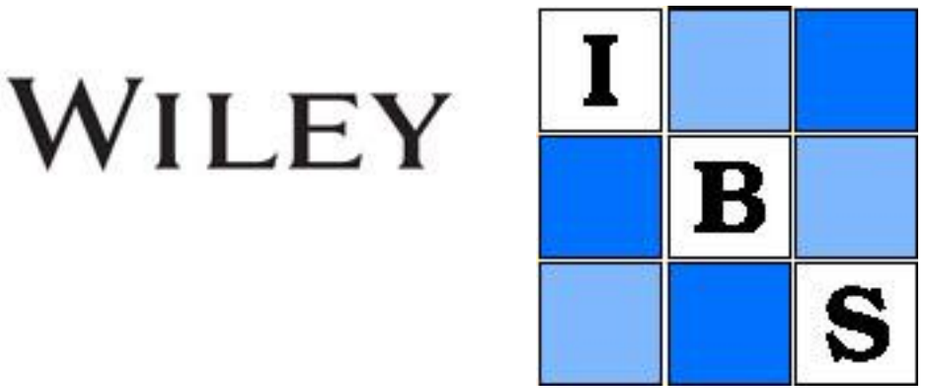


BIOMETRICS 50, 1088–1101
December 1994

Operating Characteristics of a Rank Correlation Test for Publication Bias

Colin B. Begg and Madhuchhanda Mazumdar

Department of Epidemiology and Biostatistics,
Memorial Sloan-Kettering Cancer Center,
1275 York Avenue, New York, New York 10021, U.S.A.

SUMMARY

An adjusted rank correlation test is proposed as a technique for identifying publication bias in a meta-analysis, and its operating characteristics are evaluated via simulations. The test statistic is a direct statistical analogue of the popular "funnel-graph." The number of component studies in the meta-analysis, the nature of the selection mechanism, the range of variances of the effect size estimates, and the true underlying effect size are all observed to be influential in determining the power of the test. The test is fairly powerful for large meta-analyses with 75 component studies, but has only moderate power for meta-analyses with 25 component studies. However, in many of the configurations in which there is low power, there is also relatively little bias in the summary effect size estimate. Nonetheless, the test must be interpreted with caution in small meta-analyses. In particular, bias cannot be ruled out if the test is not significant. The proposed technique has potential utility as an exploratory tool for meta-analysts, as a formal procedure to complement the funnel-graph.

1. Introduction

In recent years meta-analysis has become increasingly popular as a technique for synthesizing the information from a variety of similar research studies on a topic of interest. An important concern is the selection of studies for the meta-analysis. In particular, if the sampling is restricted to published studies, there is a risk that publication bias might adversely affect the reliability of the conclusions of the meta-analysis (Begg and Berlin, 1988). Selective publication may bias the results of the meta-analysis if the probability of publishing a study is influenced by the results of the study, e.g., if more significant findings increase the chances of publication.

Because of this potentially serious issue, it is important to evaluate the data for publication bias in the course of conducting the meta-analysis. A popular technique for assessing the risk of publication bias is the funnel-graph (Light and Pillemer, 1984), a simple graph of the sample sizes of the component studies versus the summary outcome measures, or effect sizes. Under this model all studies in the analysis are presumed to be estimating the same effect. Therefore the estimated effects should be distributed around the unknown true effect, with the spread of the estimates reflecting their variances. That is, the small studies at the bottom should be broadly spread, with the spread narrowing as the sample sizes increase. If publication bias is having an impact this should be reflected in the shape of the graph. For instance, if negative studies are less likely to be published, the graph will tend to be skewed, inducing a negative correlation in the graph, or expressed differently, a positive correlation between estimates of effects and their variances.

The funnel-graph has been used traditionally as an informal technique, in which skewness in the graph is identified visually. However, it is clear that a simple, formal test for publication bias can be constructed by examining the correlation between effect estimates and their variances, to exploit the fact that publication bias will tend to induce a correlation between these two factors. An obvious candidate is to use a rank correlation test, after first standardizing the effect sizes to stabilize the variances (Begg, 1994). This approach is attractive due to its conceptual and computational simplicity, although concerns have been raised about its possible lack of power. The purpose of this article is to examine its operating characteristics using simulations.

Key words: Meta-analysis; Publication bias; Rank correlation.

2. Methods

The selection process leading to publication bias can be conceptualized in the following way. A study is conducted which results in an effect size estimate t with sampling variance v. Based on these results, the probability that the study is published can be influenced by many aspects of the study. This selection probability is referred to as the weight function. In this article we will consider two general forms of selection: a weight function that depends only on t, the observed effect; a weight function that depends only on p, the observed p-value, where p is assumed to be a function of $t/v^{1/2}$. Similar studies with different observed effects and variances are generated until k studies have been selected. At this point one wishes to conduct a meta-analysis, and so a rank correlation test is performed to test for correlation between t and v among the selected studies. In the following we define the test used, the specific weight functions used to simulate publication bias, and characterize features of the meta-analysis which influence our power to detect the bias.

2.1 *The Test Statistic*

Let $\{t_i\}$ and $\{v_i\}$ be the estimated effect sizes and sampling variances from the k studies in the meta-analysis, $i = 1, \ldots, k$. To construct a valid rank correlation test, it is necessary to stabilize the variances by standardizing the effect sizes prior to performing the test. That is, we correlate $\{t_i^*\}$ and $\{v_i\}$, where

$$t_i^* = (t_i - \bar{t})/(v_i^*)^{1/2},$$

where

$$\bar{t} = \left(\sum v_j^{-1} t_j \right) \Big/ \sum v_j^{-1},$$

and where $v_i^* = v_i - (\Sigma\, v_j^{-1})^{-1}$ is the variance of $t_i - \bar{t}$. In many applications the variances will be approximately inversely proportional to the sample sizes in the studies, so that the test will be similar to correlating effect size with sample size. However, in applications where the effect is, say, a difference between two treatments or an odds ratio or a relative risk, and where there is substantial imbalance in the sample sizes of the groups being compared, the two approaches may lead to quite different results.

