Statistical Science
1992, Vol. 7, No. 2, 237-245

An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis

Keith B. G. Dear and Colin B. Begg

Abstract. A semi-parametric method is developed for assessing publication bias prior to performing a meta-analysis. Summary estimates for the individual studies in the meta-analysis are assumed to have known distributional form. Selective publication is modeled using a nonparametric weight function, defined on the two-sided p-value scale. The shape of the estimated weight function provides visual evidence of the presence of bias, if it exists, and observed trends may be tested using rank order statistics or likelihood ratio tests. The method is intended as an exploratory technique prior to embarking on a standard meta-analysis.

Key words and phrases: Meta-analysis, publication bias, weighted functions.

1. INTRODUCTION

A problem confronting any investigator who performs a meta-analysis is the risk that the collection of studies to be analyzed has been assembled via a biased sampling mechanism. A prominent concern in this regard is the risk of publication bias if the source of studies is the published literature. The evidence that publication bias is often a serious problem is based on the intuition of investigators who are involved in biomedical and social research, as well as on a variety of empirical studies, for example, White (1982); Glass, McGraw and Smith (1981); Coursol and Wagner (1986); Sterling (1959); Berlin, Begg and Louis (1989); Simes (1986); Dickersin and Meinert (1990) and Easterbrook et al. (1991). These studies, and the issue of publication bias in general, have been recently reviewed in Begg and Berlin (1988).

In the context of an individual meta-analysis one is interested in whether publication bias could have influenced the results. There are some informal techniques which can shed light on this issue. One approach, designed to determine retrospectively whether a statistically significant effect could be entirely due to the effect of selective publication, is to convert the p-values to z-scores, and calculate an average z-score including z-scores of zero for a hypothetical number of unpublished studies (Rosenthal, 1978). If the number of unpublished studies required to change a significant conclusion to a nonsignificant one is large, then we can be confident that the conclusion is not a false positive. Another approach is to plot the estimated effects against sample size, the so-called funnel plot (Light and Pillemer, 1984). If publication bias is present it should be a function of sample size, and thus will be reflected in the shape of the plot.

A more formal approach is to formulate the problem in the framework of a selection model, using weighted distributions (Patil and Rao, 1977), where the weight function is proportional to the probability that a study is published, as a function of a characteristic of the study which influences the decision to publish. Previous work on this model has been based on the assumption that the p-value is the factor that determines the chances of publication. Hedges (1984) has studied the implications of the extreme model in which it is assumed that studies which are significant at the 5% level are all published, and Iyengar and Greenhouse (1988) have modified this approach to incorporate parametric weight functions for the chances of the nonsignificant studies being published. In both of these papers it was assumed that the estimated effects had a normal distribution, a reasonable assumption for most applications unless the sample sizes in the studies are small.

A problem with the parametric weight functions of Iyengar and Greenhouse (1988), and the indicator weight function of Hedges (1984), is their monotonicity, in addition to their lack of flexibility for accommodating different shapes of selection functions. In

Keith B. G. Dear is Senior Lecturer in the Department of Statistics, University of Newcastle, Rankin Drive, Newcastle, NSW 2308, Australia. Colin B. Begg is Professor and Chairman of the Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10021.

practice we have little intuition, a priori, regarding the nature of the selection mechanism. Common sense, and the various empirical studies referenced earlier, suggest that a monotonic function which is at its maximum when the p-value is small may be a reasonable model in many cases. However, in some circumstances there may be prejudice in favor of the null hypothesis (Kotelchuk, 1974). So in general it is desirable that we should not predefine the shape of the weight function in any way, but rather should permit the data to dictate this.

Our goal in this paper is to develop an approach which achieves this objective, by allowing the shape of the weight function to vary in as unconstrained a manner as possible. In so doing we can develop a more convincing test of the presence of publication bias prior to embarking on a more conventional meta-analysis. Although semi-parametric models of weighted distributions have been previously studied (Vardi, 1985), these have focused on nonparametric estimation of the outcome distribution, rather than on nonparametric estimation of the weight function.

The subsequent paper in this volume by Hedges (1992) is similar in intent to our approach, and similar in practice in many respects also. The distinction is that Hedges chooses to pre-specify the regions of the p-value scale within which the weight function is assumed to be constant. The points of demarkation are selected to be those critical values commonly used in practice, such as 0.05, 0.01, 0.001, etc. The rationale is that because of the historical interpretive significance that these numbers have imparted, there is a reasonable expectation that they may reflect real points of discontinuity in the weight function. In practice this will lead typically to a weight function with fewer "steps," and as a result Hedges' method is probably more robust but less flexible than the method we develop in this article.

2. A SEMI-PARAMETRIC SELECTION MODEL

The basic idea of modeling selection bias is to define a function of a chosen aspect of a trial's outcome (a "weight function") which gives the probability that the trial is published. There are several outcomes which might be chosen, including the observed treatment effect, denoted by y, the absolute value of y or the corresponding one- or two-sided p-values.

In this article we have chosen the p-value scale as the dimension which influences publication bias. Throughout we use two-sided p-values, although the methodology can be adapted easily to the one-sided setting. That is, once a study is completed, the probability that it will be published is determined by the two-sided p-value. In the event that all the studies in the meta-analysis had the same sample size, this would be equivalent to assuming that the probability of publication is determined by the absolute treatment effect. A similar equivalence exists between one-sided p-values and unadjusted treatment effects.

