Comment

Nan Laird, G. P. Patil and C. Taillie

Publication bias has long been assumed to be a limiting factor for the utility of data synthesis of results from the published literature. As the authors so clearly demonstrate, the early attempts by meta-analysis methodologists to develop techniques for handling publication bias have been of limited utility. By applying selection models in this setting, the authors have expanded the set of available techniques for dealing with the publication bias problem.

As the authors note, the selection problem is closely related to the missing data problem; one interpretation of the weight function $\omega(t; \alpha)$ is the probability of publication given study results t. Viewing the selection model problem as a missing data problem enhances the possible approaches for analysis. In particular, the sample survey literature on methods for handling nonresponse is extensive; we feel that many of the approaches developed for sample survey can be used with advantage in the meta-analysis setting. Of course, a major distinction between the publication bias problem in meta-analysis and the standard missing data problem is the fact that in the latter, the number of nonrespondents is generally known, whereas in the former, the number of unreported studies is not. As a result, the analysis of Iyengar and Greenhouse is based on the likelihood of the observed data, conditional on the observations coming from reported studies. In contrast, in missing data problems the observed data consist not only of the observations on the respondents, but the number of nonrespondents as well, giving rise to a likelihood of the form (abusing the notation slightly)

$$\prod_{i=1}^{n_0} f_i(t_i \mid \Theta)\omega(t_i; \alpha) \prod_{i=1}^{n_m} \int f_i(t \mid \Theta)(1 - \omega(t; \alpha))\, dt, \tag{1}$$

where the total sample size is n, and n_0 and n_m are the number of respondents and nonrespondents, respectively.

Because of the similarity of the likelihood in (1) to that used by Iyengar and Greenhouse, it is worth drawing attention to some features of the maximum likelihood analysis with (1). First is that parameter estimates can be much more sensitive to the parametric form assumed for $f(t \mid \Theta)$ than in the corresponding complete data case where $n_m = 0$ (Little, 1983). This has been illustrated by Glynn, Laird and Rubin (1986) for estimating a sample mean when uncertainty exists about the shape of the error distribution. We suspect that the same is true for the selection model likelihood. This sensitivity is potentially important in meta-analysis where the distributional properties of observed study statistics may be only approximately specified, especially when study sizes are small. Sensitivity to specification of $f(t \mid \Theta)$ can be greatly reduced if responses are obtained by subsequent follow-up of initial nonrespondents, using a classical double-sampling approach (Hansen and Hurwitz, 1946). Letting n_f denote the number of follow-ups, the likelihood with followups expands to

$$\prod_{i=1}^{n_0} f_i(t_i \mid \Theta)\omega(t_i \mid \alpha) \prod_{i=1}^{n_f} f_i(t_i \mid \Theta)(1 - \omega(t_i \mid \alpha)) \prod_{i=1}^{n_m - n_f} \int f_i(t \mid \Theta)(1 - \omega(t_i \mid \alpha))\, dt. \tag{2}$$

Clearly, the simple analysis proposed by Iyengar and Greenhouse is likely to be the most applicable in meta-analysis where only n_0 is known; yet it will be simplistic and/or inefficient in some situations. For example, in some instances, results of unpublished studies are available (Peto, 1987), and this may become more common in the future, as clinical trial registries expand (Simes, 1987). One method for incorporating the known unpublished study results is to treat them as follow-ups. Another common technique in the sample survey literature is to assume that length of time to response is related to data values, and use late responders as "follow-ups" for nonrespondents. A similar approach could be used in meta-analysis by characterizing the nature of the search process, which is often quite complex and may involve the use of several computerized databases as well as writing to researchers in the field. Of course one still has the difficulty that n_m is usually known (at least precisely) in meta-analysis, but here again, capture-recapture

Nan Laird is Professor of Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115. G. P. Patil is Professor of Mathematical Statistics and Director, and C. Taillie is Senior Research Associate at the Center for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802. G. P. Patil is at present Visiting Professor of Biostatistics, Harvard School of Public Health.

techniques might be tried as a way of estimating the total number of studies.

