude of small-study effects, with a two-sided t-test of the null hypothesis of zero intercept giving a formal test for small-study effects. This is identical to a test of non-zero slope in a regression of Z/V against $1/\sqrt{V}$ with weights V.

Copyright © 2005 John Wiley & Sons, Ltd.

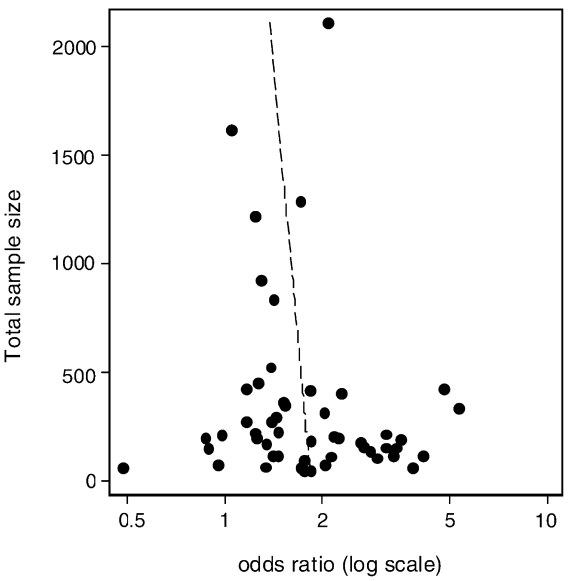

Figure 3. Funnel plot with total sample size as vertical axis, with fitted line for linear regression of log odds ratio on sample size with weights as proposed by Macaskill *et al*. Note that the figure is orientated according to the usual convention for funnel plots, rather than the convention for regression.

Using standard likelihood theory [27] it can also be shown that when θ is small and n is large, $\hat{\theta} \approx Z/V$ and $\mathrm{Var}(\hat{\theta}) \approx 1/V$. It follows that the modified test becomes equivalent to the original Egger test when all trials are large and have small effect sizes. A plot of $Z/\sqrt{V}$ against $\sqrt{V}$ is therefore similar to Galbraith's radial plot of $\hat{\theta}/\mathrm{SE}(\hat{\theta})$ against $1/\mathrm{SE}(\hat{\theta})$, as noticed by Galbraith himself [22].

When the parameter of interest is the log odds ratio θ, the efficient score is [26]

$$Z = d_1 - dn_1/n$$

and the score variance evaluated at $\theta = 0$ is

$$V = \frac{n_0 n_1 dh}{n^2(n-1)}$$

Z is familiar from the Pearson chi-squared test as the difference between the observed and expected events in one cell of the 2×2 table, where the expected number of events is calculated under the null hypothesis of no association. V is the variance of Z under the same null hypothesis. The formula for V given above is obtained by using conditional likelihood, conditioning on the marginal totals S and F in Table I. If profile likelihood is used instead, the score variance has the slightly different form $V' = n_0 n_1 dh/n^3$. The two forms differ only by a factor $(n-1)/n$ and so the difference is unimportant in trials of reasonable size. The familiar Pearson χ^2 statistic for the test of association is equal to Z^2/V' [26] and is therefore very close to Z^2/V in trials of reasonable size. The Z and V statistics are also used in the method of meta-analysis proposed by Yusuf and Peto [28].

Copyright © 2005 John Wiley & Sons, Ltd.

