http://bmj.com/collections/ma.htm

Meta-analysis

Education and debate

---

http://bmj.com/archive/7119/7119ed.htm

BMJ No 7119 Volume 315

Education and debate Saturday 22 November 1997

Meta-analysis

Potentials and promise

Matthias Egger, George Davey Smith

This is the first in a series of six articles examining the procedures in conducting reliable meta-analysis in medical research

Summary points
Well conducted meta-analysis allows for a more objective appraisal of the evidence, which may lead to resolution of uncertainty and disagreement
Meta-analysis may reduce the probability of false negative results and thus prevent undue delays in the introduction of effective treatments into clinical practice
Meta-analysis of a large number of individual studies or of individual patient data allows testing of a priori hypotheses regarding treatment effects in subgroups of patients
Heterogeneity between study results may be explored and sometimes explained
Promising research questions to be addressed in future studies may be generated, and the sample size needed in future studies may be calculated accurately

The number of papers published on meta-analyses in medical research has increased sharply in the past 10 years (fig 1). The merits and perils of the somewhat mysterious procedure of meta-analysis, however, continue to be debated in the medical community.(1-3) What, then, is meta-analysis? A useful definition was given by Huque: "A statistical analysis that combines or integrates the results of several independent clinical trials considered by the analyst to be 'combinable.'"

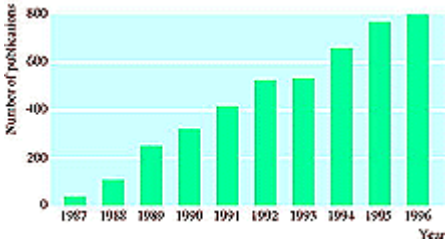


Fig 1: Number of publications about meta-analysis, 1987-96 (results from Medline search using text word and medical subject heading "meta-analysis")

(4) The terminology, however, is still debated, and expressions used concurrently include "overview," "pooling," and "quantitative synthesis." We believe that **the term meta-analysis should be used to describe the statistical integration of separate studies, whereas "systematic review" is most appropriate for denoting any review of a body of data that uses clearly defined methods and criteria** (box). **Systematic reviews can include meta-analyses, appraisals of single trials, and other sources of evidence**.(6) In this article we examine the potentials and promise of meta-analysis of randomised controlled trials. In later articles of this series we will consider the practical steps involved in meta-analysis,(7) examine various extensions beyond the calculation of a combined estimate,(8) address potential biases and discuss strategies to detect and minimise the influence of these in meta-analysis of randomised trials(9) and of observational studies.(10) We will conclude with a discussion of unresolved issues and future developments.(11) Details of relevant software will appear on the BMJ's website at the end of the series.

What's in a name? The case for "meta-analysis"

The term meta-analysis for statistically combining and analysing data from separate studies is appropriate because:

- The term makes sense. **"Meta"** implies something **occurring later, more comprehensive**, and is often used to name a new but related discipline designated to deal critically with the original one
- The alternative terms are less specific or less poignant - for example, "overview" is also used for traditional reviews, and "pooling" incorrectly implies that the source data are merged
- "Meta-analysis" has recently been included as a Medical Subject Heading (MeSH) and publication type within the Medline indexing system of the National Library of Medicine.(5)
- **"Systematic review" denotes any type of review that has been prepared using strategies to avoid bias and that which includes a material and methods section.(6) Systematic reviews may or may not include formal meta-analyses**
- "Meta-analysis" is a useful term for describing a possible component of systematic reviews, and distinguishing between the two terms contributes to methodological clarity

Historical notes

Efforts to pool results from separate studies are not new. In his account on the preventive effect of serum inoculations against enteric fever, statistician Karl Pearson, was in 1904 probably the first researcher reporting the use of formal techniques to combine data from different samples. The rationale put forward by Pearson for pooling studies is still one of the main reasons for undertaking meta-analysis today: "Many of the groups ... are far too small to allow of any definite opinion being formed at all, having regard to the size of the probable error involved." (12)

The first meta-analysis assessing the effect of a therapeutic intervention was published in 1955; interestingly, the treatment being evaluated was the placebo.(13) A simple average was calculated of the effectiveness of placebos in such diverse conditions as postoperative wound pain, cough, and angina pectoris: the placebo was apparently effective in 35% of patients. The development of more sophisticated statistical techniques, however, took place in the social sciences, in particular in education research, in the 1970s. The term meta-analysis was coined in 1976 by the psychologist Glass.(14) Meta-analysis was rediscovered by medical researchers to be used mainly in randomised clinical trial research, particularly in the fields of cardiovascular disease,(15) oncology,(16) and perinatal care.(17) Meta-analysis of observational studies(18) and "cross design synthesis" (the integration of observational data with the results from meta-analyses of randomised clinical trials(19,20)) have also been advocated.

More recently, a network of clinicians, epidemiologists, and other health professionals has been set up. The Cochrane Collaboration (named after Archie Cochrane, a pioneer in the field of evaluation of medical interventions) aims to prepare, maintain and disseminate comprehensive and systematic reviews of the effects of health care.(21,22) Since the foundation of the Cochrane Centre in Oxford in October 1992, this initiative has been growing rapidly, with the foundation of 15 further centres in Europe, North and Latin America, Africa, and Australia and hundreds of individuals from all over the globe collaborating in review groups.

The unacceptable face of "statisticism"?

Despite its widespread use, meta-analysis continues to be a controversial technique. While some exponents feel that meta-analysis should "replace traditional review articles of single topic issues whenever possible,"(23) others think of it as a "a new bête noir," which represents "the unacceptable face of statisticism" and "should be stifled at birth." (24) This mixed reception is not surprising. **The pooling of results from a particular set of studies may be inappropriate from a clinical point of view, producing a population " average" effect, while the clinician wants to know how to best treat his or her particular patient**. **Meta-analyses of the same issue may reach opposite conclusions**, as shown by assessments of low molecular weight heparin in the prevention of perioperative thrombosis(25,26) and of second line antirheumatic drugs in the treatment of rheumatoid arthritis.(27,28) It is nevertheless clear that for maximum benefit to be obtained from prior research, sound reviewing strategies must become more accessible and highly valued.

