CLINICAL PHARMACOLOGY & THERAPEUTICS
VOLUME 58 NUMBER 6 DECEMBER 1995

COMMENTARY

Meta-analysis: A method for synthesizing research

Ralph B. D'Agostino, PhD, and Michael Weintraub, MD *Boston, Mass., and Rockville, Md.*

Meta-analysis has become a widely used technique in clinical and epidemiologic research. Although there are a number of references available to the researcher and clinician to explain its objectives and methods and to review its applications,[1-14] there still remains a lack of clarity of the role of meta-analysis in research, the steps necessary to perform one, and the proper criteria to evaluate it. Further, some commentators have been very skeptical[15,16] and others have found performed meta-analyses to be unconvincing.[17,18] The objectives of this commentary are to give the reader an appreciation of the capabilities and limitations of meta-analysis and the tools necessary to evaluate and, if desired, to perform one. We discuss how to perform a meta-analysis, we review methodologic issues both in design and analysis that arise in performing a meta-analysis, and we present criteria useful for evaluating a meta-analysis. Portions of a recent meta-analysis of antihistamines in common cold preparations to treat runny nose and sneezing are used for illustration. A statistical Appendix is also provided that supplies common procedures used in the performance of a meta-analysis.

From the Mathematics Department, Boston University, Boston, and the Office of Over-the-Counter Drugs, Food and Drug Administration, Rockville.

Supported in part by a grant RO1-HL40423-06 from the National Heart Lung and Blood Institute (Bethesda, Md.).
Presented at the joint meeting of the Nonprescription Drug Advisory Committee and the Pulmonary-Allergy Drug Advisory Committee, November 15, 1994.
The views expressed in this paper are those of the authors and not necessarily those of the U.S. Food and Drug Administration.
Received for publication July 5, 1995; accepted Aug. 17, 1995.
Reprint requests: Ralph B. D'Agostino, PhD, Boston University, Statistics and Consulting Unit, 111 Cummington St., Boston, MA 02215.
CLIN PHARMACOL THER 1995;58:605-16.
13/1/68636

WHAT IS A META-ANALYSIS?

A meta-analysis is a systematic review of studies that uses quantitative statistical procedures to combine, synthesize, and integrate information across these studies. When performed appropriately, it also incorporates qualitative evaluations of the studies. Its objective is to reach conclusions concerning an issue, such as the effect of a drug therapy or an epidemiologic association between risk factor levels and mortality. If all of the studies involved have been published, a meta-analysis is a quantitative literature review. However, it is a more explicit and structured approach than a traditional literature review, and it is complementary to the narrative review that often accompanies such a literature review.

WHY PERFORM A META-ANALYSIS?

One performs a meta-analysis to understand the current state of knowledge concerning a topic such as the quantification of a comparison of a new treatment and a control, a comparison that might involve differences in means, odds ratios, or relative risks. Such a quantification will be referred to in this commentary as a *treatment effect*. The quantification of an *epidemiologic association* between a risk factor and an outcome, for example, total serum cholesterol level and myocardial infarction, is another common reason for a meta-analysis. A meta-analysis is especially useful when results from several studies lack statistical significance yet appear to have effects in the same direction, or when study results seem to disagree with regard to magnitude or direction of effect and a resolution is needed, or when a large single trial is too costly and time-consuming to perform. More precisely, a meta-analysis is performed primarily for the following reasons:

- To estimate quantitatively the current state of knowledge concerning some issue such as a *treatment effect* or an *epidemiologic association*.
- To improve the precision of an estimated measure of treatment effect or epidemiologic association.
- To obtain sufficient power to test statistically the significance of a treatment effect or an epidemiologic association.
- To resolve controversies when studies appear to disagree.
- To answer new questions that have not been posed in the individual studies and for which the existing studies individually may not have sufficient precision or power to answer.

A meta-analysis performed for any of these *primary* reasons should utilize all available information concerning that topic and should be performed in a systematic careful formal manner. We now turn to issues that need to be addressed to ensure this.

STUDY DESIGN

A detailed protocol should be developed before a meta-analysis is undertaken. This protocol should contain a complete and careful description of the study to be undertaken. The protocol should include the following:

- A statement of the objectives, including a delineation of primary and secondary objectives.
- A complete description of the strategies that will be used to identify and locate relevant studies. This includes the literature search for relevant articles, abstracts, and chapters, as well as identification of unpublished studies.
- The rules for inclusion and exclusion of a study in the meta-analysis.
- A plan to evaluate the quality of each study and the methods to be used to include this evaluation in the analysis.
- A listing of the summary descriptive measures that will be obtained from each individual study. These include efficacy measures and their standard errors for the treatment effects in clinical trials and risk assessments measures for epidemiologic associations.
- A complete description of the methods to be used to extract the summary measures from each study.
- A detailed statistical analysis plan that describes the procedures that will be used to analyze the data. This should include inferential statistical testing and estimation procedures to be applied to the summary measures from the individual studies, tests for evaluation of the homogeneity (or poolability) of these measures, methods for combining these across studies, and inferential procedures to be applied to these combined measures.

Not only is a rigorous complete protocol development important in designing a meta-analysis, any evaluation of a performed meta-analysis should weight heavily the completeness of the protocol produced *before* the data were obtained. An evaluation should also consider the influences the obtained data had on changes to the protocol and on the methods and analyses used in the meta-analysis. We now comment briefly on each of the above design issues.

Objectives. The aim of a meta-analysis in medical research is to obtain a general conclusion about a topic. For example, do antihistamines in common cold preparations reduce significantly runny nose and sneezing during the first few days of the cold? Specific objectives to this end must be listed and primary and secondary objectives should be delineated. As with all clinical and epidemiologic research the objectives of the endeavor should be stated before initiation of the activity and not be influenced by examination of the data. The objectives and all components of the meta-analysis should be stated before obtaining the data.

