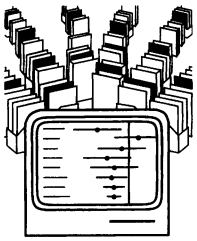
Checklists for review articles

Andrew D Oxman



This paper was presented at a meeting on systematic reviews organised jointly by the BMJ and the Cochrane Centre and held in London in July 1993

Departments of Clinical Epidemiology and Biostatistics and of Family Medicine, McMaster University, Hamilton, Ontario, Canada L8S 4L8
Andrew D Oxman, *assistant professor*

BMJ 1994;309:648-51

Preparing a review entails many judgments. The focus of the review must be decided. Studies that are relevant to the focus of the review must be identified, selected for inclusion and critically appraised. Information must be collected and synthesised from the relevant studies, and conclusions must be drawn. Checklists can help prevent important errors in this process. Reviewers, editors, content experts, and users of reviews all have a role to play in improving the quality of published reviews and promoting the appropriate use of reviews by decisionmakers. It is essential that both providers and users appraise the validity of review articles.

Why checklists?

When we think about flying, it is obvious why a checklist is used before take off. Airplanes are complex machines. Things can go wrong with them, and it is preferable that problems are discovered on the ground. However brilliant a pilot and crew might be, most of us would prefer that they use a checklist when preparing for take off, rather than relying on memory.

The need for checklists for review articles is less obvious, but the rationale is much the same. Preparing a review is a complex process entailing many judgments. The focus of the review must be decided. Studies that are relevant to the focus of the review must be identified, selected for inclusion, and critically appraised. Information must be collected and synthesised from the relevant studies, and conclusions must be drawn. Many decisions must be made throughout this process.

It is important to go through this process systematically to avoid errors. Explicitness about how decisions were made enables others to assess how well the process protected against errors. Checklists can help those doing and using reviews to avoid important errors.

Faulty reviews may not seem as perilous as faulty airplanes. However, if people are going to use reviews to guide decisions about health care, misleading reviews can indeed be deadly. On the other hand, if people are not going to use reviews to guide decisions, why bother with them?

Before deciding that we should not bother with reviews, it is important to remember that there is little choice. Whether we rely on published, formal reviews or reviews done inside our heads, or the heads of experts, the risks remain. The same judgments must be made, whether explicitly or implicitly. The advantage of using carefully done, systematic reviews becomes clear when we observe how often mistakes are made when research is reviewed non-systematically, whether by experts or others. The costs of mistaken conclusions based on non-systematic reviews can be high.

Life saving treatments, such as thrombolytic therapy and aspirin for patients with myocardial infarctions, can go unused. Other treatments that do not have proved benefits and may even be harmful, such as calcium channel blockers and antiarrhythmic agents for patients with myocardial infarctions, can be used inappropriately.[1]

Box 1—Sources of bias and methods of protecting against bias

Problem formulation
- Is the question clearly focused?

Study identification
- Is the search for relevant studies thorough?

Study selection
- Are the inclusion criteria appropriate?

Appraisal of studies
- Is the validity of included studies adequately assessed?

Data collection
- Is missing information obtained from investigators?

Data synthesis
- How sensitive are the results to changes in the way the review is done?

Interpretation of results
- Do the conclusions flow from the evidence that is reviewed?
- Are recommendations linked to the strength of the evidence?
- Are judgments about preferences (values) explicit?
- If there is "no evidence of effect" is caution taken not to interpret this as "evidence of no effect"?
- Are subgroup analyses interpreted cautiously?

What should be checked?

The most dangerous errors in reviews are systematic ones (bias) rather than ones that occur by chance alone (random errors). Consequently, the most important thing for doers and users of a review to check is its "validity": the extent to which its design and conduct are likely to have protected against bias.

Random errors can also be deadly. However, if a review is done systematically and quantitative results are presented, the confidence interval around the results provides a good indication of "precision": the extent to which the results are likely to differ from the "truth" because of chance alone.[2] A confidence interval does not provide any indication of the likelihood of bias.

Other attributes of a review are also important, including choice of focus, degree of innovation in the approach, potential impact on future scientific developments, literary quality, and handling of pertinent ethical issues. It may or may not be appropriate to include items related to these attributes in a checklist. This depends on the purpose of the particular checklist.