Throughout we have used the rank correlation test based on Kendall's tau. This involves enumerating the number of pairs of studies that are ranked in the same order with respect to the two factors (i.e., t^* and v). If this number is x, and if y is the number of pairs of studies ranked in the opposite order, then the normalized test statistic is

$$z = (x - y)/[k(k - 1)(2k + 5)/18]^{1/2}.$$

[If there are tied observations, the denominator should be modified. However, the modifications are negligible unless there are substantial groups of tied observations. For further details see Armitage and Berry (1987, pp. 410–417).]

This test procedure is proposed on empirical grounds, based on the fact that the patterns of skewness we expect to see in the funnel-graph should lead to correlation between $\{t_i\}$ and $\{v_i\}$. Formally, we will relate the operating characteristics of the test to the bias that is induced by the procedure for selecting studies for publication, described in the next section. We will use as a measure of bias the difference between the expectation of the weighted average of effect sizes under selective publication and the true effect size.

2.2 *Selection Models*

We assume throughout that the sampling distribution of t is normal, i.e., $t \sim N(\delta, v_i)$. That is, the studies are designed to estimate a common effect size δ, with variances that depend on the sample sizes in the individual studies. The normality assumption is reasonable in view of the fact that t is invariably a summary estimate of a parameter, and as such will possess an asymptotic normal distribution in most circumstances. By making this assumption, we confer generality on our results. That is, the results are applicable regardless of whether the effect under study is a difference in response rates, a hazard ratio, an odds ratio, or any measure that has an asymptotic normal distribution.

After t_i is generated, the study is selected for inclusion in the meta-analysis with probability determined by the appropriate selection model characterized by the weight function. Specifically, for the ith of the k selected studies, the probability density function of the observed effect t_i is $g_i(t_i)$, where

$$g_i(t_i) = \frac{f_i(t_i)w_i(t_i)}{\int_{-\infty}^{\infty} f_i(x)w_i(x)\,dx}.$$

In this formulation $f_i(t_i)$ is the underlying probability density of t_i prior to selection, and $w_i(t_i)$ is the weight function, i.e., the probability that the study is published. This is the standard formulation for a selection model (see Patil and Rao, 1977). Notice that the weight function can be study-dependent, and this is the case for the model in which we assume that the selection probability is a function of the observed p-value. When we employ a selection model in which the probability of publication is a function of the observed effect, t_i, the subscript on the weight function can be dropped. In either case the probability density function, $g_i(\cdot)$, is study-specific unless all the studies in the meta-analysis have the same variance. However, since the funnel-graph approach is dependent on a range of study variances, we will not be considering this special case.

In our simulations we employ different selection models, the first being dependent on the p-value. We have chosen to use the smooth exponential function:

$$w_i(t_i(p_i)) = s(p_i) = \exp\{-bp_i^a\}, \tag{1}$$

where $s(p_i)$ is the weight function evaluated at $p_i = \Phi(-t_i/v_i^{1/2})$ (one-sided test) or $p_i = 2\Phi(-t_i/v^{1/2})$ (two-sided test). We have chosen two examples of this model, strong selection bias (characterized by $a = 1.5$ and $b = 4$), and moderate selection bias (characterized by $a = 3$ and $b = 4$). These curves are displayed in Figure 1. In the moderate-bias setting there is a gradual decrease in the chance of publication as the p-value increases, dropping to 50% at $p = .56$. In the strong-bias setting, the probability of publication declines much more steeply as the p-value increases. The actual bias induced by these selection models will depend on the methodology used for analyzing the meta-analysis, in particular the summary statistics used in the analysis. We will assume that the primary objective is to estimate δ using the standard weighted average of the effect sizes, viz. $\bar{t}$. Therefore the bias, β, induced by the selection model is as follows:

$$\beta = E_g(\bar{t}) - \delta.$$

In effect our test of publication bias is a test of the hypothesis that $\beta = 0$. As we shall see in our simulations (Section 3), there are configurations where the weight function induces selective publication but the induced bias is zero, namely for two-sided selection when $\delta = 0$.

The preceding weight function is based on the assumption that the probability of publication is a function of the observed p-value. An alternative construct is to assume that it is the observed estimate of the effect size that determines the chances of publication. In this case the selection function must be expressed on the effect size scale, i.e., as a function of t. The correspondence

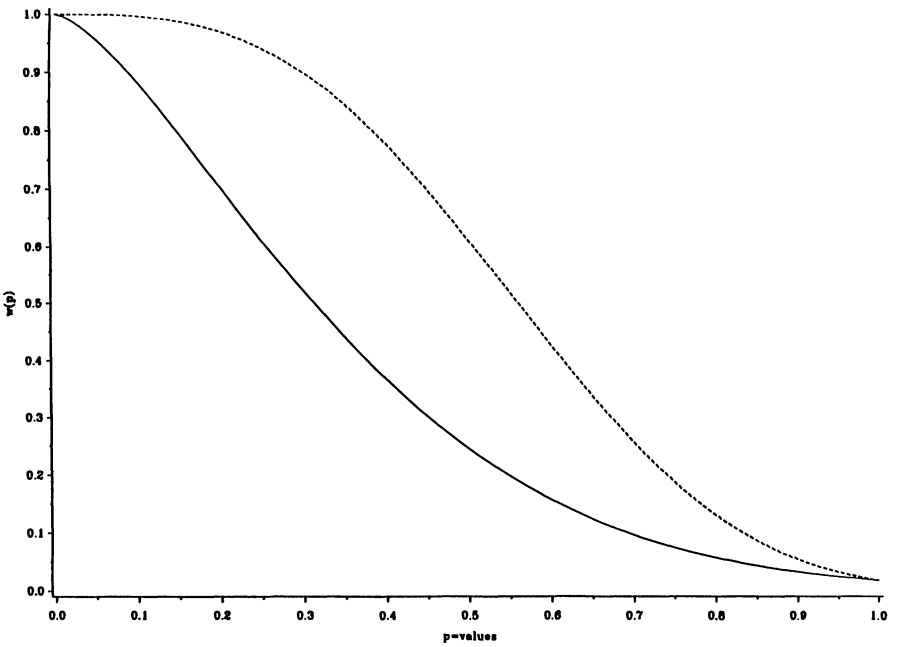


Figure 1. Selection mechanism.

