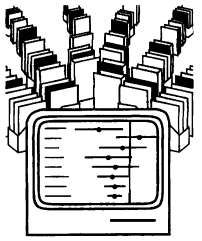
Meta-analysis and its problems

H J Eysenck



This paper was presented at a meeting on systematic reviews organised jointly by the BMJ and the UK Cochrane Centre and held in London in July 1993; it is the fourth in the series on systematic reviews

Institute of Psychiatry, London SE5 8AF
H J Eysenck, *professor emeritus of psychology, University of London*

BMJ 1994;309:789-92

Including all relevant material—good, bad, and indifferent—in meta-analysis admits the subjective judgments that meta-analysis was designed to avoid. Several problems arise in meta-analysis: regressions are often non-linear; effects are often multivariate rather than univariate; coverage can be restricted; bad studies may be included; the data summarised may not be homogeneous; grouping different causal factors may lead to meaningless estimates of effects; and the theory-directed approach may obscure discrepancies. Meta-analysis may not be the one best method for studying the diversity of fields for which it has been used.

Why do we undertake systematic reviews of a given field? The most important reason is perhaps that we are concerned about a particular theory and wish to know how the evidence for and against stacks up. There are also practical reasons; single studies often use small numbers of subjects, and basing our estimates of effect sizes on large numbers of studies drastically lowers the fiducial limits around our estimates. Systematic reviews can be of several different kinds: traditional reviews, often not very systematic, and frequently biased; meta-analyses, including (we hope) all relevant material, good, bad, and indifferent, and leading to an estimate of effect size[1-3]; best-evidence synthesis[4]; and the hypothetico-deductive approach,[5] in which the effort is directed at evaluating the evidence for and against a given theory, in an attempt to solve the problem of why contradictory results appear, rather than simply averaging often incompatible data.

Inclusiveness of meta-analysis

Critics may object to my statement that meta-analysis involves material good, bad, and indifferent, but consider the study by Smith *et al* (discussed in more detail later), which numbered among its authors the originator of the term.[6] The authors complained about the subjectivity that had led previous reviewers of studies assessing the effects of psychotherapy to exclude certain studies because of alleged design faults. This is what they claim to have done to overcome this subjectivity: "the method used is called meta-analysis. It is the statistical summary of the numerical outcomes of each study. We attempted to find and include all the controlled studies of psychotherapy outcome; that is, all the research in which one group of persons was treated for psychological conditions and compared with another, roughly equivalent untreated group. Studies were not excluded from consideration on arbitrary grounds; for example, because they used relatively inexperienced therapists or clients who had volunteered for the experiment, or had crude outcome measures. We suspected that contradictory conclusions of previous reviewers were largely the result of the arbitrary imposition of criteria for deciding which studies constituted valid evidence. These criteria had often been applied so as to favour a favourite hypothesis or vested ideological interest."

In other words, the crucial feature of this study was a statistical summary of the results of all relevant studies, however bad; indeed, later on the authors compared the results of what they considered good and bad studies, demonstrating that they were aware that many of the studies included were subject to damaging criticism. As noted by Knipschild, "it does not make sense to combine the top [best] with the other [not so good articles] and do a statistical precision ... meta-analysis."[7]

A good review is based on intimate personal knowledge of the field, the participants, the problems that arise, the reputation of different laboratories, the likely trustworthiness of individual scientists, and other partly subjective but extremely relevant considerations. Meta-analysis rules out any such subjective factors. It can be done by simply feeding the published results to a computer and coming up with an effect size. The computer avoids the bias of the subjective approach but simply adds together the biases of the authors of the original reports—which may or may not balance out.

I have pointed out problems that arise in the use of meta-analysis[5 8 9]; although the same problems may arise in connection with other methods of systematic reviewing, they are particularly likely to apply to meta-analysis. Let me list the problems that arise in this mechanical process.

Problems of meta-analytical process

REGRESSIONS ARE OFTEN NON-LINEAR

Glass and Smith carried out a meta-analysis of research on class size and achievement and concluded that "a clear and strong relationship between class size and achievement has emerged."[10] The study was done and analysed well; it might almost be cited as an example of what meta-analysis can do. Yet the conclusion is very misleading, as is the estimate of effect size it presents: "between class-size of 40 pupils and one pupil lie more than 30 percentile ranks of achievement." Such estimates imply a linear regression, yet the regression is extremely curvilinear, as one of the authors' figures shows: between class sizes of 20 and 40 there is absolutely no difference in achievement; it is only with unusually small classes that there seems to be an effect. For a teacher the major result is that for 90% of all classes the number of pupils makes no difference at all to their achievement. The conclusions drawn by the authors from their meta-analysis are formally correct, but they are statistically meaningless and particularly misleading. No estimate of effect size is meaningful unless regressions are linear, yet such linearity is seldom investigated, or, if not present, taken seriously. A simple traditional review would not have made such an obvious error.

