BIOMETRICS 56, 455–463
June 2000

Trim and Fill: A Simple Funnel-Plot–Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis

Sue Duval

Department of Preventive Medicine and Biometrics,
University of Colorado Health Sciences Center, Denver, Colorado 80262, U.S.A.

and

Richard Tweedie

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
email: tweedie@biostat.umn.edu

SUMMARY. We study recently developed nonparametric methods for estimating the number of missing studies that might exist in a meta-analysis and the effect that these studies might have had on its outcome. These are simple rank-based data augmentation techniques, which formalize the use of funnel plots. We show that they provide effective and relatively powerful tests for evaluating the existence of such publication bias. After adjusting for missing studies, we find that the point estimate of the overall effect size is approximately correct and coverage of the effect size confidence intervals is substantially improved, in many cases recovering the nominal confidence levels entirely. We illustrate the trim and fill method on existing meta-analyses of studies in clinical trials and psychometrics.

KEY WORDS: Data augmentation; File drawer problem; Funnel plots; IQ; Malaria; Meta-analysis; Missing studies; Publication bias.

1. Introduction

Since its introduction in the social sciences by Glass (1976), there has been enormous increase in the use of meta-analysis as a statistical technique for combining the results of many individual analyses. While the combined analysis may have increased inferential power over any individual study, there are several drawbacks to meta-analysis (cf., Thompson and Pocock, 1991; NRC Committee on Applied and Theoretical Statistics, 1992; Mengersen, Tweedie, and Biggerstaff, 1995), and one such concern is the need to collect all studies, both published and unpublished, relevant to the meta-analysis if the subsequent inferences are to be valid (Rosenthal, 1979; Begg and Berlin, 1988; Iyengar and Greenhouse, 1988; Dear and Begg, 1992; Hedges, 1992; Begg, 1994; Begg and Mazumdar, 1994; Gleser and Olkin, 1996; Egger et al., 1997). This is because the use of a nonrepresentative proportion of significant studies or studies differentially giving results in, say, a positive direction will lead to a nonrepresentative set of studies in the meta-analysis data set. A standard meta-analysis model will then result in a conclusion biased toward significance or positivity.

This is particularly problematic for a meta-analysis whose data come solely from the published scientific literature. It is a common belief, backed by several empirical assessments, that studies are not uniformly likely to be published in scientific journals (Cooper, 1998, pp. 54–55; Dickersin, Min, and Meinert, 1992). Easterbrook et al. (1991) suggest that statistical significance is a major determining factor of publication since some researchers (e.g., students with masters' or Ph.D. theses) may not submit a nonsignificant result for publication, and editors may not publish nonsignificant results even if they are submitted (British Medical Journal Editorial Staff, 1983). Studies may also be suppressed for many other reasons than failure to be published (see Givens, Smith, and Tweedie [1997] and the discussion thereof, Cooper [1998], and Misakian and Bero [1998]). This phenomenon has become known as publication bias, or the file-drawer problem (Rosenthal, 1979; Iyengar and Greenhouse, 1988).

Evaluating the effect of publication bias is difficult since the missing studies influence the overall mean that is estimated in the meta-analysis. Perhaps the most common method that has been proposed to detect the existence of publication bias in a meta-analysis is the funnel plot (Light and Pillemer, 1984). The solid circles in Figure 1 depict a typical funnel plot, using data from 19 studies of IQ scores and teacher expectancy (see Section 6 for more details). Each of the studies supplies an estimate Y_i of the effect in question in the ith study and an estimate of the variance σ_i^2 within that study.

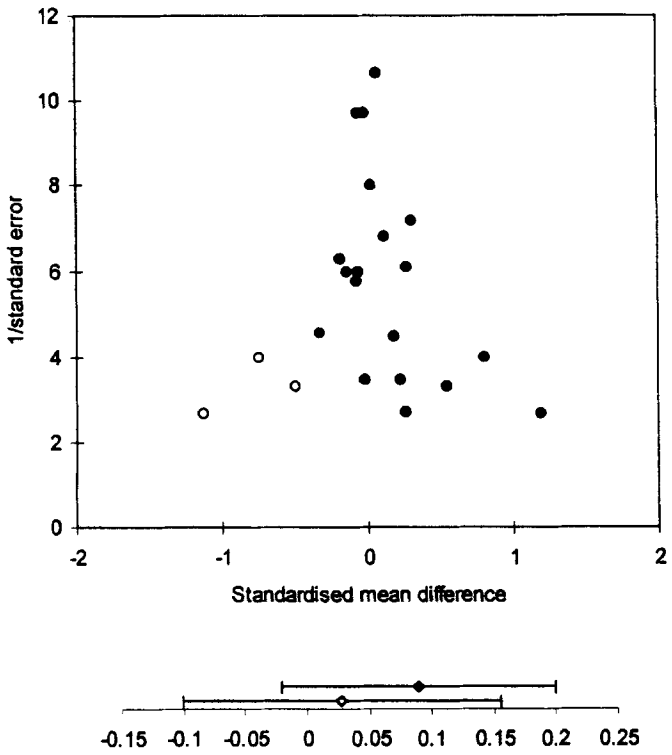

Figure 1. Top panel is a funnel plot of standardized mean differences of teacher expectancy of IQ from Raudenbush (1984). Solid circles are original data, open circles are imputed filled values. Bottom panel shows overall mean and 95% CI of standardized mean differences before and after allowing for publication bias.