ASKING THE QUESTIONS

For most, if not all purposes, the first question that should be addressed by a checklist is, is the focus of this review relevant? This question can only be answered relative to a specific context—for example, whether it is relevant to a particular patient, practice setting, or readership. The next question should be, are the results likely to be valid? If there are important

concerns about validity, any other considerations are largely irrelevant.

A number of "checklists" have been published suggesting what should be examined when assessing the validity of a review.[3-10] There are some differences in the items included in these lists and in how each item is addressed, but they all focus on the same sources of bias (box 1): how the problem was formulated; how studies were identified, selected for inclusion, and critically appraised; how data were collected and synthesised; and how the results were interpreted. Some questions relating to each potential source of bias are listed in box 1.

Several authors have examined issues pertaining to the validity of reviews,[5-13] and many of these issues are considered in the other articles in this series. The logic behind the questions in the box, and other questions that can be asked about how well a review has protected against bias, is straightforward. Preparing a review of research is in itself a research process. It is a type of survey, and the scientific principles underlying reviews and epidemiological surveys are the same. In a review a question must be posed, a target population of information sources identified and accessed, appropriate information obtained from that population in an unbiased fashion, and conclusions derived. Often statistical analysis (meta-analysis) can help in reaching conclusions.

The starting point for any research project is to ask a good question and develop a protocol laying out the methods that will be used to answer the question. Without a clearly focused question there is little point in going further. The types of people, interventions, and outcomes of interest should all be clearly specified.

INCLUDING STUDIES

The criteria used to select studies for inclusion in a review should be consistent with the focus. They should be explicit to protect against biased selection of studies. Similarly, the criteria that are used to assess the validity of the studies that are included should be explicit to minimise biased assessments and weighting of the included studies.

Variation in quality can explain variation in the results of the included studies. Statistical summaries (meta-analyses) of results from studies of variable quality can result in a "false positive" conclusion (concluding that there is an effect when the truth is that there is not) if the less rigorous studies that are included are biased towards overestimating the effectiveness of the intervention being evaluated. Such summaries might also result in a "false negative" conclusion (concluding that there is not an effect when in truth there is) if the less rigorous studies provide less precise or biased estimates of the effects of the intervention, thereby obscuring the true effect.[14] The methodological quality of the included studies is important even if the results or quality of the included studies do not vary. If the evidence is consistent but all the studies are flawed, the conclusions of the review would not be nearly as strong as if consistent results were obtained from a series of high quality studies.

Poor quality studies and insufficient reporting are, unfortunately, common in the medical literature.[15-20] Information obtained through personal communication can strengthen the results of a review. To avoid introducing bias, unpublished information that is obtained should be unambiguous; it should be obtained in writing and coded in the same fashion as published information.

ANALYSING THE DATA

The analysis or synthesis of the data that are collected for a systematic review entails the whole process of evaluating and synthesising the results of the included studies. Statistical techniques may or may not be used. The statistical techniques used in meta-analyses do not differ in principle from those used in primary research,[21] and the logic behind their use is the same. Statistical analysis is a tool which, when used appropriately, can help us to derive meaningful conclusions from data and to avoid analytic errors. Like any tool, it can also be misused.

Because there are different approaches to conducting a systematic review, it is important to ask how sensitive the results are to changes in the way the review is done. Such "sensitivity analyses" provide an approach to testing how robust the results of a review are relative to key decisions and assumptions that were made in the process of conducting the review. The types of decisions and assumptions that might be examined in sensitivity analyses include:

- Changing the inclusion criteria
- Including or excluding trials where there is some ambiguity as to whether they meet the inclusion criteria
- Excluding unpublished studies
- Excluding studies of lower methodological quality
- Reanalysing the data by using a reasonable range of results for trials in which there may be some uncertainty about the results
- Reanalysing the data imputing a reasonable range of values for missing data
- Reanalysing the data using different statistical approaches.

Even when the results of a review are robust, it is possible to reach erroneous conclusions if the results are misinterpreted. The last five questions in the table all focus on this source of bias. They need little elaboration but deserve close attention.

