

Assessing Sensitivity to Data Problems in Epidemiological Meta-analyses

R.L. Tweedie

Department of Statistics, Colorado State University,
Fort Collins CO 80523, USA

1. Introduction

In an ideal world, meta-analysis in epidemiology would be simple. We would identify a possible relationship between an "outcome" (usually a disease) and "exposure" to a possible cause of the disease; collect all of the studies on the subject; combine their results; and this would give us two invaluable outcomes: an overall measure of the relationship based on much more information than in any one study, and a statistical assessment of the significance of the relationship taking into account all the studies.

It is easy to see what might be wrong with this idealised notion, and with the huge increase in the use of meta-analysis in epidemiology (as well as other areas), especially in areas where individual studies are equivocal, there has come a large number of books and discussion papers which assess the benefits, drawbacks and problems of these techniques, such as [1, 2, 3, 4, 5, 6, 7, 8].

In this paper we restrict comment to the following specific issues, although there are many more to which similar approaches will apply:

- (i) the problem of comparability of data and study design, since for the meta-analysis to be meaningfully interpreted, we must not combine "apples and oranges";
- (ii) the effect of "publication bias", recognising that failure to obtain all relevant studies, both published and unpublished, may result in a quite distorted meta-analysis;
- (iii) the possible existence of systematic errors in individual studies, since these flow to bias in the overall analysis, and some account must be taken of them.

Clearly all of these (and many other problems) are of concern in principle in any meta-analysis, but it is not obvious whether they will cause real problems in any one specific application.

Our theme in this paper is that, by using sensitivity analyses based on the collection of real data and comparison of more and less sophisticated models, one can develop useful information on the extent and effect of such problems, rather than merely expressing concern about their existence.

We shall see that for each of (i)–(iii), we can develop approaches based on data which describe the best and worst case scenarios (and the intermediate ones also), and use these to quantify whether further steps must be taken to protect the meta-analysis from incorrect inferences.

Although other measures of relationship are possible, in epidemiology it is common to consider the effect of the relationship to be measured by the relative risk (RR), given conceptually as the ratio

$$RR = \frac{\text{Pr}[\text{disease given exposure}]}{\text{Pr}[\text{disease given no exposure}]},$$

and we shall assess whether a given problem is real rather than theoretical by considering the effect on the combined RR through use of a “what-if” approach as we vary models and parameters.

We illustrate these issues through a brief assessment of the overall association between incidence of lung cancer in female never-smokers and exposure to passive smoking, or environmental tobacco smoke (ETS), from spousal smoking. There have been many analyses of the studies of this association (see [9, 8, 10, 11] and others). Space precludes us from giving details of the data-sets involved, which can be found in these references. For the relationship of lung cancer and spousal ETS, the RR , unadjusted for any problems such as we discuss, is around 1.10-1.30 depending on the approach taken [9, 10].

The ETS studies seem appropriate to illustrate the problems in meta-analysis for several reasons. The association between lung cancer and ETS is an issue of public and legal concern, so there is a real benefit to proper evaluation; the individual studies are equivocal, and the results of the meta-analyses are also contentious; and the ETS studies exhibit the various problems above, so that we are able to quantify the effects of these problems in some detail in this example.

2. Combining heterogeneous data

Meta-analysis is designed to enable combination of results from studies which are *comparable in outcome and exposure*. The interpretation of comparability is a subjective and often difficult one. In order to paint an honest picture of the aims and applicability of any meta-analysis, we must first define the relevant measures of outcome and exposure with which we are concerned.

This can be non-trivial. In the ETS example, the clinical *outcome* assessed in the studies is death from “lung cancer”. However, several studies concentrate on one specific form of this disease (e.g. adenocarcinoma), and although some studies give data for different types of cancer, many others do not make such distinctions. Thus we must choose whether to combine RR estimates for all lung cancer types, or not: if we combine them, then we do so knowing that the overall RR estimate may be based on individual RR 's associated with several different diseases; if we do not, we potentially lose the power that meta-analysis would provide if the combination were valid.

The same problem occurs with ensuring that we are combining results of studies with comparable *exposures*: in most of the literature we find the meta-analysis is primarily restricted to the subset of all studies of adults asserted to be never-smokers, with exposure to ETS coming from living with a smoking spouse; but others consider exposure from workplace smoking or other sources, and we must decide whether these should be meta-analysed separately or together.