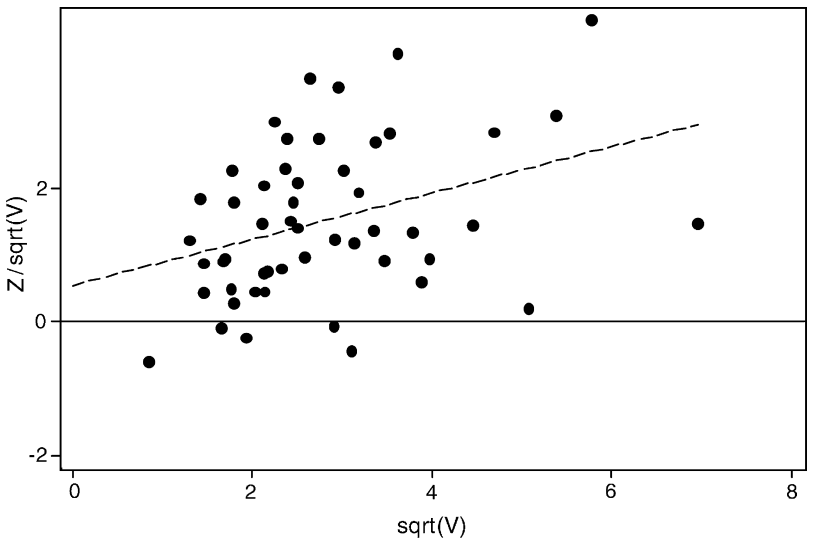

Figure 4. Modified Galbraith plot for meta-analysis of nicotine replacement gum *versus* placebo, with fitted linear regression line (dashed line).

As V depends only on the marginal totals of Table I, whereas $\mathrm{SE}(\hat{\theta})$ depends on the individual cell entries, $1/V$ is less influenced by sampling variation than is $\mathrm{Var}(\hat{\theta})$. Using simulations of a single 2×2 table with fixed group sizes and probabilities of success in both groups, it is easily demonstrated that the correlation between Z/V and $1/V$ due to sampling variation is much lower than that between $\hat{\theta}$ and $\mathrm{Var}(\hat{\theta})$. However, there is still an association between Z/V and $1/V$, and similarly between estimated effect size $\hat{\theta}$ and V, if the probabilities in one or both groups are allowed to vary, i.e. if there is heterogeneity in the true log odds ratio θ.

When the group sizes are roughly equal and event rates are low, $V \approx d/4$; thus the score variance is proportional to the total number of events, in agreement with the intuition that the total number of events (rather than the total sample size) is the main determinant of the amount of information about θ that a trial provides. The weight dh/n used in the test suggested by Macaskill *et al.* (Section 3.2 above) is proportional to V' when the group sizes n_0 and n_1 are equal.

Figure 4 shows a modified Galbraith plot for the nicotine gum example (based on plotting $Z/\sqrt{V}$ against $\sqrt{V}$ rather than $\hat{\theta}/\mathrm{SE}(\hat{\theta})$ against $1/\mathrm{SE}(\hat{\theta})$). The similarity to the Galbraith plot in Figure 2 is clear, although close examination reveals slight differences in the position of some of the smaller trials towards the left of the figure. The estimated intercept is 0.539 with standard error 0.387, giving a p-value of 0.170.

5. SIMULATIONS

For the purposes of these simulations of the type I error and power of the tests we shall use a 10 per cent significance level, in accordance with most previous work [9, 14, 15]. Note

Copyright © 2005 John Wiley & Sons, Ltd.

that this does not imply that 0.1, or indeed any other value, is an appropriate threshold for determining whether there is evidence for small-study effects. The conclusion of our simulation studies would be similar regardless of the choice of significance level.

5.1. Simulations using a full factorial design

In a first set of simulations, we replicated a subset of the scenarios and selection mechanism used by Macaskill *et al.* [14]. This provides a useful cross-check against programming error [29]. We also extended the design by including varying degrees of between-study heterogeneity.

Two configurations of study sizes were used, each with 21 component studies. Configuration A has modest variation in study sizes, containing 11 studies of 100 per group, 6 of 200/group and 4 of 300/group, while configuration B has greater variation in study sizes and one large study, containing 10 studies of 100/group, 5 of 200/group, 3 of 300/group, 2 of 500/group and 1 of 1000/group. The true probability of success in the control group was chosen at random from a uniform distribution between 0.1 and 0.5. The probability of success in the treatment group was calculated from that in the control group using odds ratios of $\frac{1}{4}, \frac{1}{2}, \frac{2}{3}$ and 1.

To assess the properties of the tests in the presence of between-study heterogeneity in treatment effects, we simulated the log odds ratio from a normal distribution with mean θ and variance τ^2. Values of $0, 0.01$ and 0.15 were used for the variance τ^2, the same values used by Terrin *et al.* [30]. To give a clearer idea of the amount of heterogeneity implied, note that 95 per cent of studies have true odds ratios within a factor of $\exp(1.96\tau)$ of the mean odds ratio. This equates to factors of 1.22 and 2.14 when τ^2 is 0.01 and 0.15, respectively.