CONCLUSIONS AND RECOMMENDATIONS

The conclusions of a review should not exceed the evidence that is reviewed. So far as possible, recommendations should be linked to the strength of the evidence. Preferably, this should be done with an explicit approach to specify levels of evidence—for example, like the one used by the Antithrombotic Therapy Consensus Conference (box 2).[22]

Because health care interventions entail costs and risks of harm as well as expectations of benefit, practice recommendations require judgments about preferences (the values attached to different outcomes) in addition to judgments about evidence.[23] When conclusions involve judgments about preferences this should be clearly stated. For example, women considering hormone replacement therapy must consider the tradeoffs between the potential benefits (prevention of hip fracture and cardiovascular disease)

Box 2—Levels of evidence for treatment

Level I	The lower limit of the confidence interval for the effect of treatment from a systematic review of randomised controlled trials exceeded the clinically significant benefit.
Level II	The lower limit of the confidence interval for the effect of treatment from a systematic review of randomised controlled trials fell below the clinically significant benefit (but the point estimate of its effect was at or above the clinically significant benefit)
Level III	Non-randomised concurrent cohort studies
Level IV	Non-randomised historical cohort studies
Level V	Case series

Detailed definitions for these levels of evidence and corresponding grades of recommendations are provided by Cook *et al*[22]

and the possible harms (breast and endometrial cancer and vaginal bleeding).[24,25] The relative value attached to these outcomes varies from woman to woman. Before drawing conclusions for clinical practice, a systematic review of the effects of preventive hormone therapy must consider all of the potentially important outcomes. Assumptions about the relative value of these should not be hidden.

ERRORS

The last two questions in box 1 refer to two common types of errors that are found in reviews (and elsewhere). One is to confuse "no evidence of effect" with "evidence of no effect." For example, no evidence that love causes heartbreak is not the same as evidence that love causes no heartbreak.

The other type of error is misinterpretation of subgroup analyses. It is frequently of interest to examine a particular category of participants in a review—for example, women, a certain age group, or those with a specific pattern of disease. These examinations, or subgroup analyses, are exceedingly common, but they are also often misleading.[26,27] Conclusions based on subgroup analyses can do harm both when a particular category of people is denied effective treatment (a false negative conclusion), and when ineffective or even harmful treatment is given to a subgroup of people (a false positive conclusion). Subgroup analyses can also generate misleading recommendations about directions for future research that, if they are followed, can waste scarce resources. Because of these risks and the frequency of their occurrence, it is important to be cautious when enticed to perform and interpret subgroup analyses.

Who should check?

Traditionally, reviews have been written by experts on the topic. When seeking critiques of review articles (peer review), editors have looked to other experts in the field for help. These policies may seem intuitively reasonable and appropriate. However, there are reasons for serious scepticism. Experts may lack the objectivity desirable in preparing or critiquing a review article. For example, personal experience in primary research is highly salient and considerably more vivid than the research of others, and therefore likely to be given undue weight in judgments.[28] This is also true for personal clinical experience.

In examining the relation of expertise to the quality of review articles, Guyatt and I found that expertise in an area was inversely related to methodological quality: the greater the expertise of the author, the poorer the quality of the review.[29] This might have been related to the strength of prior opinions and the amount of time spent preparing a review article. Experts tended to have stronger opinions about the topic of a review and to spend less time preparing a review.

We also found poor agreement about the methodological quality of reviews among experts.[29] At least two possible explanations for this are possible: lack of training, or the effects of expertise. In either case, the results cast doubt on the wisdom of relying exclusively on experts in a clinical area who do not have specific methodological training to check the quality of reviews.

Mulrow, in her classic study of the quality of review articles in the medical literature, has shown the failure of traditional review articles to describe their methods.[6] Of even greater concern, Antman and colleagues in another enlightening study showed the extent to which the conclusions of experts can differ from those based on the results of systematic reviews.[1] An extreme interpretation of findings such as these might be that experts should occupy themselves with the task of producing new data or else retire from the topics of their expertise[30] and leave the task of preparing and critiquing review articles to those who have specific training in the science of research reviews. A more reasonable interpretation might be to acknowledge the importance of expertise while at the same time recognising the need for others to play a role in ensuring the quality of reviews.

ROLES FOR EXPERTS

Experts often belong to an "invisible college" of people who interact with each other because of a common interest in an area of research. Because of these connections, experts may be aware of studies that reviewers might otherwise miss. Experts often have extensive knowledge of related evidence outside the specific focus of a review and practical experience in their area of expertise. They may be more aware of nuances than others. Hence they can bring important perspectives to the interpretation of the results of a review.

