[30 March 1994]

SECTION VI:

PREPARING AND MAINTAINING SYSTEMATIC REVIEWS

(`The Cochrane Collaboration Tool Kit')

Editor: Andy Oxman

Department of Health Services Research
National Institute of Public Health
Geitmyrsveien 75
0462 Oslo
NORWAY
Tel: +47 22 04 22 00
Fax: +47 22 04 25 95
E-mail: GENERAL@COCHRANE.CO.UK

Associate Editors:
Iain Chalmers, Mike Clarke, Murray Enkin, Ken Schulz, Mark Starr

Acknowledgements:

Kay Dickersin, Andrew Herxheimer and Chris Silagy were members of the editorial team that developed this Section of the Handbook and share whatever credit is due (but none of the blame) for the progress that has been made in developing the 'Tool Kit'. The Tool Kit has been reviewed by the members of the Steering Group who have made many helpful suggestions. The Advisory Board for the Tool Kit is composed of the coordinators of the relevant Cochrane Methods Groups. Doug Altman and Ken Schulz (statistics), Alastair Gray and Ron Akehurst (health economics), and David Moher (trial quality), together with the members of their groups have helped guide the Tool Kit's development up to now. Newly forming Methods Groups, (such as the one being organized by Lesley Stewart and Mike Clarke on reviews using individual patient data) will help guide the further development of the Tool Kit. We are very grateful for the help of all those who have contributed, and would particularly like to acknowledge the help of Sally Hunt for her administrative assistance. Financial support for developing the first edition of the Tool Kit was provided by the Nuffield Provincial Hospitals Trust, the Oxford Regional Health Authority, and the NHS R&D Programme (England).

The Cochrane Collaboration Handbook, Section VI
Preparing and Maintaining Systematic Reviews

EVALUATION FORM

From: _____ _____
Name Date

_____
Name of collaborative review group

What about this section of the Handbook do you like?

What could be improved and how?

Other comments or suggestions:

Please complete this evaluation form and return to Andy Oxman, Canadian Cochrane Centre, Health Information Research Unit, Rm 3H7, 1200 Main Street W, Hamilton, Ontario L8N 3Z5, Canada (Fax: +1 905 546 0401) by June 16, 1994. Thanks!

The Cochrane Collaboration
Review Manager Software

EVALUATION FORM

From: _____ _____
Name Date

_____
Name of collaborative review group

What about the software do you like?

What could be improved and how?

Other comments or suggestions:

Please complete this evaluation form and return to Monica Fischer, UK Cochrane Centre, Summertown Pavilion, Middle Way, Oxford OX2 7LG, UK (Fax: +44 865 516311) by June 16, 1994. Thanks!

SECTION VI: PREPARING AND MAINTAINING SYSTEMATIC REVIEWS

CONTENTS

APPENDICES

SECTION VI: PREPARING AND MAINTAINING SYSTEMATIC REVIEWS

1. THE SCIENCE OF REVIEWING RESEARCH

The scientific principles that apply to epidemiological surveys apply also to systematic reviews: 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 can help in reaching conclusions.

As with any research study, a number of methodological decisions must be made when undertaking a systematic review, and there are potential threats to validity associated with each decision. These include: how studies are identified and selected for inclusion, how their quality (validity) is assessed, how much weight each one is given, and how their results are combined and interpreted.

If reviewers do not say how these decisions were made, there is no way of knowing how much faith to put in their conclusions. Unfortunately, reviews of the same topic often disagree, and neither the expertise of the reviewer nor peer review can guarantee the validity of a review.[1,2,3] It is possible that experts in an area might actually be less objective than non-experts.[4] For instance, experts are likely to overweight research that they have undertaken themselves relative to research done by others, and they often have strong opinions about their area that can lead them to judge evidence differently according to whether it supports their beliefs.

This Section of the Handbook contains standards and guidelines for methodological decisions that reviewers must make. These guidelines are intended to help reviewers to be systematic (not mechanistic!) and, thereby avoid bias (systematic errors) and being misled by the play of chance (random errors). These guidelines are not a substitute for good judgement.

The Cochrane Collaboration, and this Section of the Handbook, focuses on systematic reviews of RCTs, because they are likely to provide more reliable information than other sources of evidence on the differential effects of alternative forms of health care. Systematic reviews of other types of evidence can also help those wanting to take better decisions about health care, particularly for forms of care where RCTs have not been done and may not be possible or usual. As the Collaboration evolves and comes to grips with the enormous number of RCTs that have not yet been systematically reviewed, this focus will broaden to include other appropriate types of evidence.

As far as possible the recommendations made here are based on empirical evidence and the conclusions of Cochrane Methods Groups (see Section I of the Handbook). The Collaboration, and the science of reviewing research, are still in the early stages of their development,[5,6,7,8] and so for some problems only general suggestions can be offered. As part of their shared responsibility for promoting research to improve reviews, the Cochrane centres are developing a directory of completed and ongoing methodological studies. Directions for accessing this directory will be added to the Handbook when the directory becomes available for general use.

The Cochrane centres are jointly responsible for coordinating development of the standards and guidelines reflected in the Handbook. Each collaborative review group is responsible for helping to ensure the reliability of the reviews it produces, and may choose to further standardize the approach that it uses. For example, a review group may choose to specify standard criteria for assessing the quality of trials or standard statistical techniques to synthesize results (within the limitations of the software used to prepare and display reviews).

The quality of a review is limited by the availability of good evidence and the resources needed to synthesize the evidence. Reviewers are responsible for finding the resources they need. In most cases the main cost will be their own time. However, some efforts to ensure the reliability of a review can entail substantial effort, such as tracking down missing information or obtaining individual patient data. How far such efforts are warranted varies from question to question, and reviewers must sometimes weigh the costs of additional efforts to improve the methodological rigour of their review against the potential gain in information.

Reviewers are also responsible for keeping their reviews up to date and for amending their reviews in response to valid criticisms. This can be a burden, but it can also be a comfort. Because Cochrane Reviews are published electronically and regularly updated, the initial effort of preparing a systematic review can be built upon. Instead of quickly becoming obsolete like most printed reviews, Cochrane Reviews can be amended as needed and kept up to date. When valid criticisms are raised, reviewers can make amendments so that any mistakes that they might have made will not remain in print to haunt them forever.

2. FORMAT OF A COCHRANE REVIEW

The format of a Cochrane Review has several objectives. It helps readers to find the results of research quickly and to assess the validity, applicability and implications of those results. It guides reviewers to report their work explicitly and concisely, and minimizes the effort required to do this. The format is also suited to electronic publication and updating, and it generates reports that are informative and readable when viewed on a computer monitor or printed.

Mike Clarke, Murray Enkin, Chris Silagy, and Mark Starr developed the format of a Cochrane Review, with input from many others. The format is flexible enough to fit different types of reviews, including those making a single comparison, those making multiple comparisons and those prepared by collaborative trialists' groups using individual patient data. Modifications of the format of Cochrane Reviews may be desired for a variety of reasons. However, because of the huge effort it can take to change the structure of reviews in the Cochrane Database, the format must be well defined and fixed. The Review Manager software included with the Handbook is designed to help authors in constructing reviews in the appropriate format and to prepare files required to transfer reviews electronically to the Cochrane Database.

Each review consists of:

- **a cover sheet** - giving the title, citation details and contact addresses
- **an optional abstract** - using a structured format
- **the text of the review** - consisting of an introduction (background and objective), materials (selection criteria and search strategy) and methods, results (trials included, methodological quality and other characteristics of the trials, and results of the review), discussion and conclusions.
- **standard tables and figures** - showing characteristics of the included trials, specification of the interventions that were compared, the results of the included trials, and a log of the trials that were excluded.
- **references**

Standard headings and tables guide reviewers preparing a report and make it easier for readers to identify information that is of particular interest to them. The headings are listed below and the rest of this chapter describes the content that should follow each of these. The headings are shown in **UPPER CASE BOLD TYPE**. An example of a Cochrane Review is attached (Appendix 1).[9]

Outline of a Cochrane Review

Cover sheet:
- TITLE
- SHORT TITLE
- REVIEWERS
- DATE ENTERED IN COCHRANE DATABASE
- DATE OF MOST RECENT AMENDMENT
- DATE OF MOST RECENT SUBSTANTIVE AMENDMENT
- THIS REVIEW SHOULD BE CITED AS
- CONTACT ADDRESS
- MODULE EDITORIAL ADDRESS
- SOURCES OF SUPPORT TO THE REVIEW
- SOURCES OF SUPPORT TO THE MODULE
- COCHRANE COLLABORATION REVIEW NUMBER

Abstract (optional):
- OBJECTIVE
- DATA SOURCES
- STUDY SELECTION
- DATA EXTRACTION
- DATA SYNTHESIS
- CONCLUSIONS

Text:
- BACKGROUND
- OBJECTIVE
- CRITERIA FOR CONSIDERING TRIALS FOR THIS REVIEW
  - TYPES OF PARTICIPANTS
  - TYPES OF INTERVENTION
  - TYPES OF OUTCOME MEASURES
- SEARCH STRATEGY FOR IDENTIFICATION OF TRIALS
- METHODS OF THE REVIEW
- TRIALS INCLUDED *(if there is no table of characteristics of trials)*
- METHODOLOGICAL QUALITY OF INCLUDED TRIALS
- OTHER CHARACTERISTICS OF INCLUDED TRIALS
- RESULTS
- DISCUSSION
- CONCLUSIONS
  - IMPLICATIONS FOR PRACTICE
  - IMPLICATIONS FOR RESEARCH
- ACKNOWLEDGEMENTS

T*ables and figures*:
- CHARACTERISTICS OF TRIALS CONTRIBUTING DATA TO THE REVIEW
- COMPARISONS
- DATA
- TRIALS EXCLUDED

References:
- REFERENCES TO TRIALS INCLUDED IN THIS REVIEW
- REFERENCES TO TRIALS EXCLUDED FROM THIS REVIEW
- REFERENCES TO TRIALS AWAITING ASSESSMENT
- ONGOING TRIALS
- ADDITIONAL REFERENCES
- PREVIOUSLY PUBLISHED VERSIONS

2.1 Cover sheet

The cover sheet includes the following headings and content (see the Appendix 1 for an example):

TITLE: The title should succinctly state the focus of the review. It should make clear the intervention(s) reviewed and the problem at which the intervention is directed. Someone scanning the title should be able to decide quickly whether the review addresses a question of interest.

SHORT TITLE: This can be up to 36 characters.

REVIEWERS: This can be the name of an individual (surname followed by initials), several individuals or a collaborative group (eg. Antiplatelet Trialists' Collaboration).

ENTERED IN COCHRANE DATABASE: This date will be entered by the module editors.

MOST RECENT AMENDMENT: This date will be entered automatically.

MOST RECENT SUBSTANTIVE AMENDMENT: The reviewer(s) must decide whether an amendment is substantive or not. Non-substantive amendments include corrections of typographical errors.

THIS REVIEW SHOULD BE CITED AS: The standard citation includes the names of the reviewer(s), the module, the module editors, the review number, date of most recent substantive amendment, and publication details (eg. for the Update Software publication of the Cochrane Database or a specialized database of Cochrane Reviews). (See, for example, reference **Error! Bookmark not defined.**.)

CONTACT ADDRESS: An address for direct contact with the reviewer(s) should be provided with the country on its own line, the full international version for telephone and fax numbers, and an e-mail address where possible.

MODULE EDITORIAL ADDRESS: An address for direct contact with the module editors should also be provided, also with telephone, fax and e-mail.

SOURCES OF SUPPORT TO THE REVIEW: Reviewers should give details of grants that supported the review and other forms of support, such as support from their university or institution in the form of a salary.

SOURCES OF SUPPORT TO THE MODULE: Details of grants that support the module group should be given here. The editorial team is responsible for ensuring that this information is correct.

COCHRANE COLLABORATION REVIEW NUMBER: This will be assigned by the module editors.

2.2 Optional abstract

Many Cochrane Reviews are not much longer than a structured abstract. It would be redundant to include an abstract for these. For longer reviews, reviewers can prepare a structured abstract of not more than 250 words. Reviewers preparing an abstract should follow the detailed instructions published in the *Annals of Internal Medicine*.[10,11] The content under each heading in the abstract should be as follows. Detailed instructions and an example can be found in Appendix 2.

OBJECTIVE: This should be a precise statement of the primary objective of the review. It should include information about the specific population, intervention and outcomes that are reviewed.

DATA SOURCES: A succinct summary of data sources (the search strategy used to identify trials) should be given including any time restrictions or other constraints (eg. English language publications only).

STUDY SELECTION: The number of studies included in the review and the criteria used to select studies for detailed review should be stated.

DATA EXTRACTION: This should describe how data were obtained and how validity was assessed.

DATA SYNTHESIS: The main results of the review, whether qualitative or quantitative, should be stated. Methods used to obtain these results should be outlined.

CONCLUSIONS: The primary implications for practice should be clearly stated, avoiding overgeneralization. Important implications for further research can also be stated.

2.3 Text

The text of the review should be as succinct as possible. It should be written so that someone who is not an expert in the area can understand it.

BACKGROUND: The review should begin with a brief synthesis of the underlying biology and health care of the topic being reviewed. This background should make clear the motivation and rationale for the review. It should be presented in a fashion that is understandable to the consumers of that health care.