The number of successes in each group of each trial was then generated using a binomial pseudo-random number generator [31, 32]. Each scenario was repeated with and without study selection. Each meta-analysis was simulated 10 000 times, giving a maximum standard error in any percentage of 0.5 per cent.

The empirical type I error was obtained by simulating without study selection. To simulate the effect of publication bias in order to obtain the empirical power of the tests, the one-sided p-value dependent selection model of Begg and Mazumdar [12] was used. The probability of a study being selected for inclusion in a meta-analysis is given by the weight function $w(p) = \exp(-4p^{3/2})$, corresponding to strong selection bias [12]. We chose to take $p = \Phi(Z/\sqrt{V'})$ (where Φ is the cumulative standard normal distribution), i.e. p is the one-sided p-value for testing $\theta < 0$ corresponding to the Pearson χ^2 test for the study. (We use this in place of the one-sided p-value from the Wald test used by Macaskill *et al.* [14] as it avoids the issue of continuity corrections.) For each of the 21 studies in the meta-analysis, sufficient studies were simulated to ensure that at least 10 000 were selected, the first 10 000 of which were used for subsequent analysis. In this way the distribution of sample sizes specified by the configuration (A or B) was maintained despite the selection. Each simulated meta-analysis was then analysed using the Egger test, the Macaskill test, and the proposed modified test. The statistical software package Stata 8 [33] was used both for generation of the simulations and analysis of the results.

The results are shown in Figure 5, with the percentage of studies censored by selection in the power simulations given in Table II. Let us first consider the left column of Figure 5, which shows the results with zero between-study heterogeneity ($\tau^2 = 0$). Consistent with previous

Copyright © 2005 John Wiley & Sons, Ltd.