In summary, content experts have an important role in ensuring the quality of reviews, but we cannot rely on their expertise alone (table). Reviewers and editors, especially, must check whether the methods used in a review are likely to protect against bias. This is essential if the quality of published review articles is to improve. In addition, users of reviews must be able to judge whether a review is clearly focused—and whether that focus is relevant to their situation. They must be able to judge whether the studies included in a review are appropriately selected relative to the focus. Finally, they must be able to ensure that the conclusions that are drawn on the basis of a review, whether the reviewers' or their own, are supported by the evidence that is reviewed.

Who should check what?

	Reviewers	Editors	Experts	Users
Focus	✓	✓		✓
Missing studies	✓	✓	✓	
Selection criteria	✓	✓		✓
Quality of studies	✓	✓		
Data collection	✓	✓		
Synthesis of data	✓	✓		
Interpretation	✓	✓	✓	✓

To take well informed decisions about health care, people at all levels of the health services need access to systematic reviews of the relevant evidence. Reviewers, editors, content experts, and users of reviews all need to check that review articles are valid if we want to make decisions based on evidence rather than on authority.[31-33]

The author received support for this work from the Nuffield Provincial Hospitals Trust as a Cochrane Centre Visiting Fellow.

1 Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatments for myocardial infarction. *JAMA* 1992;**268**:240-8.
2 Altman DG. Confidence intervals in research evaluation. *Ann Interen Med* 1992;**116**(suppl 2):A28.
3 Jackson GB. Methods for integrative reviews. *Review of Education Research* 1980;**50**:438-60.
4 Cooper H. Scientific guidelines for conducting integrative research reviews. *Review of Education Research* 1982;**52**:291-302.
5 Light RJ, Pillemer DB. *Summing up: the science of reviewing research.* Cambridge, MA: Harvard University Press, 1984:160-73.
6 Mulrow CD. The medical review article: state of the science. *Ann Intern Med* 1987;**106**:485-8.
7 Sacks HS, Berrier J, Reitman D, Ancona-Bank VA, Chalmers TC, *et al.* Meta-analyses of randomized controlled trials. *N Engl J Med* 1987;**316**:450-5.
8 L'Abbé KA, Detsk AS, O'Rourke K. Meta-analysis in clinical research. *Ann Intern Med* 1987;**107**:224-33.
9 Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *Can Med Assoc J* 1988;**138**:697-703.
10 Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. *JAMA* (in press).
11 Glass GV, McGaw B, Smith M. *Meta-analysis in social research.* Newbury Park: Sage, 1981.

12 Ysuf S, Simon R, Ellenberg S, eds. Proceedings of "Methodologic Issues in Overviews of Randomized Clinical Trials." *Stat Med* 1987;6:217-409.
13 Cooper H, Hedges LV, eds. *The handbook of research synthesis.* New York: Russell Sage Foundation, 1993.
14 Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'Abbé KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. *J Clin Epidemiol* 1992;45:255-65.
15 Williamson JW, Goldschmidt PG, Colton T. The quality of medical literature: an analysis of validation assessments. In: Bailar JC, Mosteller F, eds. *Medical uses of statistics.* Waltham, MA: NEJM Books, 1986:370-91.
16 Fletcher RH, Fletcher SW. Clinical research in general medical journals: a 30 year persective. *N Engl J Med* 1979;301:180-3.
17 Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials: agenda for the editor. *Controlled Clin Trials* 1980;1:37-57.
18 Gøtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials* 1989;10:31-56.
19 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. *Control Clin Trials* 1990;11: 339-52.
20 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Failure to conceal treatment allocation schedules in controlled trials influences estimates of treatment effects: an analysis of 250 trials in 33 meta-analyses. Atlanta: Centers for Disease Control and Prevention, 1993. (Manuscript.)
21 Laird NM, Mosteller F. Some statistical methods for combining experimental results. *Int J Technology Assess Health Care* 1990;6:5-30.
22 Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Antithrombotic Therapy Consensus Conference. *Chest* 1992;102:305-11S.
23 Eddy DM. Anatomy of a decision. *JAMA* 1990;263:441-3.
24 Grady D, Rubin SM, Petitti DB, Fox CS, Black D, Ettinger B, *et al.* Hormone therapy to prevent disease and prolong life in postmenopausal women. *Ann Intern Med* 1992;117:1016-37.
25 American College of Physicians. Guidelines for counselling postmenopausal women about preventive hormone therapy. *Ann Intern Med* 1992;117: 1038-41.
26 Ysuf S, Wittes J, Probstfiel J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266:93-8.
27 Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84.
28 Cooper HM. On the social psychology of using research review: the case of desegregation and the black achiever. In: Feldman RS, ed. *Social psychology of education.* Cambridge: Cambridge University Press, 1986:341-63.
29 Oxman AD, Guyatt GH. The science of reviewing research. *Ann NY Acad Sci* (in press).
30 Sackett DL. Proposals for the health sciences. I. Compulsory retirement for experts. *J Chron Dis* 1983;36:545-7.
31 Evidence-based Medicine Working Group. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA* 1992;268:2420-5.
32 Guyatt GH, Rennie D. Users' guides to the medical literature. *JAMA* 1993;270:2096-7.
33 Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. *JAMA* 1993;270:2093-5.

