J Clin Epidemiol Vol. 44, No. 2, pp. 127–139, 1991
Printed in Great Britain

0895-4356/91 $3.00 + 0.00
Pergamon Press plc

META-ANALYSIS IN EPIDEMIOLOGY, WITH SPECIAL REFERENCE TO STUDIES OF THE ASSOCIATION BETWEEN EXPOSURE TO ENVIRONMENTAL TOBACCO SMOKE AND LUNG CANCER: A CRITIQUE

JOSEPH L. FLEISS[1] and ALAN J. GROSS[2]

[1]Columbia University, School of Public Health, 600 West 168 Street, New York, NY 10032 and
[2]Medical University of South Carolina, Charleston, SC 29425, U.S.A.

(Received in revised form 29 August 1990)

Abstract—Meta-analysis, a set of statistical tools for combining and integrating the results of independent studies of a given scientific issue, can be useful when the stringent conditions under which such integration is valid are met. In this report we point out the difficulties in obtaining sound meta-analyses of either controlled clinical trials or epidemiological studies. We demonstrate that hastily or improperly designed meta-analyses can lead to results that may not be scientifically valid. We note that much care is typically taken when meta-analysis is applied to the results of clinical trials. The Food and Drug Administration, for example, requires strict adherence to the principles we discuss in this paper before it allows a drug's sponsor to use a meta-analysis of separate clinical studies in support of a New Drug Application.

Such care does not always carry over to epidemiological studies, as demonstrated by the 1986 report of the National Research Council concerning the purported association between exposure to environmental tobacco smoke and the risk of lung cancer. On the basis of a meta-analysis of 13 studies, 10 of which were retrospective and the remaining 3 prospective in nature, the Council concluded that non-smokers who are exposed to environmental tobacco smoke are at greater risk of acquiring lung cancer than non-smokers not so exposed. In our opinion, this conclusion in unwarranted given the poor quality of the studies on which it is based.

1. INTRODUCTION

A working definition of meta-analysis is given by Huque [1]: "... the term 'meta-analysis' refers to a statistical analysis which combines or integrates the results of several independent clinical trials, considered by the analyst to be 'combinable'." As indicated by this characterization of meta-analysis, its key application is to be found in the analysis and synthesis of data from clinical trials.

The question then remains, can meta-analytic techniques be applied in the analysis of other kinds of data such as those that arise in cohort and case–control studies found in epidemiology? The answer to this question is a guarded "yes." The criteria for reaching this affirmative answer are now considered.

In applications of meta-analysis to clinical trials, the following questions, among many other that must be addressed, arise.

- Are all studies to be included in the meta-analysis, or only the published ones?
- Are all published studies to be included in the meta-analysis, or only the "good" ones?
- When the studies' results are heterogeneous, how may they be included in a meta-analysis, or should they be meta-analyzed at all?
- Within each study, should all subjects in a treatment group be considered in a

meta-analysis, or only those subjects who were compliant with the treatment? The same question applies to subjects in the control group.

Similar issues are of concern in epidemiological studies. However, when case–control studies are under consideration, the issue of "intention-to-treat," which is the final question just listed with regard to meta-analysis in clinical trials, is not of direct concern. Instead, the following additional question needs to be addressed in a meta-analysis of case–control epidemiological studies.

- Has proper control or adjustment been made for the biases that frequently occur in epidemiological studies, such as sociodemographic or clinical differences between study populations, misclassification of subjects with regard to case–control status and to levels of exposure, factors other than the level of exposure that may affect whether a subject is a case or control (i.e. confounding variables), and the publication bias/file drawer phenomenon wherein studies that fail to show a positive association tend not to be published and are thus not candidates for inclusion in the meta-analysis?

Meta-analysis was first applied to the study of psychotherapy and to the study of educational interventions (see Hedges and Olkin [2], for example), and is now widely used to provide overviews of randomized controlled clinical trials. It is also applied in the synthesis of data from epidemiological case–control studies, but, as will be covered in Section III, with uncertain theoretical justification.

Among the principal uses of a properly performed meta-analysis are:

- To increase statistical power for important endpoints and subgroups.
- To resolve controversy when studies disagree.
- To improve estimates of effect size.
- To answer new questions that were not previously posed in the individual studies.

How meta-analysis is applied in both randomized clinical trials and epidemiological case–control studies are the topics of Sections II and III of this paper. Special emphasis will be placed on epidemiological studies of the hypothesized association between exposure to environmental tobacco smoke and lung cancer.

II. APPLICATIONS OF META-ANALYSIS TO CLINICAL TRIALS

Before reviewing and criticizing the application of meta-analysis to epidemiological studies, it is worthwhile to review and critique its application to a methodologically stronger kind of study, the randomized controlled clinical trial. We shall identify a number of areas of uncertainty and controversy concerning such applications, present the points on which consensus seems to exist, and then use the results of this review as a template for our critique of meta-analyses of studies in epidemiology.

Analyze all published studies or only the "good" ones?

In their review of published meta-analyses, Sacks *et al.* [3] found that nearly 30% of them combined results from both randomized and non-randomized studies. If there is unanimity among meta-analytic methodologists on any issue, however, it is on the requirement that only randomized clinical trials be included in a meta-analysis [4–8]. These experts take it as axiomatic that the potential for bias in the non-randomized assignment of patients to treatment groups is too great for the results of such studies to be trusted. There exist examples of non-randomized studies that are, in other respects, superior in quality to randomized studies [9], but the concern about the quality of non-randomized studies in general is a valid one. The principle that meta-analyses be restricted to randomized studies is by-and-large appropriate.

Having agreed on the criterion of non-randomized treatment assignment for excluding a study from a meta-analysis, the experts disagree on other possible criteria for excluding studies (absence of double-blinding, efficacy rather than intention-to-treat analysis, study is out of date, etc.), and indeed disagree on whether any randomized trial, no matter how poorly designed, should ever be excluded [7,8,10]. Hedges, for example, suggests that it is standard procedure in meta-analyses in the physical sciences to delete experiments that are deemed to be flawed [11]. H. J. Eysenck, a British psychologist and philosopher of science, labels as "mega-silliness" the practice of including methodologically inadequate research in a meta-analysis [12].

Chalmers and his colleagues have developed rigorous, reproducible and unbiased methods for measuring the quality of a study [13]. One

may use the derived measurements to decide whether to accept or reject the study for a meta-analysis, to determine in a formal [14] or informal [15] way whether a study's quality and its estimate of treatment effect are correlated, or to weight studies differentially according to their measured qualities (such a suggestion was recently made by Jenicek [16]). One way to carry out this latter strategy is to modify the meaning of the weight to be assigned to a study's estimated treatment effect, e, in the weighted average

$$\bar{e} = \Sigma we/\Sigma w.$$

In most applications of meta-analysis, w is roughly proportional to the total number of patients in the study. If q is the study's value on the measure of quality, perhaps scaled to vary from 0 to 1, the proposed overall measure of the effect of treatment is

$$\bar{e} = \Sigma (qw)e/\Sigma (qw).$$

A large study whose quality is good will receive all or nearly all of the weight it is entitled to, whereas a study of comparable size but measurably poorer quality will have its contribution to the weighted average correspondingly reduced.

We are not aware of any meta-analyses in which the measures of quality have actually been formally incorporated into the analysis. The proposed method is novel, and worth a critical evaluation.

Analyze all studies or only the published ones?

When planning a meta-analysis, the investigator must decide whether to analyze only those studies that had been published or to analyze as well studies that had not been published. Apparently, no consensus has yet been reached concerning which strategy to adopt. Yusuf [8] recommends that the results of both published and unpublished controlled trials be considered for analysis; he and his colleagues were successful in locating several unpublished studies in an influential meta-analysis of the effects of beta blockers after myocardial infarction [17]. Abstracts of papers presented at meetings and abstracts of master's theses and doctoral dissertations were perused in the pioneering meta-analyses of studies evaluating the efficacy of psychotherapy [18]. Chalmers *et al.* [4] have warned, however, that even a systematic and rigorous attempt at obtaining the results of all unpublished studies may produce bias or unduly decrease precision.