OBJECTIVE: This should begin with a precise statement of the primary objective of the review, including the intervention(s) reviewed and the targeted problem. It might also mention why this review was undertaken and how it might relate to a wider review of a general problem. Any prior hypotheses should be stated and the comparisons that are made in the review should be consistent with these. The comparisons made should be listed in the **COMPARISONS** table. If a review addresses more than one hypothesis and includes several comparisons, the comparisons should be grouped for each hypothesis and referenced in brackets, as illustrated in the following example:[12]

```
OBJECTIVE:

To determine the effectiveness of nicotine replacement therapy
(NRT) (including gum, transdermal patch, intranasal spray and
inhaled preparations) in achieving long-term smoking cessation.

We wished to test the following hypotheses:

1) Use of NRT is more effective than placebo or "no NRT"
intervention in promoting smoking cessation (Comparison 1).

2) NRT is more effective when offered to smokers who are
motivated to quit and will, therefore, be more effective in
clinical settings which selectively recruit motivated smokers
(Comparison 2).

3) NRT is more effective in highly dependent smokers compared
with those who are only moderately dependent (Comparison 3).

4) The provision of high-intensity support, in addition to the
use of NRT, will be more effective in producing abstinence than
addition of low-intensity support programmes (Comparison 4).
```

CRITERIA FOR CONSIDERING TRIALS FOR THIS REVIEW: The criteria used to select trials for inclusion in the review should be stated. **TYPES OF PARTICIPANTS**, **TYPES OF INTERVENTION** and **TYPES OF OUTCOME MEASURES** are subheadings in this section. The methodological criteria that were used should also be clear (eg. "all randomised controlled comparisons" or "all double blind randomised controlled trials"), for example:**Error! Bookmark not defined.**

```
All randomised controlled comparisons of nicotine replacement
```

```
therapy (including nicotine chewing gum, transdermal nicotine
patches, nicotine nasal spray, and nicotine inhalers) versus
placebo or no nicotine replacement therapy control.
```

SEARCH STRATEGY FOR IDENTIFICATION OF TRIALS: The data sources used to identify trials should be summarized, including bibliographic databases, reference lists from pertinent articles and books, conference proceedings and personal contact with experts or institutions active in the area. The databases searched and the terms used should be stated, including any constraints, such as language. Any hand-searched journals should be listed. If a collaborative review group has developed a specialized register of trials, reviews contributed by that group can include or refer to a standard description of how trials were identified, noting any additional data sources used.

METHODS OF THE REVIEW: This should include the method used to apply the selection criteria (eg. if they were applied independently by more than one reviewer), the criteria used to assess the quality of studies and how they were applied, how data were obtained (eg. if individual patient data were sought, if the number of events was calculated from published survival curves), how the data were synthesized, any statistical techniques used and sensitivity analyses performed. If a collaborative review group uses a standard approach for all of the reviews prepared by that group, the Methods section can reference a description of those methods (see Section 0); eg. reviews contributed by the Pregnancy and Childbirth collaborative review group can reference *Effective Care in Pregnancy and Childbirth*[13]. Similarly, if a Cochrane Methods Group has recommended a standard approach and a review uses that approach, the Methods section can reference the Group's Report; eg. for assessing the quality of RCTs or for statistical analysis. Brief descriptions of these standard approaches will be added to future editions of the Handbook and reviewers will have the option of inserting these in the Methods section of manuscripts prepared for publication in journals or elsewhere.

TRIALS INCLUDED: This section should be included only if there is not a table of **CHARACTERISTICS OF TRIALS CONTRIBUTING DATA TO THE REVIEW**, eg. for reviews prepared by collaborative trialist groups if resources are insufficient for preparing a table. Reviewers can decide the order of trials in this list, but it must be the same order used in the review's tables and figures. Reviewers should give each trial a short-name (maximum 20 characters, typically the first author or a group name and year of publication; eg. Abelin 1989) to help the reader to identify it later in the review, and state the number of participants in each trial. Reviewers should mark trials to show if data were obtained from published reports only (with "P"), if additional unpublished data were obtained from the trialists (with "M" for mixed), if additional unpublished data were sought but not obtained for a trial (with "S" for sought), and if only unpublished data were used in the analyses (with "U").

METHODOLOGICAL QUALITY OF INCLUDED TRIALS: This should describe the general quality of the included trials and any important flaws in individual trials. If the quality of each trial was assessed using explicit criteria, the criteria that were used should be described or referenced under **METHODS OF THE REVIEW**. How each trial scored on each criterion can be summarized in this section or, preferably, included in the table of

CHARACTERISTICS OF TRIALS CONTRIBUTING DATA TO THE REVIEW.

OTHER CHARACTERISTICS OF INCLUDED TRIALS: This should include a summary of the key characteristics of the types of participants, interventions and outcome measures used in the included trials and any important differences among the trials. The main characteristics of each trial should be reported in the table of **CHARACTERISTICS OF TRIALS CONTRIBUTING DATA TO THE REVIEW**. The sex and age range of participants should be stated here unless it is obvious (eg. if all the participants are pregnant). Reviewers must decide what other characteristics of the trials it is important for readers of the review to know.

RESULTS: This should be a summary of the main findings of the review and any sensitivity analyses that were undertaken. Subheadings can be used if they make reading easier (eg. for each prior hypothesis if a review addresses more than one). The results of individual trials, and any statistical summary of these, should be included in the **DATA** table. Reviewers should be cautious about reporting subgroup analyses. They should avoid making inferences in this section. A common mistake to avoid (both in describing the results and in drawing conclusions) is the confusion of 'no evidence of an effect' with 'evidence of no effect'.

DISCUSSION: This should include brief comments on any methodological limitations of the included trials and the review that are important for decisions about practice or future research. Comments on how the included trials fit into the context of other evidence might be included here, stating clearly whether the other evidence was systematically reviewed. Comments on how the results of the review fit into the context of current clinical practice might be included here, although reviewers should bear in mind that current clinical practice might vary internationally.

CONCLUSIONS: IMPLICATIONS FOR PRACTICE and IMPLICATIONS FOR RESEARCH are subheadings. The implications for practice should be as practical and unambiguous as possible. They should not go beyond the evidence that was reviewed.

The implications for research should not include statements like "more research is needed." Reviewers should state exactly what research is needed, why and how urgently. Opinions on how the review might be improved with additional data or resources might also be included here.

For reviews prepared by collaborative trialists' groups, if it was not possible to reach a consensus regarding implications for practice or research, the following statement can be included under **CONCLUSIONS**:

> *This review is the result of an ongoing process involving the collaborative effort of many researchers worldwide. It has not been possible to achieve a consensus among all of the collaborators on the implications for . . .* (*practice* **or** *research*). *This review presents all of the evidence from RCTs on the effects of . . . , in the hope that this will help anyone involved in the treatment of this disease.*

ACKNOWLEDGEMENTS: Credit should be given where credit is due. This could include all the members of a collaborative trialists' group. Reviewers should ensure that anyone personally acknowledged is informed.

2.4 Tables and figures

CHARACTERISTICS OF TRIALS CONTRIBUTING DATA TO THE REVIEW: This is a standard table. Its design represents a compromise between ease of use and flexibility. The display limitations of typical monitors (80 characters across) constrains the amount of data that can be viewed for each trial at one time. The table has six columns: **Trial, Methodology, Participants, Comparison groups, Outcomes, and Notes**. There is also a column in which to record queries. It is possible to use codes so that each column can include several sub-columns; eg. a reviewer could include country, setting and sex under **Participants**. Alternatively, it is possible to include several rows of information for each trial in each column (eg. with one row for each of the sub-categories of information in the example that was just given). Reviewers must decide what characteristics of the included trials are likely to interest readers of the review.

Reviewers should indicate whether published or unpublished data were used for each trial using the codes defined under "**TRIALS INCLUDED**" in Section 0 above.

COMPARISONS: The comparisons made in the **DATA** tables should be listed here. A review can include more than one comparison and a trial can be included in more than one of these. (See, for example, Appendix 1.)

DATA: The data must be entered in a standardized format from which tables and figures for each comparison can be generated.

TRIALS EXCLUDED: Any trials meeting the inclusion criteria that were excluded should be identified and the reason for exclusion should be given (eg. data not available). (See, for example, Appendix 1).

2.5 References

The references are organized under five standard headings: **REFERENCES TO TRIALS INCLUDED IN THIS REVIEW**, **REFERENCES TO TRIALS EXCLUDED FROM THIS REVIEW**, **REFERENCES TO TRIALS AWAITING ASSESSMENT**, **ONGOING TRIALS** and **ADDITIONAL REFERENCES**.

Reviewers should check their references for accuracy![14,15] Examples of the correct forms of references for a journal article and book, respectively, are:

```
Liggins GC, Howie RN. A controlled trial of antepartum
glucocorticoid treatment for prevention of the respiratory
distress syndrome in premature infants. Pediatrics 1972;
50:515-25.
```

```
Crowley P. Promoting pulmonary maturity. In: Chalmers I, Enkin
MW, Keirse MJNC, eds. Effective Care in Pregnancy and
Childbirth. Oxford: Oxford University Press, 1989: 746-64.
```

Examples of the correct form for other references can be found in Appendix 3.[16] The titles of journals should be abbreviated according to the style used in *Index Medicus*.[17] If the number of authors exceeds six, list six followed by 'et al'.

REFERENCES TO TRIALS INCLUDED IN THIS REVIEW: Each trial should be given a short-name; eg. the last name of the first author and the year of publication. The trials should be listed alphabetically by short-name with the references (if there are more than one for a trial) arranged chronologically for each trial.

REFERENCES TO TRIALS EXCLUDED FROM THIS REVIEW: These should be listed alphabetically by a short-name with the reference(s). If there are more than one reference for a trial, they should be listed chronologically.

REFERENCES TO TRIALS AWAITING ASSESSMENT: Potentially relevant trials that have been identified, but cannot be assessed for inclusion until additional data or information are obtained, should be listed here. These need not have been cited in the text of the review.

ONGOING TRIALS: Trials which are ongoing that meet the inclusion criteria should be listed here with contact addresses, and information about whether recruitment is still open if this information is known and whether the investigators are seeking other centres to join the trial.

ADDITIONAL REFERENCES: Other references cited in the text should be listed here. If a report of a trial is cited in the text for some reason other than referring to the trial (eg. because of some background or methodological information in the report), it should also be listed here.

Trials should be identified by their short-names in square brackets in the text; eg. [Liggins 1972]. Other references should be identified by the last name of the author and the year of publication in parenthesis; eg. (Crowley 1989). The last name of the first author followed by 'et al' should be used if there are multiple authors; eg. (Antman et al 1992). If no author is given, the title of the journal or publication should be used; eg. (Lancet 1993) or (International Register of Perinatal Trials 1991). A semicolon should be used between references; eg. (Crowley 1989; Liggins et al 1977).

PREVIOUSLY PUBLISHED VERSIONS: References to any previously published versions of the review in a journal or textbook should be listed here.

3. DEVELOPING A PROTOCOL

As in any scientific endeavour, the methods to be used should be established beforehand. Since reviews are retrospective analyses of data collected by others, it is important to make the process as rigorous and well defined as possible.[7] However, it is also necessary to maintain a practical perspective. For each task in the protocol the expenditure of resources must be balanced against the magnitude of the associated threat to the review's validity.

Just as protocols for RCTs must sometimes be changed to adapt to unanticipated circumstances (such as problems with participant recruitment, data collection or unexpected event rates), changes in a review protocol are sometimes necessary. While every effort should be made to adhere to a predetermined protocol, it should be recognized that this is not always possible or appropriate. Changes in the protocol should not be made on the basis of the results. *Post hoc* decisions (such as excluding selected studies) that are made when the impact on the results of the review is known are highly susceptible to bias and should be avoided. As a rule, all changes in the protocol should be documented and reported, and "sensitivity analyses" (see Section 0) of the impact of such decisions on the results of the review should be made when possible.

Because the Materials and Methods sections in the structured report of a Cochrane Review contain the same information that is needed for the protocol, the protocol for a Cochrane Review should consist of drafts of the appropriate sections of the final report. This reduces the work needed to prepare the protocol and the report of the completed review. The sections of the report that should be (at least partially) completed to register a planned review are:

- the cover sheet
- the objectives
- the criteria for considering trials for the review
- the search strategy for identification of trials
- the methods of the review

The content of these sections for a completed report is described in Chapter 0. The content to be included in a protocol is described below after listing the resources that might be required for a systematic review. Guidelines for specific methodological decisions regarding the identification, selection and assessment of trials, data collection and analysis are given in Chapters 0 - 0.

An example of a protocol prepared using the Cochrane Review Manager is attached (Appendix 4).

3.1 Resources for a systematic review

Individual Cochrane Reviews are prepared by reviewers working in collaborative review groups. Each collaborative review group has an editorial team responsible for

producing a module of edited reviews for dissemination through the Cochrane Database of Systematic Reviews, as described in Sections I and II of the Handbook. Although the UK Cochrane Centre is responsible for managing the Cochrane Database, it has no editorial responsibility for the reviews.

Because the Cochrane Collaboration is built around collaborative review groups, it is important that each reviewer is linked with a review group. Besides ensuring that Cochrane Reviews are appropriately edited, this structure reduces the burden placed on individual reviewers since the editorial teams are responsible for providing most or all of the following types of support:

- conducting systematic searches for relevant studies and coordinating the distribution of potentially relevant studies to reviewers
- establishing specific standards and procedures for the review group
- ensuring that reviewers receive the methodological support they need

The main resource required by reviewers is their own time. The majority of reviewers will contribute their time "freely" because it will be viewed as part of their existing efforts to keep up to date in their areas of interest. In some cases, reviewers may need additional resources or, at least, be able to justify the amount of time required for a systematic review to colleagues who do not yet understand either what systematic reviews entail, or their importance.