Narrative reviews

Traditional narrative reviews have several disadvantages that meta-analyses appear to overcome. The classic review is subjective and therefore prone to bias and error.(29) Without guidance by formal rules, reviewers can disagree about issues as basic as what types of studies it is appropriate to include and how to balance the quantitative evidence they provide. **Selective inclusion of studies that support the author's view is common**: the frequency of citation of clinical trials is related to their outcome, with studies in line with the prevailing opinion being quoted more frequently than unsupportive studies.(30,31) **Once a set of studies has been assembled, a common way to review the results is to count the number of studies supporting various sides of an issue and to choose the view receiving the most votes. This procedure is clearly unsound, as it ignores sample size, effect size, and research design. It is thus hardly surprising that reviewers using traditional methods often reach opposite conclusions**(32) and **miss small, but potentially important, differences**.(33) Clinical medicine is riddled with controversies, with reviews often being commissioned to end an argument. However, in controversial areas the conclusions drawn from a given body of evidence may be associated more with the specialty of the reviewer than with the available data.(23) By integrating the actual evidence, meta-analysis allows a more objective appraisal, which can help to resolve uncertainties when the original research, classic reviews, and editorial comments disagree.

Limitations of a single study

A single study often cannot detect or exclude with certainty a modest, albeit clinically relevant, difference in the effects of two treatments. A trial may thus show no significant treatment effect when in reality such an effect exists - that is, it may produce a false negative result. In this case we are dealing with a **type II error**, whose probability of occurrence can be calculated for a given difference in treatment effect, study size, and significance level. Generally better recognised is the type I error - when a trial produces a significant difference due to chance - whose probability corresponds to the probability (P) value. An examination of clinical trials that reported no significant differences between experimental and control treatment has shown that type II errors in clinical research are common: for a clinically relevant difference in outcome, the a priori probability of missing this effect (given the trial size) was greater than 20% in 115 of the 136 trials examined.(34) The number of patients included in clinical trials is thus often inadequate, a situation that has changed little over recent years. In some cases, however, the required sample size may be

difficult to achieve. A drug that reduces the risk of death from myocardial infarction by 10% could, for example, delay many thousands of deaths each year in Britain alone. To detect such an effect with 90% certainty (that is, with a type II error of no more than 10%) over 10,000 patients in each treatment group would be needed.(35)

The meta-analytic approach seems to be an attractive alternative to such a large, expensive, and logistically problematic study. Data from patients in trials evaluating the same or a similar drug in several smaller, but comparable, studies are considered. In this way the necessary number of patients may be reached, and relatively small effects can be detected or excluded with confidence.

Meta-analysis can also contribute to considerations about the generalisability of study results. The findings of a particular study may be valid only for a population of patients with the same characteristics as those investigated in the trial. If many trials exist in different groups of patients, with similar results in the various trials, then it can be concluded that the effect of the intervention under study has some generality. By putting together all available data, meta-analyses are also better placed than individual trials to answer questions about whether an overall study result varies among subgroups - for example, among men and women, older and younger patients, or subjects with different degrees of severity of disease. As discussed later in this series,(8) these questions can be addressed in the analysis and often lead to insights beyond what is provided by the calculation of a single combined effect estimate.

Epidemiology of results

Meta-analysis thus not only consists of the combination of data but includes the epidemiological exploration and evaluation of results - the "epidemiology of results," whereby the findings of an original study replace the individual as the unit of analysis.(36) New hypotheses that were not posed in the single studies can thus be tested in meta-analyses. However, although the studies included may be controlled experiments, the meta-analysis itself is subject to many biases inherent in observational studies.(37) Meta-analysis can, nevertheless, lead to the identification of the most promising or urgent research question, and may permit a more accurate calculation of the sample sizes needed in future studies. This is illustrated by an early meta-analysis of four trials that compared different methods of monitoring the fetus during labour.(38) The meta-analysis led to the hypothesis that, compared with intermittent auscultation, continuous fetal heart monitoring reduced the risk of neonatal seizures. This hypothesis was subsequently confirmed in a single randomised trial of almost seven times the size of the four previous studies combined.(39)

A more transparent appraisal

One benefit of meta-analysis is that it renders an important part of the review process transparent. In traditional narrative reviews it is often not clear how the conclusions follow from the data examined. In an adequately presented meta-analysis readers should be able to replicate the quantitative component of the argument. To facilitate this, it is valuable if the data included in meta-analyses are either presented in full or made available to interested readers by the authors.

The increased openness required by meta-analysis leads to the replacement of unhelpful descriptors such as "no relation," "some evidence of a trend," "a weak relation," and "a strong relation," with reproducible numerical values.(40) Furthermore, performing a meta-analysis may lead to reviewers moving beyond the conclusions that authors present in the abstract of papers, to a thorough examination of the actual data. As research assistants cannot be sent away with file cards to return with abridged conclusions, Rosenthal has suggested that this will lead to a "decrease in the splendid detachment of the full professor"(40) - in other words to a stronger involvement of the reviewers in the individual study results. As meta-analysis becomes a standard procedure, however, the splendid detachment may soon be restored.