If statistical significance increases the chances of publication, as we would generally expect if selective publication was occurring, then the observed (published) treatment effect sizes of the smaller studies will be increasingly extreme as the sample size decreases. Therefore if we were to construct the classical funnel-plot of sample size versus effect size we would observe a deficit of studies in the center of the plot (if selection depends on two-sided p-values) or a deficit among the negative small studies (if selection depends on one-sided p-values), as indicated by Light and Pillemer (1984). That is, our assumption that selective publication is determined by the p-value is consistent with the premise underlying the funnel-graph, although in practice selective publication may be influenced both by the observed p-value and by the magnitude of the observed effect, as well as by other features of study design and the scientific milieu at the time the study is completed (Berlin, Begg and Louis, 1989a).

Formalizing these ideas, we assume that there are n independent observed studies with normally distributed observed treatment differences y_i, $i = 1, \ldots, n$, where $E(y_i) = \theta$ and $\text{var}(y_i) = v_i^2 = u_i^2 + \sigma^2$, where u_i^2 is the known sampling variance in the ith study (largely determined by the sample size in the ith study), and σ^2 is a random effects component of variance representing the degree of heterogeneity present in the source population of effects. If we believe that the treatment effects are homogeneous we can constrain σ^2 to be zero.

We assume that the weight function, the probability that the study is published given the data, is a left-continuous step-function, operating on the scale of p-values, with discontinuities at alternate individual observed values of p. The assumption of left-continuity turns out to be of little consequence. Since the estimation of such a function using each observed p-value separately leads to nonidentifiability, we have arbitrarily grouped them in batches of two, after ordering them. The theory is easily altered to accommodate any grouping that leaves enough degrees of freedom to estimate θ and σ^2.

If the ordered p-values, ranked from the largest (p_1) to the smallest (p_n), are denoted $p_1, p_2, \ldots, p_n$, then the weight function is:

$$w(p) = \begin{cases} w_1, & \text{if } 1 \geq p > p_2, \\ w_j, & \text{if } p_{2j-2} \geq p > p_{2j}, \\ w_k, & \text{if } \begin{matrix} p_{n-1} \geq p > 0 \ (n \text{ odd}) \\ p_n \geq p > 0 \quad (n \text{ even}) \end{matrix} \end{cases},$$

where the number of weights, k, is $1 + \text{int}(n/2)$.

Thus w_1 covers the interval from $p = 1$ down to but not including p_2; w_2 covers the interval from p_2 down

to but not including p_4, and so on until w_k covers the interval from p_{n-1} down to zero or p_n down to zero, depending on whether n is odd or even, respectively.

In the scale of y, the outcome scale, the weight function depends on the individual study, since the sampling variances are different between studies. For the ith study, the weight function is

$$w_i(y) = \begin{cases} w_k, & \text{if } \infty > |y| \geq -u_i\Phi^{-1}(p_{2k-2}/2), \\ w_j, & \text{if } -u_i\Phi^{-1}(p_{2j}/2) > |y| \geq -u_i\Phi^{-1}(p_{2j-2}/2), \\ w_1, & \text{if } -u_i\Phi^{-1}(p_2/2) > |y| \geq 0, \end{cases}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

The weighted likelihood function is constructed as follows:

$$\begin{aligned} L(\theta, \sigma^2, \{w_j\}) &= \prod_{i=1}^{n} Pr(y_i | i\text{th study published}) \\ &= \prod_{i=1}^{n} \frac{f_i(y_i; \theta, \sigma^2) w_i(y_i)}{A_i(\theta, \sigma^2, \{w_j\})} \end{aligned} \tag{1}$$

where $A_i(\theta, \sigma^2, \{w_j\}) = \int_{-\infty}^{\infty} f_i(y; \theta, \sigma^2) w_i(y) dy$, and $f_i(\cdot)$ is the density function of y_i. Since $w_i(y)$ is piecewise constant, we can write

$$A_i = \sum_{j=1}^{k} H_{ij} w_j$$

where

$$H_{ij}(\theta, \sigma^2) = \int_{y: w_i(y) = w_j} f_i(y; \theta, \sigma^2) dy$$

The H_{ij}'s are the probabilities of study i yielding an outcome whose significance level is such as to cause the study to be published with relative probability w_j. See the Appendix for details.

It is computationally convenient to maximize the likelihood for (θ, σ^2) and $\{w_j\}$ separately. The MLE for $\{w_j\}$ given (θ, σ^2) is obtained as follows. From (1), the log likelihood $\mathscr{L}$ is

$$\mathscr{L} = \sum_{j=1}^{k} \lambda_j \log w_j + \sum_{i=1}^{n} \left(\log f_i(y_i; \theta, \sigma^2) - \log \sum_{j=1}^{k} H_{ij} w_j \right) \tag{2}$$

where $\lambda_1 = 1$; $\lambda_j = 2$ $(j = 2, \ldots, k-1)$; $\lambda_k = 1$ if n is even, 2 if n is odd.