We note that Rubin (1977) has suggested the mixture model as an alternative to the selection model for missing data that appears to be somewhat more robust to model misspecification (Glynn, Laird and Rubin, 1986). It is a natural model for the publication bias problem as well. The mixture model posits a two-population problem: the reported studies with mean parameter μ_r and data distribution $f_r(t \mid \mu_r)$ and the unreported studies with parameter μ_u and data distribution $f_u(t \mid \mu_u)$. The parameter of interest is still Θ:

$$\Theta = \pi\mu_r + (1 - \pi)\mu_u,$$

where π is the proportion of studies that are reported. Data will normally only be available for estimating μ_r. Use of a Bayesian approach with subjective priors on π and μ_u allows one to examine the sensitivity of results to assumptions about the magnitude and nature of the selection bias, i.e., π and μ_u. If information about π and/or μ_u is available either from the use of capture-recapture techniques and/or double-sampling schemes, this can be readily incorporated into the mixture model analysis.

With regard to the choice of weight functions, those used by Iyengar and Greenhouse represent a substantial improvement over the Hedges-Olkin scheme that imposes the unrealistic assumption that only statistically significant studies will be published. It is quite encouraging that the two different weight functions give such similar results for inferences about Θ in this example; but it is difficult to tell if this is a special case or general rule.

We suspect that the lack of sensitivity to choice of weight function is probably a result of the fact that both of the weight functions have two special features that may not be completely realistic: 1) symmetry about $t = 0$ and 2) uniform weight in the tails. In fact, the significant T values in Table 4 argue against constant tail weights. Under H_0: $\Theta = 0$, the odds in favor of $2 \le |T| < 3$ over $|T| \ge 3$ are about 16 to 1, whereas only 1 in 4 of the published T values fall in the first range. In fact, intuition suggests that the significant T scores are so extreme (under the hypothesis of constant effects) as to require unreasonably small selection probabilities for small T and unreasonably large numbers of unpublished studies. Some back of the envelope calculations bear this out. In Table 4, put $\Theta = 0$, ignore the degrees of freedom and approximate the t scores by Z scores. Also approximate the selection probabilities by

$$\begin{aligned} \omega(T) &= \omega_1 \quad \text{if} \quad |T| < 2, \\ &= \omega_2 \quad \text{if} \quad 2 \le |T| < 3, \\ &= \omega_3 \quad \text{if} \quad 3 \le |T|, \end{aligned}$$

where $\omega_1, \omega_2, \omega_3$ are constants. If N is the total number of studies, we get

$$\begin{aligned} 3 &\approx N\omega_3(.0027), \\ 1 &\approx N\omega_2(.0428), \\ 6 &\approx N\omega_1(.9545). \end{aligned}$$

The numbers on the left are the observed frequencies for the three intervals, and those on the right are the corresponding normal probabilities. Taking $\omega_3 = 1$, the equations can be solved to yield

$$\begin{aligned} N &= 1{,}111, \\ \omega_2 &= 0.02, \\ \omega_1 &= 0.006. \end{aligned}$$

Even allowing for variability, these estimates strongly suggest that equal tail weights are not realistic.

With regard to symmetry, possible ideological preferences on the part of investigators, editors and meta analysts would argue for allowing a directional component to the publication bias. We wonder if the authors have done any work with nonsymmetric weight functions. In particular, are the estimates numerically stable? Do the standard errors become large? Are the estimated effects parameters and weight parameters strongly correlated?

We were struck by the small correlations (−0.09 and 0.0) in Table 5, which are a statistical manifestation of the claimed insensitivity to the weight function. But we wonder if this is not simply a consequence of the absence of any overall effect. We would expect the correlations to be large if Θ were substantially different from zero and the estimated Θ to exhibit a marked sensitivity to the choice of weight function—especially if the weight function were nonconstant in the tails or asymmetric.

Finally we note that the assumption of constant effect size Θ, although greatly simplifying the analysis, seems suspect in view of the data. One plausible explanation for the T scores in Table 4 is between study heterogeneity, perhaps with some superimposed publication bias. In many meta-analyses, between study heterogeneity is likely to be at least as important as publication bias, and it may be difficult to distinguish the two. We are currently working on models incorporating both heterogeneity and publication bias.