There are two main reasons for including unpublished studies in a meta-analysis. One is to overcome "publication bias," the acknowledged tendency of reviewers to recommend against and of editors to decide against publishing studies that failed to show an effect of treatment. The other is to overcome the bias due to the related "file drawer phenomenon," the tendency on the part of the author not even to bother submitting for publication an article that fails to show an effect [19]. If either of these two sources of bias operates, then a meta-analysis only of results reported in published articles will tend to overstate the degree of statistical significance of the treatment's effect, and to overestimate that effect. An alternative to the difficult task of searching for unpublished studies is to attempt to undo this bias by applying statistical adjustments to the data [20,21]. These statistical procedures are still too new for them to have been theoretically and empirically evaluated. Thus, effectively, the publication bias/file drawer issue remains a serious problem in performing a meta-analysis. The careful and thorough search for unpublished studies is an expensive and time consuming endeavor, and may uncover studies of uncertain quality, but no validated alternative is currently available.

Analyze only homogeneous studies?

Some statistical reviewers at the U.S. Food and Drug Administration have strongly criticized the pooling of results from controlled clinical trials in which there is *heterogeneity of treatment effect*— i.e. sizable differences exist between studies in their estimates of the effect of treatment—and have suggested that it is valid to combine results only from studies in which the estimates are sufficiently close one to another [22,23]. Stein, in fact, denigrated as a mere "computational exercise" the meta-analysis of studies in which the estimated treatment effects were heterogeneous [23]. Sacks *et al.* refer to this criterion as *combinability* [3].

Fairly straightforward statistical methods exist to test the hypothesis of heterogeneity for both continuous measurements [24] and categorical data [25] (see the statistical Appendix), although not all FDA reviewers are in agreement as to how strict the statistical criteria should be for deciding that different studies are combinable [1]. Furthermore, it is not clear that these reviewers would accept as evidence for efficacy the finding of a statistically significant

pooled effect even if the meta-analysis was restricted to studies that were combinable.

Some justification may be granted to the tough stand taken by these FDA reviewers, as they are responsible for interpreting and applying regulations handed down to them. In settings other than regulatory ones, however, it is not obvious that the criterion of combinability must always be satisfied before a meta-analysis may be applied. DeMets, for example, questions the meaning that attaches to the overall results of a meta-analysis when there is heterogeneity across studies [26]. Others, however, suggest that it is precisely when studies differ with respect to the magnitude and perhaps even the direction of treatment effect that the formal methods of meta-analysis are needed to summarize in an unbiased manner all of the information available to date [6,27]. With respect to the possibility that the effect of a treatment is strongly positive in one study and strongly negative in another, Peto states that "(this) situation... would be unusual, although certainly not impossible" [6, p. 233]. The frequency with which such a *qualitative interaction* occurs may be greater than he and others (including the two authors of this paper) have believed. A recent randomized controlled trial of the post-infarction effect of a calcium channel blocker, for example, found this very kind of interaction [28].

Whether and how to carry out a meta-analysis in the presence of heterogeneous effects are still unanswered questions. There appears to be only one point on which there is agreement: it is invalid to delete from the set of studies to be meta-analyzed those whose results are in the "wrong direction," for the opportunity for bias in identifying the "deviant" studies is too great. Furthermore, one would be left drawing the inane conclusion that "in those studies in which the treatment effects were in the same direction (all positive, all negative, or all close to zero), the overall effect of treatment was also in that direction."

Intention-to-treat vs efficacy analysis

The controversy that exists concerning the appropriate samples of patients to be analyzed within a single trial carries over into the realm of meta-analysis. According to the intention-to-treat principle, patients are to be analyzed within the treatment groups they were randomly assigned to, no matter how much or how little treatment they actually received [29].

In an efficacy analysis, on the other hand, only data from compliant patients are analyzed [30]. Sack *et al.* found, in their review, that 9 of 19 meta-analyses that considered this issue restricted their analyses to studies that employed the intention-to-treat principle, and 9 analyzed data from either kind of study [3]. Only one meta-analysis restricted attention to studies in which efficacy analyses were performed. Because efficacy analyses tend to produce overestimates of a treatment's effect, and intention-to-treat analyses tend to produce underestimates, caution suggests that, when sufficient information is provided to ascertain which approach was used, only studies that analyzed data from the more conservative intention-to-treat perspective be included in a meta-analysis. When studies that performed only efficacy analyses are included in a meta-analysis, one may expect that some degree of bias toward greater significance and toward an overestimation of the effect of treatment is present.

Data presentation

Most theorists and many practitioners of meta-analysis agree that a graphic display of the individual studies' results and of the overall, pooled result is invaluable. Most often, each study's estimated treatment effect (expressed as a relative risk when the outcome is morbidity or mortality) is marked by a circle or tick mark, and 95 or 99% confidence limits about the estimate are displayed as straight lines extending to the left and to the right of the point estimate (see Fig. 1). The several studies' lines appear one above the other, and the last line indicates the value of the summary estimate pooled across all individual studies, along with its confidence

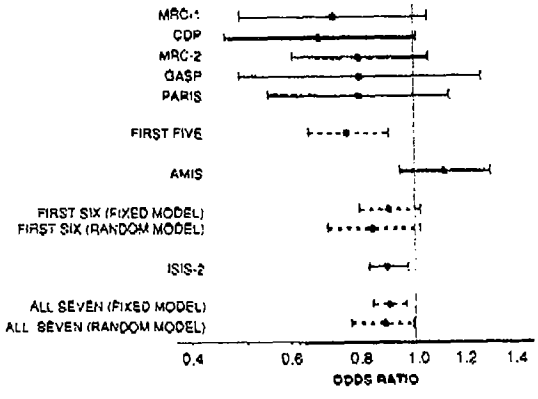

Fig. 1. Odds ratios and 95% confidence intervals for seven randomized trials of the effectiveness of aspirin after myocardial infarction.

limits. Important examples from overviews of clinical trials of the effectiveness of beta-blockers in reducing the risk of mortality after myocardial infarction have been provided by Baber and Lewis [31] and by Yusuf *et al.* [17]. Studies may be separated by the time during which they were carried out (early or late during the development of a treatment [31]), or by other characteristics of the trials.

Peto has proposed an alternate method for the graphical presentation of the individual studies' results [32], one that is less intuitively understandable but that lends itself more directly to sophisticated statistical analysis. For each study, the values of two statistics are calculated. One is the difference, $O-E$, between the observed number of treated patients who experienced the outcome event, O, and the number expected to have done so under the hypothesis that the treatment is no different from the control, E. The second statistic, V, is the variance of the difference $O-E$.

With the ordinate of a pair of ordinary rectangular axes representing $O-E$ and the abscissa representing V, each study is represented by a point. If a straight line passing through or near the origin provides an adequate fit to these points, the meta-analyst may conclude that the studies' odds ratios are approximately equal, with their average value estimable, roughly, as the antilogarithm of the slope of that line. A more precise estimator of the summary odds ratio is

$$\mathrm{OR} = \exp\left(\frac{\Sigma(O-E)}{\Sigma V}\right),$$

the antilogarithm of the ratio of the sum of all the $O-E$ values to the sum of all the V values. (When OR is outside of the interval from 0.2 to 5.0, this formula may be inaccurate [33], and should be replaced by the more accurate formulas due to Mantel and Haenszel [34] or by those presented in the Appendix). The statistical significance of the estimated odds ratio may be tested by referring the quantity,

$$Z = \frac{\Sigma(O-E)}{\sqrt{\Sigma V}}$$

to the standard normal distribution and declaring the summary odds ratio to be statistically significant if Z is sufficiently large. The lower and upper limits of an approximate 95% confidence interval for the population odds ratio, finally, are given by

$$\exp\left(\frac{\Sigma(O-E) \pm 1.96\sqrt{\Sigma V}}{\Sigma V}\right).$$

The reader should note that the confidence interval so obtained is not symmetric about the point estimate.

Studies fixed vs studies random

A debate that is far from being resolved concerns the issue of how, if at all, interstudy differences in the magnitudes of the estimated treatment effects are to be taken into account in the meta-analysis. With only a limited number of exceptions [35], virtually all meta-analyses have ignored differences in estimated effects between studies (except to describe them qualitatively), and have used in the analysis only within-study measures of precision. Thus, as an example, if in one meta-analysis there are two published studies with ORs of 1.0 and 6.0, if in another there are two published studies with ORs of 2.0 and 3.0, and if all four values of V (the variance of the logarithm of the OR) are equal to 0.01, then in both studies the value of the pooled OR will be 2.45 and in both studies the approximate 95% confidence intervals extend from 2.13 to 2.81. No cognizance is taken of the obvious fact that the two studies in the first meta-analysis are much further apart than the two studies in the second. If each study were so large that the sampling variances were all equal to zero, both confidence intervals would degenerate to the single value of 2.45. The conclusion in both cases would be that the odds ratio was known with certainty to equal 2.45, although this cannot possibly be correct given that the individual ORs do not equal one another.