Therefore

$$w_j \frac{\partial \mathscr{L}}{\partial w_j} = \lambda_j - \sum_{i=1}^{n} (w_j H_{ij} / A_i), \quad j = 1, \ldots, k \tag{3}$$

and

$$w_j^2 \frac{\partial^2 \mathscr{L}}{\partial w_j^2} = -\lambda_j + \sum_{i=1}^{n} (w_j H_{ij} / A_i)^2, \quad j = 1, \ldots, k.$$

The fact that $\log w_1$ is not multiplied by 2, in equation (2), is entirely due to the arbitrary indexing of the inequalities defining the weight function. This leads to MLE's that will be biased. To see this, consider that for given θ and σ^2 the estimate of a weight will depend on the number of studies falling within its range compared with what would be expected given the width of that range. As defined, most of the weights span two intervals and are applied to two studies in the likelihood function. When n is even, w_k is applied to only one study and spans only one interval. But w_1, although applied only to the study which generated p_1, still is defined to span two intervals, from $p = 1$ to $p = p_1$ and from $p = p_1$ to $p = p_2$. This will cause w_1 to be estimated with negative bias, since it will appear that too few studies were published with p-values in that range. Alternatively, consider that the estimating equations (3) are conceptually like score equations in which the λ_j's play the role of the random variables. Then H_{ij}/A_i is the conditional probability that study i is associated with weight j, and the sum of these terms is the expected value of λ_j. For $j = 1$, the equation has expectation zero only for $\lambda_1 = 2$. In reality, however, it is the H_{ij}'s that are random, while the λ_j's are determined as part of the model formulation.

That this bias in w_1 is important can be shown by comparing the results from the left- and right-continuous models. We found, in analyzing the 17 real data sets discussed below, that the estimates of θ differed considerably between the two models, with the estimate from the left-continuous model being consistently closer to zero than the estimate from the right-continuous model. This suggests that a downward bias in the estimated weight close to $p = 1$ (the left-continuous model) biases $\hat{\theta}$ toward zero, while conversely a similar bias in the estimated weight at $p = 0$ (in the right-continuous model) results in a positive bias in $|\hat{\theta}|$. We therefore modified the likelihood by setting $\lambda_1 = 2$. With this modification, the estimates from the two models are identical when n is odd and very close when n is even (in which case the steps of the weight function are located differently in the two models).

It is important to clarify that this procedure will always lead to a set of estimates of $\{w_j\}$, the largest of which will be 1. However, the actual selection probabilities will typically all be less than 1, since some selective publication may occur for all p-values. Therefore, the procedure estimates only the relative weights. To estimate the absolute weight function would require information on the p-values of the unpublished studies. In addition to providing a useful visual display of the relative weight function for the purpose of identifying publication bias, the estimated relative weights can be used to test formally for bias, using either a rank correlation test (e.g., Kendall's Tau) or a likelihood ratio test comparing the fitted model with the sub-model in which the weights are all constrained to be equal such as the method-of-moments model of DerSimonian and Laird (1986), or restricted maximum-likelihood (REML), in which the equations of DerSimonian and

Laird are iterated to convergence. However, a likelihood ratio test is likely to be considerably less powerful for detecting monotonic trends in the weight function.

3. COMPUTATIONAL ISSUES

The model discussed in the previous section requires optimization of a considerable number of parameters. As well as the estimated mean treatment difference θ and the between-studies variance component σ^2, there are roughly $n/2$ "weight" parameters for a meta-analysis of n studies. We fitted the model using nested iterations, optimizing the weights iteratively for each pair of values of θ and σ^2 before recomputing the derivatives of the log-likelihood with respect to θ and σ^2.

For iterative estimation of σ^2, it is necessary to perform Newton-Raphson steps in $\log \sigma^2$. This is because as σ^2 approaches zero in cases where the maximum log-likelihood is at zero, the second derivative with respect to σ^2 is often positive. This causes the iterative steps to move away from, rather than towards, zero. Using instead $\log \sigma^2$ ensures that the second derivative is negative (for sufficiently large negative $\log \sigma^2$), so that the iterative steps move in the correct direction. The derivatives used in the program were calculated in terms of σ^2, then transformed numerically.

The procedure presented here is intended primarily as a means of informally exploring the degree of publication bias which may have operated in the selection of studies contributing to a meta-analysis. Inference about θ and σ^2 should be considered secondary at this stage. However, if desired, approximate standard errors for θ and σ^2 can be obtained numerically, by inverting the 2×2 matrix of second derivatives of the profile log-likelihood for these parameters, at the local MLE for $\{w_j\}$. These estimated standard errors were found generally to be only slightly larger than those obtained from unweighted models. However, the use of the profile log-likelihood for this purpose will usually lead to underestimation of the standard error, since the variation in the estimate of $\{w_i\}$ is not fully accommodated.

One effect of setting $\lambda_1 = 2$ to obtain score equations of approximately zero mean is that the implied modified likelihood function is no longer scale-invariant in the weights. If we set $w_j' = \alpha w_j$, $j = 1 \ldots k$, then we find $\mathcal{L}(\{w'\}) = \mathcal{L}(\{w\}) + \log(\alpha)$, so that the likelihood increases indefinitely as the weights increase together. This conflicts with the definition of the weights as being *relative* probabilities of publication, which would imply that their estimates can be identified only apart from an arbitrary scale factor. By imposing the natural constraint that $w_j \leq 1$, a unique solution to the estimating equations is achieved. Relative to the MLE, this solution tends to have more uniform weights, corresponding to relatively conservative assessment of the degree of selection bias. This method of estimation has the added benefit of avoiding the computational problem of numerical underflow in the A_i when, as occasionally happens, all but one of the weights have MLE's of zero.