between selection on the p-scale and selection on the t-scale is dependent on the variance of the study, v. We have chosen to standardize the two cases in the following way. The relevant selection function from (1) is transformed to the t-scale using

$$w(t_i) = \exp\{-b\Phi(-t_i)^a\}. \tag{2}$$

This corresponds to one-sided selection for studies in which the variance is standardized, i.e., $v = 1$ and $p = \Phi(-t)$. In our simulations we always use a range of variances in which the average (median) selected variance is the standard variance.

2.3 *Factors Influencing the Power of the Test*

There are several factors that are likely to influence the power of the test. In our simulations in Section 3 we examine the power in relation to each of these factors. In the following we outline the rationale for considering each one. The factors are: the number of component studies in the meta-analysis, k; the underlying effect size parameter, δ; the range of study variances, $\{v_i\}$; the strength of the selection function; the presence of one-sided or two-sided selection pressures.

Clearly, the test makes use primarily of the distribution of the individual effect size estimates, and so the number of component studies is likely to be a principal determinant of power (as opposed to the sample sizes of the individual studies). The underlying treatment effect parameter is especially relevant in the context of selection on the basis of the statistical significance of, say, a test of the hypothesis that this parameter, δ, is zero. As δ deviates from the null value the pressure of selectivity will decrease, since all studies will be increasingly likely to be significant, regardless of their sample sizes. The test is based on exploiting differential selection effects on studies with small sample sizes (large variances) and large sample sizes (small variances), and so it is intuitive that a large variation in sample sizes (variances) will have increased power over a small variation. The strength of the selection function is an important consideration, and we use the two selection functions described in Section 2.2 in our simulations. Finally, we also consider one-sided and two-sided selection separately, in the context of selection on the p-value scale. One-sided selection does not imply that a one-sided test of $\delta = 0$ was used. It merely reflects a circumstance in which only, say, positive effects are preferentially published. More specifically, the weight functions studied are monotonically decreasing functions of the one-sided p-value.

3. Simulations

We conducted simulations using a variety of configurations of the factors outlined in the previous section, and the results are presented in Tables 1–6. We used two values for k, the number of component studies, $k = 25$ and $k = 75$. We believe that $k = 25$ is representative of meta-analyses common in medical research. In fact, we conducted a brief literature sampling to select reasonable simulation parameters. We used a computer literature search to identify the most recent 20 articles published with *meta-analysis* in the title, concerning medical or epidemiological topics, excluding letters, editorials, and meeting abstracts. The number of component studies ranged from 6 to 79, with a mean of 23. A similar search for meta-analyses in psychology, social science, or education produced a mean number of 73 component studies. The parameter reflecting the effect under study was varied from zero (the null value) through 3.0 standard deviations from the null, where the scale is in standard deviation units for the effect estimator for a study in the "middle" group, i.e., with $v = 1$. In each simulation, studies were generated in such a way that after selection for publication there were three (approximately) equal-sized groups of studies with different variances, in each case the middle group having a standardized variance of 1. [In fact, for $k = 25$, the sizes of the groups were (8, 9, 8).] Two ranges of standardized variances were used: large ($v = .1, 1.0, 10.0$) and small ($v = .5, 1.0, 2.0$). These ranges are also justified by our aforementioned sample of published medical/epidemiological meta-analyses. For each meta-analysis with sufficient data presented (17 of 20), the range from the study with the smallest variance to the study with the largest variance was characterized by the logarithm to the base 10 of the ratio of these variances, i.e., in units of orders of magnitude. Our simulated variance ranges correspond to two orders of magnitude (large) and .6 orders of magnitude (small). Our survey produced variance ranges with a mean of 1.26 orders of magnitude (approximately midway between .6 and 2.0), and a range of .12 to 2.87. [*Note:* In the survey of psychometric studies, the information in the article was almost never presented in sufficient detail to estimate variance ranges, so we have assumed that the ranges from the medical/epidemiological survey are generalizable.] Publication bias was simulated using the selection functions described in Section 2.2, and in Figures 1 and 2. That is, we first generated the simulations using a weight function that depended on the p-value in the study, i.e., using (1). Moderate bias ($a = 3, b = 4$) is displayed as the solid curve in Figure 1, and strong bias ($a = 1.5, b = 4$) is displayed

Table 1
*Power for one-sided selection: Small meta-analyses**

Selection strength:	Power [% selected, bias]			
	Strong		Moderate	
Range of variances:	Large	Small	Large	Small
Treatment effect (δ)				
.0	60% [36%, .35]	23% [37%, .75]	35% [57%, .25]	13% [57%, .54]
.5	54% [54%, .16]	23% [52%, .54]	25% [74%, .09]	12% [73%, .35]
1.0	40% [65%, .07]	19% [67%, .37]	15% [82%, .04]	9% [85%, .20]
1.5	29% [72%, .05]	14% [79%, .23]	9% [87%, .03]	7% [92%, .11]
2.0	21% [78%, .03]	10% [88%, .13]	6% [90%, .02]	5% [96%, .05]
2.5	13% [82%, .02]	6% [93%, .07]	5% [92%, .01]	4% [98%, .03]
3.0	9% [86%, .02]	5% [96%, .04]	3% [94%, .01]	4% [99%, .02]

* k = 25 studies; one-sided selection; p-value scale.