Peto, on the one hand, nevertheless asserts that this is precisely as things should be [36], whereas Meier, on the other, argues that interstudy variation is a key feature of the data and should contribute to the analysis [37]. In the jargon of the analysis of variance, Peto's perspective is that the studies represent levels of a *fixed factor* whereas Meier's is that they represent levels of a *random factor*. DeMets [26] and Bailey [38] discuss the pros and cons of these two competing statistical models, with Bailey presenting those circumstances, assumptions and research questions under which one or the other perspective is the more appropriate.

Bailey [38] suggests that, when the research question concerns whether the treatment *will* have an effect, on the average, or whether exposure to a hypothesized risk factor *will* cause disease, on the average, then the model of studies being random is the appropriate one. When the question concerns whether treatment *has* produced an effect, on the average, or whether exposure *has* caused disease, on the average, *in the studies at hand*, then the model of studies being fixed is the appropriate one. The former question implicitly assumes that there is a population of studies from which those included in the meta-analysis were sampled. It anticipates the possibility of future studies being conducted, or even previously unknown studies being uncovered. The latter question assumes that only the studies included in the meta-analysis are of interest, and that there is no interest in generalizing the result to other studies. The former question, in our opinion, is usually the more important of the two.

In the first of the two hypothetical meta-analyses, the random effects model yields an approximate 95% confidence interval extending from below 0.5 to above 10.0. In the second, it yields an approximate 95% confidence interval extending from 1.65 to 3.64. These intervals were constructed using the method of Der Simonian and Laird [39] (see the Appendix). In our opinion, the difference between the two intervals based on the random effects model accurately reflects the difference that exists between the two pairs of studies, whereas the equality of the more traditional intervals based on the fixed effects model does not.

The potential for fragility in meta-analysis

It is possible for a single study to exert a powerful influence on the results of a meta-analysis. A striking example is provided by the meta-analysis of randomized trials of the effectiveness of aspirin in preventing death after a myocardial infarction [38]. Summary results are presented in Table 1 and in Fig. 1 for seven trials carried out between 1976 and 1988. The first five [40–44] constituted a homogeneous set (the value of the chi-square statistic for homogeneity of ORs was 0.62 with 4 *df*, far from statistical significance), for which the value of the summary OR for aspirin vs control was 0.76 (statistically significant at $p < 0.01$), with a 95% confidence interval extending from 0.65 to 0.90.

The next trial, the Aspirin Myocardial Infarction Study (AMIS) [45], changed the picture radically. Its OR of 1.13, while not significantly different from unity, was significantly different from the earlier pooled OR of 0.76 ($\chi^2 = 9.31$, $df = 1$, $p < 0.01$). The value of the summary OR across the first six studies was a statistically non-significant 0.90 ($p > 0.10$), with a 95% confidence interval extending from 0.80 to 1.02.

The confidence intervals for the first five studies and for the first six demonstrate the paradox discussed earlier that as one's uncertainty as to the value of the overall OR increases, the length of the confidence interval based on the fixed effects model decreases. The estimates and confidence intervals provided by the DerSimonian–Laird random effects model [39] for the six studies are more valid given the degree of heterogeneity that exists across them. The estimated odds ratio has a borderline significant value of 0.84 ($\chi^2 = 3.05$, $df = 1$, $p < 0.10$), and the associated 95% confidence interval extends from 0.70 to 1.02. The length of the DerSimonian–Laird interval is, in logarithmic units, ln (upper limit/lower limit) = ln (1.02/0.70) = 0.38, appropriately greater than both the length of the fixed effects interval for the first five studies (ln(0.90/0.65) = 0.32) and that for the first six (ln(1.02/0.80) = 0.24). (The fixed effects and random effects analyses of the first five studies yield identical results).

There were no obvious reasons for removing AMIS from the meta-analysis, other than the invalid one that its results differed significantly from those of the first five studies. The state of

Table 1. Results of seven placebo-controlled randomized studies of the effect of aspirin in preventing death after myocardial infarction

Study	No. deaths/No. patients		OR	$y = \ln(OR)$	$w = 1/\mathrm{Var}(y)$
	Aspirin	Placebo			
MRC-1 [40]	49/615	67/624	0.720	−0.329	25.710
CDP [41]	44/758	64/771	0.681	−0.384	24.291
MRC-2 [42]	102/832	126/850	0.803	−0.219	48.801
GASP [43]	32/317	38/309	0.801	−0.222	15.440
PARIS [44]	85/810	52/406	0.798	−0.226	28.409
AMIS [45]	246/2267	219/2257	1.133	0.125	103.985
ISIS-2 [46]	1570/8587	1720/8600	0.895	−0.111	663.923

knowledge was therefore ambiguous: the first five studies collectively pointed to a statistically significant and fairly strong effect of aspirin for secondary prevention after a myocardial infarction, but the first six provided, at best, only suggestive evidence for the effectiveness of aspirin. This ambiguity was not resolved until the results of a seventh study, the Second International Study of Infarct Survival (ISIS-2) [46], were published. For the sake of comparability across all studies, the analyses below are based on the 2-year all-cause mortality rates from ISIS-2.

According to the DerSimonian–Laird random effects analysis, the overall OR across all seven studies was equal to a barely statistically significant 0.88 ($\chi^2 = 4.36$, $df = 1$, $p < 0.05$), with a 95% confidence interval extending from 0.77 to 0.99. The fixed effects analysis of these seven studies suggested (inappropriately, we believe) sharper conclusions: a highly significant point estimate for the OR of 0.90 ($\chi^2 = 10.6$, $df = 1$, $p < 0.005$), with a narrow 95% confidence interval extending from 0.84 to 0.96. We question the validity of the latter analysis because significant between-study variation remains (the OR for AMIS is not only significantly different from the average OR for the first five studies, as found before, it is significantly different from the OR for ISIS-2 ($\chi^2 = 4.95$, $df = 1$, $p < 0.05$)).

The overall substantive conclusion from this meta-analysis is that aspirin seems to be a modestly effective agent for reducing the risk of death during a period of approximately 2 years after a myocardial infarction, with a percentage reduction in the odds for dying relative to placebo equal to approximately 10%. The limits of uncertainty about this value are unsure, with the conservative random effects approach yielding a much wider confidence interval than the anticonservative fixed effects approach (in both instances, though, the upper confidence bound was less than 1.0).

The major methodological conclusion is that a single study may exert a powerful effect on one's conclusions. Here, there were two such influential studies. AMIS undid the statistically significant effect of aspirin found in the first five studies, and ISIS-2 undid the statistically non-significant effect of aspirin found in the first six. Given that one cannot know whether or when the next and potentially decisive study will be conducted, it would be prudent always to attach greater uncertainty than provided by traditional confidence intervals to the results of a meta-analysis of the studies conducted to date.

Execution and reporting

Several authors have proposed guidelines for carrying out and publishing the results of meta-analyses [3,16,47]. Although these guidelines are presented specifically for the meta-analysis of randomized clinical trials, they apply, with only minor exceptions, to meta-analyses in epidemiology as well. In our opinion, the standards and criteria offered by Sacks *et al.*[3], including, as they do, most of the others' guidelines, are the most useful. The six areas within which a meta-analysis should be evaluated, with those of their subdivisions that pertain to both clinical trials and epidemiology, follow:

(A) Study design. Just as the individual studies being meta-analyzed should be rigorously designed, with the design carefully and completely described, so should the meta-analysis itself. The meta-analysis should be carried out in accordance with a protocol prepared before the initiation of the study. The report of its results should describe the methods used by the meta-analysts to find all relevant articles, abstracts, chapters, etc.; should list the studies analyzed and enumerate those that were excluded (with reasons for their exclusion); and should provide summary data on the clinical and demographic characteristics of the subjects in the studies (subtypes of lung cancer, for example, in case–control studies).

(B) Combinability. The authors should address the statistical issue of whether the results from the separate studies should have been combined. If the estimates of treatment effect in clinical trials or of exposure–illness association in epidemiological studies differed significantly one from another, and especially if there was evidence of "qualitative interaction" (the estimates being in one direction in some studies and in the other direction in others), the authors should discuss why they proceeded to pool the results from all the studies.