Although the w_j are certainly not mathematically independent, making Newton-Raphson steps in each separately, rather than treating them as a vector and computing the inverse Hessian, proved acceptably quick in practice. The correctness of the solutions was checked by confirming that small perturbations of each w_j from its estimate reduced the likelihood, and by comparing the final estimates of θ and σ^2 with contour plots of the profile likelihood of these parameters, maximizing with respect to the weights. In all cases the contour plots showed clearly a maximum at the estimated value.

4. SIMULATIONS AND EXAMPLES

Since the intended use of the method involves a graphical display of the estimated weight function, a small simulation study was performed to illustrate the patterns to be expected with and without the presence of publication bias. We limited the simulations to three generated data sets for each configuration listed below so the results only provide a guide to what we might expect, rather than a complete study of the operating characteristics of the method.

Two underlying probability models were used to generate the p-values, prior to selective publication. Each model is based on the assumption that there is no heterogeneity. The uniform distribution was used to generate p-values under the null hypothesis of no association, that is, $\theta = 0$. For meta-analyses with a true nonzero effect we used $\theta = 1$, with a common standardized sampling variance, that is, $u_i = 1$, $\forall i$, and generated the p-values using the density function:

$$g(p) = \frac{1}{2}\{\exp[-\theta^2/2 - \theta\Phi^{-1}(p/2)] + \exp[-\theta^2/2 + \theta\Phi^{-1}(p/2)]\}$$

This is the density that is generated by a test of the hypothesis that $\theta = 0$, when $\sigma^2 = 0$ and $u_i = 1$. Note that by assuming a common sampling variance we are eliminating the impact of sample size on bias. Therefore we are dealing with a configuration in which the funnel graph would have no discriminatory power.

To simulate selective publication we first used the selection function $w(p) = \exp(-4p^3)$, which is an arbitrary but plausible function representing a moderate degree of preferential selection for studies with smaller p-values. This function is displayed in Figure 1 using a solid curve. The probability of publication is very high for p-values of 0.2 or less (e.g., 0.97 at $p = 0.2$)

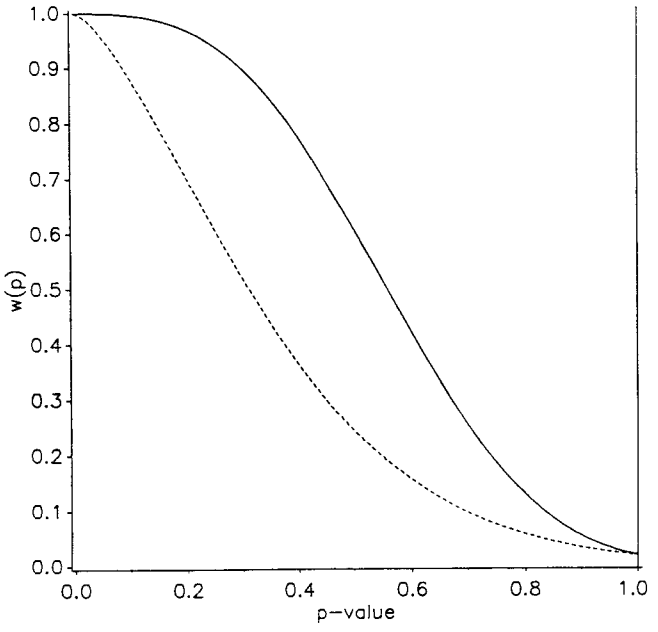

FIG. 1. *Underlying selection functions. Solid curve:* $w(p) = \exp(-4p^3)$; *dashed curve:* $w(p) = \exp(-4p^{1.5})$.

but starts to fall dramatically beyond $p = 0.3$. This function was used to induce biased samples for each of the preceding models, that is, $\theta = 0$ and $\theta = 1$. There are, therefore, four configurations of interest, denoted a–d in Figure 2 and Table 1. These are (a) no effect ($\theta = 0$) and no selection bias; (b) no effect ($\theta = 0$) with selection bias; (c) treatment effect ($\theta = 1$) and no selection bias; and (d) treatment effect ($\theta = 1$) with selection bias.

In each simulation a p-value was generated from the chosen model and then entered into the meta-analysis on the basis of a biased-coin randomization, the probability of entry being determined by the weight from Figure 1 corresponding to the sampled p-value. In configurations (a) and (c), since no studies were rejected, the meta-analysis comprised the first 25 studies generated. In configurations (b) and (d) the sampling was continued until 25 studies were actually selected for the meta-analysis. Three meta-analyses were generated for each configuration, and the results are summarized in Table 1.

Configuration (a) corresponds to sampling of p-values from a uniform distribution with no bias. Therefore the expected mean p-value is 0.50, and the expectation of $\hat{\theta}$ is 0.0. The simulated values of these quantities are as one would expect, and the rank correlation tests for publication bias are all nonsignificant. The corresponding graphs of the estimated weight functions are displayed in Figure 2a. The graphs provide visual confirmation of the lack of bias, demonstrating a seemingly random configuration of estimated weights.

Configuration (b) provides examples of biased sampling when there is no underlying treatment effect. The sampled p-values are on average substantially less than 0.5, reflecting the biased selection, but the estimation of θ seems to adjust the effects appropriately. The estimated weight functions in Figure 2b appear to display observable trends in the first and third simulations, both of which are statistically significant.