EFFECTS ARE OFTEN MULTIVARIATE RATHER THAN UNIVARIATE

Consider the effect of passive smoking on lung cancer, where several meta-analyses (and best evidence analyses) have been conducted[11-13]; these all assume a univariate relation, although they come to quite disparate conclusions. Now consider the work of Janerich *et al*, who carried out a well planned study of the effects of passive smoking on lung cancer, and concluded that "the evidence we report lends further support to the observation that passive smoking may increase the risk of subsequent lung cancer."[14] In this study, individually matched pairs (lung cancer patients and healthy controls) were compared for exposure to cigarette smoke in four different situations. For overall exposure, "no clear dose-response relationship is evident," suggesting no overall effect. For exposure in childhood and adolescence there is an overall effect. For smoking by the spouse the most widely used measure, "there was little evidence of a trend according to amount of exposure." Exposure in the workplace indicated "no evidence of an adverse effect of environmental tobacco smoke." Finally, "our analysis of exposure in social settings...showed a statistically significant inverse association between environmental tobacco smoke and lung cancer."

What would a proper summary of this work be? It would emphasise the lack of overall effects, showing no clear dose-response relation; the negative health effects of childhood and adolescent passive smoking would be contrasted with the positive health effects of smoking in social situations; and the summary would also include the lack of effect of workplace or spouse smoking. The authors concentrate on the one result out of four that is negatively significant, forgetting that statistical significance for one selected test out of four cannot be calculated as if this were the only test done (there was no Bonferroni correction), and attempt to explain it by suggesting that during childhood and adolescence probands are more likely to be responsive to passive smoking, although "we know of no specific mechanism that would explain our findings." In other words, the explanation is purely ad hoc and adds nothing to the alleged findings. The authors fail to discuss the fact that "the difference in the magnitude of the effect between exposure during childhood and adolescence and exposure during adulthood did not achieve statistical significance," a finding that would seem to disprove their own hypothesis.

Can the "unexpected" protective effect of exposure to social smoking be explained? A likely hypothesis would suggest that extraverted personality traits seem to protect against cancer and that individuals prone to cancer have personality traits usually associated with introversion.[15] Extraverted people, however, are more likely than introverted ones to attend social functions and to be exposed to cigarette smoke there. The hypotheses would explain the alleged protective function of social smoking as an artefact; such protection is due to personality characteristics shared by socially active people and people not prone to cancer. The failure of Janerich *et al* to take into account any risk factors other than smoking accounts for their failure to explain their own findings. The positive relation between environmental tobacco smoke and lung cancer in childhood may be due to genetic factors linking parents and children.[15]

M F COOTER

Effect sizes summed over heterogeneous data can hardly be accorded any validity—yet such data can be cited as proving the value of treatment

Looking at this study from the point of view of simple meta-analysis, or even best evidence summary, we would simply note an overall failure of environmental tobacco smoke to be linked with lung cancer. The hypothetico-deductive approach, however, would single out the obvious contradiction between results for social smoking and smoking in childhood, try to explain them, and suggest that in further research factors like personality and genetics should be taken into account. Indeed, these factors have been shown to be important in the effects of smoking, and no study leaving out a consideration of genetics, personality, stress, etc, is worth summarising in a meta-analysis, or any other type of analysis, because it attempts a univariate type of analysis of a clearly multivariate problem.[15]

RESTRICTION OF COVERAGE

Meta-analysis always specifies the nature of the material to be included; in the case of passive smoking and health this would normally be studies comparing health records of individuals exposed or not exposed to passive smoking and suffering or not suffering from various diseases believed to be related to smoking. This enables us to obtain an estimate of the size of putative effects of exposure. Looking at the problem from the point of the hypothetico-deductive methodology, however, such a procedure leaves out vitally relevant evidence. I will give an example. Lee carried out a meta-analysis on about 100 studies in an effort to discover the extent of misclassification when smokers pretend to be non-smokers.[16] He found that in smoking cessation studies, percentages in excess of 15-20% were commonplace, ranging up to 40% of misclassified non-smokers. Again, the percentage of true smokers found among self reported non-smokers tended to be higher in studies of men and women with lung cancer than in studies of those without lung cancer. In general, Lee suggests that of self reported never smokers, 2·5% are actually current smokers and 10% have smoked in the past. These figures may seem to suggest a rather modest level of deception, but it is sufficient to cause Lee to conclude "that the epidemiologically observed association between passive smoking and lung cancer arose from bias due to misclassification of a proportion of smokers as non-smokers."