In choosing which studies to combine, we may also need to consider the plausibility of combining results from studies in more obviously different strata, such as those with different gender mixes or from different geographic areas.

There are two different ways we suggest for handling such heterogeneity. The first is by using models that specifically allow for such an effect. It is becoming standard to combine estimates via a “random effects” model [7], which attempts to allow for inter-study variation, such as those due to rather subtle covariates such as type of disease above. This can be argued to be preferable to an earlier “fixed effects” model which essentially assumes statistical homogeneity between studies.

Formally, we are interested in estimating the overall effect $\mu = \log RR$ where we consider

for the i^{th} study the model

$$\begin{aligned} Y_i &= \theta_i + e_i \\ \theta_i &= \mu + \varepsilon_i; \end{aligned}$$

here Y_i is the observed log RR_i for each study, θ_i is the corresponding true log RR_i , and it is assumed that e_i are independent $N(0, \sigma_i^2)$ random variables, that the ε_i are independent $N(0, \tau^2)$ random variables, and that the e_i and ε_i are mutually independent. The fixed effects model takes $\tau^2 = 0$; by allowing $\tau^2 > 0$, the random effects model enables us to capture some of the inhomogeneity since it assumes different studies have values θ_i which may differ from μ .

The random effects model can be analysed both in a frequentist or a Bayesian context [7], and extends logically to hierarchical models [12]. To allow for the types of inhomogeneity we have described, we suggest that Bayesian methods might best be used. In this context the priors on the various parameters can perhaps be thought of, not as describing ‘‘prior information’’ in any strong sense, but rather as describing in more detail the way in which the studies might be heterogeneous. However, if we do have real prior information on the range of input parameters this can clearly be captured by the Bayesian approach.

Using these models as sensitivity tools, one way in which we can see if heterogeneity is of practical importance is to consider the results from all three approaches. As shown in Table 1 (taken from [10, 13]), in our ETS example we see that (before allowing for publication and other biases) the RR values for fixed effects, random effects and Bayesian models vary from 1.19 to 1.22.

Table 1: Relative Risks and Confidence/Credibility Intervals for Lung Cancer associated with Spousal Smoking

| Method | RR | 95% CI |
|--|------|------------|
| Fixed Effects Model | 1.19 | 1.02, 1.38 |
| Random Effects Model | 1.20 | 1.07, 1.34 |
| Bayesian Hierarchical Model | 1.22 | 1.08, 1.37 |
| Asian Studies (Bayes) | 1.25 | 1.03, 1.50 |
| US Studies (Bayes) | 1.13 | 0.95, 1.34 |
| Bayesian after Publication Bias Correction | 1.14 | 1.00, 1.28 |

Thus, although there is evidence of heterogeneity, it is not of major practical importance: perhaps 10% of the excess risk might be missed if the random effects models are not used.

The other (very simple) method which we advocate is to give results separately for different subsets of the data where strata are easily identified. The geographic question seems appropriately studied through such a sensitivity analysis. As seen in Table 1, for example, Asian studies seem associated with higher RR than US studies. The difference is enough, in practical terms, that in its final report [9] the EPA chose to use the US result rather than the global one in its evaluations of effects in the US: and here, based on this ‘‘sensitivity’’ evaluation, we might well conclude that there is a genuine practical problem of heterogeneity across geographic regions and the results should not be combined.

Similarly, the effects of different types of exposure (to spousal ETS or to workplace ETS, say), or for males and females, seem best handled by giving results for different exposure or gender strata separately, and then considering if they are effectively identical [10].

3. Publication bias

It is important in principle in meta-analysis to attempt to collect all published and unpublished studies relevant to the relationship in question [13]. The problem here is that unpublished studies, by their nature, are likely to differ from published studies: in particular, they are likely to be less significant and hence their omission from a meta-analysis will bias the combined result away from the null value.

Missing studies due to publication bias are not easy to work with. Unlike traditional missing data problems, there is an unknown number of them. Their effect could be huge, or it could be minute, and developing a sensitivity analysis that accounts for them is not trivial.