Cumulative meta-analysis

Cumulative meta-analysis is defined as the repeated performance of meta-analysis whenever a new trial becomes available for inclusion. Such cumulative meta-analysis can retrospectively identify the point in time when a treatment effect first reached conventional levels of significance. For example, Lau and colleagues showed that for the trials of intravenous streptokinase in acute myocardial infarction, a significant (P=0.01) combined difference in total mortality had been achieved by 1973(41) (fig 2). At that time, 2,432 patients had been randomised in eight small trials. The results of the subsequent 25 studies, which included the large GISSI-1 and ISIS-2 trials(42,43) and enrolled a total of 34,542 additional patients, reduced the significance level to P=0.001 in 1979, P=0.0001 in 1986, and to P0.00001 when the first very large trial appeared, narrowing the confidence intervals around an essentially unchanged estimate of about 20% reduction in the risk of death. Interestingly, at least one country licensed streptokinase for use in myocardial infarction before GISSI-1(42) was published, whereas many national authorities waited for this trial to appear, and some waited a further two years for the results of ISIS-2(43) (fig 2).

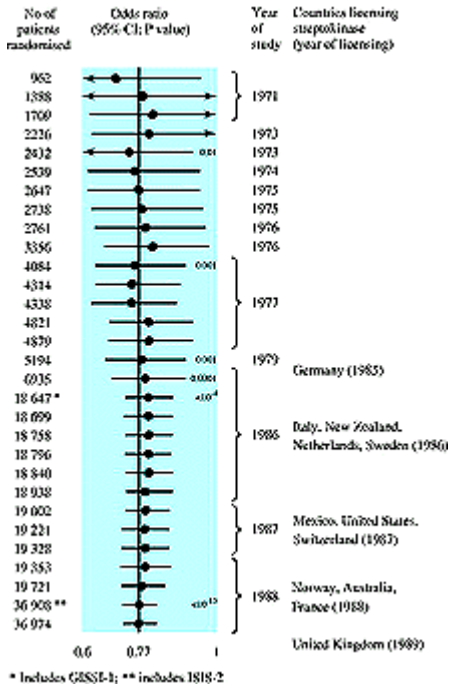


Fig 2: Cumulative meta-analysis of total mortality results from randomised controlled trials of intravenous streptokinase in myocardial infarction

A similar picture is apparent in the case of ß blockade in secondary prevention of myocardial infarction. In 1981 an influential editorial stated that "despite claims that they reduce arrhythmias, cardiac work, and infarct size, we still have no clear evidence that ß blockers improve long-term survival after infarction despite almost 20 years of clinical trials." (44) Retrospective cumulative meta-analysis, however, shows that a significant beneficial effect (P=0.02) was evident by 1977, and that the combined effect estimate was already both clinically important and highly significant (odds ratio 0.71 (95% confidence interval 0.59 to 0.84), P=0.0001) in 1981 (fig 3). Subsequent trials in a further 13,113 patients only confirmed this result.

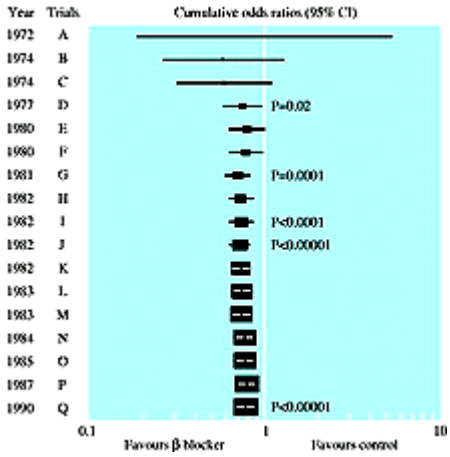


Fig 3: Cumulative meta-analysis of total mortality results from randomised controlled trials of oral ß blockers after myocardial infarction. The size of the square reflects the amount of statistical information available at a given point in time

Another application of cumulative meta-analysis has been to correlate the accruing evidence with the recommendations made by experts in review articles and textbooks. Antman and colleagues showed for thrombolytic drugs that recommendations for routine use first appeared in 1987, 14 years after a significant (P=0.01) beneficial effect became evident in cumulative meta-analysis.(45) Conversely, the prophylactic use of lidocaine continued to be recommended for routine use in myocardial infarction despite the lack of evidence for any beneficial effect and the possibility of a harmful effect being evident in the meta-analysis.

These examples have been taken to suggest that further studies in large numbers of patients may be at best superfluous and costly, if not unethical,(46) once a significant treatment effect is evident from meta-analysis of the existing smaller trials. There are several other examples, however, of meta-analyses showing benefit of statistical significance and clinical importance that were later contradicted by large randomised trials.(47) Meta-analysis clearly has advantages over conventional narrative reviews and carries considerable promise as a tool in clinical research and health technology assessment. Meta-analysis is not an infallible tool, however, as will be discussed later in this series.

We thank Dr T Johansson and G Enocksson (Pharmacia, Stockholm) and Drs A Schirmer and M Thimme (Behring, Marburg) for providing data on licensing of streptokinase in different countries. The department of social medicine at the University of Bristol is part of the Medical Research Council's health services research collaboration.

Funding: ME was supported by the Swiss National Science Foundation.

Department of Social Medicine,
University of Bristol,
Bristol BS8 2PR
Matthias Egger, reader in social medicine and epidemiology

George Davey Smith, professor of clinical epidemiology

Correspondence to: Dr Matthias Egger

email: m.egger@bristol.ac.uk

References

1 Naylor C D. Meta-analysis and the meta-epidemiology of clinical research. **BMJ** 1997;315:617-9.

2 Bailar J C. The promise and problems of meta-analysis [editorial]. **N Engl J Med** 1997;337:559-61.

3 Meta-analysis under scrutiny [editorial]. **Lancet** 1997;350:675.

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

5 Dickersin K, Berlin J A. Meta-analysis: state-of-the-science. **Epidemiol Rev** 1992;14:154-76.

6 Chalmers I, Altman D G. Foreword. In: Chalmers I, Altman D G, eds. **Systematic reviews**. London: BMJ Publishing, 1995.