Gathering the studies. A good meta-analysis should contain all available data relevant to the topic. This includes all published and unpublished materials. For the published data a literature search is often under-

taken. This can pose problems, even with the availability of computerized searches.[19-21] Unpublished studies are even more difficult to obtain. The main problem usually mentioned here is "publication bias" due to negative studies not being published and not being available.[16] Simes[22] discusses ways of addressing this problem. In some fields, computerized databases of unpublished studies do exist, as do also registries and inventories of clinical trials. Any evaluation of a meta-analysis must consider the completeness of the identification of all the relevant data.

Inclusion and exclusion criteria. Standardized criteria for the inclusion or exclusion of studies in a meta-analysis *do not* exist, and this is another serious problem with all meta-analyses.[16] Inclusion and exclusion criteria should be stated a priori and explicitly in the meta-analysis protocol. Depending on the data, they may need to be modified. However, substantial justification is needed for the changes.

One consideration for inclusion and exclusion is whether only published studies should be analyzed. Other considerations involve study design, type of experimental and control therapies, quality of the study, length of the study, and the outcomes of interest. The more similarities that exist across studies with regard to these items, the more appropriate is the meta-analysis. Ideally, the studies will have many similar features that justify the pooling processes involved in meta-analysis.

For studies that evaluate drug treatments, it is often stated that only randomized studies should be included.[23] The argument for this position is that there are great potentials for bias in nonrandomized studies. The counterargument is that the judgment to decide that a study is flawed is too subjective.[24] However, this latter view is itself flawed. Nonrandomized studies are suboptimal and, to include them in a meta-analysis, the question of bias adjustment must be addressed. Entering a set of defective studies into a meta-analysis will not correct their deficiencies.[25]

Another issue relates to the relevance of the available studies to the objectives of the meta-analysis. For example, studies that include doses of drugs that are no longer in use may not be relevant in a meta-analysis. Also, available published studies that include summary statistics from a per-protocol analysis rather than an intention-to-treat analysis may be considered biased and not appropriate. Further, studies with patient follow-up of too short a duration to be relevant may be excluded.

Other exclusion criteria relate to study quality. For example, only studies that have protocols should be considered to be sufficient quality for inclusion. Quality considerations concerning controls and compliance should also be appraised. In general, if a study was seriously flawed it can be excluded. The topic of study quality is discussed further below.

Another major problem with the inclusion and exclusion task is the possibility of including too frequently studies that "agree" with the investigators' biases, while excluding those that do not. Chalmers and Lau[26] present an elaborate scheme to reduce this source of bias. The success of the process must always be questioned and examined.

In general, protocol inclusion and exclusion criteria of the original protocol and those actually implemented should be part of the presentation of a meta-analysis. All changes from the original criteria should be justified.

Quality assessment. For those studies that have passed the inclusion and exclusion criteria, a quality assessment is often recommended. Standardized procedures for this exist.[27] The process involves a group of six to eight reviewers reading independently the methods and results sections of the studies and rating the studies with a quality score ranging from 0.0 to 1.0. Information such as the authors and institution are withheld from the reviewers. Some consensus process to obtain a final quality score per study may be needed. The quality score can be used to exclude some studies. It can also be used to weight studies in the analysis. A sensible procedure often is to take all studies above some quality threshold into the meta-analysis and not to weight them any further by quality in the statistical analysis. Whatever is done needs to be documented and justified.

Data extraction. At this point the actual data for the meta-analysis is extracted. In the ideal case the original data from each study are available and all necessary summary statistics and estimates of treatment effects or epidemiologic associations such as differences between groups, odds ratios, or relative risks, are computed for the meta-analysis directly from the data. However, usually the data are extracted from published or unpublished reports in the form of summary statistics, such as sample sizes, means and standard deviations for continuous variables, and sample sizes and the number of subjects in different categories for categorical or ordinal data. Often data for meta-analyses are dichotomous with the number of successes (e.g., desired clinical effect attained) and number of failures as the data elements. From these data or summary statistics, estimates of treatment effects or epidemiologic associations are computed separately

Table I. Meta-analysis on treatment effects (day 1 incremental reduction in runny nose; analysis on a continuous variable)

	Antihistamine			*Placebo*					
Study	n_1	$\bar{x}_1$	S_1	n_2	$\bar{x}_2$	S_2	*Pooled (S)*	*Effect (g)*	*Variance (v)*
1	11	0.273	0.786	16	−0.188	0.834	0.815	0.564	0.160
2	128	0.932	0.593	136	0.810	0.556	0.574	0.213	0.015
3	63	0.730	0.745	64	0.578	0.773	0.759	0.200	0.032
4	22	0.350	1.139	22	0.339	0.744	0.962	0.012	0.091
5	16	0.422	2.209	15	−0.017	1.374	1.853	0.237	0.130
6	39	0.256	1.666	41	0.537	1.614	1.640	−0.171	0.050
7	21	2.831	1.753	21	1.396	1.285	1.537	0.933	0.106
8	13	2.687	1.607	8	1.625	2.089	1.800	0.590	0.211
9	194	0.490	0.895	193	0.264	0.828	0.862	0.262	0.010

Test of homogeneity (from equations 1, 3, 6, and 8), χ^2 statistics of equation 3 = 9.84, df = 8, and p level = 0.28. Estimate of pooled treatment effect of equation 1 is 0.234; 95% confidence interval is 0.111 to 0.358. n_i, $\bar{x}_i$, S_i, S, g, and v are defined in the Appendix.

for each study. In addition, data on the study characteristics such as age, gender, racial characteristics, and severity of conditions should also be extracted. More is said about the data and the analysis in the statistical analysis section below.