Corresponding simulations for the model in which there is a substantial treatment effect are displayed in Figures 2c and 2d and summarized in Table 1. None of the tests for bias show any evidence of bias. This is not surprising when we consider the fact that the presence of a strong treatment effect leads to a natural preponderance of relatively small p-values, mostly less than $p = 0.30$. However, Figure 1 shows that the selection function is relatively flat in this region so that the force of selectivity is relatively small in this configuration. This is confirmed by the fact that relatively few of the trials which were generated were rejected (third column of Table 1). Another way of expressing this is to say that, in general, if there is a strong treatment effect most of the resulting p-values from individual studies will be small, and these will mostly tend to be published if the weight function is uniformly high for small p-values.

To clarify this issue we repeated the simulations for the treatment effect of $\theta = 1$, this time using a weight function with a more pronounced selection effect in the region of p-values generated by this model. This was generated using the selection function $\exp(-4p^{1.5})$, represented by the dashed line in Figure 1. The results are summarized in the last section of Table 1, and graphed in Figure 2e. These simulations indicate that the model has the capacity to distinguish the effects of selective publication from the effect of a strong underlying treatment effect, in circumstances where selective publication has a strong effect on the number of studies published.

As a further test of the method we have analyzed the results of 17 published meta-analyses. This includes a meta-analysis of 10 studies comparing experimental with traditional education on creativity which was originally reanalyzed by Iyengar and Greenhouse (1988) using their parametric weight function, and 16 meta-analyses of medical interventions which are described and analyzed using conventional methods in Berlin et al. (1989). It is probably reasonable to assume that these latter studies are representative of typical meta-analyses in the medical literature, with respect to the numbers of component studies, etc. However, it is likely that meta-analyses in the social sciences will typically have more component studies.

The results are summarized in Table 2. The number of component studies ranges from 8 to 26, so that our simulations are on the high end of this range. We found that the test for publication bias was statistically sig-

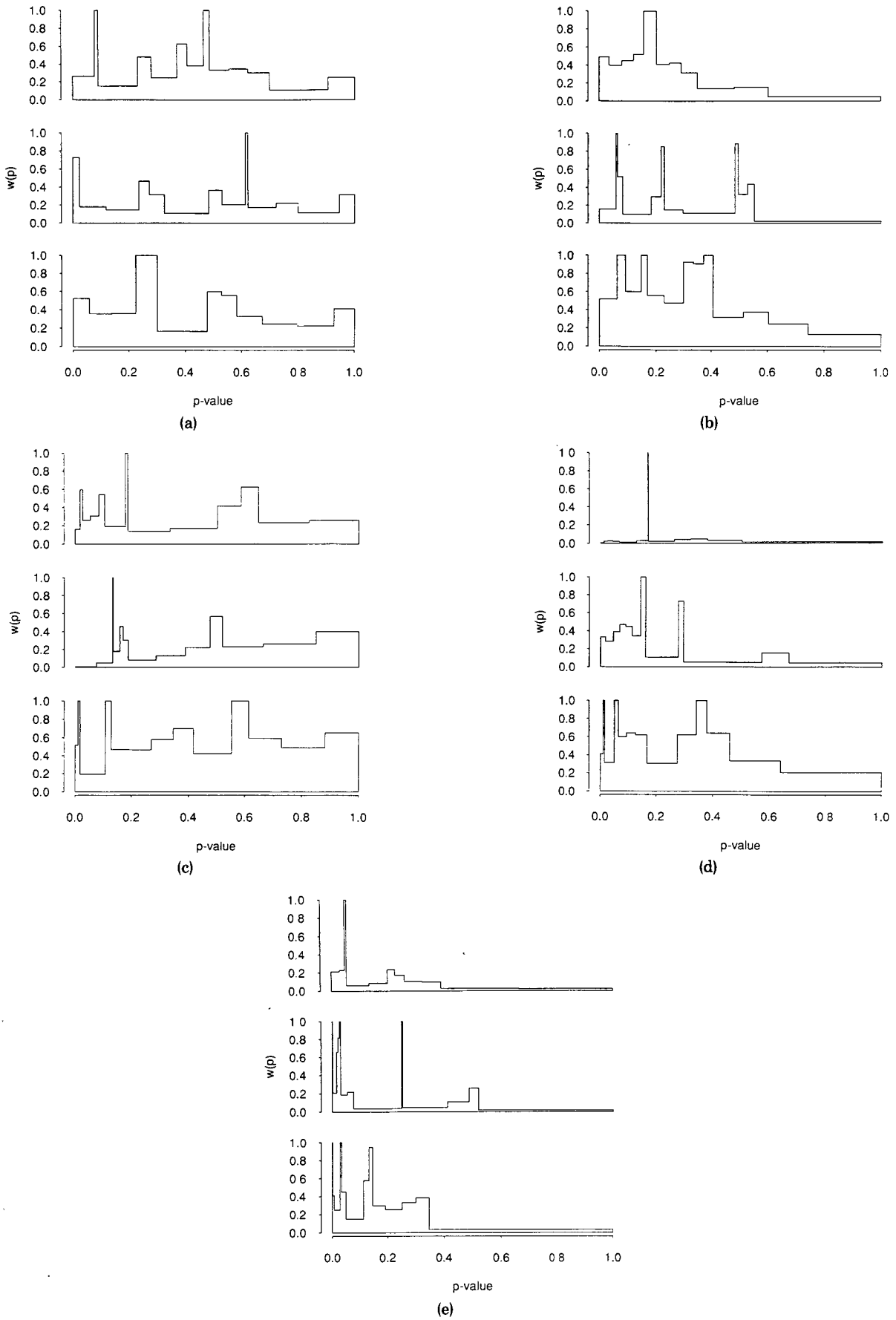

FIG. 2. *Estimated weight functions corresponding to the simulations in Table 1.*