Thus, the most precise estimates (typically, those from the largest studies) are at the top of the funnel and those from less precise or smaller studies are at the base of the funnel. (The funnel plots we use graph σ_i^{-1} as the measure of size against Y_i, although other authors use other indicators of study size.)

Funnel plots give a subjective evaluation of bias since, in the models used in meta-analysis, it is assumed that, at any fixed level of σ_i^2, studies are symmetrically distributed around the true mean Δ. The fact that there may be publication bias in the data set in Figure 1 might be subjectively inferred since the funnel shape is asymmetrical: There is an appearance of missing studies in the bottom left-hand corner and the assumption is that, whether because of editorial policy or author inaction or other reasons, these missing papers were not published because they showed, say, no significance or perhaps the reverse effect from the positive effect expected.

There is nothing special about studies being omitted on the left side rather than the right, but in what follows, we will be considering missing left studies for convenience of notation. If studies are missing on the right, then merely by reversing the sign of the effect size, we reach the same conclusions as those discussed here, as illustrated in the second example in Section 6.

The drawback to using the funnel plot is that it is purely subjective. There are a number of more quantitative methods that attempt to detect publication bias in the literature. The rank correlation test of Begg (1994) and a regression-based test (Egger et al., 1997) are most often used. However, these typically have low power to detect bias when using a standard hypothesis testing framework, and indeed Begg (1994) suggests using a very liberal significance level and notes that any evidence of publication bias should make us cautious about proceeding with the analysis. There also exist several quantitative methods that estimate the number of missing studies and provide estimates of the effect of the missing studies (Dear and Begg, 1992; Hedges, 1992; Givens et al., 1997). However, all are complex and highly computer intensive to run.

As DuMouchel and Harris note in the comments on Givens et al. (1997, p. 244), "attempts to assess publication bias beyond simple graphs like the funnel plot seem to involve a *tour de force* of modeling, and as such are bound to run up against resistance from those who are not statistical modeling wonks."

In Duval and Tweedie (2000), we described a new approach that requires very simple computation (the trim and fill algorithm) for accounting for the magnitude of the publication bias problem. Based only on symmetry assumptions, we develop a number of estimators (denoted R_0, L_0, and Q_0 below) of the number of missing studies. These are simple to implement, in contrast to, say, Dear and Begg (1992), Hedges (1992), or Givens et al. (1997); they can be used in an iterative manner when the effect size is unknown, in contrast to methods of Gleser and Olkin (1996); and they appear in practice to pick up the missing studies indicated visually by funnel plots.

Once we have these estimates, in Duval and Tweedie (2000), we showed how to impute the missing values using an iterative but still computationally simple algorithm that gives estimates of the effect on the inferences in the meta-analysis due to the publication bias.

In this paper, we investigate three properties of this approach. (a) We consider which of the estimators has the best mean square error (MSE) properties. We show in Section 3 that the estimators R_0 and L_0 are both better than Q_0 but that there are values of the number of observed and missing studies for which each is better than the other. (b) We use the distributional properties of the estimators to formulate tests for the existence of publication bias and compare these with recent methods of Begg (1994) and Egger et al. (1997) (the resulting tests, in Section 5, appear to be quite powerful if there are more than 5–6 missing studies). And (c) we compare the properties of the iterated version of the algorithm (used in practice) with the analytic results in Duval and Tweedie (1998). Using simulations, we see in Section 4 that the iteration does not adversely affect accuracy in general and so the analytic descriptions of the tests and estimators can be used in the iteration.

In Section 6, we apply the iterative algorithm and the test methods to two existing meta-analyses: a psychometric meta-analysis of IQ scores and teacher expectancy, collected by Raudenbush (1984) and analyzed by Begg (1994) and Gleser and Olkin (1996), and a collection of clinical trials on antimalarial drugs (McIntosh and Olliaro, 1998).

2. Trim and Fill: A Simple Estimation Approach

The trim and fill algorithm is based on a formalization of the qualitative approach using the funnel plot. Simply put, we trim off the asymmetric outlying part of the funnel after

estimating how many studies are in the asymmetric part. We then use the symmetric remainder to estimate the true center of the funnel and then replace the trimmed studies and their missing counterparts around the center. The final estimate of the true mean, and also its variance, are then based on the filled funnel plot.

This is illustrated in Figure 1, where we have estimated, using this algorithm, that the number of missing studies is 2–6; have replaced three symmetrically as indicated by the open circles using the estimator L_0^+ defined in Section 2; and have recovered a visually symmetric funnel plot. Rather more importantly, after filling the funnel plot using L_0^+, we obtain an overall estimate of $\hat{\Delta}$, which is considerably reduced from that estimated from the original data (see Section 6), as shown in the bottom panel of Figure 1.

The key to this method lies in estimating the number of missing studies. We now describe the nonparametric approach to this that we shall use.

In the standard structure for a meta-analysis (in the absence of publication bias), we assume we have n individual studies, all of which are addressing the same problem, and that there is some global effect size Δ that is relevant to the overall problem and that each study attempts to measure. For each $j = 1, \ldots, n$, study j produces an effect size Y_j, which estimates Δ, and an estimated within-study variance σ_j^2.

Our work could equally apply to effect size measures such as log relative risks, risk differences, or log mortality ratios in clinical or epidemiological trials or to differences in performance in sociological experiments.