The amount of time required will vary, depending on the topic, the number of trials, the methods used (eg, the extent of efforts to obtain unpublished information), the experience of the reviewers, and the types of support provided by the editorial team. The workload associated with undertaking a review is thus very variable. However, consideration of the tasks involved and the time required for each of these might help a reviewer to estimate the amount of her or his time that will be required. Each of the following tasks may take more or less time, depending on the topic and the reviewer. Estimates for an individual patient data review are very different and will be given in Chapter 0.

Task	Time
Training	Reviewers may want to budget time to attend one or more workshops on how to do a systematic review (organized by one of the Cochrane centres)
Meetings	This might include meetings of the collaborative review group and, if more than one person is working on the review, meetings of the team working on the review.

Protocol development	This might take a week (40 hours) or more, depending on the extent to which a collaborative review group has standardized the approach taken to preparing reviews. It can require an investment of time similar to that needed to prepare a grant application.
Searching for trials	This might include searching bibliographic databases, personal communication with experts in the area and, in some cases, hand-searching highly relevant sources, in addition to checking reference lists, to identify additional trials.
Assessing citations and full text reports of trials for inclusion in the review	Citations can be assessed for relevance quickly, but it may be necessary to review a large number of citations to ensure that as many relevant trials as possible are identified and included in the review. Similarly, it may be necessary to retrieve and review a large number of full text reports. For most Cochrane Reviews the main costs of identifying, retrieving and distributing trials will be assumed by the editorial team. Nonetheless, reviewers should anticipate reviewing lists of citations distributed to them by the editorial team, full text reports of any trials that might be relevant, and the reference lists of relevant full text reports that are retrieved. Additional costs might include the cost of ordering copies of articles not locally available, photocopying costs, and help with translating articles published in languages unfamiliar to the reviewer.
Assessing the quality of included trials and obtaining data	Assessing the quality (validity) of trials that meet the inclusion criteria for the review can take much longer than assessments of relevance. The time required can vary substantially from trial to trial. Depending on the criteria that are used, the average time per trial may range from 10 to 45 minutes.[18] It may be necessary to obtain additional help with translating reports of included trials published in languages unfamiliar to the reviewer.
Pursuing missing data and unpublished studies	Efforts to obtain data missing in published reports and unpublished trials may require resources for postage, facsimiles, telephone and secretarial help.
Analyzing the data	In some cases, help may be needed with statistical analyses that are not easily accomplished using the Cochrane Review Manager software.
Interpreting the results and preparing a	The amount of time needed to prepare a report will vary depending on the scope of the review

report	and the ability of the reviewer, but may be similar to that required to prepare a report of a primary study for publication.
Keeping the review up to date	The amount of time required to keep a review up to date will depend largely on the number of relevant ongoing or new trials. Additional time may be required to respond to criticisms of the review and to revise it in light of valid criticisms.

In summary, resources that might be required, in addition to the reviewers' time, include:

- searching (identifying trials is primarily the responsibility of those involved in developing registers of trials, usually the editorial teams of the review group as described in Section V of the Handbook. However, reviewers may share this responsibility and it may be appropriate to search additional databases for a specific review.)
- help for library work and photocopying
- a second reviewer, possibly a student or research assistant, to assess trials for inclusion, assess the quality of included trials, obtain data and conduct analyses
- statistical support for synthesizing the results of the included trials
- supplies and services (long distance telephone charges, facsimiles, paper, printing, photocopying, audiovisual and computer supplies)
- office space for support staff
- travel funds

Many organizations currently provide funding for systematic reviews and additional agencies are likely to recognize the importance of supporting this type of work in the future. These include research funding agencies, those organizations that provide or fund health care services, those responsible for health technology assessment and those involved in the development of clinical practice guidelines. Although applications for funding need to adhere to the requirements of the funding organization to which one is applying, a general outline of an application for funding for a systematic review should contain the following elements:

- Objectives
- Rationale
- Design of the review
  - General approach

```
        !  Identification of trials
        !  Selection of trials for inclusion
        !  Assessments of the validity of included trials
        !  Obtaining data for the included trials
        !  Analysis
        !  Inferences and presentation of results
  !  Time-chart for major activities
  !  Budget
```

The objectives and design of a review are addressed below. Describing the rationale for a systematic review is analogous to describing the "present state of knowledge" in a grant application for a primary study and may include a review of prior reviews on the same topic. The same scientific principles that apply to a review of trials can be applied to a review of reviews. The fundamental difference is the unit of analysis, which for a review of reviews is the review article.

Methodological issues that might need to be considered in reviewing reviews include how they will be identified, selected for detailed review and assessed.**Error! Bookmark not defined.**,**Error! Bookmark not defined.**,[19,20,21] Reasons for undertaking a review of prior reviews, in addition to providing the rationale for an application for funding, include learning from what earlier reviewers have done, avoiding unnecessary duplication of efforts, identifying trials (including unpublished trials), and collecting background information that may be important in interpreting the results of individual trials and drawing conclusions from the results of the review.

A time chart with target dates for accomplishing key tasks (such as those listed above) can help with scheduling the time needed to complete a review, for example:

Possible time-chart for major activities

Month*	Task	Notes
1 - 6	Additional searches for published and unpublished trials	For most Cochrane Reviews the individual reviewer's responsibility for identifying relevant trials begins with reviewing the reference lists of relevant reports. This responsibility is ongoing. However, it may be useful to set a target date for completing reasonable efforts to ensure that as many relevant trials as possible have been identified before submitting the review to the editorial group. (See Chapter 0.)

Month*	Task	Notes
1	Pilot test of inclusion criteria	It may be important to test the reproducibility of decisions about which reports to retrieve and which trials to include in a review; eg. by having two or more reviewers each apply the inclusion criteria to a sample of citations and a sample of full-text reports to test inter-reviewer agreement. (See Chapter 0.)
1 - 6	Relevance assessments	Individual reviewers are responsible for assessing whether trials distributed by the editorial team and trials identified through other means meet the inclusion criteria for the review. This responsibility begins with the first set of reports distributed by the editorial team and continues after the review has been entered in the *Cochrane Database.* (See Chapter 0.)
1	Pilot test of validity criteria	Because assessments of validity tend to require judgement, it is important to test the reproducibility of these judgements; eg. by having two or more reviewers each apply the validity criteria to a sample of relevant trials to test inter-reviewer agreement. (See Chapter 0.)
1 - 8	Validity assessments	The validity of each relevant trial that is identified should be assessed, including new trials identified after the review has been entered in the *Cochrane Database.* (See Chapter 0.)
1	Pilot test of data collection	Reviewers may want to develop paper forms for data collection (and validity assessments) or they may choose to enter data directly using the Review Manager software. In either case, it is desirable to, test the reliability of data collection by having two or more reviewers collect data for a sample of relevant trials and, if possible, to do this for all of the relevant trials. (See Chapter 0.)
1 - 8	Data collection	
1 - 8	Data entry	
2 - 8	Missing information	It may be important to continue efforts to collect missing information (eg. by contacting investigators by mail, facsimile or telephone) after a review is entered in the *Cochrane Database*, if the missing data could contribute substantively to the conclusions that can be drawn from the review. (See Chapter 0.)
6 - 8	Analysis	(See Chapter 0.)
1 - 9	Preparation of	This should begin with the preparation of the protocol and be completed by preparing

Month*	Task	Notes
	report	the Results, Discussion and Conclusions sections (Chapter 0), and the standard tables and figures (Chapter 0).
10 -	Keeping the review up to date	(See Chapter 0.)

*As noted above, these targets may vary widely from review to review. Reviewers, together with the editorial team for the review group, must determine an appropriate time frame for a specific review.

3.2 Cover sheet

The following information should be included in the cover sheet (see Section 0) when a protocol for a Cochrane Review is to be registered:

- the title
- a short title
- the name(s) of the reviewer(s)
- the date of registration (will be assigned when entered in the Database)
- contact address for the reviewer(s)
- contact address for the module editors (provided by the module editors)
- sources of support for the review (including salary support for reviewers, any applications for funds and funds already granted)
- Sources of support for the module (provided by the module editors)
- The Cochrane Collaboration review number (assigned by the module editors)

3.3 Background and objective

The protocol should begin with a brief synthesis of the underlying biology and health care issues that provide the motivation and rationale for the review. The Background section should, in general, be brief and focus on information that will help users of the review to understand and interpret the results of the review.

As with any research project, the first and most important step in preparing a review is to ask a clear question. Poorly focused questions spawn fuzzy decisions about what research to include and how to summarize it. There are three elements, which taken together, define the topic being reviewed in a Cochrane Review: the population (types of participants), the

types of interventions and the types of outcome measures of interest.

This simple, yet powerful model for framing questions can be used to guide each stage of a review, including: formulating a strategy to locate and select relevant trials, critically appraising their quality and analyzing variation among their results.

In order to protect their conclusions from threats to validity in formulating the problem that a review will address, reviewers should begin with a broad conceptual definition in mind, and subsequently focus attention on differences in the study populations, interventions and endpoints (as well as differences in study design) that might be associated with differences in results. An example of this sort of approach is a systematic review of secondary prevention of vascular disease by antiplatelet treatment that included patients with either cardiovascular or cerebrovascular disease treated with any antiplatelet agent and examined both fatal and nonfatal endpoints.[22] As it turned out, no statistically or clinically important differences in results were found among this apparently heterogeneous group of trials. This strongly supports the general conclusion that antiplatelet agents are effective for the secondary prevention of vascular disease.

3.4 Criteria for considering trials for a review

The criteria used to decide which trials to include in the review should be explicit and should specify the types of people, interventions and outcomes of interest. In addition, methodological criteria should be specified. An example is provided below:**Error! Bookmark not defined.**

CRITERIA FOR CONSIDERING TRIALS FOR THIS REVIEW:

Types of participants

Smokers of either gender were included irrespective of the setting from which they were recruited or their initial level of nicotine dependency. Studies which randomised therapists, rather than smokers, to offer NRT or a control were included provided that the specific aim of the study was to examine the effect of NRT on smoking cessation. Trials which randomized physicians or other therapists to receive an educational intervention, which included encouraging their patients to use NRT, were not included but are being handled as part of a separate review.

Types of intervention

All randomised controlled comparisons of nicotine replacement therapy (including nicotine chewing gum, transdermal nicotine patches, nicotine nasal spray, and nicotine inhalers) versus placebo or no nicotine replacement therapy control.

Randomised trials of different doses of nicotine replacement therapy were also included.

Types of outcome measures

```
We confined the review to a comparison of the effects of NRT
versus control on smoking cessation, rather than withdrawal
symptoms. Trials in which follow-up was brief (less than 6
months), or which did not include measurement of smoking
cessation, were also excluded.

In each study the strictest available criteria to define
abstinence were used. For example, in studies where biochemical
validation of cessation was available, only those participants
who met the criteria for biochemically confirmed abstinence
were regarded as being abstinent. Wherever possible a sustained
cessation rate, rather than point prevalence, was used. In
trials where patients were lost to follow-up they were regarded
as being continuing smokers.
```

3.5 Search strategy for identification of trials

The strategies used to develop and maintain a specialized register of RCTs should be specified by the editorial team. These can be referenced rather than described in detail in each review that utilizes the register. (See Section V of the Handbook.) Additional strategies used to ensure that as many relevant trials as possible are identified for the review should be described by the reviewer (see Chapter 0).

3.6 Methods of a review

A **target date for completion of the review** should be stated at the beginning of this section. The section should include a description of:

- the methods used to select trials for inclusion (see Chapter 0)
- the criteria and methods used to assess the methodological quality of the included trials (see Chapter 0)
- the methods used to collect data from the included trials (see Chapter 0)
- the methods used to synthesize the data (see Chapter 0)

Any of the methods that are standardized, either within a collaborative review group or within the Collaboration, can be referenced rather than described in detail in each review. The editorial team of each collaborative review group is responsible for preparing a description of any standard methods that are used by all the reviewers in that particular group.

3.7 Registering a protocol

Once a protocol has been completed it should be sent to the module editors to consider. When the editors are satisfied with the protocol they will register it in the *Cochrane Database*. Registered protocols will be entered into the *Cochrane Database of Systematic Reviews* and disseminated. Because this will raise expectations and may discourage others from undertaking a review on the same topic, editors and reviewers should not register a protocol unless there is a firm commitment to complete the review within a reasonable time

frame.

3.8 Methods of a collaborative review group

The editorial team of each collaborative review group is responsible for documenting the methods used by the team for editing the module (see Section II of the Handbook). The editorial team is also responsible for documenting any specific methods used by the group beyond the standard methods specified in the Handbook, including:

- any standard methodological criteria for including trials in reviews that differ from those for inclusion in a register (described in Section V of the Handbook)
- the search strategies used to develop and maintain the register of trials used by the group and method of distributing potentially relevant citations or full-text reports to reviewers
- any additional search strategies that reviewers are instructed to use routinely
- any standard methods used to select trials for reviews
- any standard criteria or methods used to assess the methodological quality of included trials
- any standard methods used for data collection
- any standard methods used for synthesizing data
- any standard methods used for deriving conclusions or indicating the strength of the evidence on which the conclusions are based
- any decision rules used to categorize interventions (as described in Section 0)
- any specific rules used for preparing the standard tables and figures
- the methods used to keep reviews up to date and respond to criticisms
- methods used to review protocols

Descriptions of specific methods used by each collaborative review group will be published electronically with the *Cochrane Database of Systematic Reviews*.