TABLE 1
Summary of simulated meta-analyses

Configuration	Simulation	Number of trials generated	Mean of p-values	$\hat{\theta}$ (±2 s.e.)	Test for publication bias
(a) No effect (θ=0) No biased selection	1	25	0.43	0.07 (±0.40)	p = 0.15
	2	25	0.48	−0.10 (±0.34)	p = 0.27
	3	25	0.44	−0.06 (±0.37)	p = 0.10
(b) No effect (θ=0) Moderately biased selection [exp(−4p^3)]	1	42	0.24	0.16 (±0.30)	p = 0.02
	2	53	0.30	−0.24 (±0.34)	p = 0.20
	3	44	0.33	−0.14 (±0.34)	p = 0.02
(c) Underlying effect (θ=1) No biased selection	1	25	0.30	1.24 (±0.83)	p = 0.55
	2	25	0.34	2.29 (±0.96)	p = 0.93
	3	25	0.36	1.14 (±0.64)	p = 0.62
(d) Underlying effect (θ=1) Moderately biased selection [exp(−4p^3)]	1	37	0.18	1.87 (±0.12)	p = 0.89
	2	35	0.22	0.63 (±0.43)	p = 0.18
	3	31	0.23	1.24 (±0.72)	p = 0.21
(e) Underlying effect (θ=1) Strongly biased selection [exp(−4$p^{1.5}$)]	1	51	0.21	0.59 (±0.38)	p = 0.01
	2	49	0.19	0.67 (±0.40)	p = 0.04
	3	56	0.14	0.91 (±0.51)	p = 0.06

nificant at the 5% level in two of the 17 studies, with a p-value of 0.09 in one of the remaining studies. Interestingly, the data set of Iyengar and Greenhouse (1988) was part of one of the two significant studies. The estimated weight function for this study is plotted in Figure 3, and a marked trend is evident even though there are only five separate categories in the histogram. This plot is fairly typical in that the high spikes can be very "thin," relative to portions of the weight function which are correspondingly "fat." Since the value of the plots is as a visual display rather than a formal analysis, the presentation could be altered to highlight the trends in the histogram more clearly. In Figure 4 the scale of the data is adjusted to provide equally spaced intervals. The disadvantage of this plot is that one must carefully study the horizontal axis to determine the ranges of p-values represented by the columns.

TABLE 2
Summary of examples of meta-analyses

Study	Number of trials	Mean of p-values	$\hat{\theta}$ (±2 s.e.)	Test for publication bias
Iyengar and Greenhouse (1988)	10	0.23	0.14 (±0.32)	p = 0.02
Berlin et al. (1989)				
1	26	0.54	−0.009 (±0.14)	p = 0.96
2	15	0.47	0.003 (±0.016)	p = 0.77
3	11	0.41	−0.015 (±0.014)	p = 0.13
4	10	0.34	−0.018 (±0.040)	p = 0.09
5	8	0.08	−0.21 (±0.30)	p = 0.77
6	14	0.53	−0.008 (±0.020)	p = 0.81
7	16	0.55	−0.004 (±0.010)	p = 0.77
8	15	0.38	−0.019 (±0.034)	p = 0.19
9	15	0.09	−0.40 (±0.15)	p = 0.34
10	17	0.47	−0.08 (±0.05)	p = 0.96
11	25	0.26	0.10 (±0.07)	p = 0.05
12	11	0.17	0.16 (±0.14)	p = 0.23
13	12	0.28	−0.09 (±0.21)	p = 0.15
14	12	0.36	−0.10 (±0.08)	p = 0.62
15	10	0.41	−0.10 (±0.07)	p = 0.50
16	20	0.05	−0.43 (±0.15)	p = 0.56

5. DISCUSSION

We have developed a technique for characterizing the nature of publication bias in a meta-analysis, by