7 Egger M, Davey Smith G, Phillips A N. Meta-analysis: principles and procedures. **BMJ** 1997 (in press).

8 Davey Smith G, Egger M, Phillips AN. Meta-analysis: beyond the grand mean? **BMJ** 1997 (in press).

9 Egger M, Davey Smith G. Meta-analysis: bias in location and selection of studies. **BMJ** 1997 (in press).

10 Egger M, Schneider M, Davey Smith G. Meta-analysis: spurious precision? Meta-analysis of observational studies. **BMJ** 1997 (in press).

11 Davey Smith G, Egger M. Meta-analysis: unresolved issues and future developments. **BMJ** 1997 (in press).

12 Pearson K. Report on certain enteric fever inoculation statistics. **BMJ** 1904;3:1243-6.

13 Beecher H K. The powerful placebo. **JAMA** 1955;159:1602-6.

14 Glass G. Primary, secondary and meta-analysis of research. **Educ Res** 1976;5:3-8.

15 Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. **Progr Cardiovasc Dis** 1985;17:335-71.

16 Early Breast Cancer Trialists' Collaborative Group. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. **N Engl J Med** 1988;319:1681-92.

17 Chalmers I, Enkin M, Keirse M. **Effective care during pregnancy and childbirth.** Oxford: Oxford University Press, 1989.

18 Greenland S. Quantitative methods in the review of epidemiologic literature. **Epidemiologic Reviews** 1987; 9:1-30.

19 General Accounting Office. **Cross design synthesis: a new strategy for medical effectiveness research**. Washington D.C. G.O.A. 1992.

20 Cross design synthesis: a new strategy for studying medical outcomes [editorial]? **Lancet** 1992;340:944-6.

21 Chalmers I, Dickersin K, Chalmers TC. Getting to grips with Archie Cochrane's agenda. **BMJ** 1992;305:786-8.

22 Bero L, Rennie D. The Cochrane Collaboration. Preparing, maintaining, and disseminating systematic reviews of the effects of health care. **JAMA** 1995;274:1935-8.

23 Chalmers T C, Frank C S, Reitman D. Minimizing the three stages of publication bias. **JAMA** 1990;263:1392-5.

24 Oakes M. **Statistical inference: a commmentary for the social and behavioural sciences.** Chichester: Wiley, 1986.

25 Nurmohamed M T, Rosendaal F R, Bueller H R, Dekker E, Hommes D W, Vandenbroucke J P, et al. Low-molecular-weight heparin versus standard heparin in general and orthopaedic surgery: a meta-analysis. **Lancet** 1992;340:152-6.

26 Leizorovicz A, Haugh M C, Chapuis F-R, Samama M M, Boissel J-P. Low molecular weight heparin in prevention of perioperative thrombosis. **BMJ** 1992;305:913-20.

27 Felson D T, Anderson J J, Meenan R F. The comparative efficacy and toxicity of second-line drugs in rheumatoid arthritis. **Arthritis Rheum** 1990;33:1449-61.

28 Götzsche P C, Podenphant J, Olesen M, Halberg P. Meta-analysis of second-line antirheumatic drugs: sample size bias and uncertain benefit. **J Clin Epidemiol** 1992;45:587-94.

29 Teagarden J R. Meta-analysis: whither narrative review? **Pharmacotherapy** 1989;9:274-84.

30 Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. **BMJ** 1992;305:15-9.

31 Götzsche P C. Reference bias in reports of drug trials. **BMJ** 1987;295:654-6.

32 Mulrow C D. The medical review article: state of the science. **Ann Intern Med** 1987;106:485-8.

33 Cooper H M, Rosenthal R. Statistical versus traditional procedures for summarising research findings. **Psychol Bull** 1980;87:442-9.

34 Freiman J A, Chalmers T C, Smith H, Kuebler R R. The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial. In: Bailar JC, Mosteller F, eds. **Medical uses of statistics**. Boston, MA: NEJM Books, 1992:357.

35 Collins R, Keech A, Peto R, Sleight P, Kjekshus J, Wilhelmsen L, et al. Cholesterol and total mortality: need for larger trials. **BMJ** 1992;304:1689.

36 Jenicek M. Meta-analysis in medicine. Where we are and where we want to go. **J Clin Epidemiol** 1989;42:35-44.

37 Gelber R D, Goldhirsch A. Interpretation of results from subset analyses within overviews of randomized clinical trials. **Stat Med** 1987;6:371-8.

38 Chalmers I. Randomised controlled trials of fetal monitoring 1973-1977. In: Thalhammer O, Baumgarten K, Pollak A, eds. **Perinatal medicine**. Stuttgart: Thieme, 1979:260.

39 MacDonald D, Grant A, Sheridan-Pereira M, Boylan P, Chalmers I. The Dublin randomised controlled trial of intrapartum fetal heart rate monitoring. **Am J Obstet Gynecol** 1985;152:524-39.

40 Rosenthal R. An evaluation of procedures and results. In: Wachter KW, Straf ML, eds. **The future of meta-analysis.** New York: Russel Sage Foundation, 1990:123.

41 Lau J, Antman E M, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. **N Engl J Med** 1992;327:248-54.

42 Gruppo Italiano per lo Studio della Streptochinasi nell'Infarto Miocardico (GISSI). Effectiveness of intravenous thrombolytic treatment in acute myocardial infarction. **Lancet** 1986;397-402.

43 ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. **Lancet** 1988;ii:349-60.

44 Mitchell J R A. Timolol after myocardial infarction: an answer or a new set of questions? **BMJ** 1981;282:1565-70.

45 Antman E M, Lau J, Kupelnick B, Mosteller F, Chalmers T C. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. **JAMA** 1992;268:240-8.