Table 2
*Power for one-sided selection: Large meta-analyses**

Selection strength:	Power [% selected, bias]			
	Strong		Moderate	
Range of variances:	Large	Small	Large	Small
Treatment effect (δ)				
.0	99% [36%, .35]	60% [36%, .74]	89% [57%, .25]	36% [56%, .54]
.5	99% [53%, .17]	59% [52%, .54]	77% [73%, .10]	33% [72%, .34]
1.0	95% [64%, .08]	51% [67%, .37]	55% [82%, .04]	21% [84%, .20]
1.5	85% [71%, .05]	37% [79%, .23]	35% [86%, .03]	12% [92%, .10]
2.0	71% [77%, .03]	23% [88%, .13]	22% [90%, .02]	7% [96%, .05]
2.5	53% [81%, .02]	12% [93%, .08]	17% [92%, .01]	5% [98%, .03]
3.0	40% [85%, .02]	8% [96%, .07]	8% [97%, .01]	4% [99%, .01]

* k = 75 studies; one-sided selection; p-value scale.

as the dotted curve. These choices are justified by recent empirical evidence of the magnitude of selective publication. In the null case (i.e., true effect size of zero), the area under the selection curve is the overall probability of publication. This is approximately 36% for our strong selection curve and 56% for the weak selection curve. In a follow-up study of Phase II trials of melanoma registered on the National Cancer Institute registry of cancer trials, only 40% were published (Begg and Berlin, 1989). A more recent study by Easterbrook et al. (1991), in which studies identified from the records of hospital ethics committees were evaluated, demonstrated publication rates of 74% for studies with statistically significant outcomes and 45% for nonsignificant outcomes. A similar study by Dickersin, Min, and Meinert (1992) had corresponding rates of 82% and 66%.

Table 3
*Power: Effect-size scale: Small meta-analyses**

Selection strength:	Power [% selected, bias]			
	Strong		Moderate	
Range of variances:	Large	Small	Large	Small
Treatment effect (δ)				
.0	58% [35%, .25]	26% [36%, .72]	48% [57%, .18]	20% [57%, .53]
.5	47% [52%, .18]	20% [52%, .55]	33% [71%, .10]	14% [73%, .34]
1.0	35% [66%, .11]	13% [68%, .38]	20% [81%, .05]	10% [85%, .20]
1.5	21% [77%, .07]	10% [80%, .25]	11% [86%, .03]	6% [92%, .11]
2.0	13% [84%, .04]	7% [89%, .14]	7% [90%, .02]	5% [96%, .05]
2.5	8% [89%, .02]	5% [95%, .07]	4% [93%, .01]	4% [98%, .03]
3.0	5% [92%, .01]	5% [97%, .04]	3% [94%, .01]	3% [99%, .01]

* k = 25 studies; one-sided selection; effect-size scale.

Table 4
*Power: Effect-size scale: Large meta-analyses**

Selection strength:	Power [% selected, bias]			
	Strong		Moderate	
Range of variances:	Large	Small	Large	Small
Treatment effect (δ)				
.0	99% [35%, .25]	70% [36%, .73]	98% [56%, .18]	55% [57%, .53]
.5	98% [51%, .19]	55% [51%, .55]	93% [71%, .10]	38% [72%, .35]
1.0	93% [66%, .12]	40% [67%, .38]	76% [80%, .05]	23% [84%, .21]
1.5	79% [76%, .07]	24% [80%, .24]	50% [86%, .03]	12% [92%, .11]
2.0	57% [84%, .04]	13% [89%, .14]	29% [89%, .02]	6% [96%, .05]
2.5	35% [88%, .02]	7% [94%, .07]	17% [92%, .01]	4% [98%, .02]
3.0	20% [91%, .02]	5% [97%, .04]	10% [94%, .01]	5% [99%, .01]

* k = 75 studies; one-sided selection; effect-size scale.

The results are presented in Tables 1 and 2 based on a one-sided test, i.e., where the chance of publication is monotonically related to the effect size estimate. The circumstance in which selection is dependent on the effect estimate, t, was generated using (2), with the same parameters. That is, for those studies generated with $v = 1$, the selection pressure applied was identical to that of the p-value scale. These results are displayed in Tables 3 and 4. Finally, we generated simulations using selection on the p-value scale under the assumption that the probability of publication was a function of the two-sided p-value of the test that $\delta = 0$ (Tables 5 and 6).

Each simulated meta-analysis was generated in the following way. First, an effect size was randomly generated from a normal distribution $N(\delta, v)$ where v is the largest of the variances under

Table 5
*Power for two-sided selection: Small meta-analyses**

Selection strength:	Power [% selected, bias]			
	Strong		Moderate	
Range of variances:	Large	Small	Large	Small
Treatment effect (δ)				
.0	3% [12%, .00]	5% [36%, −.01]	2% [57%, .00]	4% [57%, .00]
.5	14% [16%, .30]	9% [42%, .42]	7% [65%, .10]	7% [62%, .24]
1.0	34% [22%, .19]	21% [53%, .46]	11% [73%, .05]	11% [72%, .28]
1.5	48% [27%, .12]	26% [67%, .34]	13% [78%, .03]	12% [82%, .20]
2.0	54% [33%, .09]	19% [79%, .22]	12% [82%, .02]	9% [90%, .11]
2.5	55% [38%, .07]	11% [87%, .14]	10% [86%, .02]	6% [94%, .06]
3.0	55% [43%, .06]	8% [93%, .08]	7% [88%, .01]	5% [97%, .03]

* k = 25 studies; two-sided selection; p-value scale.

