How to read a paper

Papers that summarise other papers (systematic reviews and meta-analyses)

Trisha Greenhalgh

This is the ninth in a series of 10 articles introducing non-experts to finding medical articles and assessing their value

Unit for Evidence-Based Practice and Policy, Department of Primary Care and Population Sciences, University College London Medical School/ Royal Free Hospital School of Medicine, Whittington Hospital, London N19 5NF

Trisha Greenhalgh, *senior lecturer*

p.greenhalgh@ucl.ac.uk

BMJ 1997;315:672–5

Summary points

A systematic review is an overview of primary studies that used explicit and reproducible methods

A meta-analysis is a mathematical synthesis of the results of two or more primary studies that addressed the same hypothesis in the same way

Although meta-analysis can increase the precision of a result, it is important to ensure that the methods used for the review were valid and reliable

Remember the essays you used to write as a student? You would browse through the indexes of books and journals until you came across a paragraph that looked relevant, and copied it out. If anything you found did not fit in with the theory you were proposing, you left it out. This, more or less, constitutes the methodology of the journalistic review—an overview of primary studies which have not been identified or analysed in a systematic (standardised and objective) way.

In contrast, a systematic review is an overview of primary studies which contains an explicit statement of objectives, materials, and methods and has been conducted according to explicit and reproducible methodology (fig 1).

Some advantages of the systematic review are given in box 1. When a systematic review is undertaken, not only must the search for relevant articles be thorough and objective, but the criteria used to reject articles as “flawed” must be explicit and independent of the results of those trials. The most enduring and useful systematic reviews, notably those undertaken by the Cochrane Collaboration, are regularly updated to incorporate new evidence.[2]

Many, if not most, medical review articles are still written in narrative or journalistic form. Professor Paul Knipschild has described how Nobel prize winning biochemist Linus Pauling used selective quotes from the medical literature to “prove” his theory that vitamin C helps you live longer and feel better.[3,4] When Knipschild and his colleagues searched the literature systematically for evidence for and against this hypothesis they found that, although one or two trials did strongly suggest that vitamin C could prevent the onset of the common cold, there were far more studies which did not show any beneficial effect.

Experts, who have been steeped in a subject for years and know what the answer “ought” to be, are less able to produce an objective review of the literature in their subject than non-experts.[5,6] This would be of little consequence if experts’ opinions could be relied on to be congruent with the results of independent systematic reviews, but they cannot.[7]

Evaluating systematic reviews

Question 1: Can you find an important clinical question which the review addressed?

The question addressed by a systematic review needs to be defined very precisely, since the reviewer must make a dichotomous (yes/no) decision as to whether each potentially relevant paper will be included or, alternatively, rejected as “irrelevant.” Thus, for example, the clinical question “Do anticoagulants prevent strokes in patients with atrial fibrillation?” should be refined as an objective: “To assess the effectiveness and safety of warfarin-type anticoagulant therapy in secondary prevention (that is, following a previous stroke or transient ischaemic attack) in patients with non-rheumatic atrial fibrillation: comparison with placebo.”[8]

Question 2: Was a thorough search done of the appropriate databases and were other potentially important sources explored?

Even the best Medline search will miss important papers, for which the reviewer must approach other sources.[9] Looking up references of references often yields useful articles not identified in the initial search,[10] and an exploration of “grey literature” (box 2) may be particularly important for subjects outside the medical