Data abstract forms need to be generated for this task, and the research staff should be trained for uniformity. When extracting data it is usually better to extract raw data such as counts rather than proportions. Also it is very useful to have two people extract the data independently and to compare the results.

In abstracting data the investigators must be certain that subjects are not counted more than once. At times, multiple articles are produced on a study and results on the same subjects may be included in more than one article. Another potential problem relates to incompleteness of the data. Complete data are often missing from publications. Effort should be made to obtain all relevant data from all studies deemed to be fit for the analysis.

Any evaluation of a meta-analysis must consider the rigor, completeness, and success of the data extraction.

Statistical issues and homogeneity. After the data are extracted and the appropriate summary statistics (treatment effects or epidemiologic associations) are computed from each study, the statistical analysis begins. The first analysis task is to decide if the treatment effects or epidemiologic associations are *homogeneous,* meaning that the same or similar results are obtained from the studies. This usually is done by a formal test of homogeneity[12,28,29] in conjunction with a subjective appraisal of the consistency across studies.[12] We illustrate this below with an example on the use of antihistamines in common cold preparations. The Appendix contains formulas relating to this homogeneity test and other statistical analysis issues.

If the study treatment effects (or epidemiologic associations) are not homogeneous, a decision must be made to analyze the subset(s) that does display homogeneity or to analyze all the studies jointly and attempt to take into account the lack of homogeneity in later analyses. The random effects analyses procedures[2,3,12] attempt to do this. Another approach is to examine the treatment effects (or outcome measures) as they may relate to study variables such as severity of the patient populations or the dose levels in the particular studies.

It is appropriate to view a meta-analysis as a multicenter trial and the test of homogeneity as a test for interactions of the response with study sites. If there is a lack of homogeneity in a meta-analysis (or the presence of an interaction in a multicenter trial) then the investigator must proceed with caution. All conclusions may depend on characteristics of the study sites or subjects within them. Generalizations about effects may be severely limited. As with the test of interaction in a multicenter trial, the test for homogeneity in a meta-analysis is usually not a powerful test and its results should be reinforced with informal graphical analysis that displays homogeneity across studies. See the example below for further details.

Statistical issues and summary measure of effect. If the studies are judged to be relatively homogeneous with respect to treatment effect (or epidemiologic association), a summary measure or average of the treatment effect (or epidemiologic association) can be derived by use of appropriate pooling techniques. Mathematical formulas are given for some of this material in the Appendix. A common procedure is to compute this pooled average, test it for statistical significance, and/or compute a 95% confidence interval from it. These procedures are illustrated in detail in the example given below.

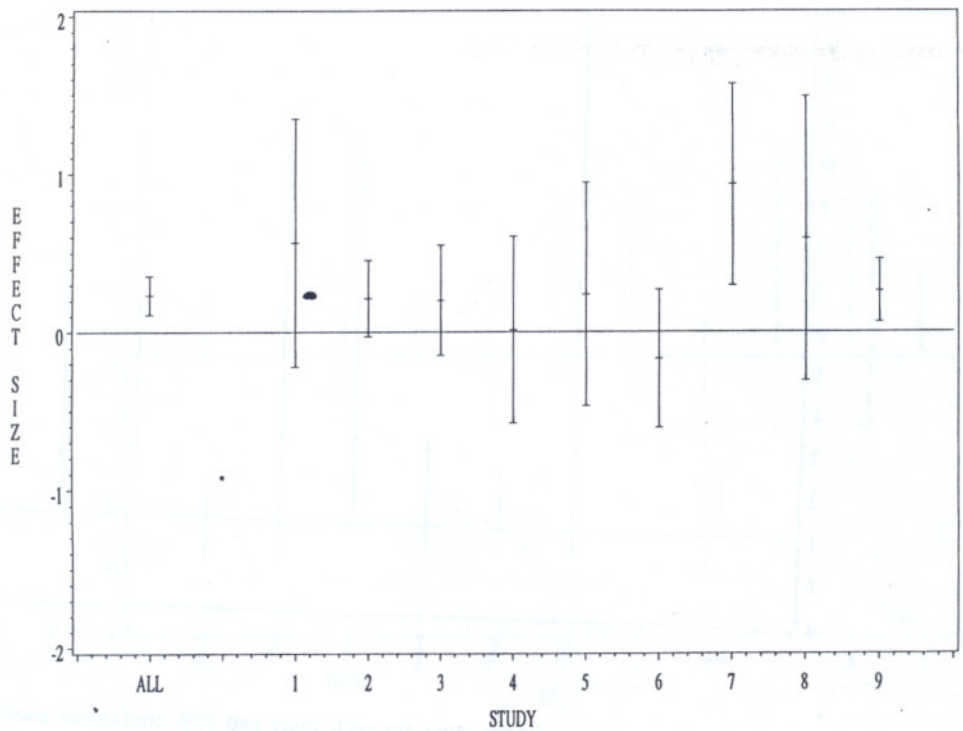

Fig. 1. Graphical comparison of effect sizes for each study and 95% confidence intervals for all studies: runny nose day 1 average increment.

In the evaluation of a meta-analysis it is essential to judge the analysis of homogeneity and the procedures used to produce and analyze statistically the pooled summary measure.

Sensitivity analysis. Because a meta-analysis involves some subjective components, it is important to judge whether the results are sensitive to changes in the procedures undertaken. One such sensitivity analysis is to analyze first all studies together, then just the published ones, and then solely the nonpublished studies. Another sensitivity analysis is to perform separate meta-analyses on other subsets of studies where the subsets have some common features such as the severity of disease or length of follow-up. Ideally, the sensitivity analyses will not produce different results and thus will justify the conclusions of the original analysis on all the studies. However, if the sensitivity analysis demonstrates major differences, then the causes need to be identified.

EXAMPLE

We now present an example of a meta-analysis. It is presented solely for the purpose of illustrating meta-analysis procedures. We plan for another publication a more detailed and complete meta-analysis of the efficacy issue involved.