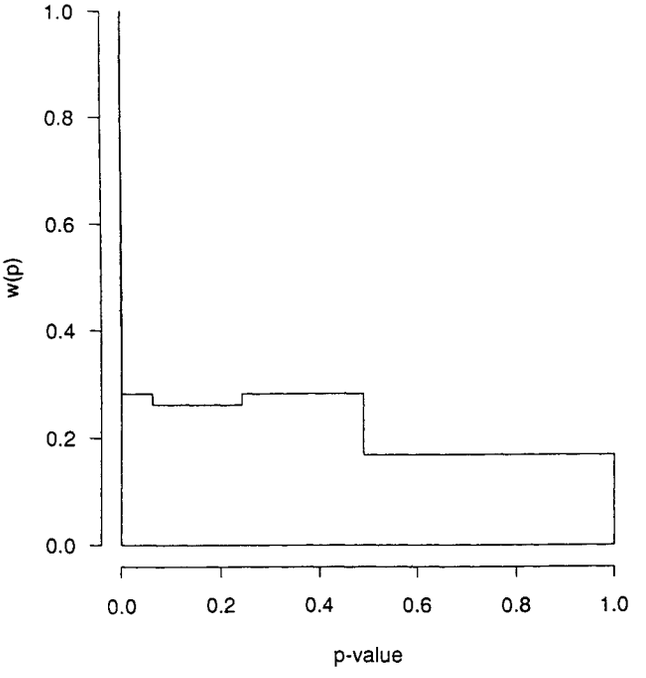

FIG. 3. *Estimated weight function for Iyengar and Greenhouse (1988) data.*

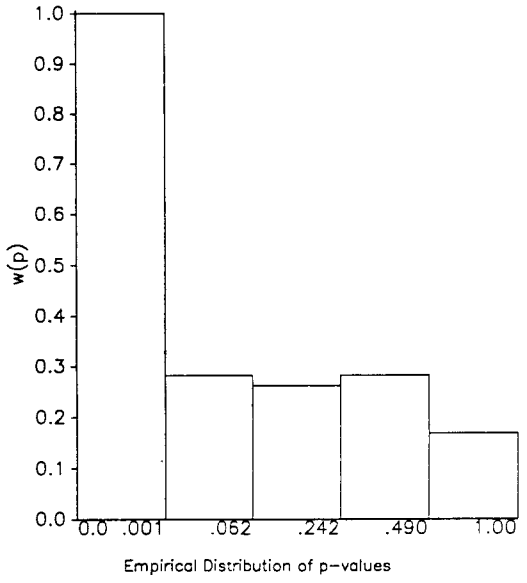

FIG. 4. *Estimated weight function for Iyengar and Greenhouse (1988) data, equally spaced intervals.*

using a semi-parametric weighted distribution model. The model is based on the premise that publication bias, if it exists, is primarily determined by the observed p-value. That is, the analysis uses the p-value scale as the metric upon which the weight function is defined. The method could be adapted to alternative metrics, such as the outcome scale, for example, if it were felt that the observed effect, rather than the p-value, was the primary determinant of bias. In principle the method could also be generalized to adjust for nonoutcome predictors of bias, such as the sample sizes of the component studies, etc., but this seems likely to be unsuitably complex for the small meta-analyses that are commonplace in medical research.

The method is complementary to the traditional approach of using a "funnel graph" to identify publication bias. The funnel-graph approach is based on the premise that bias will be identified in the relation of effect size to sample size, in that there will be a noticeable decrement in specific regions of this graph in the presence of bias, noticeably among nonsignificant, or negative small studies. This phenomenon would be a direct consequence of a selection mechanism in which the probability of publication is a function of the observed p-value. Therefore, the weight function approach advocated in this article will identify patterns of bias that are reflected in the funnel graph. It represents a more formal methodology for identifying publication bias. An important difference between the methods is that the weight function approach is determined by the "pattern" of observed p-values relative to their expectation under a normal sampling model. As such the method is, in principle, sensitive to publication bias even when the sample sizes in the component studies are very similar, a situation in which the funnel graph has no value.

The method has various unverifiable methodological assumptions. We assume that the summary outcome in each study is normally distributed, an assumption that seems reasonable if the sample sizes in the component studies are not too small. We have used likelihood theory to obtain the parameter estimates, but their asymptotic sampling properties are likely to be inaccurate if the number of studies in the meta-analysis is small. Individual estimates of the weights will always be deficient in this respect if we choose to group the outcomes in two's, although in a large meta-analysis it may be preferable to employ a coarser grouping.

In fact, the methodology proposed by Hedges (1992) employs a weight function model in which the points of demarkation separating regions of p-values with constant weight are assumed to be known in advance, and are chosen to reflect likely points of discontinuity such as commonly used critical values. Clearly the number of intervals could be chosen to be as large or as small as is convenient, but one would have to know in advance that each interval contains at least one observed p-value in the sample. Research is clearly needed in assessing and comparing the operating characteristics of these two methods. However, our intuition suggests that the Hedges model will be more suitable for meta-analyses with substantial numbers of component studies, while our method will be necessary for small meta-analyses.

Finally, we must emphasize that this is an exploratory technique. In general, as a result of the risks of sampling bias and study heterogeneity, meta-analysis must be viewed as a methodology with limited accuracy, and interpreted accordingly. Our method for assessing publication bias, while developed using formal statistical theory, is really intended as an informal tool, to assist in establishing whether there is evidence that the sampling of studies precludes the reliability of performing a standard meta-analysis.

As a result we have focused in this article on the use of the weight function to identify bias, rather than to correct it. However, there is no theoretical obstacle to using the MLE's for θ and σ^2 from the resulting model as corrected estimates of the mean study effect and the heterogeneity of the effects, with standard errors of the parameters derived from the information matrix. Our feeling is that if, in practice, the analysis identifies strong evidence of bias, one should be skeptical about the reliability of using the meta-analytic approach for making inferences about the parameters. In other words, we envisage the method as a preliminary analytic tool, which can clear the way for a conventional meta-analysis if there is no evidence of bias. However, if bias is identified one should be very cautious about using the model to correct it. Rather, attention should

be focused on the causes of bias, perhaps by initiating a search for missing studies.