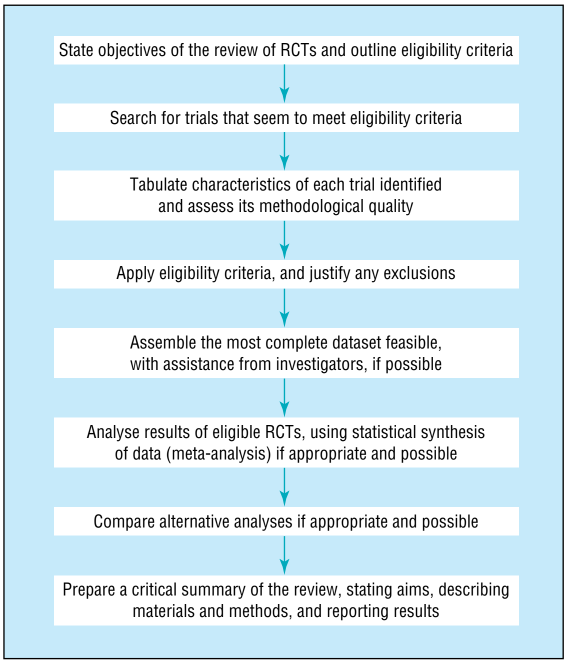

State objectives of the review of RCTs and outline eligibility criteria
Search for trials that seem to meet eligibility criteria
Tabulate characteristics of each trial identified and assess its methodological quality
Apply eligibility criteria, and justify any exclusions
Assemble the most complete dataset feasible, with assistance from investigators, if possible
Analyse results of eligible RCTs, using statistical synthesis of data (meta-analysis) if appropriate and possible
Compare alternative analyses if appropriate and possible
Prepare a critical summary of the review, stating aims, describing materials and methods, and reporting results

Fig 1 Methodology for a systematic review of randomised controlled trials[1]

Box 1
Advantages of systematic reviews[3]

- Explicit methods limit bias in identifying and rejecting studies
- Conclusions are more reliable and accurate because of methods used
- Large amounts of information can be assimilated quickly by healthcare providers, researchers, and policymakers
- Delay between research discoveries and implementation of effective diagnostic and therapeutic strategies may be reduced
- Results of different studies can be formally compared to establish generalisability of findings and consistency (lack of heterogeneity) of results
- Reasons for heterogeneity (inconsistency in results across studies) can be identified and new hypotheses generated about particular subgroups
- Quantitative systematic reviews (meta-analyses) increase the precision of the overall result

mainstream, such as physiotherapy or alternative medicine.[11] Finally, particularly where a statistical synthesis of results (meta-analysis) is contemplated, it may be necessary to write and ask the authors of the primary studies for raw data on individual patients which was never included in the published review.

Question 3: Was methodological quality assessed and the trials weighted accordingly?

One of the tasks of a systematic reviewer is to draw up a list of criteria, including both generic (common to all research studies) and particular (specific to the field) aspects of quality, against which to judge each trial (see box 3). However, care should be taken in developing such scores since there is no gold standard for the "true" methodological quality of a trial[12] and composite quality scores are often neither valid nor reliable in practice.[13 14] The various Cochrane collaborative review groups are developing topic-specific methodology for assigning quality scores to research studies.[15]

Question 4: How sensitive are the results to the way the review has been done?

Carl Counsell and colleagues "proved" (in the Christmas 1994 issue of the *BMJ*) an entirely spurious relationship between the result of shaking a dice and the outcome of an acute stroke.[16] They reported a series of artificial dice rolling experiments in which red, white, and green dice represented different therapies for acute stroke. Overall, the "trials" showed no significant benefit from the three therapies. However, the simulation of a number of perfectly plausible events in the process of meta-analysis—such as the exclusion of several of the "negative" trials through publication bias, a subgroup analysis which excluded data on red dice therapy (since, on looking back at the results, red dice appeared to be harmful), and other, essentially arbitrary, exclusions on the grounds of "methodological quality"—led to an apparently highly significant benefit of "dice therapy" in acute stroke.

If these simulated results pertained to a genuine medical controversy, how would you spot these subtle biases? You need to work through the "what ifs". What if the authors of the systematic review had changed the inclusion criteria? What if they had excluded unpublished studies? What if their "quality weightings" had been assigned differently? What if trials of lower methodological quality had been included (or excluded)? What if all the patients unaccounted for in a trial were assumed to have died (or been cured)?

An exploration of what ifs is known as a sensitivity analysis. If you find that fiddling with the data in various ways makes little or no difference to the review's overall results, you can assume that the review's conclusions are relatively robust. If, however, the key findings disappear when any of the what ifs changes,

Box 2
Checklist of data sources for a systematic review

- Medline database
- Cochrane controlled clinical trials register
- Other medical and paramedical databases
- Foreign language literature
- "Grey literature" (theses, internal reports, non-peer reviewed journals, pharmaceutical industry files)
- References (and references of references, etc) listed in primary sources
- Other unpublished sources known to experts in the field (seek by personal communication)
- Raw data from published trials (seek by personal communication)