4. IDENTIFICATION OF TRIALS

The reviewers responsible for a specific review and the editorial team for the review group share the responsibility for ensuring that as many trials as possible that meet the inclusion criteria for the review are identified. In general, the editorial team is responsible for developing and maintaining, or accessing an appropriate register of RCTs. In keeping with the principles guiding the development of the Cochrane Collaboration, each review group should aim to meet its particular needs to identify trials within its scope in a way that will minimize needless duplication of efforts among review groups (see Section V of the Handbook).

Typically, the editorial team will seek to use whichever register(s) is (are) appropriate to identify any trials that are potentially relevant for a specific review and to distribute citations or full-text reports of those trials to the people undertaking the review. If a review group is not organized in this way, the editorial team must ensure that reviewers in the group know which register(s) of trials or bibliographic databases to access and how to use them.

An international register of RCTs is expected to be available by 1995. It should then be possible to access all registered trials through the US National Library of Medicine's MEDLINE database. Further information regarding this is included in Section V of the Handbook.

Reviewers are responsible for checking reference lists to identify additional trials. In addition, they may want to use Science Citation Index (SCISEARCH) or other bibliographic databases to further ensure that as many relevant trials as possible have been identified. They may also want to communicate personally with investigators, other experts and manufacturers in the area to try to identify additional published and unpublished trials. In addition, they may want to hand-search specialist journals and other materials (eg. conference proceedings) that are particularly relevant to the focus of the review (see Section V of the Handbook).

4.1 Checking reference lists

Reviewers should check the reference lists of all relevant articles that are obtained. Additional, potentially relevant, articles that are identified should be retrieved and assessed for possible inclusion in the review. The potential for reference bias when doing this should be kept in mind.[23]

4.2 Science Citation Index (SCISEARCH)

Frequently cited articles can be identified and sought by searching forward in time (for other articles that cite them) electronically using SCISEARCH or manually using *Science Citation Index* from their year of publication. Online searches using SCISEARCH can be expensive. CD ROM publications of SCISEARCH are available at many university libraries

and can be used at no cost to reviewers. The value of using SCISEARCH or *Science Citation Index* for a particular review will depend on the nature of the topic. For topics that are poorly indexed and for which trials may have been published in sources not indexed by MEDLINE, citation searching could be an important means of identifying additional trials. However, insufficient evidence is currently available to suggest that routine use of citation searches is warranted, given the costs involved.

4.3 Other bibliographic databases

Text words that regularly appear in the title and abstract of reports of the relevant trials that have been identified should be noted. The citations can be entered into MEDLINE and the key words with which the papers are indexed, noted. These text words and key words can then be used to conduct a comprehensive MEDLINE search of the medical literature from 1966 to the present if there is concern that trials indexed in MEDLINE may have been missed. For most reviewers working in collaborative review groups with access to a register of trials this should not be necessary. It may, however, be appropriate, to use the same strategy for other computerized databases such as EMBASE (*Excerpta Medica* online). *Online Databases in the Medical and Life Sciences*,[24] a directory of 795 databases, can be used to locate and assess the relevance of additional databases. As noted above for SCISEARCH, other bibliographic databases are most likely to be helpful in identifying additional trials when a topic is poorly indexed or when trials are likely to have been published in sources not indexed by MEDLINE. For example, although EMBASE and MEDLINE overlap in coverage, EMBASE indexes non-English language journals not indexed by MEDLINE and uses different indexing terms.

4.4 Personal communication

To find additional published and unpublished trials, and further ensure that relevant trials have not been missed, it may be worth contacting investigators and other experts in an area directly. A comprehensive list of relevant articles (along with the inclusion criteria for the review) can be constructed and a letter sent to the first author of each paper, asking the authors if they know of any published or unpublished studies relevant to the area not included in the list. It may be desirable to send copies of the same letter to other experts and manufacturers in the area.

4.5 Full-text searching

If there are one or more specialist journals or "grey literature" (conference proceedings, dissertations, government documents and other forms of reports outside the main journal literature) where trials relevant to the review are likely to have been published, reviewers may be wish to hand-search these sources. Reviewers should contact the Baltimore Cochrane Center before undertaking full-text searches. The Baltimore Centre is responsible for maintaining a directory of who is searching which journal. Over 80 journals had been entered to the database by November 1993. Checking with the Center will help to avoid duplication of effort. The Center may also be able to give reviewers additional guidance on how best to proceed (see Section V the Handbook).

4.6 Registration of RCTs not already in a register

Additional trials identified by reviewers which have not already been registered should be added to relevant specialized register(s) and to the comprehensive international register of RCTs. The method of transferring eligible reports meeting the criteria for inclusion in a register is described in Section V of the Handbook.

4.7 Reviews prepared without the benefit of a register

Reviewers who have no access to a register of RCTs should work together with the editorial team of their collaborative review group to develop or obtain access to a register as quickly as possible. Because large resources are used inefficiently to identify trials without the benefit of a systematically developed and maintained register, the development of a register corresponding to a group's defined scope, or means of accessing one, should be a priority for all review groups.

4.8 Unpublished and ongoing RCTs

Unpublished studies may differ systematically from those which have appeared in peer-reviewed journals, not because their methodology is inferior, but because their results are "negative"; i.e. unpublished studies may demonstrate a smaller effect than published studies.[25,26,27,28,29,30,31,32,33] Reviewers should therefore attempt to assess the extent of "publication bias" in the area under review. Written reports of unpublished studies should be obtained if possible, and the quality of unpublished research should be evaluated using the same criteria that are used to evaluate published research. Sensitivity analyses can be used to assess the impact of unpublished research on the overall results (see Section 0).

Unpublished trials that cannot be adequately assessed can be included in the report of the review along with references to trials awaiting assessment (see Section 0) if there is substantial delay in obtaining the missing information needed to make the assessment.

Ongoing trials that meet the inclusion criteria should be listed with contact addresses, with a note indicating whether recruitment is still open and whether the investigators are seeking other centres to join the trial (if this information can be obtained).

4.9 Reference management

Reviewers already familiar with a bibliographic software program, such as Pro-Cite[34] or Reference Manager[35] (see Section V of the Handbook), may wish to use it to manage the references for their reviews and then import references into the Cochrane Review Manager. Others can enter references directly into the Review Manager software. References can only be entered for a specific review when using the Review Manager software. The software will assist reviewers in assigning names to trials, and arranging references for trials and other references under the appropriate headings: included trials, excluded trials, trials awaiting

assessment, ongoing trials and additional references (see Section 0).

5. SELECTION OF TRIALS

To ensure that as many relevant trials as possible are identified for the review the inclusion criteria should be applied liberally in deciding which articles to retrieve (when checking reference lists and other lists of citations). All the articles that are obtained must then be assessed to see whether the criteria have been met. Reviewers must decide:

- whether more than one reviewer will assess the relevance of each article
- whether the decisions concerning relevance will be made by content area experts, non-experts, or both
- whether the people assessing the relevance of trials will know the names of the authors, institutions, journal of publication and results when they apply the inclusion criteria
- what decision rule or procedure will be used when there is disagreement, if more than one reviewer applies the criteria to each article

Decisions about which trials to include in a review often involve judgement. To help ensure that these judgements are reproducible, it is desirable for more than one reviewer to apply the inclusion criteria to all the potentially relevant articles that are retrieved. However, the amount of judgement needed varies from review to review. Some reviewers may decide that the judgements that must be made about which trials to include in a review are quite straightforward. The use of multiple reviewers to assess relevance may appear frivolous in these cases. Whatever the case, the number of people assessing the relevance of each trial should be stated in the Methods section of the review (if it is not stated in a description of the methods used by all of the reviewers in a particular collaborative review group).

Experts in a particular area frequently have preformed opinions that can bias their assessments of both the relevance and validity of articles.**Error! Bookmark not defined.**,**Error! Bookmark not defined.**,[36,37,] Thus, while it is important that at least one reviewer is knowledgable in the area under review, it may be an advantage to have a second reviewer who is not an expert in the area.**Error! Bookmark not defined.**

Some reviewers may decide that assessments of relevance should be made by people who are blind to the journal from which the article comes, the authors, the institution, and the magnitude and direction of the results by editing copies of the articles.**Error! Bookmark not defined.** However, this takes much time, and may not be warranted given the resources required and the uncertain benefit in terms of protecting against selection bias.

Disagreements about whether a trial should be included can generally be resolved by discussion. Often the cause of disagreement is a simple oversight on the part of one of the reviewers. When the disagreement is due to a difference in interpretation, the issue should be resolved by consensus. Occasionally, it will not be possible to resolve disagreements about

whether to include a trial without additional information. In these cases, reviewers may choose to reference the trial as one that is awaiting assessment until the additional information is obtained.

For most reviews it will be worthwhile to pilot test the inclusion criteria on a sample of articles (say ten to twelve papers, including ones that are thought to be definitely eligible, definitely not eligible and questionable). The pilot test can be used to refine and clarify the inclusion criteria, train the people who will be applying them and ensure that the criteria can be applied consistently by more than one person.

6. ASSESSMENT OF METHODOLOGICAL QUALITY

It is important to assess the methodological quality of the trials included in a review for two reasons. Variation in quality can explain variation in the results of the included trials. More rigorous trials are more likely to yield results that are closer to the "truth". Statistical summaries (meta-analyses) of results from trials 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 to overestimating the effectiveness of the intervention being evaluated. They might also result in a "false negative" conclusion (concluding that there is not an effect when the truth is that there is) if the less rigorous studies provide less precise or biased estimates of the effects of the intervention, thereby obscuring the true effect.[38]

The second reason for assessing the methodological quality of the included trials is that it is important *per se*, even if there is not variability in the results or quality of the included trials. For example, it may be that the evidence is consistent, but all the studies are flawed. In such a case the conclusions of the review would not be nearly as strong as if consistent results were obtained from a series of high quality trials.

In the context of a systematic review, the "quality" of a trial can be defined as the extent to which its design and conduct are likely to have prevented systematic errors (bias).**Error! Bookmark not defined.**,[39] There are other aspects of quality (eg. scientific importance, clinical importance, literary quality) that may be important in other contexts (such as peer review of manuscripts submitted for publication), but which are less relevant in the context of a systematic review. Two aspects of trials that are important in the context of a systematic review, but should not be confused with "quality" (as defined here), are "precision" and "external validity".

The *precision* of the results of a trial (a measure of the likelihood of random errors) is important in the context of a systematic review. This is reflected in the confidence intervals around the estimates of effect from each trial, and in the weights given to the results of different trials when an overall estimate of effect (a weighted average of the measure of effect from each included trial) is derived using meta-analysis. More precise results are given more weight.

External validity (the extent to which the results are generalizable or applicable) is only meaningful with regard to a specified target population, intervention and outcome measure. In the context of a systematic review the reviewers' judgments about applicability or relevance are reflected in the criteria used to select trials for inclusion in the review.

Unfortunately, there are two major difficulties with assessing the methodological quality of trials. The first is inadequate reporting of trials. The second, which is in part a consequence of the first, is limited empirical evidence of the relationship between specific criteria and bias.[40]

David Moher and his colleagues have identified 22 scales and nine checklists that have been used to assess the quality of RCTs.**Error! Bookmark not defined.** The scales include anywhere from three to 35 items and take from 10 to 45 minutes to complete. Many of these scales are liable to confuse the quality of *reporting* with the quality of the *design and conduct* of a trial; i.e. by scoring whether something was reported (such as how participants were allocated) rather than whether it was done appropriately. Many also contain items that are not directly related to internal validity, such as whether a power calculation was done (an item that relates more to the precision of the results) or whether the inclusion and exclusion criteria were clearly described (an item that relates more to external validity than internal validity).

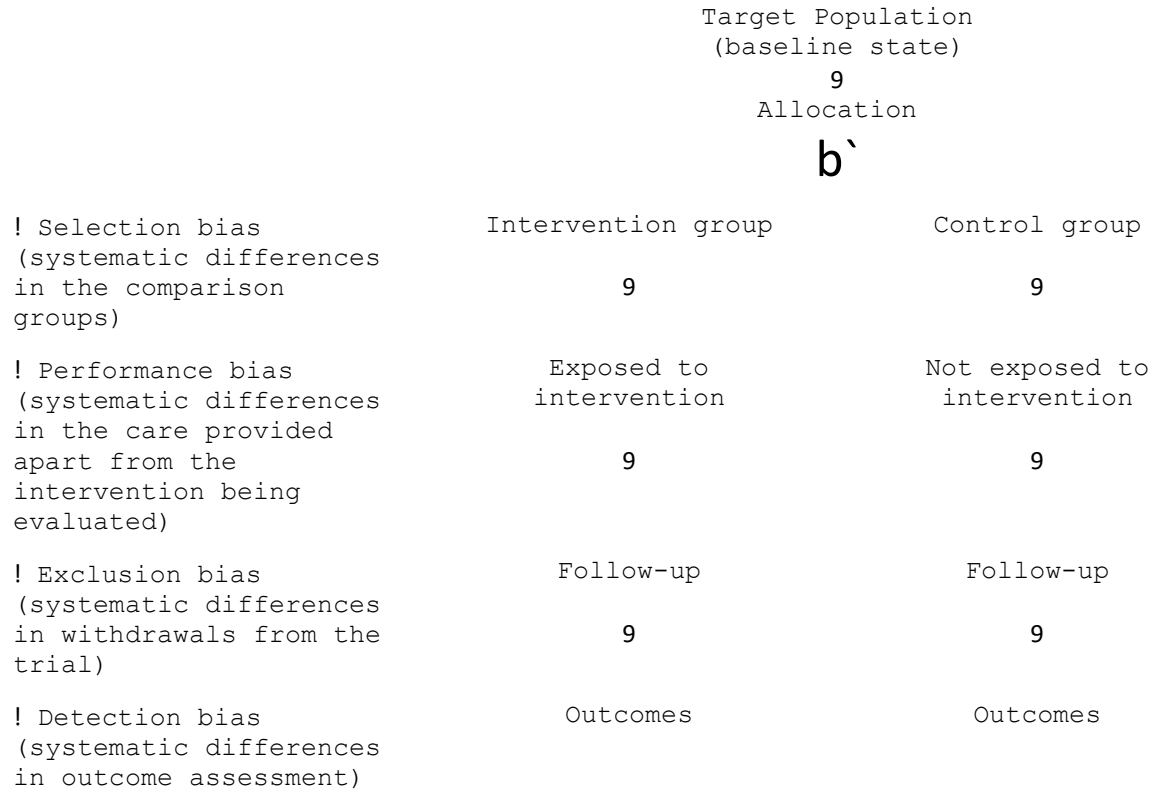
Because there is no "gold standard" for the "true" methodological quality of a trial, the possibility of validating any proposed scoring system is limited. While it is possible to apply basic principles of measurement to the development of a quality scale for RCTs,**Error! Bookmark not defined.** the relationship between a quality score and the degree to which a trial is free from bias is not obvious.**Error! Bookmark not defined.** None of the currently available scales for measuring the quality of trials can be recommended without reservation.**Error! Bookmark not defined.** If a reviewer or review group chooses to use such a scale, it must be with caution (see Section 0 below).**Error! Bookmark not defined.**

More complex scoring systems take more time to complete. They have not been shown to provide more reliable assessments of quality. They carry a greater risk of confusing the quality of reporting with the quality of the trial.**Error! Bookmark not defined.** They are more likely to include criteria that do not directly measure internal validity,**Error! Bookmark not defined.** and they are less likely to be transparent to users of the review. For these reasons, it is preferable to use simple approaches to assessing quality which can be fully reported (i.e. how each trial scored on each criterion).