APPENDIX

We have

$$H_{ij}(\theta, \sigma^2) = \int_{y:w(y)=w_j} f_i(y; \theta, \sigma^2)dy.$$

Taking $p_0 = 1$ and $p_{2k} = 0$ provides

$$H_{ij} = F_i(u_i\Phi^{-1}(p_{2j-2}/2)) - F_i(u_i\Phi^{-1}(p_{2j}/2)) + F_i(-u_i\Phi^{-1}(p_{2j}/2)) - F_i(-u_i\Phi^{-1}(p_{2j-2}/2))$$

or

$$H_{ij} = \Phi\left(\frac{-|y_{2j-2}|u_i/u_{2j-2} - \theta}{v_i}\right) - \Phi\left(\frac{-|y_{2j}|u_i/u_{2j} - \theta}{v_i}\right) + \Phi\left(\frac{|y_{2j}|u_i/u_{2j} - \theta}{v_i}\right) - \Phi\left(\frac{|y_{2j-2}|u_i/u_{2j-2} - \theta}{v_i}\right)$$
$$= \Phi(a_{ij}) - \Phi(b_{ij}) + \Phi(c_{ij}) - \Phi(d_{ij}).$$

To obtain estimates of θ and σ^2 we use the following derivatives:

$$\frac{\partial \mathscr{L}}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{v_i}\left[\frac{y_i - \theta}{v_i} + R(G_i)\right]$$

$$\frac{\partial \mathscr{L}}{\partial \sigma_2} = \sum_{i=1}^{n} \frac{1}{2v_i^2}\left[-1 + \left(\frac{y_i - \theta}{v_i}\right)^2 + R(F_i)\right]$$

$$\frac{\partial^2 \mathscr{L}}{\partial \theta^2} = \sum_{i=1}^{n} \frac{1}{v_i^2}[-1 + R(F_i) + R^2(G_i)]$$

$$\frac{\partial^2 \mathscr{L}}{\partial [\sigma^2]^2} = \sum_{i=1}^{n} \frac{1}{4v_i^4}\left[2 - 4\left(\frac{y_i - \theta}{v_i}\right)^2 + R(D_i) - 3R(F_i) + R^2(F_i)\right]$$

$$\frac{\partial^2 \mathscr{L}}{\partial \theta \partial \sigma^2} = \sum_{i=1}^{n} \frac{1}{2v_i^3}\left[2\left(\frac{y_i - \theta}{v_i}\right) - R(G_i) + R(E_i) + R(F_i)R(G_i)\right]$$

where

$$R(X_i) = \frac{\sum_{j=1}^{k} w_j X_{ij}}{\sum_{j=1}^{k} w_j H_{ij}}$$

and

$$G_{ij} = \phi(a_{ij}) - \phi(b_{ij}) + \phi(c_{ij}) - \phi(d_{ij})$$
$$F_{ij} = a_{ij}\phi(a_{ij}) - b_{ij}\phi(b_{ij}) + c_{ij}\phi(c_{ij}) - d_{ij}\phi(d_{ij})$$
$$E_{ij} = a_{ij}^2\phi(a_{ij}) - b_{ij}^2\phi(b_{ij}) + c_{ij}^2\phi(c_{ij}) - d_{ij}^2\phi(d_{ij})$$
$$D_{ij} = a_{ij}^3\phi(a_{ij}) - b_{ij}^3\phi(b_{ij}) + c_{ij}^3\phi(c_{ij}) - d_{ij}^3\phi(d_{ij}).$$

ACKNOWLEDGMENTS

The authors thank Professor Nan Laird and Dr. Stuart Lipsitz for their helpful discussions, Dr. Jesse Berlin for the use of his data and Karen Abbett and Terry Crespo for assistance in preparing the manuscript. This work was supported in part by grant CA-06516 awarded by the National Cancer Institute, DHHS.

REFERENCES

BEGG, C. B. and BERLIN, J. A. (1988). Publication bias: A problem in interpreting medical data. *J. Roy. Statist. Soc. Ser. A* **151** 419–463.

BERLIN, J. A., BEGG, C. B. and LOUIS, T. A. (1989). An assessment of publication bias using a sample of published clinical trials. *J. Amer. Statist. Assoc.* **84** 381–392.

BERLIN, J. A., LAIRD, N. M., SACKS, H. S. and CHALMERS, T. C. (1989). A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine* **8** 141–152.

COURSOL, A. and WAGNER, E. E. (1986). Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias. *Professional Psychology* **17** 136–137.

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

DICKERSIN, K. and MEINERT, C. (1990). Risk factors for publication bias: results of a follow-up study (abstract). *Controlled Clinical Trials* **11** 255.

EASTERBROOK, P. J., BERLIN, J. A., GOPALAN, R. and MATTHEWS, D. R. (1991). Publication bias in clinical research. *Lancet* 867–872.

GLASS, G. V., MCGRAW, B. and SMITH, M. L. (1981). *Meta-Analysis in Social Research*. Sage, Beverly Hills.

HEDGES, L. V. (1984). Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* **9** 61–85.

HEDGES, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statist. Sci.* **7** 246–255.

IYENGAR, S. and GREENHOUSE, J. B. (1988). Selection models and the file drawer problem. *Statist. Sci.* **3** 109–135.

KOTELCHUK, D. (1974). Asbestos research: Winning the battle but losing the war. *Health/PAC Bulletin* **61** 1–27.

LIGHT, R. J. and PILLEMER, D. B. (1984). *Summing Up: The Science of Reviewing Research*. Harvard Univ. Press.

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. North-Holland, Amsterdam.

ROSENTHAL, R. (1978). Combining results of independent studies. *Psychological Bulletin* **85** 185–193.

SIMES, R. J. (1986). Confronting publication bias: a cohort design for meta-analysis. *Statistics in Medicine* **6** 11–30.

STERLING, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Amer. Statist. Assoc.* **54** 30–34.

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178–203.

WHITE, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin* **91** 461–481.