Box 3
Assigning weight to trials in a systematic review

Each trial should be evaluated in terms of its:
- Methodological quality—the extent to which the design and conduct are likely to have prevented systematic errors (bias)
- Precision—a measure of the likelihood of random errors (usually depicted as the width of the confidence interval around the result)
- External validity—the extent to which the results are generalisable or applicable to a particular target population

PETER BROWN

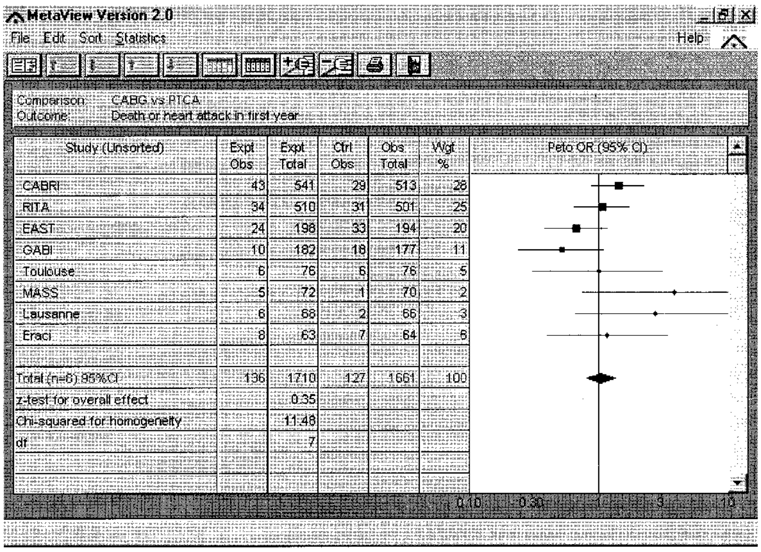
the conclusions should be expressed far more cautiously and you should hesitate before changing your practice in the light of them.

Question 5: Have the numerical results been interpreted with common sense and due regard to the broader aspects of the problem?
Any numerical result, however precise, accurate, "significant," or otherwise incontrovertible, must be placed in the context of the painfully simple and often frustratingly general question which the review addressed. The clinician must decide how (if at all) this numerical result, whether significant or not, should influence the care of an individual patient. A particularly important feature to consider when undertaking or appraising a systematic review is the external validity or relevance of the trials that are included.

Meta-analysis for the non-statistician

A good meta-analysis is often easier for the non-statistician to understand than the stack of primary research papers from which it was derived. In addition to synthesising the numerical data, part of the meta-analyst's job is to tabulate relevant information on the inclusion criteria, sample size, baseline patient characteristics, withdrawal rate, and results of primary and secondary end points of all the studies included. Although such tables are often visually daunting, they save you having to plough through the methods sections of each paper and compare one author's tabulated results with another author's pie chart or histogram.

These days, the results of meta-analyses tend to be presented in a fairly standard form, such as is produced by the computer software MetaView. Figure 2 is a pictorial representation (colloquially known as a "forest plot") of the pooled odds ratios of eight randomised controlled trials which each compared coronary artery bypass grafting with percutaneous coronary angioplasty in the treatment of severe angina.[17] The primary (main) outcome in this meta-analysis was death or heart attack within one year.



Comparison: CABG vs PTCA
Outcome: Death or heart attack in first year

Study (Unsorted)	Expt Obs	Expt Total	Ctrl Obs	Obs Total	Wgt %	Peto OR (95% CI)
CABRI	43	541	29	513	28	
RITA	34	510	31	501	25	
EAST	24	198	33	194	20	
GABI	10	182	18	177	11	
Toulouse	6	76	6	76	5	
MASS	5	72	1	70	2	
Lausanne	6	68	2	66	3	
Eraci	8	63	7	64	6	
Total (n=8) 95%CI	136	1710	127	1661	100	
z-test for overall effect		0.35				
Chi-squared for homogeneity		11.48				
df		7				

Fig 2 Pooled odds ratios of eight randomised controlled trials of coronary artery bypass grafting against percutaneous coronary angioplasty, shown in MetaView format. Reproduced with authors' permission[17]

The horizontal line corresponding to each of the eight trials shows the relative risk of death or heart attack at one year in patients randomised to coronary angioplasty compared to patients randomised to bypass surgery. The "blob" in the middle of each line is the point estimate of the difference between the groups (the best single estimate of the benefit in lives saved by offering bypass surgery rather than coronary angioplasty), and the width of the line represents the 95% confidence interval of this estimate. The black line down the middle of the picture is known as the "line of no effect," and in this case is associated with a relative risk of 1.0.