46 Murphy D J, Povar G J, Pawlson L G. Setting limits in clinical medicine. **Arch Intern Med** 1994;154:505-12.

47 Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. **BMJ** 1997;315:629-34.

---

http://bmj.com/archive/7121/7121ed.htm
BMJ No 7121 Volume 315

Education and debate Saturday 6 December 1997

Meta-analysis

Principles and procedures

Matthias Egger, George Davey Smith, Andrew N Phillips

This is the second in a series of seven articles examining the procedures in conducting reliable meta-analysis in medical research

Meta-analysis is a statistical procedure that integrates the results of several independent studies considered to be "combinable."(1) Well conducted meta-analyses allow a more objective appraisal

of the evidence than traditional narrative reviews, provide a more precise estimate of a treatment effect, and may explain heterogeneity between the results of individual studies.(2) Ill conducted meta-analyses, on the other hand, may be biased owing to exclusion of relevant studies or inclusion of inadequate studies.(3) Misleading analyses can generally be avoided if a few basic principles are observed. In this article we discuss these principles, along with the practical steps in performing meta-analysis.

Summary points
Meta-analysis should be as carefully planned as any other research project, with a detailed written protocol being prepared in advance
The a priori definition of eligibility criteria for studies to be included and a comprehensive search for such studies are central to high quality meta-analysis
The graphical display of results from individual studies on a common scale is an important intermediate step, which allows a visual examination of the degree of heterogeneity between studies
Different statistical methods exist for combining the data, but there is no single "correct" method
A thorough sensitivity analysis is essential to assess the robustness of combined estimates to different assumptions and inclusion criteria

Observational study of evidence

Meta-analysis should be viewed as an observational study of the evidence. The steps involved are similar to any other research undertaking: formulation of the problem to be addressed, collection and analysis of the data, and reporting of the results. Researchers should write in advance a detailed research protocol that clearly states the objectives, the hypotheses to be tested, the subgroups of interest, and the proposed methods and criteria for identifying and selecting relevant studies and extracting and analysing information.

As with criteria for including and excluding patients in clinical studies, eligibility criteria have to be defined for the data to be included. Criteria relate to the quality of trials and to the combinability of treatments, patients, outcomes, and lengths of follow up. Quality and design features of a study can influence the results.(4,5) Ideally, researchers should consider including only controlled trials with proper randomisation of patients that report on all initially included patients according to the intention to treat principle and with an objective, preferably blinded, outcome assessment.(6) Assessing the quality of a study can be a subjective process, however, especially since the information reported is often inadequate for this purpose.(7) It is therefore preferable to define only basic inclusion criteria and to perform a thorough sensitivity analysis (see below).

The strategy for identifying the relevant studies should be clearly delineated. In particular, it has to be decided whether the search will be extended to include unpublished studies, as their results may systematically differ from published trials. As will be discussed in later articles, a meta-analysis that is restricted to published evidence may produce distorted results owing to such publication bias. For

locating published studies, electronic databases are useful,(8) but, used alone, they may miss a substantial proportion of relevant studies.(9,10) In an attempt to identify all published controlled trials, the Cochrane Collaboration has embarked on an extensive manual search of medical journals published in English and many other languages.(11) The Cochrane Controlled Trials Register(12) is probably the best single electronic source of trials; however, citation indices and the bibliographies of review articles, monographs, and the located studies should also be scrutinised.

A standardised record form is needed for data collection. It is useful if two independent observers extract the data, to avoid errors. At this stage the quality of the studies may be rated, with one of several specially designed scales.(13,14) Blinding observers to the names of the authors and their institutions, the names of the journals, sources of funding, and acknowledgments leads to more consistent scores.(14) This entails either photocopying papers, removing the title page, and concealing journal identifications and other characteristics with a black marker, or scanning the text of papers into a computer and preparing standardised formats.(15,16)

Standardised outcome measure

Individual results have to be expressed in a standardised format to allow for comparison between studies. If the end point is continuous - for example, blood pressure - the mean difference between the treatment and control groups is used. The size of a difference, however, is influenced by the underlying population value. An antihypertensive drug, for example, is likely to have a greater absolute effect on blood pressure in overtly hypertensive patients than in borderline hypertensive patients. Differences are therefore often presented in units of standard deviation. If the end point is binary - for example, disease versus no disease, or dead versus alive) then odds ratios or relative risks are often calculated (box). The odds ratio has convenient mathematical properties, which allow for ease in combining data and testing the overall effect for significance. Absolute measures, such as the absolute risk reduction or the number of patients needed to be treated to prevent one event,(17) are more helpful when applying results in clinical practice (see below).

Statistical methods for calculating overall effect

The last step consists in calculating the overall effect by combining the data. A simple arithmetic average of the results from all the trials would give misleading results. The results from small studies are more subject to the play of chance and should therefore be given less weight. Methods used for meta-analysis use a weighted average of the results, in which the larger trials have more influence than the smaller ones. The statistical techniques to do this can be broadly classified into two models,(18) the difference consisting in the way the variability of the results between the studies is treated. The "fixed effects" model considers, often unreasonably, that this variability is exclusively due to random variation.(19) Therefore, if all the studies were infinitely large they would give identical results. The "random effects" model(20) assumes a different underlying effect for each study and takes this into consideration as an additional source of variation, which leads to somewhat wider confidence intervals than the fixed effects model. Effects are assumed to be randomly distributed, and the central point of this distribution is the focus of the combined effect estimate. Although neither of two models can be said to be "correct," a substantial difference in the combined effect calculated by the fixed and random effects models will be seen only if studies are markedly heterogeneous.(18)

Bayesian meta-analysis

Some statisticians feel that other statistical approaches are more appropriate than either of the above. One approach uses Bayes's theorem, named after an 18th century English clergyman.(21) Bayesian statisticians express their belief about the size of an effect by specifying some prior probability distribution before seeing the data, and then they update that belief by deriving a posterior probability distribution, taking the data into account.(22) Bayesian models are available under both the fixed and random effects assumption.(23) The confidence interval (or more correctly in bayesian terminology, the 95% credible interval, which covers 95% of the posterior probability distribution) will often be wider than that derived from using the conventional models because another component of variability, the prior distribution, is introduced. Bayesian approaches are controversial because the definition of prior probability will often be based on subjective assessments and opinion.