Clearly this suggestion is of vital relevance to any consideration of the theory that passive smoking causes (or is related to) lung cancer, yet the meta-analyses of this topic quoted already failed to consider it because the structure of meta-analysis is concerned with estimates of the size and significance of effect but not with the possible causes of the observed effect, which are always interpreted in terms of the original hypotheses involved without looking at evidence suggesting alternative interpretations. It is of course open to the investigator to step outside the rigid limitations of the meta-analysis format and add a discussion of alternative interpretations of the observed effect size, but this is strictly outside the rules imposed by meta-analysis and forms no part of its raison d'être.

QUALITY OF STUDIES

Proponents of meta-analysis pride themselves on the inclusiveness of the method, rejecting the notion that bad studies should be excluded as "subjective." Yet such evaluation is part and parcel of the special insight which the expert can bring to the discussion, and inclusion of bad studies may completely subvert the true outcome of a hypothetico-deductive analysis. Consider a small scale example. Schmale and Iker tested the theory that hopelessness was a predictor of cervical cancer, using a directed interviewing technique and obtaining very positive results.[17] They also administered the Minnesota multiphasic personality inventory and the Rorschach inkblot test, with completely negative results. A minuscule meta-analysis of these three sets of data (and meta-analysts encourage separate analysis of different measures of the independent variable) would show a very small effect size of doubtful significance. Yet the interview was the only procedure relevant to the theory; the tests used are all purpose instruments of doubtful reliability and validity. A hypothetico-deductive approach would say the study strongly supported the hypothesis when measures directed at the hypothesis were used to test it; both sets of test results are irrelevant (and would have been even if they had been positive).

This argument will be persuasive to anyone familiar with the critical literature concerning the Minnesota multiphasic personality inventory and the Rorschach test, yet how could a meta-analysis disregard these negative findings, other than by departing from its all inclusiveness and using what might look like subjective considerations? Undoubtedly, many investigators use multipurpose instruments like these tests to investigate a specific hypotheses for which they are quite unsuited, and negative results so achieved are usually included in meta-analyses of data allegedly relevant to the original hypotheses.

ADDING APPLES AND ORANGES

Meta-analysis is only properly applicable if the data summarised are homogeneous—that is, treatment, patients, and end points must be similar or at least comparable. Yet often there is no evidence of any degree of such homogeneity and plenty of evidence to the contrary. Consider again the study by Smith *et al* concerned with "the benefits of psychotherapy."[6] Summarising over 500 papers, these authors came to the conclusion that "psychotherapy is beneficial, consistently so and in many different ways. Its benefits are on a par with other expensive and ambitious interventions, such as schooling and medicine... the evidence overwhelmingly supports the efficacy of psychotherapy.... Psychotherapy benefits people of all ages as reliably as schooling educates them, medicine cures them, or business turns a profit." Many reviews have repeated these statements with approbation, relying on the objectivity of meta-analysis. I have expressed a contrary view related to an earlier publication by the same authors, categorising it as "an exercise in mega-silliness."[9] Why such an unparliamentary expression?

In the studies analysed by Smith *et al* neither treatments, nor patients, nor end points were remotely comparable. Patients could be severe neurotics, mild neurotics, students suffering from a specific phobic anxiety, or people suffering from some form of existentialist discomfort. Treatments were exceedingly varied; indeed, a table gives 18 different types of treatment. End points were equally diverse, consisting of elimination of objective symptoms, psychiatric opinion, answers to a questionnaire, or some projective test like the Rorschach inkblot test.

Effect sizes summed over such exceedingly heterogeneous data can hardly be accorded any validity, yet these data are often cited as proving the efficacy of psychotherapy. A proper analysis would note that many different theories are involved in the way diverse treatments are used (psychodynamic, Adlerian, client centred, Gestalt, rational-emotive, transactional, reality therapy, behaviour therapy, etc) and would also note that if it is true, as the authors suggest, that all have won and all must have prizes (that is, that all do about equally well), then surely all the theories involved must be wrong. Each would predict that only methods of treatment based on the theory proposed would have positive results, or at least would surely outperform all the others; failure to do so constitutes disproof of the theory in question. And when we note that one of the alleged methods of treatment is "placebo therapy" then we must surely conclude that the success of placebo therapy, equal to that of psychodynamic, client centred, Adlerian, Gestalt, rational-emotive, and other therapies, suggests that all the effect of psychotherapy is due to placebo effects. (Actually the effects of behaviour therapy are significantly greater than the effects of the psychotherapies mentioned, and this has also emerged from a meta-analysis of German studies,[18] so there may be some positive results.) But the resolute search for some general effect for psychotherapy appears fruitless; the data used are too heterogeneous to be analysed.