In Duval and Tweedie (2000), we modified this standard model to account for publication bias. We assume that, in addition to the n observed studies, there are an additional k_0 relevant studies that are not observed due to publication bias. The value of k_0 and the effect sizes that might have been found from these k_0 studies are unknown and must be estimated, and uncertainty about these estimates must be reflected in the final meta-analysis inference.

The key assumption behind the nonparametric method in Duval and Tweedie (2000) is that the suppression has taken place in such a way that it is the k_0 values of the Y_j with the most extreme left-most values that have been suppressed. This might be expected to lead to a truncated funnel plot such as the solid circles in Figure 1, for example. We will call this model for the overall set of studies the suppressed Bernoulli model.

This assumption differs slightly from that used in some other papers (Dear and Begg, 1992; Hedges, 1992; Gleser and Olkin, 1996), where the assumption is that the suppressed studies are exactly those with the largest p-values against the null hypothesis of no effect. A number of authors (cf., DuMouchel and Harris, 1997) have pointed out that this simple p-value suppression scenario is rather simplistic since it fails to acknowledge the role of other criteria, such as size of study, in the decision about whether a study is published. Misakian and Bero (1998) found that p-values may be the critical determinant in delaying publication but that study size does appear to have some effect also. In practice, the same papers will be suppressed under our assumption as under a strict p-value suppression rule except when the size of the study is large, and in that case, one might well expect the study to remain unsuppressed (as we assume) even if it is nonsignificant.

The trim and fill algorithm uses the ranks of the absolute values of the observed effect sizes and the signs of those effect sizes around Δ. We first describe the method assuming that Δ is known. An iterative method when Δ is not known will be described below and will be used in practice.

We write X_i for the observed value of $Y_i - \Delta$ and denote the ranks of the observed values of the $|X_i|$ as r_i^*: these ranks run from one to n. We let $\gamma^* \geq 0$ denote the length of the rightmost run of ranks associated with positive values of the observed X_i, and we also denote the Wilcoxon rank test statistic for the observed n values as $T_n = \Sigma_{X_i>0} r_i^*$. Note that because there are k_0 suppressed values, T_n does not have the usual distribution of a Wilcoxon statistic.

Based on these quantities, Duval and Tweedie (2000) defined three estimators of k_0, given by

$$R_0 = \gamma^* - 1, \tag{1}$$

$$L_0 = \{4T_n - n(n+1)\}/(2n-1), \tag{2}$$

and

$$Q_0 = n - 1/2 - \sqrt{2n^2 - 4T_n + 1/4}. \tag{3}$$

The properties of the estimators (1), (2), and (3) are developed in Theorem 1 of Duval and Tweedie (2000). Under the suppressed Bernoulli model, when the median of the original X_i is zero, the estimator R_0 has mean and variance given by

$$\mathrm{E}[R_0] = k_0, \qquad \mathrm{var}[R_0] = 2k_0 + 2 \tag{4}$$

and the estimator L_0 has mean and variance given by

$$\begin{aligned} \mathrm{E}[L_0] &= k_0 - k_0^2/(2n-1), \\ \mathrm{var}[L_0] &= 16\,\mathrm{var}[T_n]/(2n-1)^2, \end{aligned} \tag{5}$$

where $\mathrm{var}[T_n] = 24^{-1}\{n(n+1)(2n+1) + 10k_0^3 + 27k_0^2 + 17k_0 - 18nk_0^2 - 18nk_0 + 6n^2k_0\}$. The estimator Q_0 has mean and variance given (approximately) by

$$\begin{aligned} \mathrm{E}[Q_0] &\approx k_0 + 2\,\mathrm{var}[T_n]/\{(n-1/2)^2 - k_0(2n-k_0-1)\}^{3/2} \\ \mathrm{var}[Q_0] &\approx 4\,\mathrm{var}[T_n]/\{(n-1/2)^2 - k_0(2n-k_0-1)\}. \end{aligned} \tag{6}$$

Remark. In using any of these estimators, we round to the nearest nonnegative integer, and we denote these rounded estimators by R_0^+, L_0^+, Q_0^+, respectively. The means and variances of these estimators are shown using simulations in Duval and Tweedie (2000) to be acceptably close to the analytic forms given by (4), (5), and (6) for the ranges of n, k_0 that we address, and thus the use of the analytic values seems justified in practice.

3. Choosing an Optimal Estimator of k_0

We first compare the behavior of the means and MSEs of the empirical estimators R_0^+, L_0^+, and Q_0^+ using simulated data and find that L_0^+ and R_0^+ seem to perform better than Q_0^+.

To carry out the simulations we generated, for each of the cases below, 5000 sets of $N = n + k_0$ normal variates, each with mean zero and variance σ_i^2, the σ_i being taken from a $\Gamma(3, 1/9)$ distribution. This appears to give more typical funnel-plot shapes than does the choice of a uniform distribution for σ_i as in, e.g., Light and Pillemer (1984). We allowed N to range over the values 25–75 in increments of five and examined $k_0 = 0, 5, 10$.

When $k_0 = 0$, the effect on the means of truncating to ensure nonnegative forms is simple: R_0^+ produces a mean around 0.5 and both L_0^+ and Q_0^+ produce means of 1–2 over the range of N examined. Thus, one should be cautious about deducing that there is publication bias when these estimators produce such small positive values. More formal tests are developed in Section 5 below.

For all three estimators, R_0^+, L_0^+, and Q_0^+, the simulated mean behavior is quite accurate over the entire range of n when $k_0 = 5$, while at $k_0 = 10$, we find that R_0^+ is still accurate, Q_0^+ tends to overestimate a little for the smaller values of n, and L_0^+ tends to underestimate in this same range.