The Food and Drug Administration's (FDA) Over-the-Counter Drug Division established a task force to investigate the effectiveness of antihistamines in common cold preparations for reducing the severity of runny nose and sneezing. After reviewing the literature and seeking expert advice, the task force decided that a meta-analysis should be performed on studies or portions of studies that satisfied the following inclusion criteria: (1) study must be double blind, randomized, and placebo controlled, (2) the antihistamine in the common cold medication must be a single ingredient, (3) the common cold had to exist for no more than 2 days before the first application of study medication, (4) patients needed to have a runny nose of at least moderate intensity at baseline (that is, before any study medication); this was measured on an ordinal scale with at least four categories, ranging from no symptom to severe (i.e., very uncomfortable or blocked), and (5) the severity of the runny nose had to be evaluated at baseline and at least once after administration of medication during both the first and second days of medication.

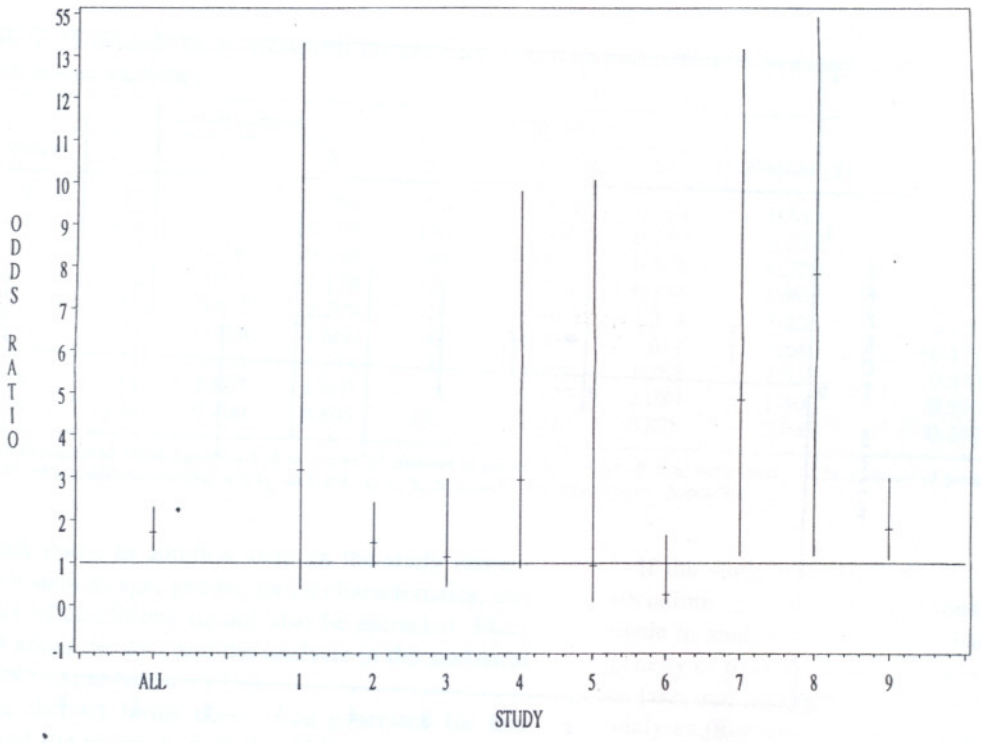

Fig. 2. Graphical comparison of odds ratios for each study and 95% confidence intervals for all studies: runny nose day 1 goal of therapy.

Table II. Meta-analysis on goal of therapy (50% reduction in runny nose severity in day 1; analysis on a dichotomous variable)

Study	*Antihistamine*				*Placebo*				*Log odds by equation 9*	*Odds ratio*
	n_1	c_{11}	c_{21}	p_1*	n_2	c_{12}	c_{22}	p_2*		
1	11	3	8	0.273	12	1	11	0.083	1.150	3.157
2	128	54	74	0.422	136	45	91	0.331	0.386	1.471
3	63	10	53	0.159	64	9	55	0.141	0.137	1.147
4	22	13	9	0.591	22	7	15	0.318	1.077	2.937
5	16	1	15	0.062	15	1	14	0.067	−0.067	0.936
6	39	1	38	0.026	41	5	36	0.122	−1.353	0.259
7	21	10	11	0.476	21	3	18	0.143	1.574	4.826
8	13	10	3	0.769	8	2	6	0.250	2.054	7.800
9	194	47	147	0.242	193	29	164	0.150	0.585	1.796

Test of homogeneity (from equation 14), χ^2 statistics of equation 14 = 12.60, df = 8 and p level = 0.13. Estimate of pooled log(odds ratio) of equation 11 is 0.538; pooled odds ratio is 1.713; 95% confidence interval for the pooled odds ratio is 1.275 to 2.301. n and c values are defined in section 3 of the Appendix.
*p_i, proportion achieving the goal of therapy.

Nine studies were identified for the meta-analysis. These constituted all the published *and* unpublished studies. The raw data were made available to the FDA for all studies. All computations were performed on these raw data. After review, all studies were considered to be of good quality for a meta-analysis.