Heterogeneity between study results

If the results of the studies differ greatly then it may not be appropriate to combine the results. How to ascertain whether it is appropriate, however, is unclear. One approach is to examine statistically the degree of similarity in the studies' outcomes - in other words, to test for heterogeneity across studies. In such procedures, whether the results of a study reflect a single underlying effect, rather than a distribution of effects, is assessed. If this test shows homogeneous results then the differences between studies are assumed to be a consequence of sampling variation, and a fixed effects model is appropriate. If, however, the test shows that significant heterogeneity exists between study results then a random effects model is advocated. A major limitation with this approach is that the statistical tests lack power - they often fail to reject the null hypothesis of homogeneous results even if substantial differences between studies exist. Although there is no statistical solution to this issue, heterogeneity between study results should not be seen as purely a problem for meta-analysis - it also provides an opportunity for examining why treatment effects differ in different circumstances. Heterogeneity should not simply be ignored after a statistical test is applied; rather, it should be scrutinised, with an attempt to explain it.(24)

Graphic display

Results from each trial are usefully graphically displayed, together with their confidence intervals. Figure 1 represents a meta-analysis of 17 trials of ß blockers in secondary prevention after myocardial infarction. Each study is represented by a black square and a horizontal line, which correspond to the point estimate and the 95% confidence intervals of the odds ratio. The 95% confidence intervals would contain the true underlying effect in 95% of the occasions if the study was repeated again and again. The solid vertical line corresponds to no effect of treatment (odds ratio 1.0). If the confidence interval includes 1, then the difference in the effect of experimental and control treatment is not significant at conventional levels (P>0.05). The area of the black squares reflects the weight of the study in the meta-analysis. The confidence interval of all but two studies cross this line, indicating that the effect estimates were non-significant (P>0.05).

The diamond represents the combined odds ratio, calculated using a fixed effects model, with its 95% confidence interval. The combined odds ratio shows that oral ß blockade starting a few days to a few weeks after the acute phase reduces subsequent mortality by an estimated 22% (odds ratio 0.78; 95% confidence interval 0.71 to 0.87). A dashed line is plotted vertically through the combined odds ratio. This line crosses the horizontal lines of all individual studies except one (N). This indicates a fairly homogenous set of studies. Indeed, the test for heterogeneity gives a non-significant P value of 0.2.

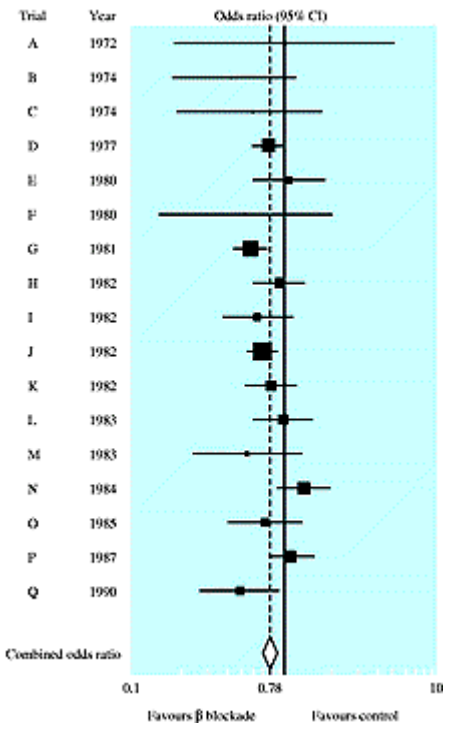


Fig 1: Total mortality from trials of ß blockers in secondary prevention after myocardial infarction. The black square and horizontal line correspond to odds ratio and 95% confidence interval for each trial. The size of the black square reflects the weight of each trial. The diamond represents the combined odds ratio and 95% confidence interval, showing 22% a reduction in the odds of death (references are available from the authors)

A logarithmic scale was used for plotting the odds ratios in figure 1. There are several reasons that ratio measures are best plotted on logarithmic scales.(25) Most importantly, the value of an odds ratio and its reciprocal - for example, 0.5 and 2 - which represent odds ratios of the same magnitude but opposite directions, will be equidistant from 1.0. Studies with odds ratios below and above 1.0 will take up equal space on the graph and thus look equally important. Also, confidence intervals will be symmetrical around the point estimate.

Relative and absolute measures of effect

Repeating the analysis by using relative risk instead of the odds ratio gives an overall relative risk of 0.80 (95% confidence interval 0.73 to 0.88). The odds ratio is thus close to the relative risk, as expected when the outcome is relatively uncommon (see box). The relative risk reduction, obtained by subtracting the relative risk from 1 and expressing the result as a percentage, is 20% (12% to 27%). The relative measures used in this analysis ignore the absolute underlying risk. The risk of death among patients who have survived the acute phase of myocardial infarction, however, varies

widely.(26) For example, among patients with three or more cardiac risk factors the probability of death at two years after discharge ranged from 24% to 60%.(26) Conversely, two year mortality among patients with no risk factors was less than 3%. The absolute risk reduction or risk difference reflects both the underlying risk without treatment and the risk reduction associated with treatment. Taking the reciprocal of the risk difference gives the "number needed to treat" (the number of patients needed to be treated to prevent one event).(17)

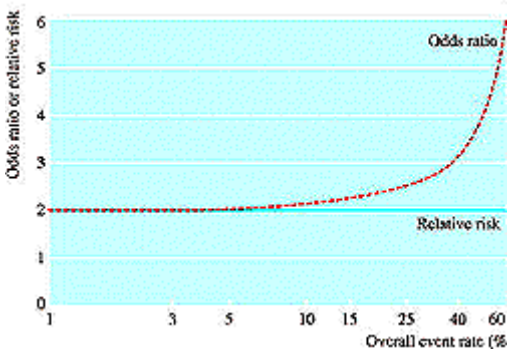
Odds ratio or relative risk?