Criteria for assessing methodological quality fall into four general categories (see figure below): protection again selection, performance, exclusion and detection biases.[41] No specific criteria have yet received wide endorsement for assessing the extent to which trials adequately guard against these biases.

At present, there is only one criterion for assessing the quality of RCTs for which there is strong empirical evidence of a relationship to the potential for bias in the results: adequate concealment of allocation.**Error! Bookmark not defined.**,42 It is recommended that reviewers should routinely use this criterion.[1]

Figure. General sources of bias



For most criteria two or three groups into which trials can be categorized are sufficient; eg. 'A' if the criterion was met and 'C' if the criterion was not met. A third category, 'B', may sometimes be used to indicate that a criterion was partially met or that it is unclear whether the criterion was met.

Reviewers should, if possible, obtain further information from the authors of a report when it is unclear whether a criterion was met. When this is not possible and many trials are in an "unclear" or "not applicable" category, reviewers should consider modifying or dropping the criterion.

[1]This recommendation and others in this chapter follow from discussions held at a Workshop on how to assess the quality of RCTs organized by David Moher in Ottawa in October 1993. The Workshop Report is not yet available. This chapter is being reviewed by the people who attended the Workshop.

The Cochrane Review Manager software can sort trials alphanumerically based on quality assessments. For trials to be displayed in order (from highest to lowest), it is recommended that letters be used to indicate quality, starting with 'A' for the highest level of quality. Confusion can arise when numbers are used to indicate quality.

Most of the available scales for assessing the quality of RCTs derive a summary score by adding the scores (with or without weights) on each item. While this approach offers appealing simplicity, it is not supported by empirical evidence.[43] An alternative approach is to explicitly define "cut-points" that relate directly to the individual criteria and the interpretation of the cut-points, such as:

		Interpretation	Relationship to individual criteria
A	Low risk of bias	Plausibly postulated bias unlikely to seriously alter the results	All of the criteria met
B	Moderate risk of bias	Plausible bias that raises some doubt about the results	One or more criteria partially met
C	High risk of bias	Plausible bias that seriously weakens confidence in the results	One or more criteria not met

These relationships to individual criteria are provided as an example only. They are only likely to be appropriate if few criteria are used, and if all of the criteria address important threats to the validity of the results.

6.1 Adequate allocation concealment

Preventing foreknowledge of treatment assignment is crucially important in trial design. When assessing a potential participant's eligibility for a trial, those who are recruiting participants should remain unaware of the next assignment in the sequence until after the decision about eligibility has been made. Then, after the assignment has been revealed, they should not be able to alter the assignment or the decision about eligibility. The ideal is for the process to be impervious to any influence by the individual making the allocation. This is most likely to be achieved securely if an assignment schedule generated using true randomisation is administered by someone who is not responsible for recruiting participants, for example, someone based in a trial office, or pharmacy. If such 'central randomisation' cannot be organised, then other precautions are required to try to prevent manipulation of the schedule of random assignment by those recruiting participants to the trial. These include, for example, using numbered or coded bottles, ampoules or other containers, and using serially numbered, sealed, opaque envelopes.[44,45,46] Simply using an open list ('table' or 'schedule') of

random numbers is as open to manipulation as is dependence on methods of assignment such as alternation or case record number.

The process of concealing assignments until treatment has been allocated has sometimes confusingly been referred to as 'randomisation blinding'.**Error! Bookmark not defined.** This term is unsatisfactory because it has often not been distinguished clearly from other forms of blinding - of patients, providers, outcome evaluators, and analysts. This is unsatisfactory for at least three reasons. First, the reason for concealing the assignment schedule is to eliminate selection bias. By contrast, other forms of blinding, used *after* the allocation of treatments, serve to reduce detection bias and performance bias (see figure above on page 30). Second, from a practical standpoint, concealing treatment assignment up to the point of assignment is always possible, regardless of the study topic, whereas blinding after allocation may be impossible, as in trials comparing surgical with medical treatments. Thirdly, control of selection bias is relevant to the trial as a whole, and thus to whatever outcomes are being compared, whereas control of detection bias is often 'outcome-specific', that is, it may be accomplished successfully for some outcomes in a trial, but not for others. Thus, blinding up to allocation and blinding after allocation of treatment are addressing different sources of bias, are inherently different in their practicability, and may apply to different parts of a trial. In light of these reasons for distinguishing the different forms of blinding clearly, we will refer to the process of concealing assignments as **'allocation concealment'**, and reserve the term 'blinding' (masking) for measures taken to reduce bias *after* treatment has been assigned.

Empirical research has shown that lack of adequate allocation concealment is associated with bias.**Error! Bookmark not defined.,Error! Bookmark not defined.** Indeed, this component has been found more important in preventing bias than the other component of allocation, the generation of the allocation sequence (e.g. computer, random number table, alternation). Trials should be judged on the reported method by which allocation was concealed. Information should be presented that provides some assurance that the allocations were not known until the point of allocation, at least. The method for assigning participants to treatments should be robust against patient- and clinician-bias and its description should be clear, concise, and explicit. The following criteria are suggested for judging concealment schemes to be **adequate**:

- Some form of centralized randomisation scheme, such as having to provide details of an enrolled participant to an office by phone to receive the treatment group allocation.
- Some form of randomisation scheme controlled by a pharmacy.
- Numbered or coded containers, such as in a pharmaceutical trial in which capsules from identical-looking, numbered bottles are administered sequentially to enrolled participants.
- An on-site computer system, given that the allocations are in a locked, unreadable file that can be accessed only after inputting the characteristics of an enrolled participant.

- If assignment envelopes were used, the report should at least specify that they were sequentially numbered, sealed, opaque, envelopes.

- Other combinations of described elements of the process that provide assurance of adequate concealment. This may include statements that imply an approach similar to one of those listed above, along with reassuring comments that the person who generated the allocation scheme did not administer it. Some schemes may be innovative and not fit any of the approaches listed above, but still seem to provide adequate concealment. This 'other' category is necessary, but it is likely to include only a small percentage of all trials deemed to have used adequate allocation concealment.

Approaches to concealment of allocation that should be considered **clearly inadequate** include:

- Alternation.

- Reference to case record numbers, dates of birth, day of the week, or any other such approach

- Any allocation procedure that is entirely transparent before assignment, such as an *open* list of random numbers or assignments

The adequacy of concealment measures for trials that do not report any concealment approach (regardless of whether a method of generating a random sequence is reported) should be considered **unclear**. This category includes:

- Just stating that a 'list' or 'table' was used for allocating assignments.

- Just specifying 'envelopes' or 'sealed envelopes' were used.

! Occasionally an adequate allocation concealment scheme may be reported, but other information is also reported that leads the reviewer to be suspicious of the scheme's success. This is likely to happen rarely, but reviewers should usually classify such a trial as having unclear allocation concealment.

Trials with unclear concealment measures have been shown to yield more pronounced estimates of treatment effects than trials that have taken adequate measures to conceal allocation schedules, but less pronounced than inadequately concealed trials.**Error! Bookmark not defined.**

6.2 Other components of trial quality

Other components of trial quality are probably important, particularly blinding and the handling of exclusions. However, empirical results are less clear. Ongoing research suggests that double-blinding may prove to be important in preventing performance bias and detection bias for some outcomes.**Error! Bookmark not defined.**

Until further empirical results are available, reviewers may want to consider the use of 'double-blinding' as a criterion for quality. This can be done with a single criterion; eg. "Was the study described as double-blind?"**Error! Bookmark not defined.** Anything beyond this simple dichotomy may not be worthwhile, although it might be worth considering blinding of data entry, verification and analysis.[47] Trials that blind outcome assessors regarding treatment allocation should theoretically be rated more highly than trials that do not institute any blinding. That they would be less biased makes sense. However, perhaps due to an artifact of reporting, this has not been observed empirically.**Error! Bookmark not defined.**,[48] Hence, the simple 'double-blinding or not' dichotomy is recommended in most instances.

Blinding is likely to be particularly important in research with "soft" outcome measures (eg. pain). Reviewers working in areas where blinding is very likely to be important may want to develop criteria for judging the appropriateness of the method of blinding that was used. For example, in drug trials, appropriate blinding of participants requires the use of control interventions (placebos) with physical characteristics similar to those of the intervention being tested.

Because of inadequacies in the reporting of how exclusions after allocation (eg. withdrawals, dropouts, protocol deviations) were handled, reviewers should be cautious about using the *reported* approaches to exclusions to assess trial quality. Certainly, approaches to exclusions by investigators have great potential to bias results. However, reporting inadequacies cloud this entire issue. What is reported, or more frequently implied in reports, on exclusions after allocation has not been found to be empirically related to bias. Thus, reviewers must be cautious about using reported approaches to exclusions as a quality criterion in most areas. This is a general recommendation, however, and may not apply to certain areas that have higher quality reporting or where it is possible to obtain missing information from the authors of reports.

Other criteria for assessing the quality of trials may be important in specific areas of research. For example, in areas where multiple outcome measures are used, as in evaluations of treatments for rheumatoid arthritis, selective reporting of significant results may cause important bias,[Error! Bookmark not defined.] suggesting that the specification of predefined primary outcomes and analyses may be useful indicators of the quality of such trials. Alternatively, selective reporting of results could be taken to suggest the need for better reporting and efforts by reviewers to obtain missing data.

Criteria other than concealment of allocation must be considered relative to each specific outcome measure. The Cochrane Review Manager software will prompt reviewers to do this when entering the results of trials.

6.3 Application of criteria for assessing trial quality

The quality of all the trials that meet the inclusion criteria for a review should be assessed using explicit criteria. Reviewers (or review groups) must make the same decisions as for selection criteria (see Chapter 0):

- whether more than one reviewer will assess the quality of each trial
- whether the assessments of quality will be made by content area experts, non-experts, or both
- whether the people assessing the quality of trials will know the names of the authors, institutions, journal of publication and results when they apply the criteria
- what decision rule or procedure will be used if there is disagreement when more than one reviewer applies the criteria to each article

Although experts may have preformed opinions that can bias their assessments,[Error! Bookmark not defined.] they may nonetheless give more consistent assessments of the quality of trials than non-experts.[Error! Bookmark not defined.] It may be desirable for both an expert and a non-expert to assess the quality of the trials included in a review. However, this tentative recommendation is unsupported by empirical evidence.

Some empirical evidence suggests that blind assessments of reports might produce lower and more consistent scores than open assessments.**Error! Bookmark not defined.** However, such assessments are very time consuming.[49] Reviewers must weigh the potential benefits of blind assessments (in terms of protection against biased assessments) against the costs involved in deciding whether or not to blind the assessors. Further research is underway comparing blind and open assessments of trial quality, the results of which may help guide this decision (Jesse Berlin, personal communication).

Many times, it will not be possible to assess the quality of a trial without additional information. In these cases, reviewers may choose to reference the trial as one that is awaiting assessment until the additional information is obtained or they may want to adopt a decision rule, such as "if what was done is not clearly stated, assume that it was not done appropriately" or "assign partial credit." Such decision rules should be reported explicitly.

For most reviews it will be important to test the quality criteria on a pilot sample of articles (say three to six papers, including ones that are thought to be of low, moderate and high quality) to ensure that the criteria can be applied consistently by more than one person.