For the example reported here, efficacy was the reduction in the severity of runny nose and was based on change from baseline achieved by the end of day 1. Two meta-analyses are presented. Formulas for these are given in the Appendix. The first is for the continuous variable "incremental change from baseline." For each of the nine studies the mean incremental change and standard deviation (SD) were computed. From these, the difference in mean values of the antihistamine subjects and the placebo subjects divided by the

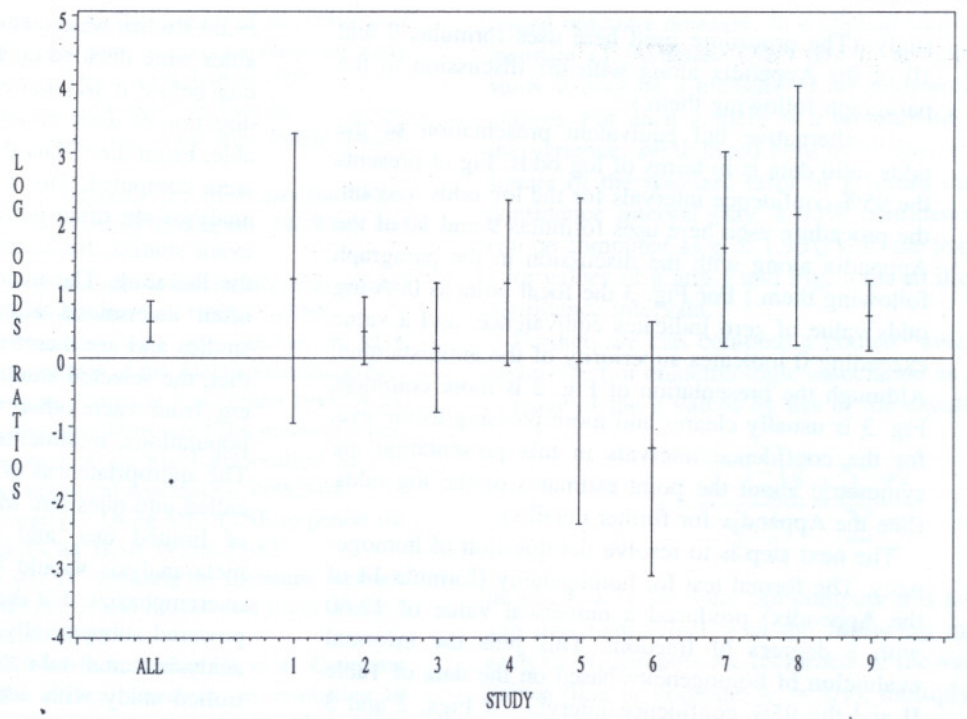

Fig. 3. Graphical comparison of the natural log of odds ratios for each study and 95% confidence intervals for all studies: runny nose day 1 goal of therapy.

pooled SD is computed to produce a quantity represented in Table I by *g* and often called "the treatment effect":

$$g = \frac{\bar{x}_1 - \bar{x}_2}{S}$$

(See formulas 6 to 8 of the Appendix for more detail.) The value g is a standardized difference of the treatment means. Values of the order of magnitude 0.2, 0.5, and 0.8 are often considered to be small, moderate, and large, respectively. Table I contains the numerical values for these treatment effects and Fig. 1 contains the 95% confidence intervals. If there were no difference between antihistamine and placebo we would expect the confidence intervals to include zero.

As stated above, the first step in the meta-analysis is to compute a formal statistical test of homogeneity. (This is achieved by use of the χ^2 test given by formula 3 as described in the Appendix. The value of this statistic is 9.84 with 8 degrees of freedom, which is not statistically significant.) This test does not give any indication of lack of homogeneity. This conclusion is reenforced by the important informal procedure of evaluating the consistency of the studies as seen by the numerical values of g in Table I and the 95% confidence intervals displayed in Fig. 1. All treatment effect estimates except 1 are larger than 0.

Given that the treatment effects display homogeneity, the next step is to estimate a pooled treatment effect. (This is achieved by use of formulas 6, 7, 8, 1, and 2 of the Appendix.) This pooled treatment effect and its 95% confidence interval are 0.234 and 0.111 to 0.358, respectively. (Formula 5 of the Appendix is used for the confidence interval.) The value 0 is not in the confidence interval, implying that we have statistically significant evidence that the antihistamine does significantly reduce the severity of the runny nose condition better than the placebo.

The second meta-analysis reported here is on the dichotomous variable "attaining goal of therapy for day 1." This goal is attained basically if the severity of the runny nose is reduced by at least 50% by the end of day 1. Table II shows, for each study, the proportion of placebo- and antihistamine-treated subjects who achieved the goal of therapy. For each study, the odds ratio was computed to evaluate the difference between the antihistamine and the placebo. A value of 1 indicates equivalence, and a value exceeding 1 indicates superiority of the antihistamine. Fig. 2 contains the 95% confidence intervals for the odds ratios for each

study. (The procedure used here uses formulas 9 and 10 of the Appendix along with the discussion in the paragraph following them.)

An alternative but equivalent presentation of the odds ratio data is in terms of log odds. Fig. 3 presents the 95% confidence intervals for the log odds. (Again, the procedure used here uses formulas 9 and 10 of the Appendix along with the discussion in the paragraph following them.) For Fig. 3 the focal point is 0. A log odds value of zero indicates equivalence, and a value exceeding 0 indicates superiority of the antihistamine. Although the presentation of Fig. 2 is more common, Fig. 3 is usually clearer and more pleasing to the eye, for the confidence intervals in this presentation are symmetric about the point estimates of the log odds. (See the Appendix for further details.)

The next step is to resolve the question of homogeneity. The formal test for homogeneity (formula 14 of the Appendix) produced a numerical value of 12.60 with 8 degrees of freedom. This plus the informal evaluation of homogeneity based on the data of Table II and the 95% confidence intervals of Figs. 2 and 3 indicate homogeneity. The estimate of the pooled odds ratio and its 95% confidence interval are 1.713 and 1.275 to 2.301, respectively. Because the interval does not contain unity, we have significant evidence that the antihistamine-treated subjects have a significantly higher odds of attaining the goal of therapy than the placebo-treated subjects. (The appropriate formulas for computing the pooled odds ratio and its confidence interval are given by equations 11, 12, 13, 16, and 17 of the Appendix.)