In [13] we approach this problem by extending the Bayesian hierarchical model to allow for missing studies. We assume studies are more likely to be missing if they have less statistical significance, since journals publish significant results and refuse insignificant results, and researchers differentially write up significant results. This is especially the case when a relationship might be one of many being examined in a cohort study: only those relationships achieving "significance" make it to press, so that there might be a considerable number of hidden instances of publication bias in any one such study.

By using this more flexible model, and then basing the priors applied to μ, τ^2 and the probabilities of studies being missing on data sources that seem relevant, we can evaluate how sensitive to such missing publications the *RR* might be.

Spousal ETS studies give an example of such bias. The funnel plot in Figure 1 of [8] shows a clear indication of the absence of small studies with negative (perhaps nonsignificant) estimates of effect, with perhaps 6-10 or so small but negative studies expected but not present.

Such graphical indications are all very well: but what difference do the missing studies make? In [13] we show that it is a small but important difference: using the Bayesian methods allowing for publication bias, we find as in Table 1 a posterior mean relative risk of 1.14 compared with the Bayesian posterior values of 1.22 ignoring publication bias.

Thus we are able to quantify the possible problem: about a third of the calculated excess risk may be due to publication bias in this case. This indicates that the observed excess is not totally an artefact of publication bias, but nor is it devoid of such bias.

4. Systematic Biases

It cannot be stressed too often that a meta-analysis is only as good as the studies that go into it. If the studies are flawed then to different extents so is the meta-analysis.

This is particularly true if there is a *systematic bias* in the underlying studies. Such biases can occur in, for example, studies that are not appropriately blinded at the point of data collection, since the problems from lack of blinding are likely to cause the *RR* to increase or decrease in the same direction in all studies. Such systematic bias must be accounted for in a meta-analysis as well as in the analysis of a single study.

One particularly striking (and perhaps surprising) example of this occurs in our ETS example and is discussed in detail in [14]. In studies of lung cancer in female non-smokers with ETS exposure due to spouses, misclassification of smokers as non-smokers spuriously and systematically elevates the observed relative risk. This occurs because, if the female is really a smoker, she has a high risk for lung cancer; she is more likely to have married a smoker than is a true non-smoking female, and thus she will be differentially in the "spousally exposed" group; and this increases the seeming *RR* for the ETS group differentially.

We show in [14] that data-based sensitivity analysis is a powerful tool in these circumstances. The extent of the "misclassification bias" depends on a number of estimated parameters: the rate of misclassification, the true RR for misclassified smokers, and above all the differential marriage rates. None of these is known accurately: our approach in [14] is to vary all of them over ranges *supported by known data*. The RR varies accordingly, and in [14] we found that the spurious excess risk might be as high as 0.2-0.5 from this one source alone if it occurs in every study.

Of course we cannot assume that such systematic bias occurs in every study. Indeed, some of the studies which give data on the sources of error (such as the rates of misclassification) are precisely those in which misclassified smokers have been identified, presumably reducing the problem to a very much lower if not non-existent level.

Such issues can only be resolved by using more detailed data than is often easily available, and for example getting well-founded data on the rate at which smokers marry other smokers can be very difficult. The point of identifying the extent of the problem through sensitivity analysis is that it pinpoints areas where such data collection might be valuable, and in [14] we stress the need to collect relevant data to support the ranges rather than taking totally hypothetical approaches to each one.

5. Conclusions

Meta-analysis is often used simply to increase statistical power: that is, in effect to narrow the confidence limits around an estimate of effect, even if results are fairly consistent and clearcut in each study. It can be used to greater advantage, however, in situations where individual outcomes are difficult to interpret, or when relative risks are small or not significant in each study alone. It is important to realise that the impact of selection of studies to be combined, and evaluation of bias, can be substantial, as we have seen in the ETS example.

It is of course established wisdom in carrying out single analyses that a small estimated RR should be treated with some caution. However, most of the guidance on what constitutes a "small estimated RR " appears to rely on folklore or accumulated experience. For example, Mantel [15] advocates that values of RR below 2.0 should not be regarded as established, and in studying the association between street oil consumption and toxic syndrome, Sir Richard Doll [16] asserts "no specific limit can be set to the size of the RR that excludes confounding, but past experience suggests that confounding is seldom likely to be the explanation if the lower 95% confidence limit of the estimated RR is greater than 3".