(C) Control and measurement of potential bias. Several sources of unconscious bias exist, each of which should be addressed in the protocol for the meta-analysis. The more important ones would be discussed in the publication reporting on the results of the meta-analysis. Bias may occur in the decision as to which studies to select and which to exclude. Ideally, the decision should be made by one or more reviewers who

concentrate on the study's methods and are kept blinded to the study's results.

Bias may exist in the process of extracting the summary estimates of effect or association from the publication of the study's results. This is especially likely to occur in epidemiological studies, in which many actual or potential confounding variables are controlled—separately or in combination—and in which many estimates of relative risk are produced. Depending on their predilections, reviewers might tend to choose the largest estimate, the smallest estimate, or, in order to be fair, the estimate closest to the average of the individual ones. A solution may be to have two or more reviewers carry out the data-extraction independently, and to resolve any disagreement by having them discuss the study and reach consensus.

Sacks and his colleagues recommend, finally, that the sources of support for the meta-analysis be identified.

(D) Statistical analysis. Depending on the nature of the response variable, quantitative or categorical, either an analysis of variance [24] or a variation of the Mantel–Haenszel procedure [34], both of which properly average within-study differences, should be employed. Point and interval estimation are desirable in addition to significance tests. If the meta-analysis fails to demonstrate a significant overall effect or association, the possibility of inadequate power should be considered. When specific effects or associations within subgroups were hypothesized *a priori*, separate meta-analyses should be performed within those subgroups.

(E) Sensitivity analysis. When possible, the studies' results should be analyzed in two or more ways in order to confirm that the final result from the meta-analysis is qualitatively the same no matter how the results are analyzed. The quality of the individual studies should be determined and incorporated into the final conclusions from the meta-analysis. The possible impact of publication bias and of the "file drawer problem" should be carefully considered.

(F) Application of results. Bringing to bear all of the above considerations, the meta-analysts should come to a decision as to whether the pooled results provide a definitive, effectively final answer to the research question, or whether the conclusions are tentative and further individual studies are needed.

III. APPLICATIONS OF META-ANALYSIS TO EPIDEMIOLOGICAL STUDIES OF EXPOSURE TO ETS AND LUNG CANCER

An important application of meta-analysis was the analysis by the National Research Council (NRC) [48] of all known epidemiological studies (through 1986) of the hypothesized association between a non-smoker's exposure at home to environmental tobacco smoke (ETS) and the risk of lung cancer. The overall OR found by the NRC was a statistically significant 1.34 ($p < 0.001$), with a 95% confidence interval extending from 1.18 to 1.53. Among the other criticisms of the NRC's meta-analysis that are to be addressed subsequently is the criticism that many biases in the individual studies that should have been accounted for were not.

Four studies were excluded from the NRC's meta-analysis, for apparently valid reasons: no reference population was given, no raw data were presented, etc. Aside from their specification of the reasons for the exclusion of these four studies, the authors of the NRC report appear not to have followed the major guidelines proposed by Sacks *et al.* [3]. For example, they did not provide a formal protocol for the meta-analysis, nor, apparently, did they give any consideration to the possibility of heterogeneous ORs across the several studies.

In addition, most of the decision points and sources of bias discussed in Section II in connection with the meta-analysis of clinical trials also apply to the meta-analysis of epidemiological studies. For example, studies that fail to show an effect of treatment are often not published either due to a "publication bias," i.e. articles that fail to show an effect of treatment are often rejected for publication, or due to the "file drawer phenomenon," i.e. there is a tendency on the authors' part not to submit for publication an article that fails to show an effect. These two related sources of bias are clearly present in epidemiological studies as well as in clinical trials. For example, it may be that a study comparing the incidence of lung cancer among non-smokers exposed to ETS against the incidence of lung cancer among non-smokers not so exposed yielded a relative risk substantially less than one. The world of science, as it is today, might well preclude the publication of such a study [49].

Furthermore, the question comes to mind whether the existing epidemiological studies of a possible association between exposure to ETS and the incidence of lung cancer in non-smokers

are of adequate quality. Indeed, there is the question whether any of these studies meets even minimal standards of quality [50, 51].

Letzel *et al.* [51] considered the effects of misclassification errors on the results of the NRC's meta-analysis. They assumed three sets of conservative rates of misclassification for both disease and exposure to ETS, and, after adjusting for misclassification, concluded that "taking all this evidence together our calculations show that the findings of all studies about female lung cancer from passive smoking are consistent with the null hypothesis." Their final statement, also worth quoting, reads: "This brings us to final conclusion that there are presently only 2 alternatives—accepting the null hypothesis or creating new empirical evidence by peforming a really good study."

We shall nevertheless meta-analyze the nine American epidemiological studies that have, to our knowledge, been performed, the five that were included in the NRC report [48] plus four more recent ones. It is important to remind ourselves beforehand that there are many biases and confounders that will tend to inflate the relative risk. An especially important one is the misclassification of actual smokers as non-smokers. As Lee [52, 53] and Letzel *et al.* [51] point out, a woman who claims to be a non-smoker is more likely to be or to have been an actual smoker if married to a smoker than if married to a non-smoker. Other sources of bias include the misclassification of disease (i.e. misclassifying a non-lung cancer patient as having lung cancer as the primary disease, and misclassifying a lung cancer patient as not having cancer [51]), differing lifestyles between households where tobacco is used and those where tobacco is not used, and differential exposure and duration of exposure to air pollution between the exposed and unexposed groups. These biases have been considered to varying degrees by researchers in the field, with generally little success in controlling them.

One major source of bias that has not been analyzed sufficiently thoroughly, and has not yet been adequately controlled, is the misclassification of the spouse's smoking history (in this discussion the spouse is the husband or wife of the patient or of the control). It is possible that a non-smoking woman with lung cancer will overestimate the amount or duration of her husband's smoking in an attempt to find a causal explanation for her disease. The same tendency might be expected to exist when it is a surrogate for the patient—a child or sibling, or the spouse himself—who is being asked to report on the spouse's smoking history. The latter point is important because the proportion of patients reported on by a surrogate exceeds 50% in some studies.

U.S. studies of ETS and lung cancer

There are many reasons for restricting attention to American studies of whether there is an elevated risk of lung cancer to non-smokers exposed to ETS relative to non-smokers not so exposed. One is that this is the population to whom policy decisions will apply and on whom those decisions should be based. Another is that the summary ORs in the individual studies are derived from distributions of smoking amounts and durations, and of brands of cigarettes and other tobacco products, that pertain to populations within the U.S., and may thus be expected to be relatively homogeneous. Odds ratios from studies in other countries, on the other hand, are derived from distributions that may differ markedly from those in the U.S., and thus the ORs themselves may not be relevant to the American experience. Genetic and lifestyle differences between the U.S. population and the populations studied elsewhere (mainly in east Asia) also argue for a meta-analysis only of U.S. studies.

The first U.S. study, by Garfinkel [54], was a prospective follow-up study of more than 175,000 women who reported themselves to be non-smokers. All types of cancer of the lung were taken as end points. A "non-smoker" in this study was not only a women who reported that she never smoked, but also one who reported that she smoked only occasionally but not regularly. Little if any attempt seems to have been made to verify these women's self-reports.

The remaining U.S. studies were all case-control studies comparing patients with lung cancer against one or another kind of comparison group. The first was the study in New Orleans by Correa *et al.* [55]. Controls were patients with other diseases, from the same hospitals as the lung cancer cases, who were matched to the cases on age, sex and race. Specially trained interviewers were relied on to obtain exposure data for the cases and controls, although it is not clear whether the interviewers were blinded to whether a patient was a case or a control. The next of kin served as a proxy for the patient in 24% of the cases and 11% of the controls.

In the study by Kabat and Wynder [56], lung cancer patients in six cities were identified, and controls were matched to the cases on age, sex, race, date of interview and hospital. More care seems to have been taken in this study than in the others to ensure that subjects classified as non-smokers were truly such. Interviewers used a standardized questionnaire, but it is unclear whether they were blinded to the status of the patient as a case or a control.

The study of Buffler *et al.* [57] was conducted in six coastal counties in Texas. Little information is provided about such key features of the study as the criteria for classifying the spouse as a "regular smoker" or not, whether ex-smokers were included or excluded, whether the interviewers were blinded, and the number of patients for whom a surrogate interviewer was required.

Garfinkel *et al.* [58] studied cases and controls from hospitals in New Jersey and Ohio. Controls were patients with colorectal cancer, matched to the cases on age and hospital. The interviewers were kept blinded to the status of each patient. Women were counted as unexposed to ETS even if their husbands smoked cigarettes "only occasionally." There was extensive reliance on surrogate interviewees, many with questionable knowledge about the patient: approximately one quarter of all interviews were with someone other than the patient, the spouse or a child, about 60% were with the spouse or a child, and only 12% were with the patient herself.