If the confidence interval of the result (the horizontal line) crosses the line of no effect (the vertical line), that can mean either that there is no significant difference between the treatments or that the sample size was too small to allow us to be confident where the true result lies. The various individual studies give point estimates of the relative risk of coronary angioplasty compared with bypass surgery of between about 0.5 and 5.0, and the confidence intervals of some studies are so wide that they do not even fit on the graph. Now look at the tiny diamond below all the horizontal lines. This represents the pooled data from all eight trials (overall relative risk of coronary angioplasty compared with bypass surgery = 1.08), with a new, much narrower, confidence interval of this relative risk (0.79 to 1.50). Since the diamond firmly overlaps the line of no effect, we can say that there is probably little to choose between the two treatments in terms of the primary end point (death or heart attack in the first year). Now, in this example, every one of the eight trials also suggested a non-significant effect, but in none of them was the sample size large enough for us to be confident in that negative result.

Note, however, that this neat little diamond does not mean that you might as well offer coronary angioplasty rather than bypass surgery to every patient with angina. It has a much more limited meaning—that the average patient in the trials presented in this meta-analysis is equally likely to have met the primary outcome (death or myocardial infarction within a year), whichever of these two treatments they were randomised to receive. If you read the paper by Pocock and colleagues[17] you would find important differences in the groups in terms of prevalence of angina and requirement for further operative intervention after the initial procedure.

Explaining heterogeneity

In the language of meta-analysis, homogeneity means that the results of each individual trial are mathematically compatible with the results of any of the others. Homogeneity can be estimated at a glance once the trial results have been presented in the format illustrated in figures 2 and 3. In figure 2 the lower confidence limit of every trial is below the upper confidence limit of all the others (that is, the horizontal lines all overlap to some extent). Statistically speaking, the trials are homogeneous. Conversely, in figure 3 some lines do not overlap at all. These trials may be said to be heterogeneous.

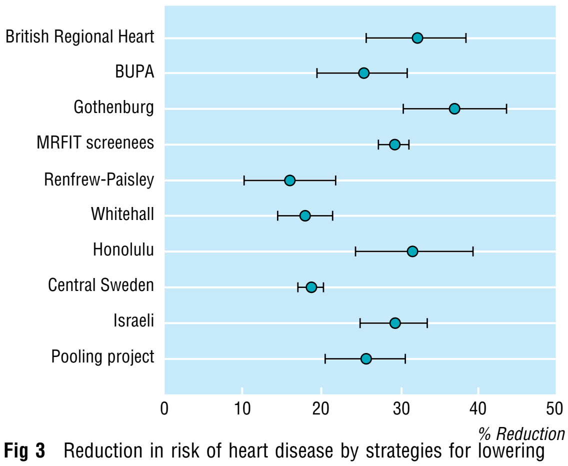


Fig 3 Reduction in risk of heart disease by strategies for lowering cholesterol. Reproduced with permission from Chalmers and Altman[18]

The definitive test for heterogeneity involves a slightly more sophisticated statistical manoeuvre than holding a ruler up against the forest plot. The one most commonly used is a variant of the χ^2 (chi square) test, since the question addressed is whether there is greater variation between the results of the trials than is compatible with the play of chance. Thompson[18] offers the following rule of thumb: a χ^2 statistic has, on average, a value equal to its degrees of freedom (in this case, the number of trials in the meta-analysis minus one), so a χ^2 of 7.0 for a set of eight trials would provide no evidence of statistical heterogeneity. Note that showing statistical heterogeneity is a mathematical exercise and is the job of the statistician, but explaining this heterogeneity (looking for, and accounting for, clinical heterogeneity) is an interpretive exercise and requires imagination, common sense, and hands-on clinical or research experience.

Figure 3 shows the results of ten trials of cholesterol lowering strategies. The results are expressed as the percentage reduction in risk of heart disease associated with each reduction of 0.6 mmol/l in serum cholesterol concentration. From the horizontal lines which represent the 95% confidence intervals of each result it is clear, even without knowing the χ^2 statistic of 127, that the trials are highly heterogeneous. Correcting the data for the age of the trial subjects reduced this value to 45. In other words, much of the "incompatibility" in the results of these trials can be explained by the fact that embarking on a strategy which successfully reduces your cholesterol level will be substantially more likely to prevent a heart attack if you are 45 than if you are 85.