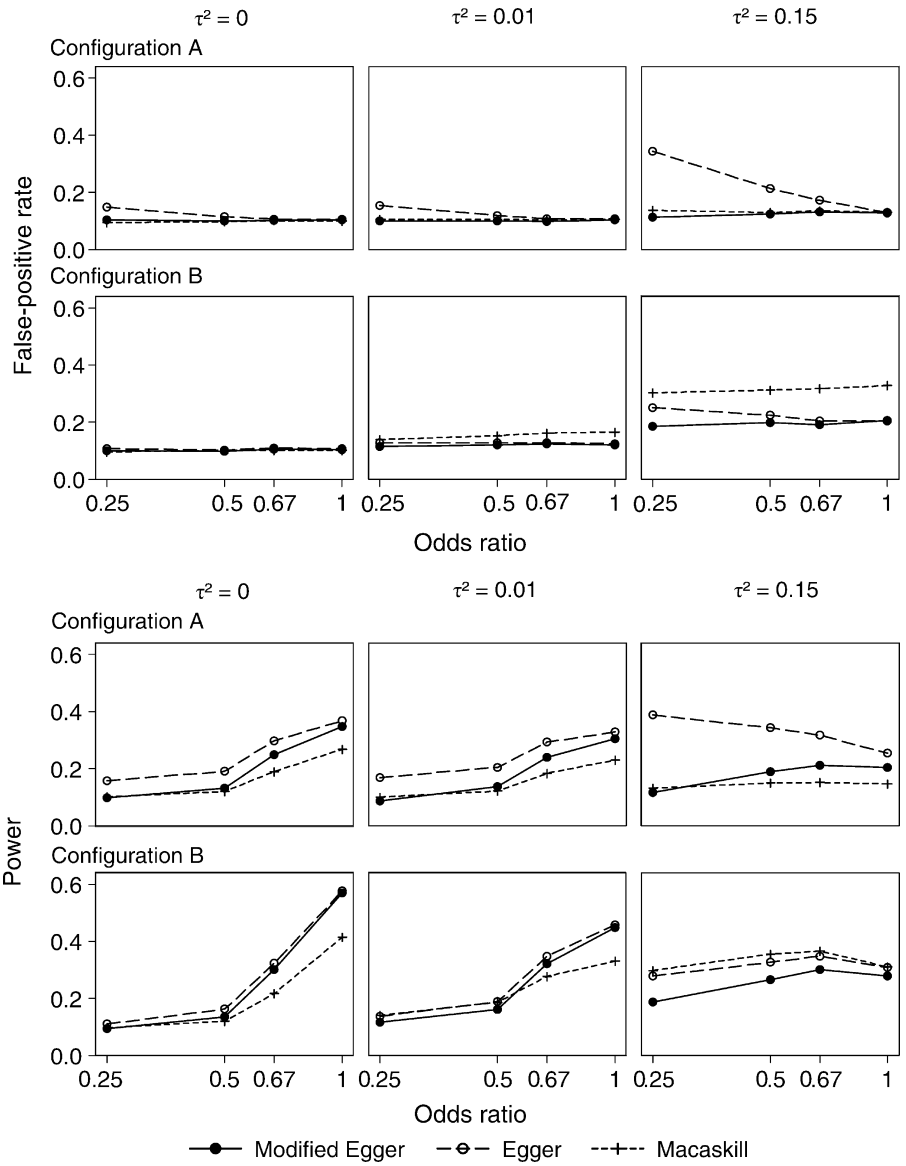

Figure 5. False positive rate and power for simulations. Configuration A: 21 studies—11 × 100/grp, 6×200/grp, 4×300/grp. Configuration B: 21 studies—10×100/grp, 5×200/grp, 3×300/grp, 2×500/grp, 1 × 1000/grp. Maximum standard error of type I error rate or power is 0.5 per cent.

reports [9, 14] the type I error rate of the Egger method is noticeably inflated only when the effect size is large (odds ratio of 0.25) and the studies show little variation in size (configuration A). In this scenario, the Macaskill and modified Egger tests show type I error rates closer to the nominal level of 10 per cent. The power of the tests is limited in all

Copyright © 2005 John Wiley & Sons, Ltd.

Table II. Percentage of studies censored in each set of power simulations.

	$\tau^2 = 0$				$\tau^2 = 0.01$				$\tau^2 = 0.15$			
	Odds ratio				Odds ratio				Odds ratio			
Configuration	0.25	0.50	0.67	1.00	0.25	0.50	0.67	1.00	0.25	0.50	0.67	1.00
A	0.7	7.2	21.1	64.5	0.7	7.7	22.3	63.7	1.6	14.8	30.4	59.3
B	0.6	6.5	19.1	64.3	0.7	7.0	20.2	63.6	1.5	13.9	29.2	58.9

cases, particularly when the effect size is large as few studies are then censored by selection (Table II). As expected, it is greater for scenario B than A as there is more variation in study size. The Macaskill test has the lowest power. When the Egger test has type I error rate close to the nominal 10 per cent level, the power of the modified Egger test is slightly lower than that of the Egger test.

In the presence of between-study heterogeneity, all three tests have noticeably inflated type I error rates, often grossly inflated when the heterogeneity is large ($\tau^2 = 0.15$). The Egger test generally has the most inflated type I error rate in configuration A (similar sized studies) and the Macaskill test in configuration B (greater variation in study sizes). The modified Egger test has the least inflated type I error rates, though it still performs poorly in configuration B when $\tau^2 = 0.15$, with false positive rates around 0.2. When the false positive rate is well controlled, the power is similar to that in the absence of heterogeneity. It is important to interpret the power in the context of the type I error when the latter is inflated, as it is the power in excess of the type I error that is of practical importance.

5.2. *Simulations based on published meta-analyses*

In order to further investigate the type I error rate of the tests in a range of heterogeneity and combination of parameters known to occur in practice, we performed an additional set of simulations based on the same set of 78 published meta-analyses discussed by Sterne *et al.* [9]. These simulations did not include study selection and so did not assess power. From each of the published meta-analyses we derived a summary log odds ratio from random-effects meta-analysis using the DerSimonian and Laird method [34]. We also extracted the treatment and control group sizes from each trial, and estimated the control group event rate in each trial using the continuity correction of Section 2.1 as $(s_C + \frac{1}{2})/(n_C + 1)$. We used values of between-study variance τ^2 of $0, 0.01, 0.04$ and 0.15. These simulations then proceeded in the same way outlined above for the factorial-designed simulations. Again, 10 000 sets of results were simulated for each meta-analysis.