Odds and odds ratio

The odds is the number of patients who fulfil the criteria for a given endpoint divided by the number of patients who do not. For example, the odds of diarrhoea during treatment with an antibiotic in a group of 10 patients may be 4 to 6 (4 with diarrhoea divided by 6 without, 0.66); in a control group the odds may be 1 to 9 (0.11) (a bookmaker would refer to this as 9 to 1). The odds ratio of treatment to control group would be 6 (0.66d0.11).

Risk and relative risk

The risk is the number of patients who fulfil the criteria for a given end point divided by the total number of patients. In the example above the risks would be 4 in 10 in the treatment group and 1 in 10 in the control group, giving a risk ratio, or relative risk, of 4 (0.4 divided by 0.1).



The odds will be close to the relative risk if the end point occurs relatively infrequently, say in less than 20%. If the outcome is more common (as in the diarrhoea example) then the odds ratio will considerably overestimate the relative risk

For a baseline risk of 1% a year, the absolute risk difference shows that two deaths are prevented per 1,000 patients treated (table). This corresponds to 500 patients (1 divided by 0.002) treated for one year to prevent one death. Conversely, if the risk is above 10%, less than 50 patients have to be treated to prevent one death. Many clinicians would probably decide not to treat patients at very low risk, given the large number of patients that have to be exposed to the adverse effects of ß blockade to prevent one death. Appraising the number needed to treat from a patient's estimated risk without treatment and the relative risk reduction with treatment is a helpful aid when making a decision in an individual patient. A nomogram that facilitates calculation of the number needed to treat at the bedside has recently been published.(27)

ß Blockade in secondary prevention after myocardial infarction - absolute risk reductions and numbers needed to treat for one year to prevent one death for different levels of mortality in control group		
One year mortality risk among controls (%)	**Absolute risk reduction**	**No needed to treat**
1	0.002	500
3	0.006	167
5	0.01	100
10	0.02	50
20	0.04	25
30	0.06	17
40	0.08	13
50	0.1	10
Calculations assume a constant relative risk reduction of 20%.		

Meta-analysis using absolute effect measures such as the risk difference may be useful to illustrate the range of absolute effects across studies. The combined risk difference (and the number needed to treat calculated from it) will, however, be essentially determined by the number and size of trials in patients at low, intermediate, or high risk. Combined results will thus be applicable only to patients at levels of risk corresponding to the average risk of the trials included. It is therefore generally more meaningful to use relative effect measures for summarising the evidence and absolute measures for applying it to a concrete clinical or public health situation.

Sensitivity analysis

Opinions will often diverge on the correct method for performing a particular meta-analysis. The robustness of the findings to different assumptions should therefore always be examined in a thorough sensitivity analysis. This is illustrated in figure 2 for the meta-analysis of ß blockade after myocardial infarction. Firstly, the overall effect was calculated by different statistical methods, by using both a fixed and a random effects model. The figure shows that the overall estimates are virtually identical and that confidence intervals are only slightly wider with the random effects model. This is explained by the relatively small amount of variation between trials in this meta-analysis.

Secondly, methodological quality was assessed in terms of how patients were allocated to active treatment or control groups, how outcome was assessed, and how the data were analysed.(6) The maximum credit of nine points was given if patient allocation was truly random, if assessment of vital status was independent of treatment group, and if data from all patients initially included were analysed according to the intention to treat principle. Figure 2 shows that the three low quality studies (7 points and under) showed more benefit than the high quality trials. Exclusion of these three studies, however, leaves the overall effect and the confidence intervals practically unchanged.

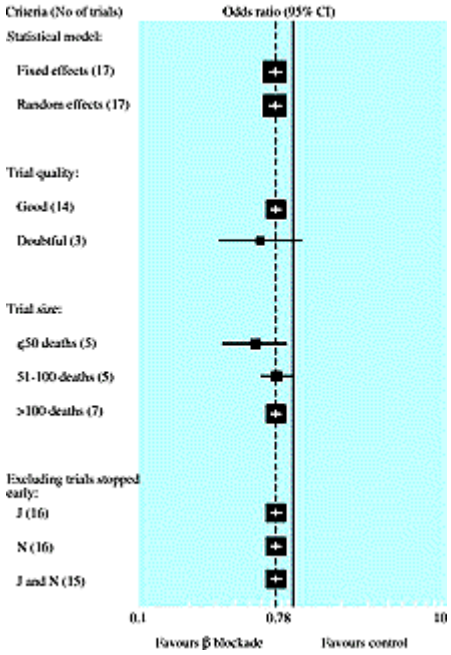


Fig 2: Sensitivity analysis of meta-analysis of |gb blockers in secondary prevention after myocardial infarction (see text for explanation)

Thirdly, significant results are more likely to get published than non-significant findings,(28) and this can distort the findings of meta-analyses. The presence of such publication bias can be identified by stratifying the analysis by study size - smaller effects can be significant in larger studies. If publication bias is present, it is expected that, of published studies, the largest ones will report the smallest effects. Figure 2 shows that this is indeed the case, with the smallest trials (50 or fewer deaths) showing the largest effect. However, exclusion of the smallest studies has little effect on the overall estimate.