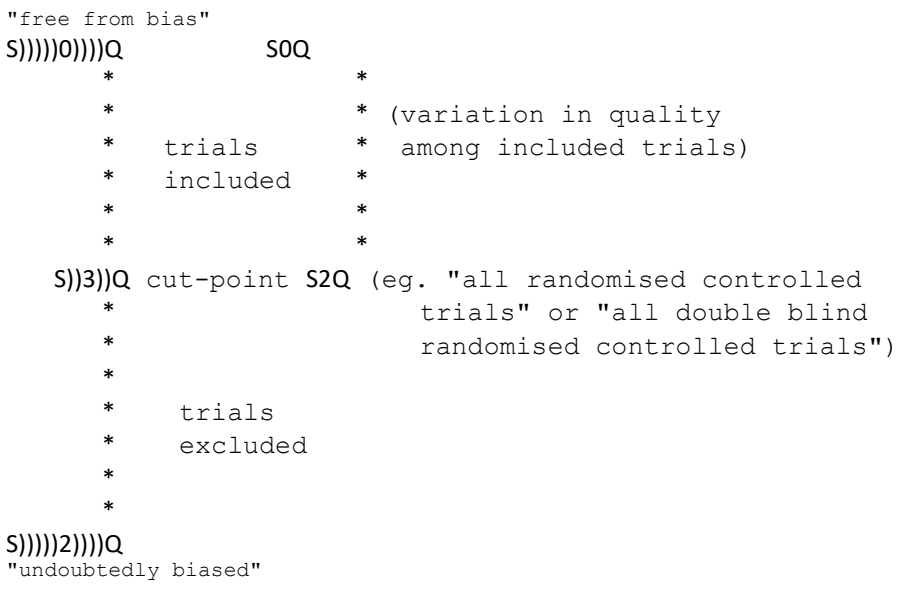
6.4 Incorporation of quality assessments in a review

There are several possible ways in which quality assessments can be used in a review:**Error! Bookmark not defined.**

- as a threshold for inclusion
- as a possible explanation for differences in results between trials
- in sensitivity analyses
- as weights in statistical analysis (meta-analysis) of the results

Failure to meet one or more criteria (in addition to random allocation) may indicate such a high risk of bias in some reviews that it constitutes grounds for exclusion of those trials. For example, reviewers may decide to include only double blind trials. The decision about where to set the cut-point for inclusion can be conceptualized as existing on a continuum between "free from bias" and "undoubtedly biased", as illustrated below.

Figure. Continuum of methodological quality



If reviewers raise the methodological cut-point for including trials, there will be less variation in quality among the included trials. Assessments of methodological quality categorize trials by the risk of bias within the range of quality above the cut-point for inclusion. If the cut-point is sufficiently high, the variation in quality among the included trials may be immaterial.

Differences in methodological quality may explain variation in the results of trials included in a review. When there is important heterogeneity in results, reviewers should explore whether differences in the control of bias among trials explains the heterogeneity (see Section 0). Ideally, hypotheses regarding methodological quality as a possible explanation for heterogeneity should be stated before inspecting the results (for the reasons explained in Section 0 on subgroup analyses). Generally, it can be hypothesized that more rigorous trials yield smaller estimates of treatment effects than less rigorous trials, but this is not always the case.**Error! Bookmark not defined.**

Several approaches can be taken to examining the relationship between methodological quality and the results of the trials included in a review:

- Visual plots of the results arranged in order of their quality can be used.**Error! Bookmark not defined.**,50

- Subgroups of studies above a methodological cut-point can be analyzed. (This approach can be used whether or not the results are heterogeneous, as a sensitivity analysis to determine if the overall results are the same when only studies with little risk of bias are included in the analysis.)

- The results of each trial can be combined sequentially in order of methodological quality (similar to "cumulative meta-analysis" but in methodological rather than chronological order), examining the impact on the overall results as trials of decreasing quality are included.[51] (The Cochrane Review Manager software does not include an option for performing cumulative meta-analysis. Future editions of the software may include this.)

Statistical techniques that are used to combine the results of several trials generally weight the influence of each trial by the inverse of the variance for the estimated measure of effect; i.e. trials with more precise results are given more weight. It is also possible to weight studies according to methodological quality so that studies of higher quality have more influence on the overall results. The main objection to this approach is that there is no empirical basis for determining how much weight to assign to different criteria or for quantitatively relating differences on any of the available scales to differences in the risk of bias. (The Cochrane Review Manager software does not currently include an option for weighting studies by methodological quality.)

If all the trials included in a review are of uniform methodological quality, variation in quality is obviously not a concern, but it is still important to note the quality of the trials since this affects how far the results of the review can be relied upon.

7. DATA COLLECTION

It is necessary to extract data for each trial regarding the study populations, the interventions and outcome measures that were used, and the results. As with any research project, this should ideally be done using pretested data collection forms and, if possible, at least two coders should be used to ensure that the data are extracted reliably. However, because of inconsistencies in how the results of trials are reported and the retrospective nature of a review, it can be difficult to design a satisfactory data-extraction form. Because of this, and for reasons of expediency, reviewers may decide to enter data directly from published reports into the Review Manager software. The software includes electronic data-entry forms that can be adapted, within certain limitations, to accommodate a particular review. Entering data directly from published reports into Review Manager might reduce transcription errors and save time; two coders can enter the same data independently or a second coder can check the accuracy of the entered data. Future editions of the Cochrane Review Manager software will offer the option of double data entry.

Insufficient reporting is common in the medical literature. Missing information should be obtained from investigators whenever possible. Information obtained through personal communication can strengthen the results of a review and often provides a valuable service to users of the review who cannot routinely do this. To avoid introducing bias, unpublished information that is obtained should be unambiguous, it should be obtained in writing, and it should be coded in the same fashion as published information.

The extent to which blinding is necessary for data extraction is uncertain. While it is theoretically justifiable, blinding may not be warranted for data-extraction for the same reasons that are described above for assessments of relevance and validity. Sensitivity analyses around important or questionable judgments or assumptions, on the other hand, can be performed relatively easily, and should be done.

7.1 Characteristics of the included trials

The format of a Cochrane Review includes a standard table for noting key characteristics of the trials included in a review (see Section 0), including:

- methodological quality
- characteristics of the participants
- characteristics of the interventions (comparison groups)
- characteristics of the outcome measures
- notes

Reviewers must decide which characteristics of the included trials are most important for users of the review to know and ensure that these characteristics are reliably extracted and recorded, and displayed in a way that is understandable and informative to users of the review.

7.2 Results of the included trials

The results of the included trials must be entered in a standardized format from which tables and figures for each comparison can be generated (see Section 0). Reviewers must decide which results to extract (i.e. for which outcomes and which subgroups) and ensure that they are reliably extracted and recorded. As a general rule, results for all practically important outcomes should be recorded (regardless of the number of trials in which they are reported). Reviewers should, if possible, distinguish between data that were not collected in a trial and collected data that were not reported. This is particularly important for trials reporting multiple outcomes, because of the potential for biased reporting.**Error! Bookmark not defined.**

The number of subgroups for which data are extracted should be kept to a minimum. The selection of subgroups that are of interest should be driven by the plausibility of there being a true and practically important difference in the effects of the interventions of interest across different subgroups; i.e. there should be reasonable indirect evidence to suggest that there might be true differences in subgroup response. Collection and analysis of data for large numbers of subgroups is risky because it may reduce the reliability of data extraction, use resources that might better be spent elsewhere, and lead to spurious conclusions.[52,53]

7.3 Information for subsequent cost-effectiveness analyses

Often decisions about health care entail making tradeoffs between the estimated benefits and the estimated harms and costs of the intervention. Sometimes there is sufficients uncertainty about these tradeoffs that formal economic analyses can be called for to guide decisions about whether the expected benefits are worth the potential harms and costs. In these situations it can be very helpful for reviewers to collect appropriate data about costs from the included trials, when these are available. Guidelines for reviewers on collecting information for subsequent cost-effectiveness analyses, developed by a Methods Group coordinated by Alastair Gray and Ron Akehurst, can be found in Appendix 5.

8. DATA SYNTHESIS

The analysis, or data synthesis, for a systematic review entails the whole process of evaluating and synthesizing the results of the included trials. Statistical techniques (meta-analysis) may or may not be used. The statistical techniques used in meta-analyses do not differ in principle from those used in primary research,[54] 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.

When statistical methods are not used in reviews a common error is to compare the number of "positive" and "negative" studies. With such a "vote counting" approach it is possible for a study to be counted as "positive" in one review and "negative" in another, depending on how the reviewers interpret the results. Also, small but clinically important effects tend to be overlooked when vote counting is used, particularly if studies with statistically "nonsignificant" results are counted as "negative".**Error! Bookmark not defined.**,[55] Another error is inappropriate weighting of the results of the individual studies. For example, vote counting gives studies equal weight, ignoring differences in size and quality. Reviewers may also inappropriately stress the results of one study over another.

Decisions about whether it makes sense to use statistical methods to combine the results of the trials included in a review rely on good sense and clinical judgement. *Before considering the results*, it is important to consider whether the participants, interventions and outcomes in each of the included trials are sufficiently similar that combining them is meaningful.

It might make sense for some reviews to use broad inclusion criteria (eg. to get an overview of the interventions that are used for a particular problem). This may result in a set of trials which it makes no sense to combine quantitatively. Such a broad review could be done in combination with more focused reviews of subsets of the trials included in the larger review (where it would make sense to combine the results of the included trials). For example, a broadly focused review of RCTs of continuing medical education (CME) interventions could be useful for obtaining a broad overview of the types of interventions that have been tested in trials and help to derive some general conclusions about the effectiveness of CME in changing physician performance.[56] More focused reviews of specific types of CME interventions (such as audit and feedback, reminders, outreach visits and use of opinion leaders) could permit quantitative assessments of the impact of these different types of interventions on physician performance and patient outcomes.

Reviewers might also choose to combine subsets of trials within a single review. For example, Silagy et al. in their review of nicotine replacement therapy (NRT) for smoking cessation made several different comparisons within the same review addressing different hypotheses about the effectiveness of different forms of NRT.**Error! Bookmark not defined.**

Meta-analysis may be misleading in a systematic review of an intervention for which no good quality trials can be identified.

When meta-analysis is used, a frequent complaint is that "apples and oranges" are being combined because there are almost always differences in the participants, interventions, outcomes or methods among the trials that are combined. This is true, but there is nothing wrong with combining apples and oranges, if one is interested in fruit. For example, trials of secondary prevention with beta-blockers after a myocardial infarction have differed in terms of the patient characteristics, type and dose of beta-blocker and how outcomes were measured. Nonetheless, the best estimate of the effectiveness of beta-blockers after a myocardial infarction comes from the overall results of all the valid studies.[57] Differences in the results of these trials are more likely to be due to chance than to differences in the effectiveness of the drug regimens that were used or the responsiveness of the patients studied.**Error! Bookmark not defined.**

The remainder of this chapter addresses issues relating to meta-analysis, although the underlying principles also apply to non-quantitative syntheses. Whether or not statistical methods are used, reviewers must decide how to summarize the results of each included trial, how to examine differences in the results between trials, how to derive an overall estimate of what was observed across trials (if this is a sensible thing to do), how to present the results and how to test their robustness. The recommendations in this chapter regarding statistical methods are based, as far as possible, on the results of a Workshop on statistical methods for data synthesis organized by Ken Schulz and Doug Altman.[58]

8.1 Choice of summary statistic

Reviewers must first decide how the effect of the intervention observed in each trial will be summarized. For binary outcomes (proportions), summary statistics that can be used to summarize the effect of an intervention include:

- the odds ratio (OR)
- the relative risk (RR)
- the relative risk reduction (RRR), sometimes called the percent risk reduction
- the risk difference (RD), sometimes called the absolute risk reduction
- the number needed to treat (NNT),[59] the inverse of the risk difference (1/RD)

These measures can all be related to one another in a simple 2x2 table (see Appendix 6).

The choice of which measure to use is not always straightforward, and the measure used in the analysis may not be the best measure for clinical decision-making. For instance, the odds ratio is commonly used in meta-analyses because of advantages in the statistical techniques available for combining odds ratios, but it can be difficult to interpret the clinical importance of an odds ratio. Conversely, although it is easy to interpret the clinical importance of a risk difference, the statistical methods available for combining risk differences can have disadvantages.

The statistical methods used should as far as possible be consistent across the Reviews contained in the Cochrane Database. The Review Manager software provides default

methods of analysis to encourage this, but it is possible for reviewers to override these. The default summary statistic that is used is the odds ratio, because of its statistical advantages. However, the odds ratio may not be the best summary statistic for presenting the results of a review (because of problems in its interpretation), as discussed below in Section 0.

It may be useful to analyse the data using alternative outcome measures if it is uncertain which is the most appropriate.

The analysis of trials with continuous outcome measures (eg. blood pressure) is discussed below in Section 0.

8.2 Test for heterogeneity

Having decided that it makes sense to combine a group of studies and what summary statistic to use for the analysis, reviewers should first check whether the differences among the results of the trials are greater than could be expected by chance alone. One way of doing this is to look at a graphical display of the results. If the confidence intervals for the results of the studies (represented by horizontal lines) do not overlap, the differences are likely to be "statistically significant".[60] Tests of homogeneity are formal statistical analyses for estimating the probability of observed differences in results having occurred due to chance alone. The more significant (closer to zero) the test, the less likely it is that the observed differences occurred by chance alone.

"Statistically significant" heterogeneity suggests that the observed differences in results are due to factors other than chance. However, when looking for possible explanations for the differences, such as differences in the interventions that were evaluated, it is important to be cautious.**Error! Bookmark not defined.,Error! Bookmark not defined.** Even in a review that includes only high quality RCTs, participants are not randomized to one trial versus another. Because trials usually differ in a number of ways, it is risky to attribute differences in results to any one factor. If this is done at all, it should be considered "hypothesis generating" rather than hypothesis testing.