The results are shown in Figure 6 in the form of histograms showing the distribution of the false positive rate over the 78 meta-analyses. When there is no heterogeneity ($\tau^2 = 0$), the tests all have false positive rates close to the nominal level of 0.1, with the Egger test performing slightly more poorly than the other two tests. The performance of all the tests degrades markedly as τ^2 increases, with false positive rates generally higher but also sometimes lower than the nominal level. The modified Egger test retains a narrower distribution than the other two tests, with acceptable performance for $\tau^2 = 0.01$ and arguably tolerable performance

Copyright © 2005 John Wiley & Sons, Ltd.

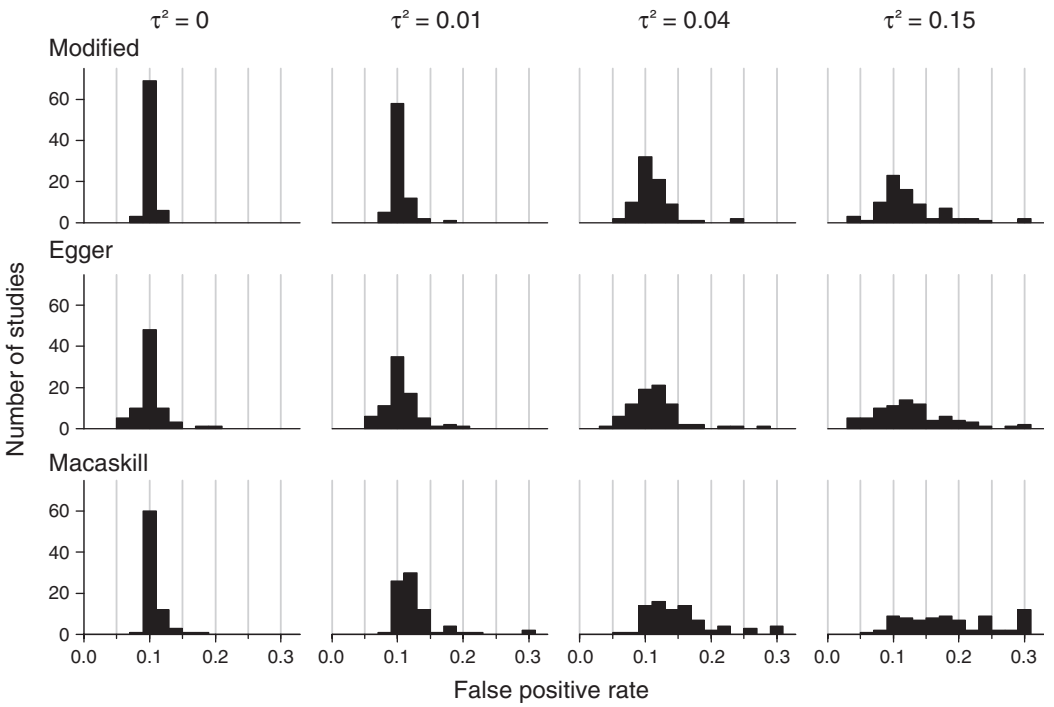

Figure 6. Histograms showing distribution of false positive rate of the proposed modified test for simulations based on the characteristics of 78 published meta-analyses with different values of the between-study variance τ^2. False positive rates have been truncated at 0.3.

for $\tau^2 = 0.04$. None of the tests show acceptable distributions of false positive rates when $\tau^2 = 0.15$.

We also examined how the false positive rate varied with the median value of I^2, the proportion of variance due to heterogeneity [19, 20], for each meta-analysis (results not shown). The pattern of results (not shown) was much less clear than when plotted for varying τ^2 as in Figure 6. This suggests that τ^2 rather than I^2 is the chief determinant of the properties of these tests.