We also compared MSEs and found that, based on these, the two estimators of choice are likely to be L_0^+ and R_0^+, with the former being slightly preferable in circumstances when n is smaller. Both seem to give more robust performance over virtually all ranges than does Q_0^+. More details are given in Duval (1999).

Another way of considering which of these two estimators is preferable is to compare the analytic forms of their MSEs, based on (4), (5), and (6). We find that, for any fixed n, the MSE of L_0 is smaller than that of R_0 for larger values of k_0. This can be quantified, and the region in which L_0 is better is very closely approximated by $k_0 \geq n/4 - 2$. Thus, if the number of missing studies is estimated to be more than about 25% of those observed, we suggest use of L_0 might be preferable.

In practice, we advise using both these estimators before making a judgment on the actual number that might be suppressed, and indeed we advocate considering the value of Q_0^+ also. If the three agree, then conclusions are obvious, and if there is disagreement, one should perhaps use the resulting range of values as a basis for a sensitivity analysis.

4. Iterative Methods and Filling Funnel Plots

The estimators of k_0 we have used above depend on knowing the value of Δ because they rely on knowledge of whether a given observation is to the left or the right of Δ. This is clearly not the case in practice, and assuming that $\Delta = 0$ will lead to an obvious bias if in fact $\Delta > 0$.

In Duval and Tweedie (2000), we formally described how to carry out an iterative algorithm using the estimators above, and we now investigate the properties of this method.

The iteration is simple in concept. We first estimate Δ using a standard fixed or random effects model as described below. Using this value, we use one of the estimators of k_0 (say L_0^+) to decide how many unmatched values there might be around the initial estimate of Δ. We trim off this many values, and this leaves a more symmetric funnel plot.

On this trimmed set of data, we then re-estimate Δ, typically getting a value to the left of our previous value due to the studies we have trimmed. Using this new central value, we re-estimate the number of missing studies and then repeat the trimming process. We found in practice that this stabilizes on the real examples below after only 2–3 iterations.

When we have a final estimate of $\hat{\Delta}$ and a final estimate of $\hat{k}_0$, we fill the funnel with the imputed missing studies. We do this simply by taking the rightmost $\hat{k}_0$ studies, symmetrically reflecting their values of Y_i around the mean $\hat{\Delta}$ and using their values of σ_i for the imputed new studies. As a final step, we then re-estimate Δ using the observed and imputed studies and also use the observed and imputed studies to estimate a standard error for the effect size corresponding to that we would have seen if all these studies had been observed.

The random effects (RE) model that we use to combine the Y_j is $Y_j = \Delta + \beta_j + \epsilon_j$, where $\beta_j \sim \mathrm{N}(0, \tau^2)$ is introduced to account for heterogeneity between studies and $\epsilon_j \sim \mathrm{N}(0, \sigma_j^2)$ represents the within-study variability of study j. The RE approach has been argued (NRC Committee on Applied and Theoretical Statistics, 1992) to be preferable to the fixed effects (FE) model, which assumes that $\tau^2 = 0$, i.e., that any heterogeneity between studies is purely random.

Standard theory (Cooper and Hedges, 1994) then gives the meta-analyzed estimate of Δ as $\hat{\Delta} = \Sigma Y_j w_j / \Sigma w_j$, where $w_j = (\sigma_j^2 + \tau^2)^{-1}$ and $\mathrm{var}[\hat{\Delta}] = 1/\Sigma w_j$. In fitting this estimator, it is usually assumed that the σ_j^2 are known. In the RE model, there are various moment-based and maximum likelihood approaches giving estimates of τ^2 (Biggerstaff and Tweedie, 1997). The most common is the DerSimonian–Laird estimator (DerSimonian and Laird, 1986), and we use this method throughout.

Again we use simulations to evaluate the behavior of the iterated form of the estimators. The simulations are carried out for $N = n + k_0$ at values of 25, 50, 75, and the three values of $k_0 = 0, 5, 10$. We generate, for each of these combinations, 1000 sets of funnel plots, with each individual plot generated as in Section 3. These simulations are generated under a fixed effects model, although we analyze them using RE methods, as would typically be done in practice. Thus, the confidence intervals (CIs) we generate are conservative.

In Table 1, we first give the means and standard errors of each estimator for each of the parameter combinations. We also compare them with the means and standard errors for the noniterated versions, where Δ is known. It is clear that there is very little difference between the two except in the extreme case where $n = 15$ and $k_0 = 10$. Here the iterated version was more conservative than the version where we assumed that Δ is known.

This is precisely the outcome we might hope for. It indicates that the properties of the iterated version, which we must use in practice, are essentially exactly those of the noniterated version, which we are able to assess analytically.

Second, using these simulations, we consider the effect of the filling mechanism on recovering estimates of the true mean and on the inference concerning whether the mean is positive or zero. In Table 2, we give the average of the estimates of Δ and its 95% CI in four scenarios, i.e., when there are suppressed data and when the data are augmented by filled values based on R_0^+, L_0^+, and Q_0^+. It is clear that the trim and fill algorithm improves the accuracy when there are missing studies.