The Cochrane Review Manager software will automatically test the homogeneity of the results of the individual trials being combined for each comparison in a review. The software currently does not include a test for homogeneity between groups of studies.

8.3 Estimate of overall effect

The central aim of most Cochrane Reviews is to provide a reliable estimate of the effectiveness of an intervention or type of intervention, based on a weighted average of the results of all the available good quality trials. Typically, the weight given to each trial is the inverse of the variance; i.e. more precise estimates (from larger studies with more events) are given more weight.[61] It is possible to give trials more or less weight based on their methodological quality as well as their precision, but this is rarely done (see Section 0). More often a "sensitivity analysis" (see Section 0) in which the robustness of the results is tested by excluding subgroups of trials, such as those of poorer methodological quality, is done.

If it makes sense to combine the results of a group of trials and the observed differences between the results of the trials are not statistically significant or practically important, it is relatively straightforward to combine the results. Each trial is summarized using a measure of effect (usually the odds ratio) that represents the within study comparison of the intervention and control groups. In this way participants in each study are only compared to people in the same study.

Occasionally a meta-analysis is published where an "effect size" is calculated for each study group based on the difference in outcomes before and after treatment and groups from different studies are directly compared with each other. When this is done, the results of a meta-analysis must be interpreted with a great deal of caution. Even if all of the data come from RCTs, the power of randomization is completely lost when this approach is taken, and the data have been reduced to the equivalent of before-after studies. This approach to synthesizing data is not supported by the Review Manager software.

The confidence interval around an estimate derived from an appropriate statistical synthesis indicates how precise that estimate is. It is, in principle, no different than a confidence interval around an estimate derived from primary research.[62] However, one important consideration that arises in meta-analyses is whether to incorporate between study variation in estimating the confidence interval (a "random effects" approach). If there is little between study variation (i.e. the test for homogeneity results in a large *p* value, say greater than 0.10), this will make little difference. If there is significant between study variation, an analysis that ignores this (a "fixed effects" approach) will give a narrower confidence interval than one that does not. If there is significant heterogeneity reviewers should be cautious about interpreting the overall estimate derived from a meta-analysis, as well as attributing between-study differences to any one factor.

There is disagreement on how to summarize the results of trials, and whether they should be combined when there is substantial heterogeneity.**Error! Bookmark not defined.** However, some common ground exists on at least four points. First and foremost, reviewers should alert readers of their review when there is substantial heterogeneity, and they should encourage cautious interpretation of the aggregated results. Reviewers should attempt to explore the reasons for the heterogeneity and to explain it, but the interpretations must be cautious because the analyses are usually *post hoc*. Subgroup analyses in general are often useful but difficult to interpret (see Section 0).**Error! Bookmark not defined.**

One potential explanation for heterogeneity that can be investigated in an *a priori* fashion is methodological quality. Differing methodological quality in the component trials could produce both bias and accompanying heterogeneity. If there are important differences in the quality of trials, reviewers should consider excluding poorer trials.

Secondly, it is generally agreed that the fixed effects approach for a test of significance for the overall null hypothesis ("no effect in all trials") is statistically valid. A statistically significant result indicates that a treatment effect exists in at least one of the trials.

Thirdly, whether heterogeneity is present or not, the overall typical effect measure from a fixed effects approach is an informative average measure of treatment effect.

Fourthly, the random effects methods rely on assumptions that the trials are a random sample from a hypothetical population of trials and that the heterogeneity between trials can be represented by a single variance. Moreover, a pragmatic concern is that the "random effects" methods give more weight to the smaller studies than the "fixed effects" methods and smaller trials are often of poorer quality and may be more susceptible to publication bias.

That leaves as the primary dispute the different approaches to the calculation of a confidence interval around a typical effect measure when there is clear heterogeneity. Even given the concerns about the assumptions behind random effects models, it may be worth examining the confidence interval that a random effects model affords.

Various methods are available for analyses using fixed and random effects models. Generally, not much difference exists between the results from the various fixed effect approaches.[63] Less is known about the various random effects approaches.

The statistical methods currently available for synthesizing data in the Review Manager software are identified and referenced in the Software Notes. Additional statistical methods may be added to the software in the future.

8.4 Presentation of results

For the most part, reviewers should be more concerned about how they interpret the summary statistic that was used than about which summary statistic was used in the analysis. For example, when results are summarized as odds ratios they usually represent the odds of an adverse outcome occurring in the group that received the intervention compared to the odds of an adverse event in the control group. If the proportion of people who have the outcome is very low the odds ratio and the relative risk are approximately the same, and the percent risk reduction is equal to one minus the odds ratio. However, as the proportion of people who have the outcome increases, the odds ratio becomes progressively *smaller* than the relative risk. Thus, in situations where the baseline risk is high, the assumption that the odds ratio equals the relative risk does not hold true, and can lead to overestimating the relative risk reduction.

The Cochrane Database of Systematic Reviews is designed to consistently display the odds ratio, or another summary statistic, for an unfavourable outcome. Thus, odds ratios and relative risks less than one (and risk differences less than zero) always indicate that treatment is better. Both reviewers and users of the Reviews can be confused when this convention is not maintained and there is a greater risk of errors being made. Because of this convention, reviewers must always use *bad* outcomes when entering data, even if results are typically reported for *good* outcomes. For example, livebirth after treatment for infertility must be reported as 'no livebirth' (or something equivalent) and smoking cessation after an intervention to help people to quit smoking must be reported as 'failure to quit smoking'. Reviewers do not need to adhere to this convention in the text of the review.

As noted above, the statistical technique that is used and the presentation of results can be considered as separate issues. The number needed to treat (NNT),**Error! Bookmark not defined.** for example, is a sensible way to describe results, but this does not imply that the analysis needs to be done using the NNT or the risk difference (from which the NNT can be derived directly). Nonetheless, it is often desirable to describe results both as a relative measure (odds ratio, relative risk or relative risk reduction) and an absolute measure (NNT or risk difference). Absolute measures can be more informative than relative measures because they reflect the baseline risk as well as the change in risk with the intervention. When absolute measures, such as the NNT, are reported, it is useful to report them for a range of baseline risks (say, the range observed in the included studies). Once the data for a review have been entered, both the Review Manager software and the Cochrane Database of Systematic Reviews will allow reviewers and users of the Database to display the results in different ways.

A Cochrane Workshop on how to present the results of systematic reviews is being organized by Andreas Laupacis for 1994. The results of this Workshop will be used to further develop guidelines for reviewers. The default convention now is to present the results as odds ratios, with options to display the results as relative risks, relative risk reductions, risk differences or NNTs. A standard graphical display of the results is included in the attached example of a review (Appendix 1) and described in Chapter 0.

8.5 Sensitivity analyses

Because there are different approaches to conducting a systematic review, reviewers should ask: "How sensitive are the results of the analysis to changes in the way it was done?"**Error! Bookmark not defined.** This provides reviewers with an approach to testing how robust the results of the review are relative to key decisions and assumptions that were made in the process of conducting the review. Each reviewer must identify the key decisions and assumptions that are open to question, and might conceivably have affected the results, for a particular review. Generally, the types of decisions and assumptions that might be examined in sensitivity analyses include:

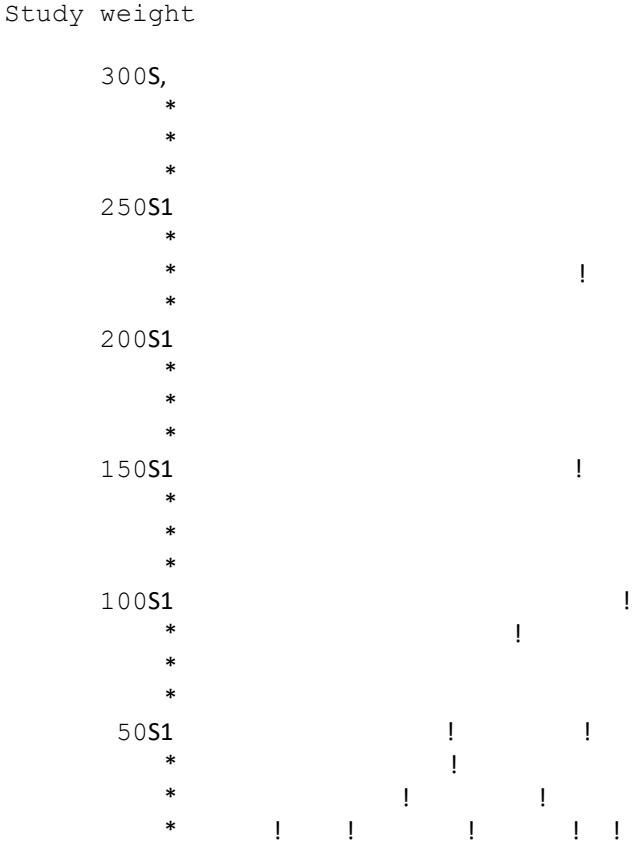
- changing the inclusion criteria (including the types of participants, interventions and outcome measures, and methodological cut-points)
- including or excluding trials where there is some ambiguity as to whether they meet the inclusion criteria
- excluding unpublished studies
- excluding studies of lower methodological quality
- reanalysing the data using a reasonable range of results for trials where there may be some uncertainty about the results (eg. because of inconsistencies in how the results are reported that cannot be resolved by contacting the investigators or because of

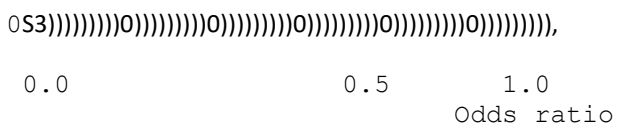
differences in how outcomes are defined or measured)

- reanalysing the data imputing a reasonable range of values for missing data
- reanalysing the data using different statistical approaches (eg. using a random effects model)

Although several ways of examining the potential impact of publication bias on the results of a systematic review have been described, all of them are problematic.[Error! Bookmark not defined.,64] There is a growing consensus that the only satisfactory way to address this bias is through prospective registration of trials.[Error! Bookmark not defined.,Error! Bookmark not defined.,65] Nonetheless, reviewers may want to consider using a "funnel graph" to look for publication bias if there is a sufficient number of trials. This can be done by plotting the study weight (one divided by the variance of the odds ratio) or sample size (on the "y" axis) against the odds ratios (on the "x" axis). Typically this will give the appearance of a funnel with larger studies (with greater weights) at the top in the middle and smaller studies spread out at the bottom. A gap at the bottom of the funnel on the right side (assuming smaller odds ratios (less than one) represent beneficial effects) indicates that small "positive studies were identified, whereas equal numbers of small studies in the opposite direction were not, as illustrated below:

Figure. Funnel graph of study weight against odds ratio





If the sensitivity analyses that are done do not materially change the results, it strengthens the confidence that can be placed in them. If the results change in a way that might lead to different conclusions, this indicates a need for greater caution in interpreting the results and drawing conclusions. It might also enable reviewers to clarify the source of existing controversies about the effectiveness of an intervention, or lead reviewers to hypothesize potentially important factors that might be related to the effectiveness of the intervention and warrant further investigation.

8.6 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.**Error! Bookmark not defined.,Error! Bookmark not defined.** 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, reviewers need to be cautious about doing subgroup analyses and about interpreting the results of the ones that they feel compelled to do. The following guidelines can be used to help decide whether to do a subgroup analysis, whether an apparent difference in subgroup response is real and, consequently, when to include a recommendation based on a subgroup analysis in a review.**Error! Bookmark not defined.** They can also be used to guide decisions about whether apparent differences in response due to differences in the intervention (such as the specific drug or dose that is used) are real.

! Is there indirect evidence in support of the difference?

Another way of asking this question is: Is the difference plausible, based on what is known about the intervention and the problem? If not, the subgroup analysis probably should not be done.

! Did the hypothesis about the difference precede rather than follow the analysis?
! Is the subgroup analysis one of a small number of hypotheses tested?

Reviewers should, so far as possible, attempt to state any hypotheses they have about potentially important differences in subgroup response *a priori* (in the protocol for the

review) and the number of planned subgroup analyses should be kept to a minimum to avoid spurious findings. The greater the number of hypotheses tested, the greater the number of subgroup differences one will find, by chance. Decisions about which analyses to do and which ones to report are much more likely to be data driven with *post hoc* analyses, and thereby more likely to be spurious. On the other hand, when a hypothesis has been clearly and unequivocally suggested by a different data set the subgroup analysis rests on stronger ground.

- Is the difference suggested by comparisons within, rather than between studies?
- Is the difference consistent across studies?

Reviewers should be extremely cautious about examining and drawing conclusions about differences in subgroup response based on differences in results between studies. There may be many other factors, aside from the most salient difference that is the basis of the inference being made, which could explain the difference in results between studies. For instance, differences in the specific drugs, doses or treatment schedules used, different patients (varying in risk of adverse outcomes, for example), varying degrees of co-intervention, or varying criteria for the outcome, could all explain differences in the results of drug trials. Stated simply, inferences based on between-study comparisons are based on comparisons between non-comparable groups. Even when all the individual studies are RCTs, patients were not randomized to one study or another.

Differences in subgroup response observed within a single trial provide a stronger basis for drawing conclusions, and if a difference is observed across a number of high quality trials it strengthens a conclusion that the difference is real.

- Is the magnitude of the difference practically important?

If not, it makes no difference whether recommendations are based on the subgroup analysis or the overall analysis.

- Is the difference statistically significant?