For a sensitivity analysis, the above meta-analyses were performed on subsets of the studies. As would be expected, given the consistency of the data, all subset analyses produced results similar to the analyses reported above, confirming the effectiveness of the antihistamine in the common cold preparation.

DISCUSSION

A meta-analysis, as presented here, is a systematic structured review of studies that incorporates qualitative evaluation and quantitative statistical procedures to reach conclusions about an issue such as drug treatment efficacy or epidemiologic association. There are a number of important steps and considerations that must be taken to ensure a good meta-analysis. When a meta-analysis is performed, the steps discussed above should be carefully applied. In the evaluation of a meta-analysis, a review to judge the adherence and the success in applying the steps should be undertaken. A unique example was given above, in which (1) all relevant studies were identified, (2) treatment effect variables were decided on before examination of the data *and* before it was known what would be the size and direction of these effects, and (3) raw data were available. From these data the relevant statistics for pooling were computed. This is not the usual situation. Meta-analyses are often performed on only a subset of relevant studies, for example, on only those available in the literature. The inclusion and exclusion criteria are often determined after examination of the existing studies and are therefore prone to serious biases. Further, the selected studies are often substantially different from each other with respect to study designs, populations, treatments, and other important features. The appropriateness of such meta-analyses is seriously called into question. Many meta-analyses are probably of limited use, and skepticism when evaluating a meta-analysis should be the rule. Finally, we cannot overemphasize that the pooling of a number of underpowered suboptimally performed studies into a meta-analysis cannot take the place of a well-designed controlled study with adequate power to detect relevant differences.

References

1. Rosenthal R. Meta-analysis procedures for social research. Beverley Hills, California: Sage Publications, 1991.
2. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177-88.
3. Berlin JA, Laird N, Sacks HS, Chalmers TC. A comparison of statistical methods for combining event rates from clinical trials. Stat Med 1989;8:141-51.
4. L'Abbe KA, Detsky AS, O'Rouke K. Meta-analysis in clinical research. Ann Intern Med 1987;107:224-33.
5. Green BF, Hall JA. Quantitative methods for literature reviews. Annu Review Psychol 1984;35:37-53.
6. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analysis of randomized controlled trials. N Engl J Med 1987;316:450-5.
7. Hedges LV, Olkin I. Statistical methods for meta-analysis. Orlando, Florida: Academic Press, 1985.
8. Peto R. Why do we need systematic overviews of randomized trials? Stat Med 1987;6:233-40.
9. Chalmers TC, Levin R. Meta-analysis of clinical trials as a scientific discipline; I: control of bias and comparison with large co-operative trials. Stat Med 1987;6:315-25.
10. Greeland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987;9:1-30.
11. Dickerson K, Berlin JA. Meta-analysis: state-of-the-science. Epidemiol Rev 1992;14:154-76.
12. Fleiss JL. The statistical basis of meta-analysis. Stat Methods Med Res 1993;2:121-45.

13. Gelber RD, Goldhirsh A. The concept of an overview of cancer clinical trials with special emphasis on early breast cancer. J Clin Oncol 1986;4:1696-1703.
14. Louis T, Fienberg HV, Mosteller F. Findings for public health from meta-analyses. Annu Rev Public Health 1985;6:1-20.
15. Oakes M. Statistical inference: a commentary for the social and behavioural sciences. Chichester: JW Wiley and Sons, 1986.
16. Oakes M. The logic and role of meta-analysis in clinical research. Stat Methods Med Res 1993;2:147-60.
17. Huque MF. Experiences with meta-analysis in NDA submissions. Proceedings of the Biopharmaceutical Section of the American Statistical Association 1988:28-33.
18. Stein RA. Meta-analysis from one FDA reviewer's perspective. Proceedings of the Biopharmaceutical Section of the American Statistical Association 1988:34-8.
19. Hewitt P, Chalmers TC. Using MEDLINE to peruse the literature. Control Clin Trials 1985;6:75-83.
20. Hewitt P, Chalmers TC. Perusing the literature: methods of accessing MEDLINE and related data bases. Control Clin Trials 1985;6:168-75.
21. Hayes RB, McKibbon KA, Walker CJ, et al. Computer searching of the medical literature: an evaluation of MEDLINE search systems. Ann Intern Med 1985; 103:306-17.
22. Simes RJ. Confronting publication bias: a cohort design for meta-analysis. Stat Med 1987;6:11-30.
23. Peto R, Collins R, Gray R. Large-scaled randomized evidence: large, simple trials and overviews of trials. J Clin Epidemiol 1995;48:23-40.
24. Fleiss JL. Discussion contribution to Light RJ. Stat Med 1987;6:221-8.
25. D'Agostino RB, Kwan H. Measuring effectiveness: what to expect without a randomized control group. Med Care 1995;33:AS95-105.
26. Chalmers TC, Lau J. Meta-analytic stimulus for change in clinical trials. Stat Methods Med Res 1993;2:161-72.
27. Chalmers TC, Smith H Jr, Blackburn B, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials 1981;2:31-49.
28. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719-748.
29. Davis W, Breslow NE, Day NE. Statistical methods in cancer research; vol 1: the analysis of case-control studies. Lyon, France: IARC Sci Publ, 1980:32-63.
30. Yusuf S, Peto R, Lewis J, Collins R, Steight P. Beta blockage during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis 1985;27:335-71.

APPENDIX I: STATISTICAL METHODS

1. General procedure

Say we have K studies available for a meta-analysis. For an individual study, let *g* represent the estimated outcome measure. In a clinical trial g may be a measure of *treatment effect* and in an epidemiologic study it may be a measure of an *epidemiologic association*. For each g there is a measure of variability, the variance, symbolized here by *v*. This v is also the square of the standard error of g. From each study a confidence interval (say, a 95% confidence interval) can be computed as $g \pm 1.96\sqrt{v}$. In the text see Table I for values of g and v and Fig. 1 for an illustration of confidence intervals.