We also give the percentage of the 1000 data sets in which the lower end of the relevant calculated CIs were less than zero. If these are true 95% CIs, this percentage should be

Table 1

Behavior of various estimators of k_0 using the iterative method for Δ unknown

	$n = 25$				$n = 50$				$n = 75$			
	Iterative		Noniterative		Iterative		Noniterative		Iterative		Noniterative	
$k_0 = 0$	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
R_0^+	0.5	1.2	0.5	1.1	0.6	1.2	0.5	1.1	0.5	1.2	0.5	1.1
L_0^+	1.1	1.8	1.3	1.8	1.7	2.6	1.7	2.5	2.0	3.0	2.0	3.0
Q_0^+	1.3	2.5	1.4	2.1	1.9	3.0	1.8	2.7	2.1	3.4	2.1	3.2
	$n = 20$				$n = 45$				$n = 70$			
	Iterative		Noniterative		Iterative		Noniterative		Iterative		Noniterative	
$k_0 = 5$	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
R_0^+	4.6	3.6	5.0	3.4	5.1	3.7	5.0	3.4	5.0	3.4	5.1	3.5
L_0^+	3.7	2.7	4.4	2.8	4.9	3.9	5.0	3.8	5.1	4.4	5.5	4.6
Q_0^+	5.8	5.6	5.7	4.3	5.9	5.6	5.5	4.5	5.7	5.3	5.9	5.2
	$n = 15$				$n = 40$				$n = 65$			
	Iterative		Noniterative		Iterative		Noniterative		Iterative		Noniterative	
$k_0 = 10$	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
R_0^+	6.4	3.3	9.6	3.9	10.0	5.3	10.1	4.8	9.9	5.4	10.0	4.7
L_0^+	4.6	2.1	6.6	1.8	7.9	4.2	8.8	4.1	8.7	5.3	9.3	5.1
Q_0^+	8.6	5.1	10.8	3.9	11.4	8.5	10.7	5.9	10.7	7.9	10.4	6.2

Table 2

Behavior of various estimators of Δ and its CI using the iterative trim and fill algorithm; % cover indicates the frequency with which the left end of the CI is less than zero

	$n = 25$		$n = 50$		$n = 75$	
$k_0 = 0$	% Cover	Δ [CI]	% Cover	Δ [CI]	% Cover	Δ [CI]
Suppressed	98.4	0.00 [−0.09, 0.08]	98.5	0.00 [−0.06, 0.06]	97.8	0.00 [−0.04, 0.05]
R_0^+	98.6	0.00 [−0.09, 0.08]	98.6	0.00 [−0.06, 0.06]	98.0	0.00 [−0.05, 0.05]
L_0^+	98.5	−0.01 [−0.09, 0.08]	99.0	−0.01 [−0.06, 0.05]	98.1	0.00 [−0.05, 0.04]
Q_0^+	98.5	−0.01 [−0.10, 0.08]	99.0	−0.01 [−0.06, 0.05]	98.1	0.00 [−0.05, 0.04]
	$n = 20$		$n = 45$		$n = 70$	
$k_0 = 5$	% Cover	Δ [CI]	% Cover	Δ [CI]	% Cover	Δ [CI]
Suppressed	89.5	0.03 [−0.05, 0.12]	92.6	0.01 [−0.04, 0.07]	95.3	0.01 [−0.03, 0.05]
R_0^+	95.7	0.00 [−0.08, 0.09]	97.3	0.00 [−0.06, 0.06]	97.7	0.00 [−0.05, 0.05]
L_0^+	94.4	0.01 [−0.07, 0.10]	96.8	0.00 [−0.06, 0.06]	97.5	0.00 [−0.05, 0.05]
Q_0^+	94.9	−0.01 [−0.09, 0.08]	97.3	0.00 [−0.06, 0.05]	97.6	0.00 [−0.05, 0.04]
	$n = 15$		$n = 40$		$n = 65$	
$k_0 = 10$	% Cover	Δ [CI]	% Cover	Δ [CI]	% Cover	Δ [CI]
Suppressed	58.8	0.09 [−0.01, 0.19]	80.0	0.03 [−0.02, 0.09]	87.9	0.02 [−0.03, 0.06]
R_0^+	83.1	0.04 [−0.05, 0.14]	95.9	0.00 [−0.06, 0.06]	97.5	0.00 [−0.05, 0.05]
L_0^+	78.9	0.06 [−0.04, 0.16]	94.2	0.01 [−0.05, 0.07]	96.8	0.00 [−0.04, 0.05]
Q_0^+	84.5	0.02 [−0.07, 0.11]	95.1	0.00 [−0.06, 0.05]	97.2	0.00 [−0.05, 0.04]

Table 3
Power and size of tests based on R_0^+

k_0	0	1	2	3	4	5	6	7	8	9	10
$P(R_0^+ \leq 1)$	0.875	0.688	0.500	0.344	0.227	0.145	0.090	0.055	0.033	0.019	0.011
$P(R_0^+ \leq 2)$	0.938	0.813	0.656	0.500	0.363	0.254	0.172	0.113	0.073	0.046	0.029
$P(R_0^+ \leq 3)$	0.969	0.891	0.773	0.637	0.500	0.377	0.274	0.194	0.133	0.090	0.059

97.5%, or half the Type I error rate. (Note that, since suppression moves the mean to the right, we would only expect the Type I error rate to increase because of noncoverage of zero on this side.)