Clinical heterogeneity, essentially, is the grievance of Professor Hans Eysenck, who has constructed a vigorous and entertaining critique of the science of meta-analysis.[19] In a world of lumpers and splitters, Eysenck is a splitter, and it offends his sense of the qualitative and the particular to combine the results of studies which were done on different populations in different places at different times and for different reasons.

Eysenck's reservations about meta-analysis are borne out in the infamously discredited meta-analysis which showed (wrongly) that giving intravenous magnesium to people who had had heart attacks was beneficial. A subsequent megatrial involving 58 000 patients (ISIS-4) failed to find any benefit, and the meta-analysts' misleading conclusions were subsequently explained in terms of publication bias, methodological weaknesses in the smaller trials, and clinical heterogeneity.[20 21]

Thanks to Professor Iain Chalmers for advice on this chapter.

The articles in this series are excerpts from *How to read a paper: the basics of evidence based medicine.* The book includes chapters on searching the literature and implementing evidence based findings. It can be ordered from the BMJ Publishing Group: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

1 *The Cochrane Centre.* Cochrane Collaboration Handbook [updated 9 December 1996]. The Cochrane Collaboration; issue 1. Oxford: Update Software, 1997.
2 Bero L, Rennie D. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995:274:1935-8.
3 Chalmers I, Altman DG, eds. *Systematic reviews.* London: BMJ Publishing Group, 1995.
4 Pauling L. *How to live longer and feel better.* New York: Freeman, 1986.
5 Oxman AD, Guyatt GH. The science of reviewing research. *Ann NY Acad Sci* 1993; 703: 125-31.
6 Mulrow C. The medical review article: state of the science. *Ann Intern Med* 1987;106: 485-8.
7 Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomised controlled trials and recommendations of clinical experts. *JAMA* 1992;268:240-8.
8 Koudstaal P. Secondary prevention following stroke or TIA in patients with non-rheumatic atrial fibrillation: anticoagulant therapy versus control. *Cochrane Database of Systematic Reviews.* Oxford: Cochrane Collaboration, 1995. (Updated 14 February 1995.)
9 Greenhalgh T. Searching the literature. In: *How to read a paper.* London: BMJ Publishing Group, 1997:13-33.
10 Knipschild P. Some examples of systematic reviews. In: Chalmers I, Altman DG. *Systematic reviews.* London: BMJ Publishing Group, 1995:9-16.
11 Knipschild P. Searching for alternatives: loser pays. Lancet 1993; 341: 1135-6.
12 Oxman A, ed. Preparing and maintaining systematic reviews. In: *Cochrane Collaboration handbook, section VI.* Oxford: Cochrane Collaboration, 1995. (Updated 14 July 1995.)
13 Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clin Trials* 1990;11:339-52.
14 Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Health Technol Assess* 1996;12:195-208.
15 Garner P, Hetherington J. Establishing and supporting collaborative review groups. In: *Cochrane Collaboration handbook, section II.* Oxford: Cochrane Collaboration, 1995 (Updated 14 July 1995.)
16 Counsell CE, Clarke MJ, Slattery J, Sandercock PAG. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *BMJ* 1994;309:1677-81.
17 Pocock SJ, Henderson RA, Rickards AF, Hampton JR, Sing SB III, Hamm CW, et al. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet* 1995;346:1184-9.
18 Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. In: Chalmers I, Altman DG. *Systematic reviews.* London, BMJ Publishing Group, 1995:48-63.
19 Eysenck HJ. Problems with meta-analysis. In: Chalmers I, Altman DG. *Systematic reviews.* London: BMJ Publishing Group, 1995:64-74.
20 Magnesium, myocardial infarction, meta-analysis and mega-trials. *Drug Ther Bull* 1995;33:25-7.
21 Egger M, Davey Smith G. Misleading meta-analysis: lessons from "an effective, safe, simple" intervention that wasn't. *BMJ* 1995;310:752-4.

Correction

Statistics for the non-statistician. I: Different types of data need different tests

An author's error appeared in this article by Trisha Greenhalgh (9 August, pp 364-6). In table 1, the χ^2 test is listed as a parametric test. In fact, both the χ^2 test and Fisher's exact test are non-parametric.