evidence of publication bias, but its impact is probably not especially strong in this meta-analysis, and as a result the apparent overall trend in the data is a small positive effect of passive smoking on lung cancer risk. However, the limitations of data quality, and the apparent weakness of the effect of passive smoking mean that the analysis is far from conclusive, and it is unlikely that additional observational studies could affect this overall conclusion. My feeling is that these are the appropriate conclusions from a relatively simple analysis of these data, comprising a plot of the data as in Figure 1, a funnel-graph as in Figure 2, some rudimentary tests for publication bias and a careful evaluation of the quality of the component studies. Givens, Smith and Tweedie have brought fashionable modern statistical techniques to bear on the issue, with the attendant jargon of Gibbs sampling, burn-in periods, suppression criteria, elaborate prior distributions and all the rest. Does this stuff really add insight to the analysis? I'm afraid my vote is no.

Comment

William DuMouchel and Jeffrey Harris

The paper by Givens, Smith and Tweedie (GST) is a fresh attempt to tackle the "file drawer problem," which at first blush seems insoluble without actually going out and finding some missing studies. For example, the attempt by Iyengar and Greenhouse (1988) seemed to fall short of a solution. The current authors use more sophisticated modeling tools, primarily in their use of a hierarchical random effects model and Gibbs sampling, and perhaps they also have a more fortunate example data set. However, all attempts to assess publication bias beyond simple graphs like the funnel plot seem to involve a *tour de force* of modeling, and as such they are bound to run up against resistance from those who are not statistical modeling wonks. After all, the present analysis is pretty hard to follow, even though the paper is well written, and readers who think they do understand the presentation of the modeling process are likely to be the type who enjoy nit-picking on the details. The following discussion is offered in this latter wonkish spirit.

The random effects model, equation (2) in GST, represents each published study effect as

$$Y_j = \Delta + \beta_j + \varepsilon_j,$$

where the standard deviation of ε_j is σ_j and these standard deviations, usually given as the nominal standard errors presented by the authors of the original studies, play a key role in the detection of publication bias. Some might object that the variance of a study effect involves more than a simple sample size calculation and that, for example, a study that carefully measured exposures and documented lung cancer cases should have a smaller within-study error than a study that did not carefully gauge exposure and relied upon undocumented cancer ascertainment. This raises the question of how and whether measures of study quality can be incorporated into a meta-analysis. If such measures are not available for specific studies, but you suspect that there is a lot of variation in study quality, then the random effect term β_j in the above model provides a handy way to represent such variation. If you desire to incorporate specific information about the quality of particular studies, there are two modeling strategies available. First, you can subjectively inflate the values of σ_j for poor-quality studies. Second, you can incorporate regression terms into the model involving study-level covariates. Both strategies were used in the meta-analysis of biological effects of diesel and related emissions reported in DuMouchel and Harris (1983).

A key assumption made by GST is that the publication selection criterion is based solely on each study's one-sided p-value for rejecting the null hypothesis $\Delta \leq 0$. Why should this be based on the one-sided p-value? Are the authors assuming that studies showing a significant protective effect of ETS would be discriminated against?

William DuMouchel is with AT&T Labs—Research, 600 Mountain Avenue, Room 2C 271, Murray Hill, New Jersey 07974 (e-mail: dumouchel@research.att.com). Jeffrey Harris is Associate Professor, Department of Economics, Massachusetts Institute of Technology, E52-252, Cambridge, Massachusetts 02139 (e-mail: jeffrey@mit.edu).

More generally, we suspect dependence of the selection criterion on more than the p-value. For example, the sample size, cost or power of a study seem natural additional selection criteria. These could be summarized in the value of σ_j. If studies with high values of σ_j are harder to publish, then of course high p-values would also be underrepresented. If the authors change their model so that (7) reads

$$\Pr[\text{a study with standard error } \sigma_j \text{ and a } p\text{-value in } I_k \text{ is published}] = w^k/\sigma_j,$$

their substantive conclusions may be very different.

Section 3.3 of GST, on the definition of the likelihood function, is the technical heart of the paper, and perhaps the hardest section to follow. For example, the authors condition on the numbers of observed studies, $n = \{n^k\}$ in each p-value interval, whereas normally one imagines that the n^k are a function of the Y_k. Maybe it's all right, but there is the appearance of circularity in (11), in that X depends on m, m depends on n and n depends on X. We have a similar difficulty understanding the role of the normalizing functions $A(\cdot)$ in (12), (13), (14) and (16). How exactly are they defined?

In the discussion of the Gibbs sampling steps in Section 3.4, it is stated that the missing p-values are drawn uniformly on the intervals I_k. But is not a uniform distribution for the p_j only appropriate if $\Delta = 0$? Could this be producing a bias in the Gibbs sampling in favor of the null hypothesis?

Considering the simulation experiments, the authors assume that the prior distributions for the selection weights in Section 3.5(b), with no suppression, are uniform on $[0.5, 1]$ for all of the p-value intervals. Yet in the analysis of the ETS data, stronger priors were used. It would be nice to have a comparison assuming identical priors.

Finally, the authors refer to the report of Bero, Glantz and Rennie (1994), who found five unpublished negative studies not cited in the EPA Report (EPA, 1992). What were the values of the σ_j for these new studies? We guess that they are larger than those for most of the first-reported studies.

To summarize our discussion, in spite of what may seem like critical comments we do assume that publication bias is a real phenomenon and that the paper under discussion is a nice contribution to the methodology of detecting and correcting for such bias. Our most serious concern is with the form of the assumed publication bias criterion, and we would like to see whether adding a factor for dependence on the σ_j, as we suggested above, would modify the results of the ETS analysis.

Comment

Annette Dobson and Keith Dear

The culture of meta-analysis has traditionally favored very simple methods, such as weighted averages and the one-step random effects method of Der Simonian and Laird. The same is true of early approaches to publication bias, such as the file drawer of null studies conceived by Rosenthal. Now that meta-analysis is taking a high-profile role in public policy-making and regulatory affairs, it is entirely appropriate that more sophisticated techniques, such as those proposed by Givens, Smith and Tweedie, be developed. In these comments we will concentrate on the methodology, and not on the specific results about the relationship between ETS and lung cancer.

The choice of prior is always an issue in Bayesian analysis and seems to us to be critical here. Three simulations are provided in Section 3.5. In the first simulation [Section 3.5(a)] mild suppression is applied and the prior on the $\mathbf{w^k}$ reflects this by preferring lower probability of publication if the p-value is greater than 0.5 than if it is less than 0.5. This is described as "not reflecting strong beliefs about the amount of publication bias present"; however, it does embody the belief that there is some. In Section 3.5(b) no suppression is applied, and the prior reflects this by having equal priors on all three regions of the p-value scale. Finally, in Section 3.5(c) strong suppression is applied, and publication bias this time is forced into the model by the use of a $U(0.2, 0.7)$ prior for $0.1 < p < 1$. In

Annette Dobson is Professor, and Keith Dear is Senior Lecturer, Department of Statistics, University of Newcastle, NSW 2308, Australia 61-249-215544 (e-mail: stajd@cc.newcastle.edu.au and dear@mail.newcastle.edu.au).