6. EXTENSIONS

We have concentrated here on the (log) odds ratio as the measure of treatment effect. Although this has been advocated as the most appropriate measure when examining funnel plots [21], other measures such as the (log) risk ratio are sometimes preferred [35] as they are more easily interpreted. Whitehead and Whitehead [26] give the formulae for Z and V for binary, ordinal, normally distributed and survival data, enabling the method proposed here to be easily extended to cover these cases, although the properties of such tests would require checking

Copyright © 2005 John Wiley & Sons, Ltd.

by further simulations. However, Whitehead and Whitehead [26] do not consider the (log) risk ratio. Following standard profile likelihood arguments [27] it can be shown that for the log risk ratio $Z = (d_1 n - d n_1)/h$ and $V = n_0 n_1 d / nh$. (This form of V is based on the expected rather than the observed information, the two differing in this case but not for the log-odds ratio as the log odds is the canonical parameter for the binomial distribution.) For normally distributed data there is no correlation between the effect estimate (the difference in means) and its variance so the original Egger test has good properties.

It is also interesting to consider meta-analysis of proportions from studies with no control group, for example the proportion of operative mortality in repair of ruptured aortic aneurysms [36]. In this case, there is no natural null hypothesis so it is necessary to choose an arbitrary fixed proportion p_0 with which to compare the proportion in each study. If we choose the parameter of interest to be the log-odds, then as shown by Whitehead [27] (Section 3.8.2), $Z = d - n p_0$ and $V = n p_0(1 - p_0)$, where d is the number of events out of a total n in each study. Note that V is proportional to n. The implication is that in this case the study size n is a true reflection of the amount of information contributed by each study, since there is no nuisance parameter. This is therefore one case where it is sensible to construct funnel plots using (some function of) n as the vertical axis.

7. DISCUSSION

The proposed modified Egger test has types I and II error rates close to or lower than both the original Egger test and the Macaskill test in all situations investigated. In the absence of between-study heterogeneity, it has a type I error close to the nominal level, with power considerably higher than the Macaskill test and close to that of the original Egger test. The power of all three tests is low in some situations (modest numbers of studies, little variation in study sizes), as is inevitably the case for tests based on the relation between treatment effect and a measure of study size. In the presence of appreciable between-study heterogeneity, none of the tests has uniformly well-controlled type I error, although the modified Egger test has better properties than the other two tests.

The original Egger test, by using a t-test of the intercept in a linear regression on the Galbraith plot, effectively assumes a multiplicative model for heterogeneity in which each within-study variance is multiplied by the same estimated overdispersion parameter to allow for between-study heterogeneity [9]. Thompson and Sharp [37] noted that the statistical rationale for such a multiplicative variance inflation factor is weak and recommended models allowing an additive component of variance. The same arguments apply to the modified Egger test proposed here. Analysis of the simulations in Section 5.2 above using such a model gives a small reduction in the type I error, but this is still inflated in the presence of substantial heterogeneity due to the correlation between effect size and V mentioned in Section 4 (results not shown). In addition, the assumption of normality of the random effects is rarely verifiable in practice as there are too few studies. On balance, we do not believe that the improvement is worth the complication of requiring specialized software. Multiplicative models of overdispersion have a long history as an approximate way to allow for small amounts of heterogeneity [38] and our simulations indicate that the performance of the modified test remains acceptable when the between-study heterogeneity is modest. When the degree of heterogeneity is larger, the focus should be on exploring reasons for the heterogeneity, such as variations

Copyright © 2005 John Wiley & Sons, Ltd.

in study quality, patient population, duration and dose of treatment [39] as it is unlikely that a large amount of heterogeneity could be due to publication bias alone.

One would ideally wish to estimate the influence small study effects have on the meta-analytic summary estimate and produce an adjusted estimate that corrects for such effects. A number of such approaches have been suggested, most of which are based on a model for the process that determines which studies are selected for publication ('selection models') [40–42]. However, such methods are complex and have not been widely used in practice. An alternative model-free (non-parametric) approach is based on adding studies to the funnel plot to make it symmetrical [43, 44]. Simulation studies have found that this 'trim and fill' method adds 'missing' studies to a substantial proportion of meta-analyses even in the absence of bias [45]. More importantly, a major limitation of these methods is that they assume that small-study effects are due to publication bias alone or at least that they can be modelled as if they are. This assumption conflicts with empirical evidence that the low quality of many small trials is more important source of bias than the biased dissemination of trials [46].