Finally, two studies (J and N; see fig 1) were stopped earlier than anticipated on the grounds of the results from interim analyses. Estimates of treatment effects from trials that were stopped early are liable to be biased away from the null value. Bias may thus be introduced in a meta-analysis that includes such trials.(29) Exclusion of these trials, however, affects the overall estimate only marginally.

The sensitivity analysis thus shows that the results from this meta-analysis are robust to the choice of the statistical method and to the exclusion of trials of poorer quality or of studies stopped early. It also suggests that publication bias is unlikely to have distorted its findings.

Conclusions

Meta-analysis should be seen as structuring the processes through which a thorough review of previous research is carried out. The issues of completeness and combinability of evidence, which need to be considered in any review,(30) are made explicit. Was it sensible to have combined the individual trials that comprise the meta-analysis? How robust is the result to changes in assumptions? Does the conclusion reached make clinical and pathophysiological sense? Finally, has the analysis contributed to the process of making rational decisions about the management of patients? It is these issues that we explore further in later articles in this series.

The department of social medicine at the University of Bristol and the department of primary care and population sciences at the Royal Free Hospital School of Medicine, London, are part of the Medical Research Council's health services research collaboration.

Funding: ME was supported by the Swiss National Science Foundation.

Department of Social Medicine,
University of Bristol,
Bristol BS8 2PR
Matthias Egger, reader in social medicine and epidemiology
George Davey Smith, professor of clinical epidemiology

Department of Primary Care and Population Sciences,
Royal Free Hospital School of Medicine,
London NW3 2PF
Andrew N Phillips, professor of epidemiology and biostatistics

Correspondence to: Dr Egger

email: m.egger@bristol.ac.uk

References

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

2 Egger M, Davey Smith G. Meta-analysis: potentials and promise. **BMJ** 1997;315:1371-4.

3 Egger M, Davey Smith G, Schneider M, Minder C E. Bias in meta-analysis detected by a simple, graphical test. **BMJ** 1997;315:629-34.

4 Sacks H, Chalmers T C, Smith H J. Randomized versus historical controls for clinical trials. **Am J Med** 1982;72:233-40.

5 Schulz K F, Chalmers I, Hayes R J, Altman D G. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. **JAMA** 1995;273:408-12.

6 Prendiville W, Elbourne D, Chalmers I. The effects of routine oxytocic administration in the management of the third stage of labour: an overview of the evidence from controlled trials. **Br J Obstet Gynaecol** 1988;95:3-16.

7 Begg C B, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized

controlled trials. The CONSORT statement. **JAMA** 1996;276:637-9.

8 Greenhalgh T. The Medline database. **BMJ** 1997;315:180-3.

9 Dickersin K, Hewitt P, Mutch L, Chalmers I, Chalmers T C. Perusing the literature: comparison of Medline searching with a perinatal clinical trial data base. **Controlled Clinical Trials** 1985; 6:306-317.

10 Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. **BMJ** 1994;309:1286-91.

11 Chalmers I, Dickersin K, Chalmers TC. Getting to grips with Archie Cochrane's agenda. **BMJ** 1992;305:786-8.

12 The Cochrane Controlled Trials Register. In: **Cochrane Library**. CD ROM and online. Cochrane Collaboration (issue 1). Oxford: Update Software, 1997.

13 Moher D, Jadad A R, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. **Controlled Clinical Trials** 1995; 16:62-73.

14 Jadad A R, Moore RA, Carrol D, Jenkinson C, Reynolds D J M, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? **Controlled Clinical Trials** 1996; 17:1-12.

15 Chalmers T C. Problems induced by meta-analyses. **Stat Med** 1991;10:971-80.

16 Moher D, Fortin P, Jadad A R, Jüni P, Klassen T, Le Lorier J, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. **Lancet** 1996;347:363-6.

17 Laupacis A, Sackett D L, Roberts R S. An assessment of clinically useful measures of the consequences of treatment. **New Engl J Med** 1988;318:1728-33.

18 Berlin J A, Laird N M, Sacks H S, Chalmers T C. A comparison of statistical methods for combining event rates from clinical trials. **Stat Med** 1989;8:141-51.

19 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;17:335-71.

20 DerSimonian R, Laird N. Meta-analysis in clinical trials. **Controlled Clinical Trials** 1986;7:177-88.

21 Carlin J B. Meta-analysis for 2x2 tables: a bayesian approach. **Stat Med** 1992;11:141-58.

22 Lilford R J, Braunholtz D. The statistical basis of public policy: a paradigm shift is overdue. **BMJ** 1996;313:603-7.

23 Eddy D M, Hasselblad V, Shachter R. **Meta-analysis by the confidence profile method. The statistical synthesis of evidence.** Boston: Academic Press, 1992.

24 Bailey K R. Inter-study differences: how should they influence the interpretation and analysis of results? **Stat Med** 1987;6:351-8.

25 Galbraith R. A note on graphical presentation of estimated odds ratios from several clinical trials. **Stat Med** 1988;7:889-94.

26 Multicenter Postinfarction Research Group. Risk stratification and survival after myocardial infarction. **New Engl J Med** 1983;309:331-6.

27 Chatellier G, Zapletal E, Lemaitre D, Menard J, Degoulet P. The number needed to treat: a clinically useful nomogram in its proper context. **BMJ** 1996;312:426-9.

28 Easterbrook P J, Berlin J A, Gopalan R, Matthews D R. Publication bias in clinical research. **Lancet** 1991;337:867-72.

29 Green S, Fleming T R, Emerson S. Effects on overviews of early stopping rules for clinical trials. **Stat Med** 1987;6:361-7.

30 Oxman A D. Checklists for review articles. **BMJ** 1994;309:648-51.