Further, we can compute a pooled "average" treatment effect or epidemiologic association as a weighted average of the g values by use of the equation:

$$\bar{g} = \frac{\sum \left(\frac{1}{v}\right) g}{\sum \frac{1}{v}} = \frac{\sum wg}{\sum w} \qquad (1)$$

in which w is 1/v. One can compute this pooled average treatment effect $\bar{g}$ in the following manner. For each study, compute the reciprocal of the variance of g values (that is, compute w = 1/v), multiply this by g (that is, compute wg) and then sum these terms over all the studies. Next, compute the sum of all the reciprocals of the variances (that is, sum the values of w). Now divide the first sum by the second sum. That is, divide the sum of the wg terms by the sum of the w terms. Table I supplies all the items needed to compute this average treatment effect for the antihistamine example.

The variance (squared standard error) of the average treatment effect is as follows:

$$\text{Var} = \text{Var}(\bar{g}) = \left(\sum \frac{1}{v}\right)^{-1} = \frac{1}{\sum w} \qquad (2)$$

In the computation of $\bar{g}$ we have already described the computation of the sum of the 1/v = w terms. Given this sum, the variance of $\bar{g}$, Var of (2), is simply the reciprocal of this sum. For example, for the data of Table I the sum of the w = 1/v terms is 257.021 and the variance, Var of equation 2, equals 1/257.021 = 0.00389. The reader should be able to verify this to within round-off error.

With the above in hand, the first step in the statistical analysis is to test whether the outcome measures, g, can be pooled legitimately. This requires a formal test of homogeneity with the χ^2 test statistic:

$$\chi^2 = \sum \left(\frac{1}{v}\right)(g - \bar{g})^2 = \sum w(g - \bar{g})^2 \qquad (3)$$

The computation of formula 3 requires four steps. First, compute the differences of the individual study

effects g from the average effect $\bar{g}$ (that is, compute $(g - \bar{g})$). Next, square each of these (that is, compute $(g - \bar{g})^2$). Third, multiply each of these by w from each study. Fourth, sum these terms.

For testing homogeneity the computed value of equation 3 is compared to the critical value obtained from the χ^2 distribution with $K - 1$ degrees of freedom, in which K is the number of studies. Nonsignificance of the test implies homogeneity of studies and the data are legitimate to pool. Formula 1 is one possible measure of the average or pooled treatment effect (or epidemiologic association). In all cases the graph of confidence intervals (such as given in Fig. 1) for the individual studies should be examined to informally judge homogeneity.

If the test of homogeneity indicates lack of homogeneity across studies, a random effects analysis may be appropriate. The reader is referred to the references for these.[2,3,12] Our preference is to understand why there is lack of consistency across the studies rather than to apply some artificial statistical model. See the comments in the above commentary for further details.

Next, a test for statistical significance of the pooled or average outcome measure can be performed by a z test with use of the equation:

$$z = \frac{\bar{g}}{\sqrt{\text{Var}}} \tag{4}$$

Formula 4 is a simple z statistic formed by dividing the average effect $\bar{g}$ of equation 1 by its standard error, which is the square root of Var of equation 2. This statistical test has a null hypothesis that there is no treatment effect (or epidemiologic association). If z exceeds 1.96 in absolute value, then there is statistically significant reason to say that there is a treatment effect (or epidemiologic association) at the 0.05 level of significance. Further, in place of equation 4 or in conjunction with it, a 95% confidence interval for the average outcome measure is as follows:

$$\bar{g} \pm 1.96\sqrt{\text{Var}} \tag{5}$$

Formula 5 is the usual formula for a confidence interval, where added and subtracted to the point estimate (here, $\bar{g}$ of equation 1) is 1.96 times the standard error of it.

2. Continuous data

If the outcome variable is continuous and there are two groups being compared, the individual study g values of equation 1 are usually presented as a standardized difference of mean values:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{S} \tag{6}$$

That is, for each individual study, g is the difference of sample means divided by S in which S is the pooled SD defined as follows:

$$S = \sqrt{\frac{(n_1 - 1)\,S_1^2 + (n_2 - 1)\,S_2^2}{n_1 + n_2 - 2}} \tag{7}$$

Formula 7 is the usual formula for the pooled SD used in the two-sample *t* test. Here, n_1 and n_2 are the sample sizes and S_1^2 and S_2^2 are the sample variances from the first and second samples, respectively.

The variance v for g in this setting is as follows:

$$v = \frac{g^2}{2\,(n_1 + n_2 - 2)} + \frac{n_1 + n_2}{n_1 n_2} \tag{8}$$

The pooled outcome measure (called the pooled effect size in this context) and its variance are now given directly by equations 1 and 2. The test for homogeneity is given by equation 3. A test of the null hypothesis that the effect size is zero is obtained from equation 4 and the 95% confidence interval from equation 5. A numerical example is given in the above commentary by use of formulas 6 to 8 for the incremental change in severity of runny nose from baseline to the end of day 1 on treatment. The data are given in Table I.

3. Dichotomous data

If the variable of interest is dichotomous, for example, achieving or not achieving the goal of therapy, the data can be displayed in a 2 × 2 table such as the following:

	T	P
EV	c_{11}	c_{12}
NEV	c_{21}	c_{22}
	n_1	n_2

Here, T can represent the treatment group in a clinical trial and P can represent the reference group, such as a placebo group. The symbol EV represents an event such as achieving goal of therapy and NEV represents not achieving the goal of therapy. One common outcome measure in such situations is the odds ratio.[3,12] In this case the g of equation 1 is usually given by an estimate of the log of the odds ratio (OR):

$$g = \log(OR) = \log\left[\frac{(c_{11} + 0.5)(c_{22} + 0.5)}{(c_{12} + 0.5)(c_{21} + 0.5)}\right] \quad (9)$$

in which log is the natural logarithm. The variance of g, v, in this setting are given by the following:

$$v = \text{Var}(\log(OR)) = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{(c_{ij} + 0.5)} \quad (10)$$

The logs are used because the log odds ratio is less influenced by outliers than the odds ratio directly; thus the analysis is more stable. The 0.5 values in formulas 9 and 10 adjust for 0 entries in the cells of the above 2 × 2 table.