7.1. Limitations

The simulations used in this paper have been based on typical meta-analyses of randomized trials, the area to which meta-analysis has been most successfully applied to date. In particular, the simulations in Section 5 had equal numbers in the treatment and control arms. We have not assessed the properties of the tests when there is large imbalance in both margins, a situation common in cohort studies and studies of diagnostic test accuracy. Diagnostic tests also commonly have (diagnostic) odds ratios much larger than those considered here. Both high imbalance and very large treatment effects may be anticipated to adversely affect the properties of the modified test we have proposed, particularly in combination, as they affect the Peto ('one-step') method of meta-analysis based on the same statistics [47]. In principle, it would be possible to improve the properties of the test by choosing to evaluate the score variance at a representative effect size rather than at zero effect size, but this is algebraically complex. Cohort studies usually include more than two exposure groups and adjust for confounders, making the modified test inapplicable. The original Egger test will have satisfactory properties in cohort studies in which the effect sizes are modest and there are many events per study [9].

7.2. Conclusions

In summary, we have proposed a modified version of the Egger test for small-study effects in meta-analysis of controlled trials with binary endpoints. This test has consistently good properties in simulations of balanced trials with little or no heterogeneity. When there is considerable heterogeneity, no test performs well and we recommend exploring possible reasons for the variation in trial results. We do not recommend the modified test be used in meta-analyses of cohort studies where there is large imbalance in the group sizes; however, in this situation the original Egger test will often have good properties. Finally, we stress that statistical tests will not resolve the problems responsible for study effects and bias in meta-analyses of clinical trials. We support the ongoing efforts to prevent publication bias through their registration at inception [48], and the recommendations to improve the quality of the reporting and conduct of trials [49].

Copyright © 2005 John Wiley & Sons, Ltd.

ACKNOWLEDGEMENTS

The Department of Social Medicine of the University of Bristol is the lead centre of the MRC Health Services Research Collaboration. RMH wishes to acknowledge the influence of the teaching of John and Anne Whitehead of the University of Reading.

REFERENCES

1. Simes RJ. Publication bias: the case for an international registry of clinical trials. *Journal of Clinical Oncology* 1986; **4**:1529–1541.
2. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society*, *Series A* 1988; **151**:419–463.
3. Dickersin K. The existence of publication bias and risk factors for its occurrence. *Journal of the American Medical Association* 1990; **263**(10):1385–1389.
4. Easterbrook PJ, Berlin J, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991; **337**: 867–872.
5. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *British Medical Journal* 1997; **315**(7109):640–645.
6. Light RJ, Pillemer DB. *Summing Up. The Science of Reviewing Research*, vol. 1. Harvard University Press: Cambridge, MA, London, England, 1984.
7. Egger M, Davey Smith G, Schneider M, Minder CE. Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* 1997; **315**:629–634.
8. Egger M, Juni P, Bartlett C, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment* 2003; **7**(1). Available at http://www.ncchta.org/minisumm/min701.rtf
9. Sterne JAC, Gavaghan DJ, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology* 2000; **53**:1119–1129.
10. Sterne JAC, Egger M, Davey Smith G. Investigating and dealing with publication and other biases in meta-analysis. *British Medical Journal* 2001; **323**(7304):101–105.
11. Deeks JJ, Higgins JPT, Altman DG. Analysing and presenting results. In *Cochrane Reviewers*' *Handbook 4.2.2*, Alderson P, Green S, Higgins J (eds), [updated March 2004]; http://www.cochrane.org/resources/handbook/handbook.pdf (accessed 17 January 2005).
12. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. *Biometrics* 1994; **50**:1088–1101.
13. 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**:470.
14. Macaskill P, Walter SD, Irwig L. A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine* 2001; **20**(4):641–654.
15. Schwarzer G, Antes G, Schumacher M. Inflation of type I error rate in two statistical tests for the detection of publication bias in meta-analyses with binary outcomes. *Statistics in Medicine* 2002; **21**(17):2465–2477.
16. Woolf B. On estimating the relation between blood group and disease. *Annals of Human Genetics* 1955; **19**(3):251–253.
17. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; **23**(9):1351–1375.
18. Silagy C, Lancaster T, Stead L, Mant D, Fowler G. Nicotine replacement therapy for smoking cessation. *The Cochrane Database of Systematic Reviews*, Issue 3, 2004.
19. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**(11):1539–1558.
20. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *British Medical Journal* 2003; **327**(7414):557–560.
21. Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis. *Journal of Clinical Epidemiology* 2001; **54**:1046–1055.
22. Galbraith R. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine* 1988; **7**:889–894.
23. Galbraith R. Graphical display of estimates having differing standard errors. *Technometrics* 1988; **30**(3):271–281.
24. Steichen TJ, Egger M, Sterne JAC. sbe19.1: tests for publication bias in meta-analysis. *Stata Technical Bulletin* 1998; **44**:3–4.
25. Tang JL, Liu JL. Misleading funnel plot for detection of bias in meta-analysis. *Journal of Clinical Epidemiology* 2000; **53**(5):477–484.
26. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine* 1991; **10**(11):1665–1677.