The case–control study by Wu *et al.* [59], conducted in Los Angeles County, was the first to use neighborhood rather than hospital-ized controls. Only cases who were still alive were interviewed (all on the telephone); i.e. no

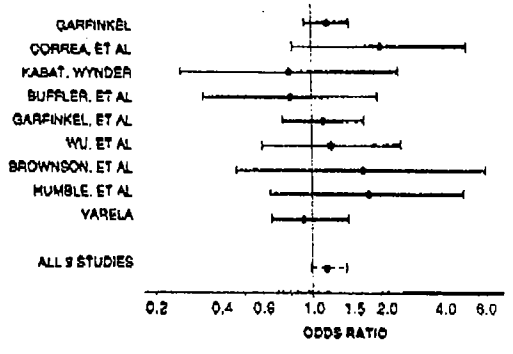

Fig. 2. Odds ratios and 95% confidence intervals for nine U.S. epidemiological studies of the hypothesized association between exposure to environmental tobacco smoke and lung cancer.

Table 2. Summary results of U.S. epidemiologic studies of the association between a non-smoking woman's exposure to environmental tobacco smoke and the risk of lung cancer

Study	OR	$y = \ln(OR)$	$w = 1/\mathrm{Var}(y)$
Garfinkel [54]	1.17	0.157	73.730
Correa *et al.* [55]	2.02	0.703	4.745
Kabat and Wynder [56]	0.79	−0.233	3.061
Buffler *et al.* [57]	0.80	−0.226	4.777
Garfinkel *et al.* [58]	1.12	0.113	22.330
Wu *et al.* [59]	1.22	0.199	7.545
Brownson *et al.* [60]	1.68	0.519	2.310*
Humble *et al.* [61]	1.78	0.577	3.826
Varela [62]	0.91	−0.090	36.268

*From personal communication from Dr Brownson.

surrogates were permitted for cases who had died or who refused to be interviewed. No information was provided as to whether the interviewers were blinded. The point estimate of the OR given by the authors in their paper's abstract and in the text on p. 748, OR = 1.2, is inconsistent with the confidence interval reported in both of those places, 0.5–3.3 (the geometric mean of the limits must equal the point estimate). Instead of working with these incorrect values, we used in Table 2 and in Fig. 2 of this paper the values for "spouse smoked" for adenocarcinoma in their Table 2: OR = 1.2, with a 95% confidence interval extending from 0.6 to 2.5.

Brownson *et al.* [60] carried out their case–control study in Denver. The controls were patients with cancer of the colon or bone marrow and were matched according to age and sex (there was approximately a 50:50 split on sex for the patients with lung cancer). The interviewer was blinded to the case or control status of the patient. The interviewee was someone other than the patient (mainly the spouse but occasionally a sibling or child) for nearly 70% of the cases and almost 40% of the controls. Exposure to ETS was not dichotomized in their Table 4 as "none" vs "any" but as "less than four hours per day" vs "four or more". The 95% confidence interval for the OR there should extend from 0.46 to 6.10 (personal communication from the senior author).

The study by Humble *et al.* [61] was a population-based case–control study in New Mexico. Controls were obtained by random digit dialing or from a randomly generated list of Medicare recipients. They were selected to match the frequency distributions of the cases on sex, ethnicity and age. The patient's status as a never-smoker was checked against the information recorded in the hospital chart. More than half of the time a surrogate was relied on

for exposure information about the case. It is unclear whether the interviewers were kept ignorant of the subject's status.

Varela's study [62] was carried out in New York State, with the controls being selected from State motor vehicle records. They were matched to the cases on age, sex, county of residence and previous smoking history. The questionnaire administered to the cases was slightly different from the one administered to the controls, so blinding was obviously impossible. Surrogate interviewees were permitted for the one-third of cases who had died or could not be interviewed for other reasons, but in such instances a surrogate respondent was interviewed for the matched control.

A final publication pertaining to experience in the U.S. is by Dalager *et al.* [63]. Rather than being a report on a new study, this paper reports the results of a summarization of data reported on earlier by Correa *et al.* [55] and by Buffler *et al.* [57], plus data that apparently have not yet been published anywhere. The results of the analysis may not be totally valid because "the data from all three study areas were merged" instead of being combined using the methods described in the Appendix. In any event, to have included the results of the quasi-meta-analysis by Dalager *et al.* [63] in our meta-analysis would have resulted in the studies by Correa *et al.* [55] and Buffler *et al.* [57] inappropriately being counted twice.

The overall quality of the American studies is obviously quite variable, just as it is for all such studies world-wide. Because we did not develop *a priori* a set of procedures for the unbiased measurement of a study's quality, it is appropriate that we include all known U.S. studies in our meta-analysis [54–62]. Most of these studies reported several values for the odds ratio. We selected for analysis one value per study, the value we believe the authors took to be their most accurate measure of association between exposure to ETS and lung cancer. These usually agreed with the values selected by the NRC [48] and by Layard [50] in their meta-analyses.

The results are presented in Table 2 and Fig. 2. There is no evidence for study-to-study heterogeneity (the value of the chi-square statistic with 8 *df* is a non-significant 5.46). The OR of 1.17 for the single prospective study, that by Garfinkel for the American Cancer Society [54], is close in value to the average OR of 1.07 for the eight case–control studies [55–62]. The overall OR across all nine studies is equal to a statistically non-significant 1.12 ($\chi^2 = 1.88$, $df = 1$), with the 95% confidence interval extending from 0.95 to 1.30. The fact that no significant association was found neither vindicates nor condemns the meta-analysis of these epidemiological studies. Given the biases that exist in each individual study, the safest conclusion from the present meta-analysis is a negative one: there is no convincing scientific evidence from the epidemiological literature of an association between exposure to ETS and the risk of lung cancer in the U.S.

IV. CONCLUSIONS

Meta-analyses, when properly performed, can be used effectively in both clinical trials and epidemiological studies for the following purposes:

- To increase the power of statistical tests for important endpoints and subgroups.
- To make sense out of studies with conflicting conclusions.
- To improve estimates of effect size.

However, uncritical use of meta-analysis can and does lead to unsubstantiated conclusions. Only when all the issues that we have discussed are considered and properly accounted for is it possible to apply meta-analysis to combine studies so that the overall result is scientifically valid. These issues include publication bias, the question of heterogeneity across studies, whether all subjects should be included in the meta-analysis or only those who are compliant with their treatment (this pertains only to clinical trials), whether proper control or adjustments have been made in epidemiological studies for sociodemographic or clinical differences between study populations, and the possible misclassification of subjects with regard to levels of exposure and case–control status.

It is very unlikely that the biases present in the epidemiological studies of the possible association between exposure to ETS and the risk of lung cancer can ever be removed. The meta-analysis performed by the NRC [48] must either be completely discounted or, as Stein [23] concluded so succinctly in another context, considered a mere "computational exercise."

Acknowledgements—This research was supported by a grant from The Tobacco Institute, Washington, D.C., U.S.A.

We thank Dr Myron Weinberg, President of the Weinberg Group/WASHTECH, for encouraging us to develop this critique.