Table 6
*Power for two-sided selection: Large meta-analyses**

Selection strength:	Power [% selected, bias]			
	Strong		Moderate	
Range of variances:	Large	Small	Large	Small
Treatment effect (δ)				
.0	2% [36%, .00]	4% [36%, .00]	2% [56%, .00]	4% [56%, .00]
.5	25% [45%, .18]	15% [41%, .41]	12% [65%, .10]	9% [61%, .25]
1.0	56% [54%, .09]	52% [53%, .45]	25% [72%, .05]	25% [71%, .27]
1.5	68% [60%, .06]	68% [66%, .35]	31% [78%, .03]	31% [82%, .20]
2.0	71% [66%, .04]	54% [78%, .22]	30% [81%, .03]	23% [89%, .12]
2.5	72% [71%, .03]	32% [87%, .14]	28% [85%, .02]	11% [94%, .07]
3.0	68% [75%, .03]	15% [92%, .08]	26% [87%, .01]	7% [97%, .04]

* k = 75 studies; two-sided selection; p-value scale.

study. The probability of selection for publication was calculated depending on the configuration under study, i.e., depending on the calculated p-value or the effect size. The decision to include or exclude this study in the meta-analysis was made based on a biased-coin randomization using the calculated probability of publication. The process was repeated until the requisite number of studies with variance v was selected. [*Note:* In fact, at each variance level the variance was altered by an arbitrarily small quantity each time a study was selected, in order to avoid tied variances.] The variance was then changed and the process repeated until k studies were selected with the desired mix of variances. The rank correlation test was then computed, in addition to the summary estimate of δ. Also, the number of studies required to generate k published studies was recorded. The entire

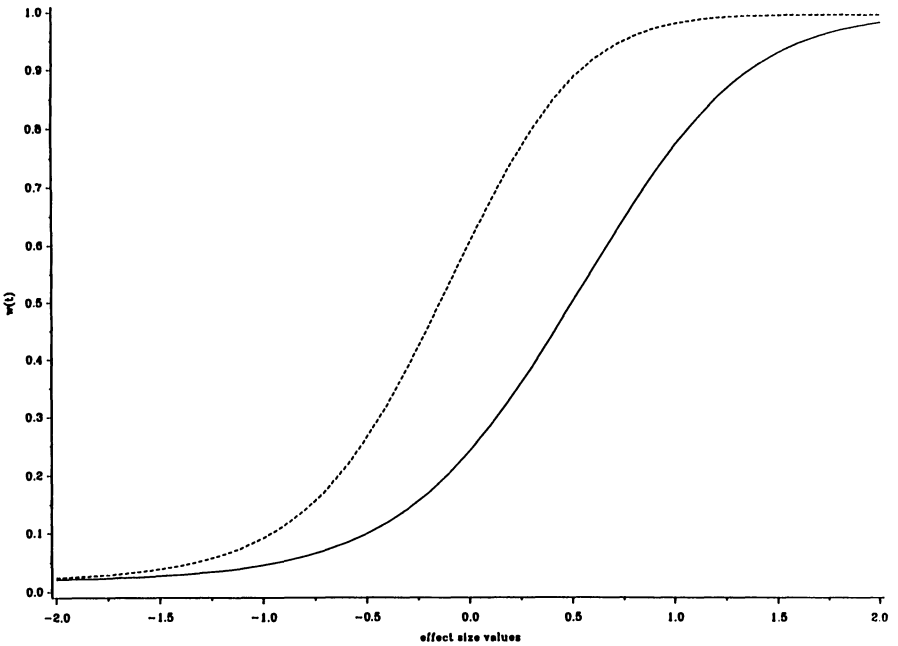


Figure 2. Selection mechanism.

process was then repeated 5,000 times. The tables contain the relative frequency with which the test was significant at the 5% level (two-sided). Consequently the estimates of the power have a (maximum) standard error of ±.7%.

In interpreting the tables, it is important to recall that δ is expressed in standard deviation units relative to the variance of the effect size estimate in the average study (viz., $v = 1$). That is, the operating characteristics depend on the configuration $\delta/v^{1/2}$, in addition to the range of variances and the other relevant characteristics, rather than on the absolute value of the effect size.

There are three statistics presented in the tables for each simulated configuration: the power, the average proportion of studies selected, and the bias induced in the estimate of δ. Consider, for example, the third entry in the first column of Table 1. This corresponds to studies sampled from a distribution with mean effect $\delta = 1.0$, under strong selection bias with a large range of variances ($v = .1, 1.0, 10.0$). The power of the test is 40%. The selection function led to, on average, 65% of the studies being included in the meta-analysis (i.e., published). In other words, in the typical simulation 38 studies were generated before the 25th was selected for inclusion. The bias in $\hat{\delta}$ was .07. That is, the average estimate of δ from the meta-analyses based on the 25 selected studies was 1.07, whereas the true value of δ was 1.00.

The first four tables address configurations in which the selection function is one-sided. That is, the probability of publication increases monotonically with the effect size. [*Note:* The results are equally applicable for one-sided preferential selection of negative effects.] In the first two tables selection was based on the observed p-value. Table 1 presents results for relatively small meta-analyses ($k = 25$), whereas Table 2 addresses larger meta-analyses ($k = 75$). For the most part, the test is only moderately powerful for small meta-analyses, with much better power for larger meta-analyses. Both the strength of the selection function and the range of variances (sample sizes) have a substantial impact on the power. It is reassuring that the best power is achieved when δ is relatively close to the null value. The absence of power for large δ is much less of a problem since the bias in $\hat{\delta}$ induced by selective publication in this case is relatively small. For example, for $\delta = 1.5$, strong selection, and a large range of variances, the power is only 29%. However, the bias induced in $\hat{\delta}$ is only .05, and so the real impact of selective publication is small. This is also reflected in the fact that a larger proportion of the generated studies are selected as δ increases (since the p-value for a test of the hypothesis that $\delta = 0$ becomes increasingly small, with a correspondingly high chance of publication).

The impact of selection based on the observed effects is presented in Tables 3 and 4. The results are generally very similar to those in Tables 1 and 2. There is a little more power for moderate strength of selection, but this is not true across the board. The range of variances and the treatment effect are again seen to be important influences.