When there are no missing studies, the error rate is actually a little less than the nominal 2.5% due to the conservative nature of the RE model. This is not much affected by the filling algorithm. When there are missing studies, the filled versions are much closer to the real values and the coverage is much improved. When there is a large number of suppressed studies ($k_0 = 10$), the coverage typically reaches reasonable ranges (often around 95–97.5%) even though the observed (suppressed) data set leads to coverages as poor as 60–80%. When there are proportionally fewer missing studies, the coverage becomes more accurate still with the filling method.

5. Formal Tests of $k_0 = 0$

Using the distributional properties of R_0^+, we are able to construct surprisingly powerful tests of the hypothesis that $k_0 = 0$, where the rejection regions are of the form $\{R_0^+ > K\}$ for suitable values of K. Theorem 2 of Duval and Tweedie (2000) shows that, from (1), R_0 has the modified negative binomial distribution,

$$P(R_0 = m) = \binom{k_0 + m + 1}{m + 1} 0.5^{k_0+m+2}, \quad \text{for } m = -1, 0, 1, \ldots. \tag{7}$$

Using this, we can explicitly construct both the size and power of tests of the form $\{R_0^+ > K\}$. Table 3 shows the cumulative distribution for R_0^+ under various values of k_0 (recall that when $R_0 = -1$ then $R_0^+ = 0$). This shows that, under the null hypothesis $k_0 = 0$, a test of size greater than 95% is given by the region $\{R_0^+ > 3\}$. This test is of power uniformly greater than 80% for alternative hypotheses of $k_0 \geq 7$. If we consider the region $\{R_0^+ > 2\}$, the size is still over 93% and the power is 75% for alternative hypotheses of $k_0 \geq 5$. Note that these regions are independent of n and so can be used for meta-analyses with arbitrary numbers of studies.

We do not have the analytical properties of the estimator R_0^+ under the iterative trimming algorithm. However, simulation studies show that the analytical size and power of the regions $\{R_0^+ > 3\}$ and $\{R_0^+ > 2\}$ hold virtually exactly for all combinations of n, k_0, even when the iterative algorithm is used.

Although we do not have a simple form for the distribution of L_0^+, based on simulations, we find that the regions $\{L_0^+ > K\}$ typically give tests of smaller size and lower power for given K than those using R_0^+. Table 4 gives some illustrative values. This does show that, even if we accept the null hypothesis based on the test using R_0^+, we might still reject based on the test using L_0^+ if we find L_0^+ is larger than R_0^+ by two or three. These configurations should thus clearly be treated with some caution.

6. Examples from Psychometrics and Clinical Trials

We now apply the method to two examples from psychometrics and clinical trials and compare the results obtained from three other methods, i.e., those of Begg (1994), Egger et al. (1997), and Gleser and Olkin (1996). The first provides an estimate of the rank correlation between the individual standardized effect sizes and their variances based on Kendall's tau. We correct for ties, so our values differ slightly from those in Begg (1994) or Begg and Mazumdar (1994), who do not appear to make this correction. The second of these calculates the intercept of a simple regression of the effect size divided by its standard error against the precision, defined as the inverse of the standard error. The third is based on an assumption that the true value of $\Delta = 0$, and this appears to make it less applicable in many data sets.

Table 4
Size of test regions based on L_0^+ using simulations with $k_0 = 0$

n	$[L_0^+ \leq 1]$	$[L_0^+ \leq 2]$	$[L_0^+ \leq 3]$	$[L_0^+ \leq 4]$	$[L_0^+ \leq 5]$	$[L_0^+ \leq 6]$	$[L_0^+ \leq 7]$	$[L_0^+ \leq 8]$
19	0.711	0.828	0.898	0.948	0.975	0.993	0.999	1.000
28	0.691	0.775	0.866	0.917	0.951	0.979	0.989	0.996
29	0.679	0.796	0.864	0.924	0.955	0.979	0.989	0.995
35	0.655	0.766	0.842	0.903	0.946	0.962	0.978	0.990
50	0.641	0.723	0.795	0.853	0.896	0.931	0.955	0.972

Example 1. *IQ and teacher expectancy.* As a first example, we consider the set of 19 randomized studies of the effects of teacher expectancy on later pupil performance on an IQ test (Raudenbush, 1984) depicted in Figure 1. In these studies, a researcher tests the IQ of a random set of students. A randomly selected treatment group is identified to their teachers as likely to experience substantial intellectual growth, and the test is readministered at a later date. The effect size for each study represents the mean increase in IQ score of the high expectancy group minus the mean increase of the control group divided by a pooled standard deviation. One might hypothesize that teacher expectancy increases performance.

In a meta-analysis ignoring publication bias, the RE overall estimate of $\hat{\Delta} = 0.089$ with 95% CI $(-0.020, 0.199)$.

These data have been analyzed for publication bias by Begg (1994) and Gleser and Olkin (1996), among others. The tests all give results consistent with some degree of publication bias: $p = 0.07$ for Begg (1994), $p = 0.06$ for Egger et al. (1997), $p = 0.13$ for R_0^+, and $p = 0.17$ for L_0^+.

Estimating the number of missing studies, the methods of Gleser and Olkin (1996) yield three possible values: 0, 59, or (under some quite strong assumptions) 82. This degree of variability seems to indicate considerable unreliability in their approach. Our method gives much more plausible results than any of the Gleser and Olkin results: $R_0^+ = 2$ and $L_0^+ = 3$ (after three iterations), with standard errors of 2.4 and 2.9, respectively, using (4) and (5). We also find $Q_0^+ = 6$, indicating the overestimate discussed above.