REFERENCES

1. Huque MF. Experiences with meta-analysis in NDA submissions. **Proc Biopharmaceutical Section of the American Statistical Association** 1988; 28–33.
2. Hedges LV, Olkin I. **Statistical Methods for Meta-analysis.** New York: Academic Press; 1985.
3. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomized controlled trials. **N Engl J Med** 1987; 316: 450–455.
4. Chalmers TC, Levin H, Sacks HS, Reitman D, Berrier J, Nagalingam R. Meta-analysis of clinical trials as a scientific discipline. I. Control of bias and comparison with large co-operative trials. **Stat Med** 1987; 6: 315–325.
5. O'Rourke K, Detsky AS. Meta-analysis in medical research: Strong encouragement for higher quality individual research efforts. **J Clin Epidemiol** 1989; 42: 1021–1029.
6. Peto R. Why do we need overviews of randomized trials. **Stat Med** 1987; 6: 233–240.
7. Pocock SJ. **Clinical Trials: A Practical Approach.** New York: Wiley; 1983: 244.
8. Yusuf S. Obtaining medically meaningful answers from an overview of randomized clinical trials. **Stat Med** 1987; 6: 281–286.
9. Gehan EA, Freireich EJ. Non-randomized controls in cancer clinlical trials. **N Engl J Med** 1974; 290: 198–203.
10. Light RJ. Accumulating evidence from independent studies: What we can win and what we can lose. **Stat Med** 1987; 6: 221–228.
11. Hedges LV. Commentary. **Stat Med** 1987; 6: 381–385.
12. Eysenck HJ. An exercise in mega-silliness. **Am Psychol** 1978; 33: 517.
13. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A. A method for assessing the quality of a randomized control trial. **Contr Clin Trials** 1981; 2: 31–49.
14. Berlin JA, Colditz GA. A meta-analysis of physical activity in the prevention of coronary heart disease. **Am J Epidemiol** 1990; 142: 612–628.
15. Glass GV, McGass B, Smith ML. Meta analysis in social research. Beverly Hills, Calif.: Sage; 1981: 220–226.
16. Jenicek M. Meta-analysis in medicine: Where we are and where we want to go. **J Clin Epidemiol** 1989; 42: 35–44.
17. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction. An overview of the randomized trials. **Prog Cardiovasc Dis** 1985; 27: 335–371.
18. Smith M, Glass G, Miller T. **The Benefits of Psychotherapy.** Baltimore: Johns Hopkins University Press; 1980.
19. Rosenthal R. The "file drawer problem" and tolerance for null results. **Psychol Bull** 1979; 86: 638–641.
20. Iyengar S, Greenhouse JB. Selection models and the file drawer problem. **Stat Sci** 1988; 3: 109–117.
21. Berlin JA, Begg CB, Louis TA. An assessment of publication bias using a sample of published clinical trials. **J Am Stat Assoc** 1989; 84: 381–392.
22. Dubey S. Regulatory considerations on meta-analysis, dentifrice studies and multicenter trials. **Proc Biopharmaceutical Section of the American Statistical Association** 1988; 18–27.
23. Stein RA. Meta-analysis from one FDA reviewer's perspective. **Proc Biopharmaceutical Section of the American Statistical Association** 1988; 34–38.
24. Fleiss JL. **The Design and Analysis of Clinical Experiments.** New York: Wiley; 1986: 176–180.
25. Fleiss JL. **Statistical Methods for Rates and Proportions,** 2nd edn. New York: Wiley; 1981: 161–164.
26. DeMets DL. Methods for combining randomized clinical trials: Strengths and limitations. **Stat Med** 1987; 6: 341–348.
27. Stampfer MJ, Goldhaber SZ, Yusuf S, Peto R, Hennekens CH. Effect of intravenous streptokinase on acute myocardial infarction. Pooled results from randomized trials. **N Engl J Med** 1982; 307: 1180–1182.
28. Multicenter Diltiazem Postinfarction Trial Research Group. The effect of diltiazem on mortality and reinfarction after myocardial infarction. **N Engl J Med** 1988; 318: 385–392.
29. Friedman LM, Furberg CD, DeMets DL. **Fundamentals of Clinical Trials,** 2nd edn. Littleton, Mass.: PSG Publishing; 1985: 246–249.
30. Sackett DL, Gent M. Controversy in counting and attributing events in clinical trials. **N Engl J Med** 1979; 301: 1410–1412.
31. Baber NS, Lewis JA. Confidence in results of beta-blockers post-infarction trials. **Br Med J** 1982; 284: 1749–1750.
32. Peto R. Discussion of DeMets DL. Methods for combining randomized clinical trials: Strengths and limitations. **Stat Med** 1987; 6: 349–350.
33. Greenland S, Salvan A. Bias in the one-step method for pooling study results. **Stat Med** 1990; 9: 247–252.
34. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. **J Natl Cancer Inst** 1959; 22: 719–748.
35. Hine L, Laird NM, Hewitt P, Chalmers TC. Meta-analysis of empirical long-term anti-arrhythmic therapy after myocardial infarction. **JAMA** 1989; 262: 3037–3040.
36. Peto R. Discussion of Peto R. Why do we need systematic overviews of randomized trials? **Stat Med** 1987; 6: 242.
37. Meier P. Commentary. **Stat Med** 1987; 6: 329–331.
38. Bailey KR. Inter-study differences: How should they influence the interpretation and analysis of results? **Stat Med** 1987; 6: 351–358.
39. DeSimonian R, Laird N. Meta-analysis in clinical trials. **Contr Clin Trials** 1986; 7: 177–188.
40. Elwood PC, Cochrane AL, Burr ML, Sweetnam PM, Williams G, Welsby E, Hughes SJ, Renton R. A randomized controlled trial of acetyl salicylic acid in the secondary prevention of mortality from myocardial infarction. **Br Med J** 1974; 1: 436–440.
41. Coronary Drug Project Group. Aspirin in coronary heart disease. **J Chron Dis** 1976; 29: 625–642.
42. Elwood PC, Sweetnam PM. Aspirin and secondary mortality after myocardial infarction. **Lancet** 1979; 2: 1313–1315.
43. Breddin K, Loew D, Lechner K, Uberla EW. Secondary prevention of myocardial infarction. Comparison of acetylsalicylic acid, phenprocoumon and placebo. A multicenter two-year prospective study. **Thromb Haemost** 1979; 41: 225–236.
44. Persantine–Aspirin Reinfarction Study Research Group. Persantine and aspirin in coronary heart disease. **Circulation** 1980; 62: 449–461.
45. Aspirin Myocardial Infarction Study Research Group. A randomized controlled trial of aspirin in persons recovered from myocardial infarction. **JAMA** 1980; 243: 661–669.
46. ISIS-2 Collaborative Group. Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. **Lancet** 1988; 2: 349–360.
47. Meinert CL. Meta-analysis: Science or religion? **Contr Clin Trials** 1989; 10: 257S–263S.
48. National Research Council. **Environmental Tobacco Smoke: Measuring Exposures and Assessing Health Effects.** Washington: National Academy Press; 1986.

49. Vandenbrouke JP. Passive smoking and lung cancer: A publication bias? **Br Med J** 1988; 296: 391–392.
50. Layard MW. Environmental tobacco smoke and cancer: The epidemiologic evidence. In: Ecobichon DJ, Wu JM, Eds. **Environmental Tobacco Smoke: Proc Int Symp.** McGill University, 1989. Lexington, Mass.: Lexington Books; 1990: 99–115.
51. Letzel H, Blumner E, Uberla K. Meta-analyses on passive smoking and lung cancer: Effects of study selection and misclassification of exposure. **Environ Technol Lett** 1988; 9: 491–500.
52. Lee PN. Lung cancer and passive smoking. **Toxicol Lett** 1987; 35: 157–162.
53. Lee PN, Chamberlain J, Alderson MR. Relationship of passive smoking to risk of lung cancer and other smoking-associated diseases. **Br J Cancer** 1986; 54: 97–105.
54. Garfinkel L. Time trends in lung cancer mortality among nonsmokers and a note on passive smoking. **J Natl Cancer Inst** 1981; 66: 1061–1066.
55. Correa P, Pickle LW, Fontham E, Lin Y, Haenszel W. Passive smoking and lung cancer. **Lancet** 1983; 2: 595–597.
56. Kabat GC, Wynder EL. Lung cancer in nonsmokers. **Cancer** 1984; 53: 1214–1221.
57. Buffler PA, Pickle LW, Mason TJ, Contant C. The causes of lung cancer in Texas. In: Mizell M, Correa P, Eds. **Lung Cancer: Causes and Prevention.** New York: Verlag-Chemie International; 1984: 83–99.
58. Garfinkel L, Auerbach O, Joubert L. Involuntary smoking and lung cancer: A case-control study. **J Natl Cancer Inst** 1985; 75: 463–469.
59. Wu AH, Henderson BE, Pike MC, Yu MC. Smoking and other risk factors for lung cancer in women. **J Natl Cancer Inst** 1985; 74: 747–751.
60. Brownson RC, Reif JS, Keefe TJ, Ferguson SW, Pritzl JA Risk factors for adenocarcinoma of the lung. **Am J Epidemiol** 1987; 125: 25–34.
61. Humble CG, Samet JM, Pathak DR, Marriage to a smoker and lung cancer risk. **Am J Publ Health** 1987; 77: 598–602.
62. Varela LR. Assessment of the association between passive smoking and lung cancer, PhD dissertation. Yale University; 1987.
63. Dalager NA, Pickle LW, Mason TJ, Correa P, Fontham E, Stemhagen A, Buffler PA, Ziegler RG, Fraumeni JF. The relation of passive smoking to lung cancer **Cancer Res** 1986; 46: 4808–4811.