Selection based on a two-sided test of the hypothesis that $\delta = 0$ is examined in Tables 5 and 6. The test for publication bias is really not designed for this situation. For example, at $\delta = 0$ the generated studies are symmetrically distributed around zero, so that any selection function that is dependent on the two-sided p-value will result in a symmetrical pattern of selected studies, i.e., a pattern in which the correlation between t and v is zero. Therefore it is not surprising that the power is shown to be the nominal significance level, i.e., less than .05. However, this symmetrical selection pattern induces no bias in the estimate of δ. As δ departs from 0 the selection pressures do induce a correlation between t and v, and this leads to respectable power for strong selection effects when δ ranges from about 1.0 to 2.5. Elsewhere the power is low.

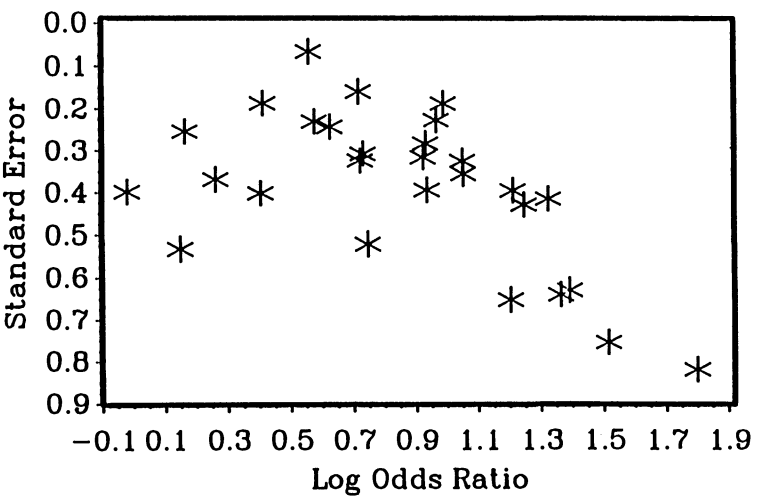
We computed corresponding simulations to those in the tables when there was no selection bias. In all cases the nominal significance level was less than 5%. Finally, we repeated these simulations to examine the validity of the rank correlation test of $\{t_i\}$ versus $\{v_i\}$, that is the unadjusted rank correlation of the elements in the funnel-graph, omitting the variance-stabilizing adjustments. It was observed that the size of this test averaged .10 for $k = 25$ with large variance ranges, and .09 for $k = 25$ with small variance ranges, demonstrating that this procedure is not a valid approach.

4. Examples

We illustrate the method using three examples from the literature. The examples are selected because of their interesting features, and therefore the performance of the test in them cannot be regarded as representative of its likely performance in practice. Two of the three examples were identified in our survey of 20 recently published meta-analyses (Cottingham and Hunter, 1992; Morris et al., 1992). The other example was presented by the author as an example of publication bias (Vandenbroucke, 1988). The raw data for these examples are listed in the Appendix, ranked by the variances of the component studies.

In the first example (Cottingham and Hunter, 1992), the authors have examined the association between Chlamydia trachomatis and oral contraceptive use, based on 29 case–control studies and two prospective studies. We have excluded the prospective studies from our analysis. The authors recognized the potential problem of publication bias, noting that the most extreme odds ratios seemed to occur in the smallest studies. The funnel-graph of these data is presented in Figure 3. Applying the adjusted rank correlation test from Section 2.1, we obtain a z-statistic of 1.76, leading to a two-sided p-value of .08, a statistic strongly suggestive of publication bias. Interestingly, in this example all but one of the 29 odds ratios are greater than 1, and the summary odds ratio is a highly significant 1.93 [confidence interval (1.77, 2.11)], so the likely effect of publication bias is to exaggerate the summary odds ratio, but not to refute the clear evidence of a positive association. Of note is the fact that an unadjusted rank correlation test of the data from the funnel-graph leads to a highly significant $p = .005$. However, as noted at the end of Section 3, such an approach is anticonservative and invalid, due to the fact that the data points are not identically distributed.

In the second example, Vandenbroucke (1988) challenged a previously published article on the risk of lung cancer due to passive smoking (Wald et al., 1986), based on a meta-analysis of 20 studies (Figure 4). Interestingly, the author's visual impression of bias based on the funnel-graph is not



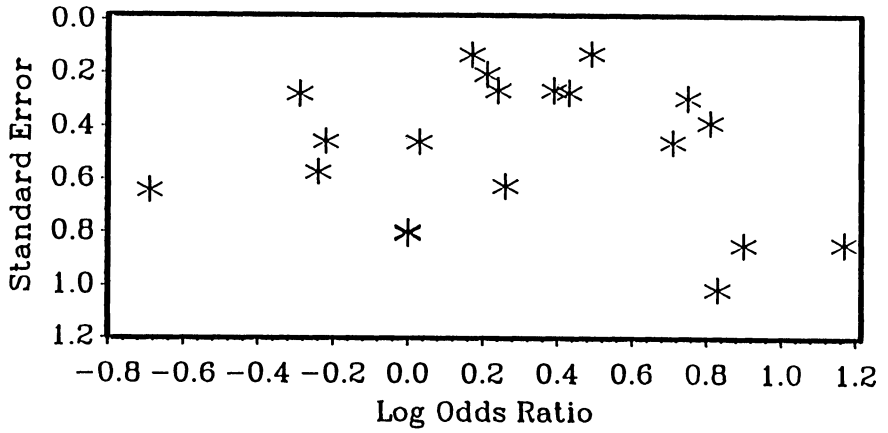


Figure 4. Funnel-graph: Vandenbroucke data.

borne out by our test statistic, which demonstrates very little rank correlation, with $z = .19$ ($p = .85$).