Filling with the three missing studies indicated by L_0^+, then, as in Figure 1, we obtain an overall estimate of $\hat{\Delta} = 0.027$ with 95% CI $(-0.100, 0.155)$. The overall effect is considerably reduced from that estimated from the original data. After allowing for even this small amount of publication bias, the estimate of Δ is reduced by 2/3, as shown in the bottom panel of Figure 1.

This shows not only that there might be missing studies but that, on this data set, publication bias could be causing problems. Certainly there is sufficient indication to warrant wider investigation in this area.

Example 2. *Antimalarial drugs.* As a second example, we consider a review of antimalarial drugs in the *Cochrane Database of Systematic Reviews* (McIntosh and Olliaro, 1998). Thirteen studies, including a total of 2448 patients, compared artemisinin derivatives with quinine. These derivatives may have advantages for treating severe malaria since they are fast acting and effective against quinine-resistant malaria parasites. The objective of this review was to assess the effects of artemisinin drugs for severe and complicated falciparum malaria in adults and children. Since negative effect sizes (reduced mortality) are positive results in this study, the sign of the reported effect size needed to be reversed to fit our model. The data are shown as the solid circles in Figure 2.

In a meta-analysis ignoring publication bias, the RE overall estimate of the log odds ratio $\hat{\Delta} = 0.482$, with 95% CI $(0.157, 0.807)$, i.e., the artemisinin derivatives appear to be significantly more effective overall than quinine. An FE analysis gave similar results, $\hat{\Delta} = 0.362$, with 95% CI $(0.141, 0.582)$.

As is visually plausible from Figure 2, tests give results consistent with some degree of publication bias, i.e., $p = 0.18$ for Begg (1994) and $p = 0.04$ for Egger et al. (1997). The simple estimator of Gleser and Olkin (1996) yields an estimated number of missing studies of one.

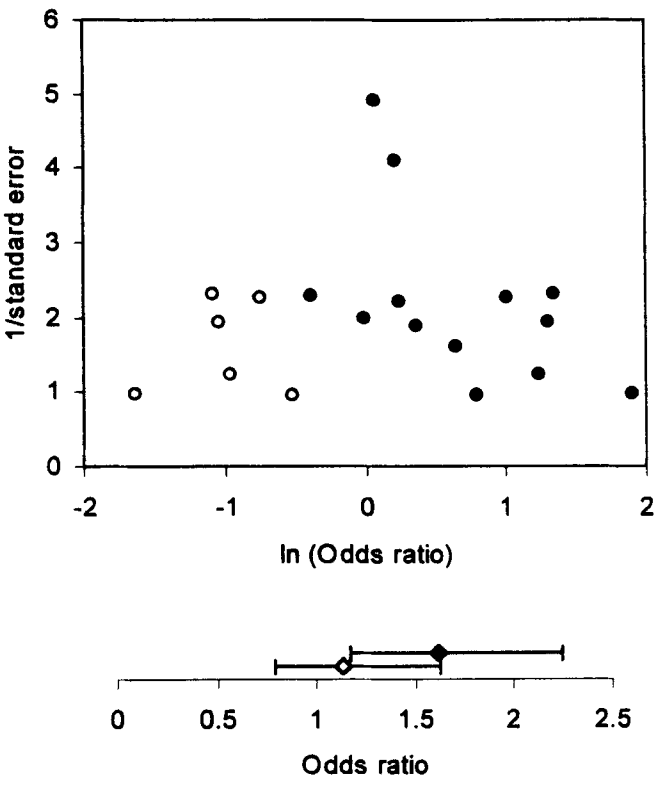

Figure 2. Top panel is a funnel plot of log odds ratios of mortality in studies of malaria from McIntosh and Olliaro (1998). Solid circles are original data, open circles are imputed filled values. Bottom panel shows overall mean and 95% CI of odds ratios before and after allowing for publication bias.

One of the interesting aspects of this example is that use of RE and FE models give different results. For the RE model, we find $R_0^+ = 0$, $L_0^+ = 6$, and $Q_0^+ = 12$, while for the FE model, we find $R_0^+ = 6$, $L_0^+ = 6$, and $Q_0^+ = 12$ (so Q_0^+ again seems to overestimate k_0). A test based on L_0^+ therefore gives significance, but that based on R_0^+ is significant only in the FE case. This difference is due to the different initial estimates of Δ in these two situations since R_0^+ can be sensitive to the distance of the leftmost value from this initial point.

Filling with the six missing studies indicated by L_0^+, then as in Figure 2 for the RE case, we obtain an overall estimate of $\hat{\Delta} = 0.123$ with 95% CI $(-0.242, 0.488)$; for the FE estimator, we obtain an overall estimate of $\hat{\Delta} = 0.114$ with 95% CI $(-0.085, 0.313)$. This is shown in terms of the odds ratio in the bottom panel of Figure 2.

This changes the inference in the *Cochrane Database of Systematic Reviews* from one of significance to one of nonsignificance, showing that again one needs to be sure that all the negative studies really have been reported in this area. A more detailed assessment of the degree of such bias in a large subset of the meta-analyses in the *Cochrane Database of Systematic Reviews* is in Sutton, Duval, Tweedie, Abrams, and Jones (unpublished manuscript).