APPENDIX

Statistical Appendix: Analysis of Log Odds Ratios

Suppose that there are, all told, S studies to be meta-analyzed. In a typical one, say study s, let n_{s1} and n_{s2} be the sample sizes in the two groups being compared and let p_{s1} and p_{s2} be the proportions having the characteristic under study. In a randomized controlled trial the two groups would be the treated and placebo samples and the characteristic under study might relapse or some other kind of failure. In an epidemiological case–control study the two groups would be the cases and controls and the characteristic under study would be exposure to the hypothesized risk factor. A review of the fixed effects analysis of the data follows.

The logarithm of the OR in study s, denoted by y_s, equal to

$$y_s = \ln\left(p_{s1}(1-p_{s2})/p_{s2}(1-p_{s1})\right).$$

where ln denotes natural logarithm. The standard error of y_s is given by

$$\text{se}_s = \left(\frac{1}{n_{s1}p_{s1}(1-p_{s1})} + \frac{1}{n_{s2}p_{s2}(1-p_{s2})}\right)^{1/2},$$

and the factor by which y_s is weighted in the classical fixed effects analysis, w_s, is

$$w_s = 1/(\text{se}_s)^2.$$

The overall OR across all S studies is equal to

$$\overline{\text{OR}} = \exp(\bar{y}),$$

where $\bar{y} = \Sigma w_s y_s / \Sigma w_s$, and the limits of a 95% confidence interval for the overall OR are given by

$$\exp(\bar{y} \pm 1.96/\sqrt{\Sigma w_s}).$$

This interval will not be symmetric about $\overline{\text{OR}}$.

The "combinability" of the S studies, i.e. the hypothesis that the S underlying ORs are equal, may be tested by referring the statistic

$$Q = \Sigma w_s (y_s - \bar{y})^2$$

to percentage points of the chi-square distribution with $S-1$ *df*. This same statistic Q plays a central role in the DerSimonian–Laird random effects analysis of the data [39]. In particular, the DerSimonian–Laird analysis is identical to the fixed effects analysis just presented if $Q \leqq S-1$, but the two methods diverge if $Q > S-1$.

Assume, therefore, the $Q > S-1$, and define

$$D = \frac{(Q-(S-1))\Sigma w_s}{(\Sigma w_s)^2 - \Sigma w_s^2}.$$

The DerSimonian–Laird weighting factor for study s is equal to

$$w_s^* = \left(D + \frac{1}{w_s}\right)^{-1},$$

and the random effects point and interval estimates of the overall OR become

$$\exp(\bar{y}^*)$$

and

$$\exp(\bar{y}^* \pm 1.96/\sqrt{\Sigma w_s^*}),$$

where $\bar{y}^* = \Sigma w_s^* y_s / \Sigma w_s^*$.

J Clin Epidemiol Vol. 44, No. 2, pp. 103–107, 1991
Printed in Great Britain

0895-4356/91 $3.00 + 0.00
Pergamon Press plc

Editorial

META-META-ANALYSIS: UNANSWERED QUESTIONS ABOUT AGGREGATING DATA

WALTER O. SPITZER*

Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Quebec, Canada H3A 1A2

(Received for publication 30 October 1990)

Some days I ask myself if investigators still do their own trials, with admissible standards and adequate power. It seems like the most frequent prelude to a discussion on efficacy or safety of an intervention is, "... We did a meta-analysis of studies on ...". Does anyone do straightforward unhyphenated analyses any more?

In this issue of the **Journal** (pp. 127–139), Fleiss and Gross review and reassess meta-analysis and report an illustrative case study of the method focusing on the association between exposure to environmental tobacco smoke (ETS) and lung cancer. The article refers to a succinct and useful definition of meta-analysis offered by Huque [1] "... a statistical analysis which combines or integrates the results of several independent clinical trials considered by the analyst to be 'combinable'". I also take Huque's characterization as the working definition for this editorial comment and join Fleiss and Gross in highlighting the fact that meta-analysis' key and almost exclusive application to date has been in the integration of data from clinical trials. I would add that a distinctive characteristic of the strategy is the derivation of a single quantitative estimate of effect of an intervention or a risk factor. Fleiss and Gross touch on most of the main issues, including consensus and controversy. I will not repeat or summarize their elegant review which enumerates many of the unanswered questions about meta-analysis of experimental trials. But it deserves emphasis that the main unresolved challenges are to settle on universal widely acceptable criteria for exclusion of trials and the development of a reproducible, valid and accepted weighting index that would enable analysts to invoke the quality of the research into a final result. I will confine my comments mostly to the application of meta-analysis in aggregation of non-experimental (observational) studies. The controversies surrounding meta-analysis of experimental trials are equally relevant to non-experimental studies which are usually epidemiological. But there are additional unanswered questions. Fleiss and Gross ask, "... can meta-analytic techniques be applied in the analysis of other kinds of data such as those that arise in cohort and case–control studies found in epidemiology? Their answer is a "guarded 'yes'". I do not know whether the question can be answered at all. The illustrative meta-analytic project reported by them goes a long way in avoiding potential pitfalls. However, some problems are not fully addressed either in the execution of the study nor in the discussion about such applications in conventional epidemiology. There are many difficulties that have not been surmounted yet, either theoretically or empirically. When phrasing the following 13 unanswered questions, I have used words such as "merge", "combine", "integrate" or "put together", in respect to data from different case series, different series of reference groups, different cohorts, etc. I have done so recognizing that one usually

*Reprint requests should be addressed to W. O. Spitzer at the above address.

combines intrastudy *differences*, not intrastudy disease or exposure rates. I ask the reader's indulgence for having avoided repetitive dense statistical technical terminology in deference to readable English prose. These then, are the questions:

(1) Operationally, what are the "stringent conditions" (Fleiss and Gross' phrase) under which both case–control studies and cohort studies may be included in one single meta-analysis? Should such analyses ever be done without access to the raw data of the component studies?
(2) When is it permissible to combine different types of cohort? For instance, for both exposed cohorts and comparison cohorts should one integrate data from a fixed cohort with an open one?
(3) Is it permissible to integrate exposed patients sampled from hospitals with those from primary care settings?
(4) For reference cohorts, *not exposed* to an intervention or risk factor, other questions arise. For example,
    (a) Is a comparison cohort from Sweden combinable with one from Italy or Japan?
    (b) Are cohorts taken from occupational sampling frames sufficiently similar to those from the corresponding general population (or another geographically-defined one) to put them together?
    (c) How separate in time must the accrual or demarcation of unexposed cohorts become to be ineligible for aggregation? (The question is also pertinent for exposed cohorts.)

Turning now to case–control designs:

(5) Is it admissible to merge hospital-based with population-based case groups? Or in Miettinen's terms, can two or more case series be combined if they are not representative of the same type of base experience? [2].
(6) Conceptually, and in execution, is a nested case–control study similar enough to a conventional case–control study for both to be included in the same meta-analysis?
(7) When there are two or more control groups in a case–control study does one merge all the control groups? If not, what criteria must one use to exclude any control group from the meta-analysis? There is no parallel between multiple arms defined by exposure in a randomized controlled trial and multiple reference samples demarcated by outcome in a case–control study.
(8) Should control groups assembled by matching be combined with independent samples of referenced populations?
(9) What constitutes "proper control or adjustment for the biases that frequently occur in epidemiological studies?" [3]. Case–control studies are the designs in common use most vulnerable to bias [4]. I believe that the difficulties in minimizing bias acceptably, alone, could vitiate the validity of meta-analyses of case–control studies. Bias is barely manageable (and seldom managed well) even in single case–control studies. As pointed out by Fleiss and Gross [3], as well as Letzl earlier [5], misclassification bias for exposure is a particularly thorny problem.
(10) Are data provided by proxy informants similar enough to data from respondents to be considered equivalent?
(11) Should one include case–control studies in which data-gatherers were unblinded with blinded studies in one meta-analysis? (Should one do so in cohort research?)
(12) How homogeneous must the outcome be? For instance, can one pool data from a study that ascertained "all cancers of the lung", with one that did so only for "oat cell Ca", or only "adenocarcinoma"?
(13) How do we interpret values and confidence intervals of single estimates derived with meta-analysis? Consider a report of a single study comparing two fixed cohorts of 2500 persons. The relative risk (RR) for the association of the putative risk factor with the incidence of a well-defined, hard, but relatively uncommon outcome is 1.45. The 95% confidence interval is 1.02–2.07 and $p = 0.04$. In a hypothetical meta-analysis of five other follow-up studies with 500 persons per cohort (two in each study assessing the same risk factor and outcome) the result is an identical relative risk, the same confidence interval and an

identical p value. In the second scenario the relative risks for the five component studies were 3.0 ($p = 0.013$), 0.6 ($p = 0.48$), 1.1 ($p = 0.80$), 1.5 ($p = 0.15$), 1.1 ($p = 0.83$). Do both sets of statistics mean the same thing? I remind the reader that it is not standard practice to incorporate interstudy variation with one's meta-analysis. Should one do so? Beyond problems of multiple comparisons which we usually recognize and correct, might there be a problem of "multiple combinations"?