In the third example (Morris et al., 1992) the authors have examined the association between chlorination by-products in drinking water and several types of cancer. In this dataset there are 12 small meta-analyses, each addressing a different cancer site. Clearly there is very little power in each subgroup to detect bias. However, in most of them the adjusted rank correlation is positive (Table 7). It seems appropriate in this setting to use a stratified version of the test as follows. Letting $x_l - y_l$ be the contribution to the numerator of the test statistic for the lth subgroup, and letting $d_l = [k_l(k_l - 1)(2k_l + 5)/18]$ be the corresponding variance, we can construct a stratified test using

$$z = \sum (x_l - y_l) \Big/ \left(\sum d_l \right)^{1/2}.$$

In the example, this procedure leads to $z = 2.02$, or $p = .04$, two-sided. Clearly the test provides strong evidence of publication bias. Moreover, unlike the first example in which the overall disease/exposure association was clearcut, in this example evidence for an association was marginal at the outset, so the evidence for bias casts serious doubt on the credibility of the conclusions.

Table 7
Contributions to stratified test (Example 2)

Cancer site	$x - y$	S.D.$(x - y)$	z	p-value
Bladder	7	6.7	1.05	.29
Brain	1	1.0	1.00	.32
Breast	2	2.9	.68	.50
Colon	−1	6.7	−.15	.88
Colorectal	0	8.1	.00	1.00
Esophagus	4	4.1	.98	.33
Kidney	2	2.9	.68	.50
Liver	2	2.9	.68	.50
Lung	6	4.1	1.47	.14
Pancreas	6	5.3	1.13	.26
Rectum	1	5.3	.19	.85
Stomach	5	5.3	.94	.35
Overall	35	17.3	2.02	.04

5. Discussion

The motivation for this research has been to develop a formal statistical approach for testing for the presence of publication bias, as a preliminary step in performing a meta-analysis, and to evaluate its properties. The method is complementary to the funnel-graph, a popular informal technique for evaluating the likelihood of bias. It is extremely accessible to meta-analysts, being easily evaluated

by hand, and also available in many statistical packages, after the variances are suitably standardized. The method, therefore, fills a gap in the methodological armamentarium.

Our simulations have shown that the power of the test is highly variable, depending on the characteristics of the meta-analysis under study, some of which are known to the analyst, such as the number of component studies and the range of variances, and some of which are unknown, such as the selection mechanism, if any, and the true effect size. Nonetheless we have gained some insights into when the test will be powerful, and when a nonsignificant result will have to be viewed with caution. In general, the test is quite powerful for large meta-analyses, except in circumstances in which the bias induced in the selection mechanism is small. For smaller meta-analyses such as those typical in medical research, the power is moderate even for quite strong selection effects. Therefore it is important not to rule out the possibility of publication bias when the test is not significant. The test is generally not powerful when both the selection mechanism is two-sided and the underlying treatment effect is close to the null value. However, this kind of selection does not induce a large bias in the estimate of the treatment effect.

Alternative approaches to this problem involving direct estimation of the selection probabilities using weighted distribution theory have been proposed. Dear and Begg (1992) have developed a method in which the empirical weight function is estimated using maximum likelihood. Publication bias would have the effect of introducing a trend into this step function, and so a rank correlation test of the estimated weights can be applied. Although the operating characteristics of this test have not been studied, the authors showed using a few examples that the test might be quite powerful for detecting strong selection effects when $k = 25$, even when the variances of the component studies are homogeneous, a configuration in which the test proposed in this paper has no leverage. Hedges (1992) has independently developed an analogous method in which the p-value scale is partitioned prior to obtaining the maximum likelihood estimates of the selection probabilities. The operating characteristics of this method have not yet been studied. Both of these approaches are attempts to improve upon the method proposed by Iyengar and Greenhouse (1988), in which various parametric weight functions were evaluated. Each of these methods has the capacity to detect publication bias when the range of variances is small, but they are much more complex computationally than the method proposed in this article.

Our method is designed solely for the purpose of detecting publication bias. We have not addressed the issue of how to proceed in the event that the evidence for bias is strong or suggestive. This is a difficult issue that is beyond the scope of this article. One could endeavor to provide a summary estimate of effect size that is adjusted for publication bias, either using one of the aforementioned weighted distribution models, or using the kind of response surface modelling approach advocated by Rubin (1990). Alternatively, one could take the view that the evidence for bias invalidates any further analyses. In fact, this decision could be based on the underlying strength of evidence in the data, as in the contrast between Examples 1 and 3 of the previous section.

Finally, we must emphasize that publication bias is not the only potential methodologic flaw in the conduct of a meta-analysis. Other important considerations include the difficulty of expressing the effect size in commensurate units, and other analytic and selection biases that may affect the component studies in different ways. Therefore it is important that the test for publication bias be regarded as merely one of a variety of necessary methodological checks, preparatory to using conventional meta-analytic methods.

Acknowledgements

We are grateful to Terry Crespo for assistance with the manuscript. The research was supported by the National Institutes of Health, Awards LM-05527 and CA-08748.

Résumé

Pour déceler les biais de publication dans les méta-analyses, les auteurs proposent un test basé sur la corrélation des rangs et en étudient les propriétés à l'aide de simulations. Ce test n'est rien d'autre qu'une formalisation statistique du fameux graphe dit "de l'entonnoir" (funnel-graph), qui confronte en abscisse les effets—et un indicateur de leurs variances—observés dans les études, et en ordonnée les tailles des échantillons. Les simulations montrent que le nombre d'études dans la méta-analyse, la nature du mécanisme de sélection à l'origine du biais, la valeur inconnue de l'effet que l'on cherche à estimer, ainsi que les valeurs des variances autour de cet effet sont autant d'éléments propres à influencer la puissance du test. Ce test, en fait, semble vraiment puissant pour de grandes méta-analyses comprenant autour de 75 études, mais s'avère assez médiocre pour des méta-analyses bâties à partir de seulement 25 études. Cependant, il est à noter que lorsque la puissance est très

faible, il s'agit dans la plupart des cas de méta-analyses où le biais dans l'estimation de l'effet est peu important. Quoiqu'il en soit, il faut être très prudent dans l'interprétation du test lorsqu'il s'applique à de petites méta-analyses: en particulier, ce n'est évidemment pas parce que le test n'est pas significatif que l'on peut écarter l'hypothèse d'un biais. En conclusion, cette technique peut être vue comme un outil exploratoire pour les méta-analystes, et comme une procédure formalisant et complétant le graphe dit "de l'entonnoir."