EFFECTS OF GROUPING

It is often the purpose of a meta-analysis to compare different causal factors (therapies, for example) in their effects. Thus Smith *et al*'s study compared behaviour therapy with different types of psychotherapy. In such a study all published accounts containing groups divided into no treatment and behaviour therapy, or no treatment and psychotherapy, would be compared. Doing so would pay no attention to the fact that different types of behaviour therapy are known to be appropriate for different types of symptoms. Thus desensitisation works with anxieties and phobias, while flooding with response prevention works well with obsessive-compulsive disorders; but vice versa, the results are quite disappointing.[19] Now consider a meta-analysis that would throw together studies that use the correct pairing with others that fail to do so. The resulting estimate of effect size would be strictly meaningless, averaging proper and improper uses of the method. Older studies were done before this difference in effectiveness became known; meta-analyses of (or including) older studies would come up with much lower estimates of effect size than summaries of more recent studies. Experts would know things like that, but not all authors of meta-analyses are experts, and in any case textbooks on meta-analysis do not give guidance on how to incorporate such knowledge in the carrying out of the analysis.

THE THEORY DIRECTED APPROACH

Concern with the truth or falsity of a given theory should make us look not at some measurement of effect size but rather at apparent anomalies and contradictions in the data, and at possible explanations of such contradictions. Consider some experiments reviewed elsewhere.[20] I had put forward a theory of introversion-extraversion which predicted that on a test of eyeblink conditioning introverts would perform much better than extraverts, and several experiments verified this prediction at a high level of significance. I also predicted that there would be no correlation with neuroticism-anxiety, and there was none. Very satisfactory.

At the same time Kenneth Spence, a well known psychologist at Iowa, predicted that neuroticism-anxiety would be related to quick and strong eyeblink conditioning, but introversion-extraversion would fail to show any relation to eyeblink conditioning. He too provided extensive evidence in favour of his theory. A typical meta-analysis would have shown very weak estimates of effect size for both our theories. Would that be a meaningful summary of the evidence? The obvious course to adopt would be to search for significant differences in the conduct of the experiment in an attempt to explain the observed differences in outcome.

Nothing in the published accounts provided a suggestion, but when an independent observer visited both laboratories the answer became clear. In our work subjects were reassured, told that there would be no electric shocks; the apparatus that might frighten them was carefully hidden and every effort was made to eliminate anything that might cause anxiety. As a consequence, differences in anxiety proneness (neuroticism) had no chance to emerge, and hence the predicted effects of high cortical arousal in introverts were observed. Spence, on the other hand, failed to reassure his subjects, and in fact made every effort to exploit their proneness to anxiety. Consequently, high and low scorers on neuroticism showed very different degrees of anxiety, and these swamped any differential effects of cortical arousal. Clarifying discrepancies is more important than averaging estimates of effect size over discrepant data; however, such averaging is what occurs in typical meta-analyses.

Summary

Newton wrote in a letter to Oldenburg in 1676: "For it is not number of Exp^{ts}, but weight to be regarded; where one will do, what need of many?" And Rutherford once pointed out that when you needed statistics to make your results significant, you would be better off doing a better experiment. Meta-analyses are often used to recover something from poorly designed studies, studies of insufficient statistical power, studies that give erratic results, and those resulting in apparent contradictions. Occasionally, meta-analysis does give worthwhile results, but all too often it is subject to methodological criticisms, some of which have been discussed above.

Careful workers can of course avoid these errors, but in doing so they will often violate the paradigms on which the whole notion of meta-analysis is built, and then will incur the accusation of subjectivity. Systematic reviews range all the way from highly subjective "traditional" methods to computer-like, completely objective counts of estimates of effect size over all published (and often unpublished) material regardless of quality. Neither extreme seems desirable. There cannot be one best method for fields of study so diverse as those for which meta-analysis has been used. If a medical treatment has an effect so recondite and obscure as to require meta-analysis to establish it, I would not be happy to have it used on me. It would seem better to improve the treatment, and the theory underlying the treatment.

[1] Hedges LV, Olkin I. *Statistical methods for meta-analysis.* New York: Academic Press, 1985.

[2] Huque MF. Experiences with meta-analysis in NDA submissions. *Proceedings of the Biopharmaceutical Section of the American Statistical Association* 1988;2:28-33.