As summarized by Fleiss and Gross, the indications for a properly conducted meta-analysis are,

(a) to increase statistical power,
(b) to deal with controversy when individual studies disagree,
(c) to improve estimates of size of effect and
(d) to answer new questions not previously posed in component studies [3]. But is it always necessary, or justified, to pursue a single estimate with its related probability qualifiers to derive conclusions from a series of research projects?

An alternative to both meta-analysis on one hand, and traditional (often haphazard) reviews is an approach proposed by Slavin and designated *best-evidence synthesis*. This approach considers that the "best evidence" in any field comes from studies having the highest internal and external validity, that use well-specified, defined, explicit *a priori* inclusion and exclusion criteria and favour size-effect data to statistical significance alone when interpreting the literature reviewed. Such syntheses emphasize numeric findings but the conclusions need not depend on a single estimation nor on statistical significance [6]. In common with properly executed meta-analyses best-evidence syntheses cannot evade the difficult challenge of deciding what to exclude and how to document the exclusions. Getting around "publication bias" [7] which means finding and judging unpublished work is particularly daunting.

What is attractive about best-evidence synthesis is that it liberates the analyst from the apparent obsession which meta-analysts have to calculate a single estimate as a necessary intermediate step to reach an opinion about an association, an effect or a casual relationship. Nevertheless, best-evidence synthesis does not exonerate the analyst from the highest attainable rigour in setting forth a protocol for the synthesis in advance. The protocol must then be followed when deciding which component studies reach a predetermined level of scientific admissibility, in establishing exclusion criteria, when implementing methods to document excluded research, when developing valid weighting schemes for the quality of papers, and when formulating predetermined explicit rules for judging effect size. The foregoing list of specifications for a best-evidence synthesis protocol is not exhaustive. I am of the opinion that even more rigour is required of the meta-analyst.

Turning now to Fleiss and Gross' illustrative case study that re-examines the association between exposure to environmental tobacco smoke and lung cancer [3], I shall highlight some of its features:

(i) They excluded non-American studies, a sensible move, given the unresolved methodological problems of pooling people of very different ethnic nature and of different culturally-determined views of smoking. The exclusion is particularly appropriate if the intention was to make inferences chiefly about the American population.
(ii) The meta-analysis incorporates one cohort study and eight case–control studies.
(iii) Hospital-based and population-based groups (both cases and controls) appear to have been considered equivalent. It is probably valid to have done so *when the purpose was to test the null hypothesis* of no association. Were one to attempt non-null inferences, it would have been a mistake to consider them equivalent.
(iv) Matched and unmatched controls were incorporated in a similar way.
(v) The overall analysis does not seem to have been adjusted by blindness status of data-gatherers nor by the extent that proxies or respondents provided information on exposure.
(vi) Outcomes were somewhat heterogeneous. Consequently, matched groups might have been different.

Lastly, they report,

(vii) "... we did not develop *a priori*, a set of procedures for the unbiased measurement of a study's quality ...".

The seven comments are not made to rebuke Fleiss and Gross' approach but to underscore the enormous difficulties that they and anyone else unavoidably confront attempting a meta-analysis of non-experimental epidemiological studies.

The single estimate they report for the nine studies is 1.12 (CI 95% 0.95–1.30), $\chi^2 = 1.88$ (1 *df*). I consider the result and the resulting conclusion plausible and I completely agree with their comment. "The fact that no significant association was found neither vindicates nor condemns the meta-analysis of epidemiological studies" [3].

My colleagues and I, in a Working Group on Passive Smoking that reported early last year [8], also examined the association between exposure to environmental tobacco smoke (ETS) and lung cancer. We used Slavin's method of best-evidence synthesis rather than meta-analysis. We also included the world literature, not just U.S. studies. It is instructive to compare the two sets of conclusions on the association:

Fleiss and Gross: "... there is no convincing scientific evidence from the epidemiologic literature of an association between exposure to ETS and the risk of lung cancer in the United States" [3].

Spitzer et al.: "The weight of evidence is compatible with a positive association between residential exposure to environmental tobacco smoke (primarily from spousal smoking) and the risk of lung cancer". "There is no evidence for an association between non-residential exposure to ETS and any form of cancer" [8].

The two conclusions are not identical. But they are not directly contradictory nor mutually exclusive. Moreover, given the restriction of the meta-analysis to U.S. studies and the inclusion of admissible studies from anywhere in the world for the best-evidence synthesis the compatibility of the "verdicts" tends to mutually buttress their validity. Admittedly, the language of the discussion of the Fleiss and Gross paper [3] favours a *non-association* interpretation while the Spitzer group's comments do the opposite [8]. For example, contrast these phrases: "... the safest conclusion from the present meta-analysis is a negative one" [3]; "The preponderance of positive studies is consistent with a causal relationship betwen exposure to ETS and lung cancer" [8]. But it is obvious that there is a lot of common ground between the conclusions of two different methodological approaches to the quantitative evidence available on the subject.

A more general conclusion of Fleiss and Gross is important: "Meta-analyses, when properly performed, can be used effectively in both clinical trials and epidemiological studies ...". In today's state of science I accept the conclusion guardedly and warily, *for clinical trials only* despite many unresolved controversies about what "properly performed" means in the method. In my opinion, however, the unanswered questions about meta-analysis in non-experimental epidemiological studies do not yet warrant widespread application except as methodological research. I view Fleiss and Gross' analysis as a courageous, honest, trailblazing early step in the development of the method. Their own caveats throughout their article lend support to my opinion and to their scientific integrity.

In the near future we need a level of international consensus about the methods of meta-analysis as high as that which prevails for unmeta-analyzed single randomized controlled trials. Perhaps a "summit" should be called for. I would find it difficult to ignore a general agreement on criteria for scientific admissibility of meta-analytic studies (both clinical trials and non-experimental studies) if it were endorsed by a group including, say, Armitage, Breslow, Cole, Day, Detsky, Gross, Feinstein, Fleiss, Meier, Miettinen, R. Peto, Sackett, Schwartz, Uberla, Vessey, Walter and Zelen. Minor miracles still happen occasionally.

Finally, let's abolish the verb "to meta-analyze" as a substitute for "to review", "to synthesize", "to interpret" or even "to read". Careless use of the technically-specific term does not do justice to the unfulfilled promise of meta-analysis, nor to the painstaking work of many colleagues who pursue excellence as they attempt to deliver the promise.

Acknowledgement—This study was supported by the National Health Resources Development Programme, Health and Welfare, Canada.

REFERENCES

1. Huque MF. Experiences with meta-analysis in NDA submissions. **Proc Biopharmaceutical Section of the American Statistical Association** 1988; 2: 28–33.
2. Miettinen OS. The "case-control" study: valid selection of subjects. **J Chron Dis** 1985; 38: 543–548.
3. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke

and lung cancer: a critique. **J Clin Epidemiol** 1991; 44: 127–139.

4. Ibrahim MA, Spitzer WO, Eds. The case–control study: consensus and controversy. **J Chron Dis** 1979; 32: 1–90.
5. Letzel H, Blummer E, Uberla K. Meta-analyses on passive smoking and lung cancer: Effects of study selection and misclassification of exposure. **Environ Technol Lett** 1988; 9: 491–500.
6. Slavin RE. Best-evidence synthesis: an alternative to meta-analytic and tradiational reviews. **Educ Res** 1986; 15: 5–11.
7. Vandenbroucke JP. Passive smoking and lung cancer: a publication bias? **Br Med J** 1988; 296: 391–392.
8. Spitzer WO, Lawrence V, Dales R *et al*. Links between passive smoking and disease: A best-evidence synthesis. **Clin Invest Med** 1990; 13: 17–42.