References

Armitage, P. and Berry, G. (1987). *Statistical Methods in Medical Research*, 2nd edition. Oxford: Blackwell Scientific Publications.

Begg, C. B. (1994). Publication bias. In *Handbook of Research Synthesis*, H. Cooper and L. Hedges (eds), 399–409. New York: Sage Publications.

Begg, C. B. and Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A* **151,** 419–463.

Begg, C. B. and Berlin, J. A. (1989). Publication bias and dissemination of clinical research. *Journal of the National Cancer Institute* **81,** 107–115.

Cottingham, J. and Hunter, D. (1992). Chlamydia trachomatis and oral contraceptive use: A quantitative review. *Genitourinary Medicine* **68,** 209–216.

Dear, K. G. B. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* **7,** 237–245.

Dickersin, K., Min, Y. I., and Meinert, C. L. (1992). Factors influencing publication of research results: Follow-up of applications submitted to two Institutional Review Boards. *Journal of the American Medical Association* **267,** 374–378.

Easterbrook, P. J., Berlin, J. A., Gopalan, R., and Matthews, D. R. (1991). Publication bias in clinical research. *Lancet* **337,** 867–872.

Hedges, L. V. (1992). Modelling publication selection effects in meta analysis. *Statistical Science* **7,** 246–255.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science* **3,** 109–117.

Light, R. J. and Pillemer, D. B. (1984). Summing up: The science of reviewing research. Cambridge, Massachusetts: Harvard University Press.

Morris, R. D., Audet, A. M., Angelillo, I. F., Chalmers, T. C., and Mosteller, F. (1992). Chlorination, chlorination by-products, and cancer: A meta-analysis. *American Journal of Public Health* **82,** 955–963.

Patil, G. P. and Rao, C. R. (1977). The weighted distributions: A survey of their applications. In *Applications of Statistics*, P. R. Krishnaiah (ed), 383–405. Amsterdam: North Holland.

Rubin, D. B. (1990). A new perspective. In *The Future of Meta-Analysis*, K. W. Wachter and M. L. Straf (eds), 155–166. New York: The Russell Sage Foundation.

Vandenbroucke, J. P. (1988). Passive smoking and lung cancer: A publication bias? *British Medical Journal* **296,** 391–392.

Wald, N. J., Nanchahal, K., Thompson, S. G., and Cuckle, H. S. (1986). Does breathing other peoples' tobacco smoke cause lung cancer? *British Medical Journal* **293,** 1217–1222.

Received December 1992; *revised April* 1993; *accepted May* 1993.

APPENDIX

Examples from the Literature

Example 1: Cottingham and Hunter (1992)		Example 2: Vandenbroucke (1988)	
Log-odds ratio	Variance	Log-odds ratio	Variance
.56	.004	.49	.017
.72	.026	.17	.018
.99	.035	.21	.043
.41	.035+	.39	.071
.97	.052	.24	.071+
.58	.053	.43	.075
.63	.059	−.29	.078
.17	.065	.75	.086
.93	.080	.81	.149
.41	.083	−.22	.209
.73	.095	.71	.211
.92	.099	.03	.211+
.72	.104	−.24	.329
1.05	.106	.26	.393
1.05	.126	−.69	.413
.26	.135	.00	.640
.94	.154	.00	.652
1.21	.156	1.17	.715
−.02	.158	.90	.721
.41	.162	.83	1.03
1.32	.172		
1.25	.185		
.75	.270		
.15	.284		
1.39	.397		
1.36	.410		
1.20	.426		
1.52	.570		
1.80	.672		

Example 3:
Morris et al (1992)

Cancer site	Log-odds ratio	Variance	Cancer site	Log-odds ratio	Variance
Bladder	.17	.0028	Kidney	.00	.0027
Bladder	.17	.011	Kidney	.40	.021
Bladder	−.02	.015	Kidney	−.01	.032
Bladder	.34	.020	Kidney	1.01	.501
Bladder	.52	.052			
Bladder	.14	.062	Liver	.18	.016
Bladder	.79	.320	Liver	.05	.027
			Liver	.00	.074
Brain	−.16	.015	Liver	1.09	.345
Brain	.76	.101			
			Lung	−.06	.00025
Breast	−.12	.00051	Lung	.11	.0083
Breast	.17	.0073	Lung	−.22	.018
Breast	.22	.014	Lung	−.04	.051
Breast	.82	.113	Lung	.58	.064
Colon	−.12	.00028	Pancreas	−.06	.00071
Colon	.10	.0034	Pancreas	.02	.0085
Colon	.01	.0080	Pancreas	.07	.0095
Colon	.42	.0096	Pancreas	.14	.031
Colon	.48	.028	Pancreas	.68	.056
Colon	−.12	.050	Pancreas	−.22	.089
Colon	−2.30	1.19			
			Rectum	−.04	.0015
Colorectal	−.08	.0015	Rectum	.67	.015
Colorectal	.12	.0028	Rectum	.20	.015+
Colorectal	.37	.0076	Rectum	.19	.038
Colorectal	.54	.018	Rectum	.66	.053
Colorectal	−.03	.018+	Rectum	.35	.125
Colorectal	−.11	.041			
Colorectal	.03	.046	Stomach[a]	−.03	.0011
Colorectal	.34	.106	Stomach	.18	.0098
			Stomach	−.03	.014
Esophagus	.08	.026	Stomach	.44	.037
Esophagus	−.03	.031	Stomach	.60	.045
Esophagus	.75	.107	Stomach	−.49	.144
Esophagus	−.05	.282			
Esophagus	.56	.360			

[a]The data for stomach cancer, not presented in the original article (Morris et al., 1992), were kindly supplied by the authors.