In general, we applaud the use of selection modeling ideas in meta-analysis, which take us far beyond the "fail-safe" sample size approach. Even the modification of the fail-safe sample size formulas suggested by the authors can be improved by incorporating the use of weight functions.

The quantity $1/p$, where p is a p-value, is the waiting time for the spurious occurrence of a significant result, and is a "man in the street" way of assessing the strength of the evidence against a null hypothesis. The fail-safe sample size can likewise be considered as measuring the weight of evidence. Thus, $n(0)$ is the number of *hypothetical* studies, conducted under the null hypothesis, that need to be added to the database to just offset a significant result. Here "offset" is used in the sense that the updated test statistic should have its expected value, *conditional on the published studies*, just equal to the critical point.

The authors' equation (4) modifies the foregoing in two respects: (i) the hypothetical studies are replaced by actual and unpublished studies and (ii) a weight function is made available for modeling the publication bias. Now we have difficulty interpreting the numerator of (4) as a conditional expectation. Granted the values reported in the published studies are noninformative for the unpublished studies, but the number k of published studies, when coupled with the weight function, is informative and should be reflected in the conditional expectation. Thus, does $n(\alpha)$ in equation (4) refer to actual studies (in file drawers) or to hypothetical studies?

As an alternative, we suggest using the weight function to estimate the number of unpublished studies. For example, letting k and k_0 be the number of published and unpublished studies and $N = k + k_0$,

$$k_0 \approx N \int (1 - \omega(x)) f(x)\, dx,$$

$$k \approx N \int \omega(x) f(x)\, dx,$$

and k_0 can be estimated as the solution of

$$\frac{k_0}{k} = \frac{E[1 - \omega(X)]}{E[\omega(x)]}. \tag{3}$$

Note that the righthand side of (3) reduces to a $(1 - \alpha)/\alpha$ in case of a dichotomous weight function. Also, $\omega(x)$ must have a phenomenological interpretation as a probability. It will not do to replace ω by a scalar multiple or to use an unbounded ω.

The solution of (3) is sensitive to the weight function, and one may prefer to look upon the estimate of k_0 as shedding light upon the reasonableness of the chosen ω instead of upon the "true" significance of the published results.

ADDITIONAL REFERENCES

GLYNN, R. J., LAIRD, N. M. and RUBIN, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples* (H. Wainer, ed.). Springer, New York.

HANSEN, M. H. and HURWITZ, W. N. (1946). The problem of nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **41** 517–529.

LITTLE, R. J. A. (1983). The nonignorable case. In *Incomplete Data in Sample Surveys* (W. G. Madow, I. Olkin and D. B. Rubin, eds.). Academic, New York.

PETO, R. (1987). Why do we need systematic overviews of randomized trials? *Statist. Med.* **6** 233–240.

RUBIN, D. B. (1977). Formalizing subjective notions about the effect of nonresponse in sample surveys. *J. Amer. Statist. Assoc.* **72** 538–543.

SIMES, R. J. (1987). Confronting publication bias: A cohort design for meta-analysis. *Statist. Med.* **6** 11–29.

Comment

M. J. Bayarri

"Meta-analysis, like rock and roll, is here to stay" claim the authors of this interesting and stimulating paper, and they are right. Similar experiments are conducted and replicated, providing informaiton about the same unknown quantity, and statisticians have to face the challenge of providing methods for pooling this information. In a sense, the problem is similar to that of combining a set of expert opinions. Unfortunately, results from experiments are not, in general, expressed as distributions of the unknown quantity. If they were, then not only would the publication bias due to statistical significance be greatly reduced, but also the techniques for combining probability distributions would be available (for an excellent summary and comprehensive annotated bibliography about these techniques see Genest and Zidek, 1986). Moreover, meta-analysis is usually based on results that got published in the scientific literature. Due to the overabuse of hypothesis testing as a statistical methodology and to the overappreciation of statistical significance, it does not come as a surprise that publications are highly biased toward studies showing statistically significant results. This publication bias should be taken into account when carrying out a

M. J. Bayarri is Titular Professor at the University of Valencia (Spain). Her mailing address is Departamento de Estadística e IO, Facultad de Matemáticas, Av. Dr. Moliner 50, 46100 Burjasot, Valencia, Spain.