7. Discussion

It is well known that there may be a nonrepresentative set of studies in the scientific literature. Dear and Dobson (1997), commenting on existing frequentist approaches (Dear and Begg, 1992) and the Bayesian approach in Givens et al. (1997) to solving this problem, noted (p. 246) 'previous methods have not been much used ... (and) ... the value of any new statistical methodology depends, in part, on the extent to which it is adopted'; they also noted (p. 245) that 'the culture of meta-analysis has traditionally favoured very simple methods'.

The trim and fill technique seems to fit this description. It uses only simple symmetry assumptions and an iterative approach, easy to implement in practice, to estimate the number of missing studies. In the real examples we have examined, the trim and fill method matches the subjective impression of bias given by the funnel plots.

Simulations indicate that our nonparametric estimators for the number of missing studies work well, and we identify appropriate ranges where using either R_0^+ or L_0^+ appears optimal. By using both, we get a good basis for judgment about the number of studies that need to be trimmed. Our evaluation of the iterative method shows that the estimates of Δ based on using the trim and fill approach are close to unbiased, and their variability is of an order that allows us to test for the existence of missing studies.

How much do we need to be concerned about publication bias? It is clear from simulations that we might wish to change the inferences made if we had the full picture. In the real examples we have examined here, in Duval (1999) and in Sutton et al. (1999), we have found the same thing. In those datasets where we detect considerable publication bias, the filled funnel plot may lead to inferences that are quantitatively different from those in the original data when one ignores possible suppressed studies.

Nonetheless, the main goal of this work should be seen as providing methods for sensitivity analyses rather than actually finding the values of missing studies. We are not interested in the exact imputed values. We are, however, interested in how much the value of Δ might change if there are missing studies, and from that perspective, the trim and fill approach does seem to give good indications of which meta-analyses do not suffer from publication bias and which need to be evaluated much more carefully.

Résumé

Nous étudions des méthodes non-paramétriques récemment développées pour l'estimation du nombre d'études manquantes pouvant exister dans une méta-analyse, et l'effet que ces études pourraient avoir sur son résultat. Ce sont des techniques simples par accroissement des données, basées sur les rangs, et qui formalisent l'utilisation des diagrammes en entonnoir. Nous montrons qu'elles fournissent des tests valides et relativement puissants pour évaluer l'existence de tels biais de publication. Après ajustement pour les études manquantes, nous montrons que l'estimateur de l'effet taille global est approximativement correct et que la qualité des intervalles de confiance de l'effet taille est nettement améliorée, égalant le niveau de confiance nominal dans de nombreux cas. Nous illustrons la méthode de coupure et remplissage à partir de méta-analyses existantes sur des études d'essais cliniques et de psychométrie.

References

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

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 Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics* **50,** 1088–1101.

Biggerstaff, B. J. and Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* **16,** 753–768.

British Medical Journal Editorial Staff. (1983). The editor regrets ... (editorial). *British Medical Journal* **280,** 508.

Cooper, H. (1998). *Synthesizing Research.* Thousand Oaks, California: Sage Publications.

Cooper, H. and Hedges, L. V. (eds). (1994). *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

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

Dear, K. and Dobson, A. (1997). Comment on Givens, G. H., Smith, D. D., and Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science* **12,** 245–246.

DerSimonian, R. and Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7,** 177–188.

Dickersin, K., Min, Y., and Meinert, C. (1992). Factors influencing publication of research results. *Journal of the American Medical Association* **267,** 374–378.

DuMouchel, W. and Harris, J. (1997). Comment on Givens, G. H., Smith, D. D., and Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science* **12,** 244–245.

Duval, S. J. (1999). Effects of publication bias in meta-analysis. Ph.D. dissertation, University of Colorado Health Sciences Center, Department of Preventive Medicine and Biometrics.

Duval, S. J. and Tweedie, R. L. (2000). A non-parametric 'trim and fill' method of assessing publication bias in meta-analysis. *Journal of the American Statistical Association* **95,** (in press).

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

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315,** 629–634.

Givens, G. H., Smith, D. D., and Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate (with discussion). *Statistical Science* **12,** 221–250.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher* **5,** 3–8.

Gleser, L. J. and Olkin, I. (1996). Models for estimating the number of unpublished studies. *Statistics in Medicine* **15,** 2493–2507.

Hedges, L. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science* **7,** 227–236.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem (with discussion). *Statistical Science* **3,** 109–135.

Light, R. and Pillemer, D. (1984). *Summing Up: The Science of Reviewing Research.* Cambridge: Harvard University Press.

McIntosh, H. M. and Olliaro, P. (1998). Artemisinin derivatives in the treatment of severe malaria (Cochrane Review). In *The Cochrane Library,* Issue 3. Oxford: Update Software.

Mengersen, K., Tweedie, R., and Biggerstaff, B. (1995). The impact of method choice in meta-analysis. *Australian Journal of Statistics* **37,** 19–44.

Misakian, A. L. and Bero, L. A. (1998). Publication bias and research on passive smoking: Comparison of published and unpublished studies. *Journal of the American Medical Association* **280,** 250–253.

NRC Committee on Applied and Theoretical Statistics. (1992). *Combining Information: Statistical Issues and Opportunities for Research.* Washington, D.C.: National Academy Press.

Raudenbush, S. W. (1984). Magnitude of teacher expectancy effects of pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 18 experiments. *Journal of Educational Psychology* **76,** 85–97.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. *Psychological Bulletin* **86,** 85–97.

Thompson, S. and Pocock, S. (1991). Can meta-analyses be trusted? *Lancet* **338,** 1127–1130.

Received October 1998. *Revised June* 1999.
Accepted July 1999.