Sensitivity analysis can help to confirm whether, with known potential problems of the type discussed in this paper, levels of at least 2.0 or 3.0 are indeed necessary to give some confidence that the association observed is in fact a true one and not an artefact of the specific identified data problems. It should be a routine step in carrying out any meta-analysis, and it would strengthen the inferences considerably in many situations if the effects of identified problems could be quantified as the ones above have been.

References

- [1] L.V. Hedges and I. Olkin. *Statistical Methods for Meta-analysis*. Academic Press, 1985.
- [2] H. Cooper and L.V. Hedges eds. *The Handbook of Research Synthesis*. Russell Sage Foundation, New York, N.Y., 1994.
- [3] T.C. Chalmers. Problems induced by meta-analysis. *Statistics in Medicine*, 10:971–980, 1991.
- [4] S.G. Thompson and S.J. Pocock. Can meta-analyses be trusted? *Lancet*, 338:1127–1130, 1991.
- [5] Frederick Mosteller and Thomas Chalmers. Some progress and problems in meta-analysis of clinical trials. *Statistical Science*, 7:227–236, 1992.
- [6] D.T. Felson. Bias in meta-analytic research. *J. Clin. Epidemiol.*, 45:885–892, 1992.
- [7] NRC Committee on Applied and Theoretical Statistics. *Combining Information: Statistical Issues and Opportunities for Research*. National Academy Press, Washington, 1992.
- [8] K.L. Mengersen, R.L. Tweedie, and B.J. Biggerstaff. The impact of method choice in meta-analysis. *Australian J. Statistics*, 37:19–44, 1995.
- [9] EPA Review. *Health Effects of Passive Smoking: Assessment of Lung Cancer in Adults and Respiratory Disorders in Children*. National Academy Press, U.S. EPA, Washington, 1992.
- [10] R.L. Tweedie, D.S. Scott, B.J. Biggerstaff, and K.L. Mengersen. Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer*, 14:S171–S194, 1996. Suppl. 1.
- [11] P.N. Lee. *Environmental Tobacco Smoke and Mortality*. Karger, Basel, 1992.
- [12] William DuMouchel. Bayesian meta-analysis. In *Statistical Methods for Pharmacology*, pages 509–529. Marcel Dekker, New York, 1990. ed. D. Berry.
- [13] G.H. Givens, D.D. Smith, and R.L. Tweedie. Publication bias in meta-analysis: a Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science*, 1997. (Accepted for Publication).
- [14] R.L. Tweedie, K.L. Mengersen, and J.A. Eccleston. Garbage in, garbage out: Can statisticians quantify the effects of poor data? *Chance*, 7(2):20–27, 1994.
- [15] N. Mantel. What is the epidemiological evidence for a passive smoking-lung cancer association? In *Indoor Air Quality*, pages 341–347. Springer-Verlag, Berlin, 1990. ed. H. Kasuga.
- [16] R. Doll. Occupational cancer: a hazard for epidemiologists. *Int J. Epid.*, 14:22–31, 1985.

SUMMARY

It is well-known that meta-analyses are subject both to problems common to single analyses (such as systematic biases, inaccuracies in question definitions, etc) and to problems specific to combining analyses (such as homogeneity of studies combined, publication bias and so on). We review some of these problems, suggest some ways to formally assess publication bias in particular, and argue for the systematic use of data-based analyses in assessing sensitivity to such potential problems. Studies on health effects of environmental tobacco smoke are used to illustrate the points made.

Il est bien connu que les méta-analyses sont sujettes à la fois aux problèmes d'analyses simples (bias systématique, imprécision dans les questions, etc.) et aux problèmes relevant de la combinaison d'analyses (homogénéité des études combinées, bias de publication, etc.). Nous proposons une revue de ces problèmes, proposant en particulier une méthode d'évaluation du bias de publication, et défendons la thèse que les analyses fondées sur les données soient utilisées systématiquement pour évaluer la sensibilité à ces problèmes. Des études sur l'effet de la tabagie passive sur la santé sont utilisées pour illustrer cette thèse.