Recent Advances

Otorhinolaryngology

Anthony Hinton, Victoria Moore-Gillon

As in all other specialties, changes in the organisation of the NHS and the purchaser-provider split have had effects on the practice of otorhinolaryngology. The most extensive change in practice over the past 12 months has not been the widespread adoption of a radical new diagnostic technique or surgical procedure but the increasing introduction of day case surgery. Procedures such as grommet insertion and reduction of fractured noses have long been performed as day cases, but many departments are now carrying out adenoidectomy, tonsillectomy, all types of nasal surgery, and even major ear surgery on a day case basis.[1] This has necessitated changes in working practices and philosophy as well as an appraisal of the safety of such a move to day case surgery. The specialty is also being changed, however, not just by financial considerations and pressures from purchasers but by technical advances in fields as diverse as molecular biology, optical fibres, computers, microelectronics, and metallurgy.

Advances in otorhinolaryngology

- Virtual reality programs for surgical training
- Otoacoustic emission testing for hearing impairment
- Advances in understanding the physiology of smell
- Better lasers for precision surgery
- New techniques for laryngeal reinnervation and palatal surgery
- Implanted aids to treat hearing loss

Virtual reality programs and surgical training

The complex and variable anatomy of the middle and inner ear, and the disastrous consequences of operative errors, means that the otorhinolaryngologist in training must spend much time operating on cadaver temporal bones before even starting to deal with patients. Computers running virtual reality programs are becoming increasingly sophisticated and widespread,[2] and a virtual cadaver system is already available for training medical students and junior surgeons.

A virtual temporal bone model is currently under development and will allow the complex three dimensional anatomy of the temporal bone and the relation between the middle ear ossicles, cochlea, vestibular labyrinth, and the facial nerve to be better understood and appreciated. The advantages of such a system are that the anatomy can be viewed from any angle and not just from those possible by operation or temporal bone dissection. It is possible, for example, to view the contents of the middle ear from within the inner ear.

Diagnostic techniques—otoacoustic emissions

External sound waves move the basilar membrane in the cochlea. The cochlea itself then produces sounds, otoacoustic emissions, which can be detected and measured by a microphone in the external ear canal.[3] Emissions are more easily recordable in younger subjects than in older people, making the technique particularly valuable in those in whom the more conventional tests of hearing function, which require a subjective response and the cooperation of the patient, are difficult. Testing the hearing of a 2 year old child may take two experienced technicians up to two hours with conventional behavioural hearing tests, but an otoacoustic emission may be completed in less than 10 minutes.[4] The technique has obvious potential as a screening test in neonates and also in the relatively rare cases of feigned deafness or of hysterical deafness.[5] Equipment is relatively large and expensive at present, but advances in microelectronics and computing power suggest that a hand held, cheap, portable unit cannot be more than two or three years away.

Basic science—understanding the sense of smell

The mechanisms underlying odour transduction—by which binding of odour molecules to the olfactory mucosa produces neuronal impulses—and those

Department of Otorhinolaryngology, St George's Hospital, London SW17 0QT
Anthony Hinton, *senior registrar*
Victoria Moore-Gillon, *consultant*

Correspondence to:
Dr Hinton.

BMJ 1994;309:651-4