Probably the best way to understand the computations required in equations 9 and 10 is by an example. Say we have the following 2 × 2 table:

	T	P
EV	5	10
NEV	95	90
	100	100

Formula 9 for the estimate of the log odds ratio is as follows:

$$g = \log(OR) = \log\left[\frac{(5 + 0.5)(90 + 0.5)}{(10 + 0.5)(95 + 0.5)}\right] = \log(0.4964) = -0.7003$$

and the estimate of the variance v is as follows:

$$v = \frac{1}{5 + 0.5} + \frac{1}{10 + 0.5} + \frac{1}{95 + 0.5} + \frac{1}{90 + 0.5} = 0.2985$$

Confidence intervals for the log odds ratios of the individual studies are computed as $g \pm 1.96\sqrt{v}$. The confidence intervals for the odds ratios are found by exponentiating these [namely, $\exp(g \pm 1.96\sqrt{v})$].

For the numerical example above the 95% confidence interval for the log odds ratio is as follows:

$$-0.7003 \pm 1.96\,(\sqrt{0.2985})$$

or

$$-0.7003 \pm 1.0709$$

which corresponds to the interval −1.7712 to 0.3705. The 95% confidence interval for the odds ratio is found by exponentiating the limits of this confidence interval for the log odds ratio. This produces exp(−1.7712) to exp(0.3705) or 0.170 to 1.449. Figs. 3 and 2, respectively, in the above commentary show 95% confidence intervals for the log odds ratios and the odds ratios for the antihistamine studies. Table II contains the raw data.

The pooled estimate and the test of homogeneity can now be obtained with use of formulas 1 to 3. The test for significance of the pooled estimate and the 95% confidence interval can be produced by use of formulas 4 and 5. These formulas refer to the *log odds ratio*. After the 95% confidence interval is produced, the confidence limits need to be exponentiated to obtain an interval for the odds ratio directly.

An alternative procedure for analysis of the pooled odds ratio is the Peto analysis[30] which uses the symbols of the above 2 × 2 table and proceeds as follows. The average measure (pooled log odds ratio [log(ÔR]) is:

$$\bar{g} = \log(\hat{OR}) = \frac{\sum(O - E)}{\sum V} \quad (11)$$

with a variance (square of the standard error) of:

$$\text{Var}(\log((\hat{OR})) = 1/(\sum V) \quad (12)$$

Here,

$$O = c_{11}$$

$$E = \left(\frac{n_1}{n_1 + n_2}\right)(c_{11} + c_{12})$$

$$V = E\left(\frac{n_2}{n_1 + n_2}\right)\left(\frac{c_{21} + c_{22}}{n_1 + n_2 - 1}\right) \quad (13)$$

For the table of numerical data given above:

$$O = 5$$

$$E = \frac{100}{100 + 100}(5 + 10) = 7.5$$

and

$$V = 7.5\frac{100}{(100 + 100)}\frac{(95 + 90)}{(100 + 100 - 1)} = 3.486$$

Further

$$O - E = 5 - 7.5 = -2.5$$

To compute formula 11, produce for each study the values of O, E, O − E, and V. The O − E values are summed and the V values are summed over all the studies. Then the value of equation 11 is computed as the ratio of these. The variance of equation 12 is the reciprocal of the sum of the V values. Note the value V in formulas 11 and 12 are not the same as the v values of formulas 1 and 2.

The test for homogeneity is given by the equation:

$$\chi^2 = \sum \frac{(O - E)^2}{V} - \frac{[\sum (O - E)]^2}{\sum V} \qquad (14)$$

in which the statistic is compared to the critical point in a χ^2 distribution with $K^* - 1$ degrees of freedom where K^* is equal to the number of non-zero values of V of equation 13. The test for significance of the pooled odds ratio is given by the following:

$$z = \frac{\log(\hat{OR})}{\sqrt{\text{Var}(\log(\hat{OR}))}} \qquad (15)$$

The reader should note that the z statistic of equation 15 is based on the estimates of the log of the odds ratios, where the numerator of equation 15 is the estimate defined by formula 11 and the denominator is the square root of the variance defined in formula 12.

The pooled odds ratio is estimated by the equation:

$$\text{pooled}(\hat{OR}) = \exp(\log(\hat{OR})) \qquad (16)$$

This is simply the exponential of the estimate of the pooled log odds ratio given by equation 11. For example, if the pooled log odds ratio is 0.88, then the estimate of the odds ratio is exp(0.88) = 2.40.

The 95% confidence interval for the pooled odds ratio is given by the following:

$$\exp[\log(\hat{OR}) \pm 1.96\sqrt{\text{Var}(\log(\hat{OR}))}] \qquad (17)$$

To compute this 95% confidence interval for the pooled odds ratio, we first calculate the 95% confidence interval for the pooled *log* odds ratio. We obtain this by adding and subtracting to the estimated log odds ratio of equation 11 1.96 times the square root of the variance from equation 12. Then we exponentiate the two limits from the confidence interval to get the upper and lower limits of the 95% confidence interval for the pooled odds ratio.

Other outcome measures can be used for dichotomous data, such as the difference between proportions and the relative risk. The reader is referred to the references for these and other discussions on related issues.

A meta-analysis that used formulas 9 to 17 on dichotomous data is presented in the above commentary for the goal of therapy variable in the antihistamine studies. The raw data are given in Table II, along with the results of some of the intermediary computations.