Copyright © 2005 John Wiley & Sons, Ltd.

27. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Wiley: Chichester, 1992.
28. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases* 1985; **27**(5):335–371.
29. Maldonado G, Greenland S. The importance of critically interpreting simulation studies. *Epidemiology* 1997; **8**(4):453–456.
30. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine* 2003; **22**(13):2113–2126.
31. Hilbe J, Linde-Zwirble W. sg44: random number generators. *Stata Technical Bulletin* 1995; **28**:20–21.
32. Hilbe J, Linde-Zwirble W. sg44.1: corrections to random number generators. *Stata Technical Bulletin* 1998; **41**:23.
33. StataCorp. *Stata Statistical Software*: *Release 8.0*. Stata Corporation: College Station, TX, 2003.
34. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
35. Sackett DL, Deeks JJ, Altman D. Down with odds ratios! *Evidence-Based Medicine* 1996; **1**:164–166.
36. Bown MJ, Sutton AJ, Bell PRF, Sayers RD. A meta-analysis of 50 years of ruptured abdominal aortic aneurysm repair. *British Journal of Surgery* 2002; **89**(6):714–730.
37. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine* 1999; **18**:2693–2708.
38. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman & Hall: London, 1989.
39. Davey Smith G, Egger M, Phillips AN. Meta-analysis: beyond the grand mean? *British Medical Journal* 1997; **315**(7122):1610–1614.
40. Iyengar S, Greenhouse JB. Selection problems and the file drawer problem. *Statistical Science* 1988; (3): 109–135.
41. Dear KBG, Begg CB. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* 1992; **7**(2):237–245.
42. Copas JB, Shi JQ. A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research* 2001; **10**(4):251–265.
43. Duval S, Tweedie R. 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.
44. Duval S, Tweedie R. A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* 2000; **95**(449):89–98.
45. Sterne JAC, Egger M. High false positive rate for trim and fill method (electronic letter). http://bmj.bmjjournals.com/cgi/eletters/320/7249/1574 2000 (accessed 17 January 2005).
46. Egger M, Ebrahim S, Davey Smith G. Where now for meta-analysis? *International Journal of Epidemiology* 2002; **31**(1):1–5.
47. Greenland S, Salvan A. Bias in the one-step method for pooling study results. *Statistics in Medicine* 1990; **9**:247–252.
48. Evans T, Gulmezoglu M, Pang T. Registering clinical trials: an essential role for WHO. *Lancet* 2004; **363**(9419):1413–1414.
49. Moher D, Schulz KF, Altman D, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Journal of the American Medical Association* (JAMA) 2001; **285**(15):1987–1991.

Copyright © 2005 John Wiley & Sons, Ltd.