[3] Spitzer WO. Meta-analysis: unanswered questions about aggregating data. *J Clin Epidemiol* 1991;44:103-7.

[4] Slavin RE. Best-evidence symptoms: an alternative to meta-analysis and traditional reviews. *Educational Research* 1986;15:9-11.

[5] Eysenck HJ. Meta-analysis: an abuse of research integration. *Journal of Special Education* 1984;18:41-59.

[6] Smith ML, Glass GV, Miller TI. *The benefits of psychotherapy.* Baltimore: Johns Hopkins Press, 1980.

[7] Knipschild P. Systematic reviews—some examples. *BMJ* 1994;309:719-21.

[8] Eysenck HJ. Meta-analysis: sense or non-sense? *Pharmaceutical Medicine* 1992;6:113-9.

[9] Eysenck HJ. An exercise in mega-silliness. *Am Psychol* 1978;33:517.

[10] Glass GV, Smith ML. Meta-analysis of research on class size and achievement. *Educational evolution and policy analysis* 1979;1:2-16.

[11] National Research Council. *Environmental tobacco smoke: measuring exposure and assessing health effects.* Washington: National Academy Press, 1986.

[12] 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, 439.

[13] Stein RA. Meta-analysis from one FOA reviewer's perspective. *Proceedings of the biopharmaceutical section of the American Statistical Association* 1988;2: 34-8.

[14] Janerich DT, Thompson WD, Varela LR, Greenwald P, Chorost S, Tucci C, *et al.* Lung cancer and exposure to tobacco smoke in the household. *N Engl J Med* 1990;323:632-6.

[15] Eysenck HJ. *Smoking, personality and stress: psychosocial factors in the prevention of cancer and coronary heart disease.* New York: Springer Verlag, 1991.

[16] Lee PN. *Misclassification of smoking habits and passive smoking.* New York: Springer Verlag: 1988.

[17] Schmale AH, Iker W. Hopelessness as a predictor of cervical cancer. *Soc Sci Med* 1971;5:95-160.

[18] Wittmann W, Matt G. Meta-analyse als Integration von Forschungsergebnissen am Beispiel deutschprachiger Arbeiten zur Effektivität von Psychotherapie. *Psychologische Rundschau* 1986;37:20-40.

[19] Eysenck HJ, Martin I. *Theoretical foundations of behavior therapy.* New York: Plenum Press, 1987.

[20] Eysenck HJ, ed. *A model for personality.* New York: Springer Verlag, 1981.

Lesson of the Week

Detection of bilateral isodense subdural haematomas

R J Davenport, P F X Statham, C P Warlow

Computed tomography may not show bilateral subdural haematomas, leading to a delay in diagnosis

Department of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU

R J Davenport, *registrar in medical neurology*

P F X Statham, *consultant neurosurgeon*

C P Warlow, *professor of medical neurology*

Correspondence to:
Dr Davenport.

BMJ 1994;309:792-4

Although cranial computed tomography is sensitive for detecting most subdural haematomas, those which are of the same density (isodense) as normal brain tissue may be difficult to identify in a scan. When the isodense haematoma is unilateral the presence of mass effect with a shift of the midline provides the diagnostic clue; with bilateral haematomas, however, the midline may remain undisturbed and the computed tomogram may be interpreted as either normal or showing generalised cerebral swelling of uncertain cause. We report on two patients with bilateral haematomas, both of whom were receiving anticoagulants, in whom the diagnosis was delayed even after computed tomography.

Case reports

CASE 1

A 71 year old man was admitted as an emergency with a two month history of progressive headache; coughing and straining exacerbated the pain but he had no nausea, vomiting, or disturbance of vision and his headache was not related to posture or time of day. He had a two year history of intermittent, non-specific headache, and an unenhanced cranial computed tomogram done 18 months previously had shown cortical atrophy only (figure (top)); no firm diagnosis had been reached at that time. He was taking warfarin for recurrent pulmonary emboli; six days before admission clotting tests showed an international normalised ratio of >6. On examination he had no abnormal physical signs; he had an erythrocyte sedimentation rate in the first hour of 32 mm and an international normalised ratio of 2·9. He was reviewed the following day by a consultant physician and discharged with a diagnosis of cervical spondylosis. Two days later he was readmitted with drowsiness and confusion. On examination he opened his eyes to speech, obeyed commands, and was oriented with no focal localising signs; he had an international normalised ratio of 2·5. A cranial computed tomogram with contrast enhancement (figure (bottom)) was reported by a radiology registrar as showing diffuse swelling of the brain of uncertain cause. A neurological opinion and a neurosurgical opinion were obtained; the scan was reviewed and reported as showing minimal effacement of the