The Review Manager software does not provide an appropriate method for testing this. Statistical techniques for conducting subgroup analysis include the Breslow-Day technique and regression approaches.[66] With the Breslow-Day technique and similar approaches,[67] it is possible to use a test for heterogeneity to estimate the probability that an observed interaction might have arisen by chance alone. However, in general, subgroup analyses based on differences between studies are discouraged as noted above.

Commonly, a number of comparisons for different subgroups are simply conducted without formally testing for interactions. This practice, together with only reporting subgroups within which sizable treatment differences are found, can lead to an overestimate of the significance, as well as of the size of the difference. It should be avoided.

Regression models, such as logistic regression,**Error! Bookmark not defined.** can also be used

for analysis of interactions if the interactions are modeled by product terms. This approach allows the significance of an interaction to be tested while controlling for other factors. However, if there are many subgroup factors, the number of product terms necessary for an adequate modelling of the interactions may be higher than the number of observations and an analysis of the interactions is impossible. An additional problem with this approach is deciding which of many possible interaction terms to enter into the model, and the potential for bias in their selection.

Reviewers without a statistical background who consider it is important to formally test the significance of an apparently important difference in response in a subgroup should consider asking a statistician for help. It is not important for reviewers to understand the details of the appropriate statistical approaches that can be used, but it is important to understand the concepts of statistical significance and power in subgroup analysis. Statistical analysis is a useful tool for assessing whether an observed difference in subgroup response might have occurred by chance alone, but it is not a substitute for good judgement.

8.7 Dose response analyses

The same principles described above for subgroup analyses can be applied to dose response and duration response analyses. Conclusions about differences in effect due to differences in dose are on strongest ground if participants are randomized to one dose or another within a trial and a consistent relationship is found across studies. While reviewers should attend to dose effects, and consider this as a possible explanation for heterogeneity of results, they should be cautious about drawing conclusions about dose or duration effects based on between study differences.

8.8 Time to event (survival) analyses

Several approaches are available for extracting and analyzing data from reports of trials with survival time (or time to an event) as the endpoint of interest. The present Review Manager software does not support this type of analysis. However, it is possible to conduct and display the results of analyzes for different follow-up periods, thereby approximating a survival analysis.

For reviews where survival time or time to an event is particularly important, individual patient data is highly desirable (see Chapter 0). Methods for extracting and analyzing survival data from published reports may be added to future editions of the Handbook and Review Manager software.

8.9 Continuous outcome variables

Meta-analysis of continuous outcome measures has received much less attention in health care research than in areas such as psychology and education. Mean differences from each trial can be combined using weights based on their variances.[68] Other approaches, such as expressing the mean difference as a multiple of the control group standard deviation ('effect size') are available, but are not supported by the Review Manager software.

Difficulties that may arise with synthesizing continuous outcome measures include the choice of effect measure, the issue of data distribution (and whether means or median might be more appropriate), how to handle studies with before-after assessments (and especially crossover trials), weighting of studies and missing data. Of perhaps greatest concern is the interpretation of the results when mean differences are analyzed. It can be difficult to determine whether an average change in a continuous outcome measure, such as birth weight, is worthwhile relative to the costs and risks of an intervention without knowing the distribution of the changes; i.e. the relative proportions of participants for whom the difference in the outcome variable was practically important.

For reviews of trials with continuous endpoints, or a mixture of binary and continuous endpoints, continuous outcomes should not automatically be transformed to binary ones. However, this might be acceptable under circumstances where there are sufficient data and a good rationale for selecting a cut-point.

The same general principles discussed above regarding aggregation under conditions of heterogeneity apply to continuous data.

8.10 Summary of recommendations

The following is a summary of recommendations on statistical synthesis of data based on the July 1993 Workshop:

- For binary outcomes the initial analysis should use the odds ratio. In addition to the default odds ratio, the Review Manager software provides optional calculations using the relative risk and risk difference.

- The analysis should begin using a fixed effects model, particularly for the overall test of significance, and an estimate of the average overall treatment effect. A confidence interval for that effect should also be provided. In addition to the default fixed effects model, the Review Manager software provides optional calculations of the effect measure and confidence interval based upon a random effects model.

- If the test of heterogeneity is statistically significant and the differences in results are practically important:

    - reviewers should consider the sources of the heterogeneity (e.g. differences in dose, timing of treatment, length of treatment, participant characteristics). In particular, reviewers should explore whether differences in the control of bias among the trials explain the heterogeneity (e.g. concealment of randomization).

    - If the heterogeneity can be explained, reviewers should consider (with due caution) the option of presenting the results of subgroup analyses.

    - If the heterogeneity cannot be explained, reviewers should consider the

following options:

- Do not aggregate the trials at all
- Use the fixed effects model with appropriate cautious interpretation
- Use the random effects model with appropriate cautious interpretation
- Use both the fixed and random effects models as an additional aid in explaining the uncertainty around an analysis with heterogeneous trials

- For continuous outcomes the approach suggested by Bracken**Error! Bookmark not defined.** should be used unless a preferable approach is forthcoming.

- Whatever is done, reviewers should clearly explain what was done and why.

9. DISCUSSION AND CONCLUSIONS

9.1 Discussion

As a rule, the Discussion section of a Review should be restricted to issues that are important to people using the review to help them with a decision about practice or future research, including:

- any *important* methodological limitations of the included trials and the review that might affect a decision
- how the included trials fit into the context of other evidence (stating clearly whether the other evidence that is referenced was systematically reviewed)
- how the results of the review fit into the context of current clinical practice (bearing in mind that 'current clinical practice' might vary widely from one setting to another)
- any important considerations about how the results are interpreted from the perspective of relevant *fields*, including different types of health care providers, users and settings (see Section III of the Handbook).

9.2 Conclusions

The conclusions should not exceed the evidence that is reviewed. So far as possible, recommendations should be linked to the strength of the evidence. A review group may elect to use a standard approach for doing this that is systematic and explicit.[69]

Because conclusions regarding the strength of inferences regarding the effectiveness of an intervention are essentially causal inferences, reviewers might want to consider guidelines for assessing the strength of a causal inferences, such as those put forward by Hill.[70] In the context of a systematic review of RCTs, these considerations might include:

- How good is the quality of the included trials?
- How large and significant are the observed effects?
- How consistent are the effects across trials?
- Is there a clear dose-response relationship?
- Is there indirect evidence that supports the inference - including evidence relating to intermediate outcomes, evidence from studies of different populations (including animals) and evidence from analogous relations (i.e. similar interventions)?

! Have other plausible competing explanations of the observed effects (eg. bias or co-intervention) been ruled out?

One common error that occurs, both in describing results and in reaching conclusions, when there is inconclusive evidence is to confuse 'no evidence of effect' with 'evidence of no effect'. Reviewers should be careful not to do this.

In addition to considering the strength of evidence underlying any conclusions that are drawn, reviewers should be as explicit as possible about any judgements about preferences (the values attached to different outcomes) that are made. Health care interventions generally entail costs and risks of harm, as well as expectations of benefit. Drawing conclusions about the practical usefulness of an intervention entails making tradeoffs, either implicitly or explicitly, between the estimated benefits and the estimated costs and harms.[71] It is beyond the scope of Cochrane Reviews to incorporate formal economic analyses (although they might well be used for such analyses).[72,73] However, reviewers should consider all of the potentially important outcomes of an intervention when drawing conclusions, including ones for which there may be no reliable data from the included trials. They should also be cautious about any assumptions they make about the relative value of the benefits, harms and costs of an intervention.

In drawing conclusions for research, reviewers should avoid statements like: "More research is needed." They should state exactly what research is needed and why. Opinions on how the review might be improved with additional data or resources can also be noted.

9.3 Classification of interventions into one of six categories

The above cautions about drawing conclusions not withstanding, review groups (and users of Cochrane Reviews) may find it useful to categorize interventions into one of six mutually exclusive categories. This has been done by the Pregnancy and Childbirth Group,[74] based on an earlier effort to classify interventions into four categories that drew a great deal of attention and praise. The first three categories of interventions, listed below, are ones for which there is sufficient evidence to reach relatively firm conclusions for practice. The last three are categories for which further research may be required before firm conclusions for practice can be drawn.

```
1     Forms of care for which there is sufficient evidence to
      provide clear guidelines for practice

      A     Forms of care that improve outcome (eg. continuity
            of care-giver during labour)
```

B Forms of care that should be abandoned in light of the available evidence (eg. routine weight measurement throughout pregnancy or elective caesarean for 'gestational diabetes')

C Forms of care that involve important trade-offs between known benefits and known adverse effects (eg. epidural analgesia in labour versus other forms of pharmacological relief)

2 Forms of care for which the evidence is insufficient to provide clear guidelines for practice, but which should influence priorities for research

A Forms of care that appear promising, but require further evaluation (eg. support for care-givers during at-risk pregnancy)

B Forms of care that have not been shown to have the effects expected from them, but which may require further attention (eg. routine iron supplementation during pregnancy)

C Forms of care with reasonable evidence that they are not effective for the purpose for which they have been used (eg. dietary regulation for 'gestational diabetes')

10. TABLES AND FIGURES

The current strategy is to only include standard tables and figures in the Cochrane Database. These include:

- a table of characteristics of trials contributing data to the review
- a table listing the comparisons that are made in a review if more than one comparison is made
- tables displaying the data from included trials and analyses of these data, from which the standard figures are generated, for each comparison
- a table listing the trials that meet the inclusion criteria but were excluded from the review, giving the reason for exclusion (such as data not available)

The standard tables and figures for displaying the results can change over time since only the data are stored in the Database. The analyses and the tables and figures used to display the results are generated by the software. Reviewers may want to prepare other tables and figures for other publications. However, for the present, there is no capacity to store these in Cochrane Reviews. Examples of tables and figures prepared using the Review Manager software are included in the attached review. Examples of standard figures are shown below. The first figure displays the overall results for all of the outcomes that were included in the review.[75] Each row in the figure represents a summary of the data from all of the trials reporting a particular outcome, such as mortality. The vertical line represents "no effect" (an odds ratio of 1). The box indicates the overall odds ratio, less than one indicating lower odds of the outcome occurring in the intervention group. The size of the box reflects the relative amount of data contributing to each estimate. The horizontal lines show the 95% confidence interval around the estimated odds ratios.

The second figure shows the odds ratio from each trial for a single outcome. In this figure each row represents a trial and the size of the box indicates the relative influence (weight) given to each trial in deriving the overall result. The overall result (a weighted average of the results from each trial) is shown as a diamond at the bottom of the graph.

Reviewers have the option of changing the order in which outcomes or trials are displayed; eg. trials can be displayed in chronological order, order of size or order of methodological quality.

Figure. Overall results of all the outcomes reported in a review

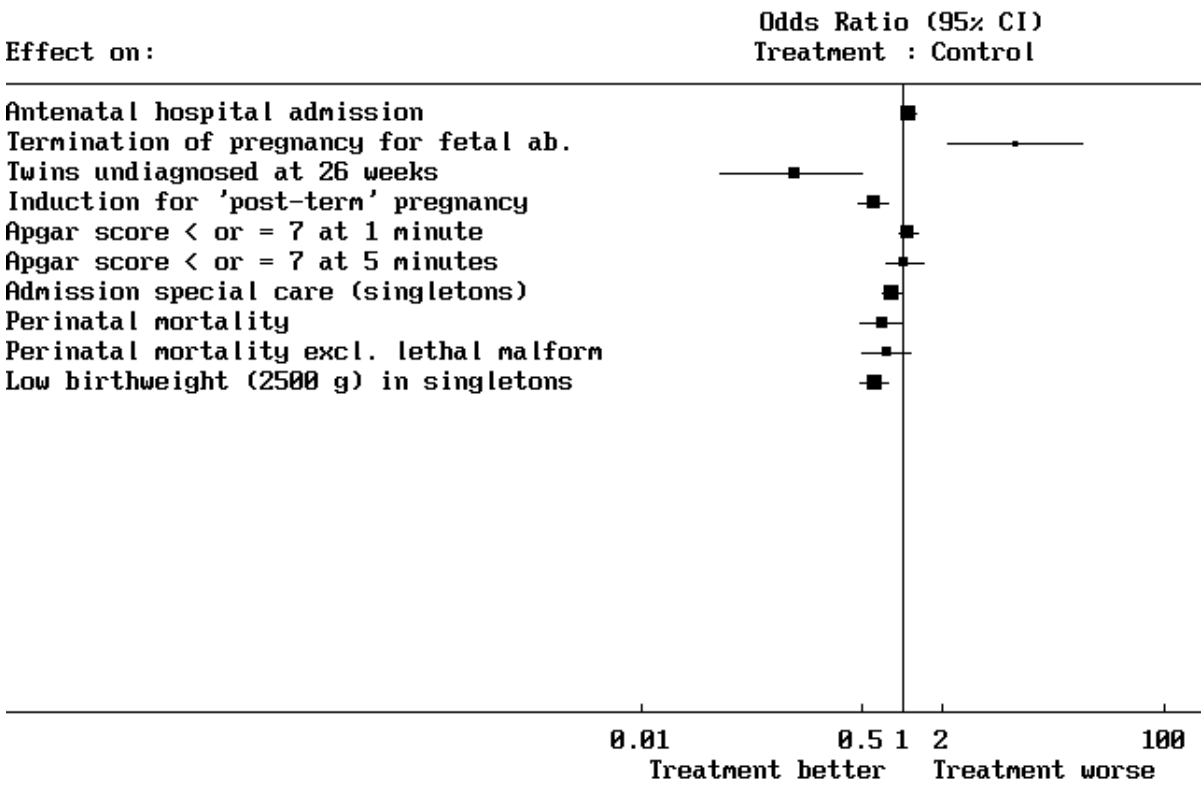
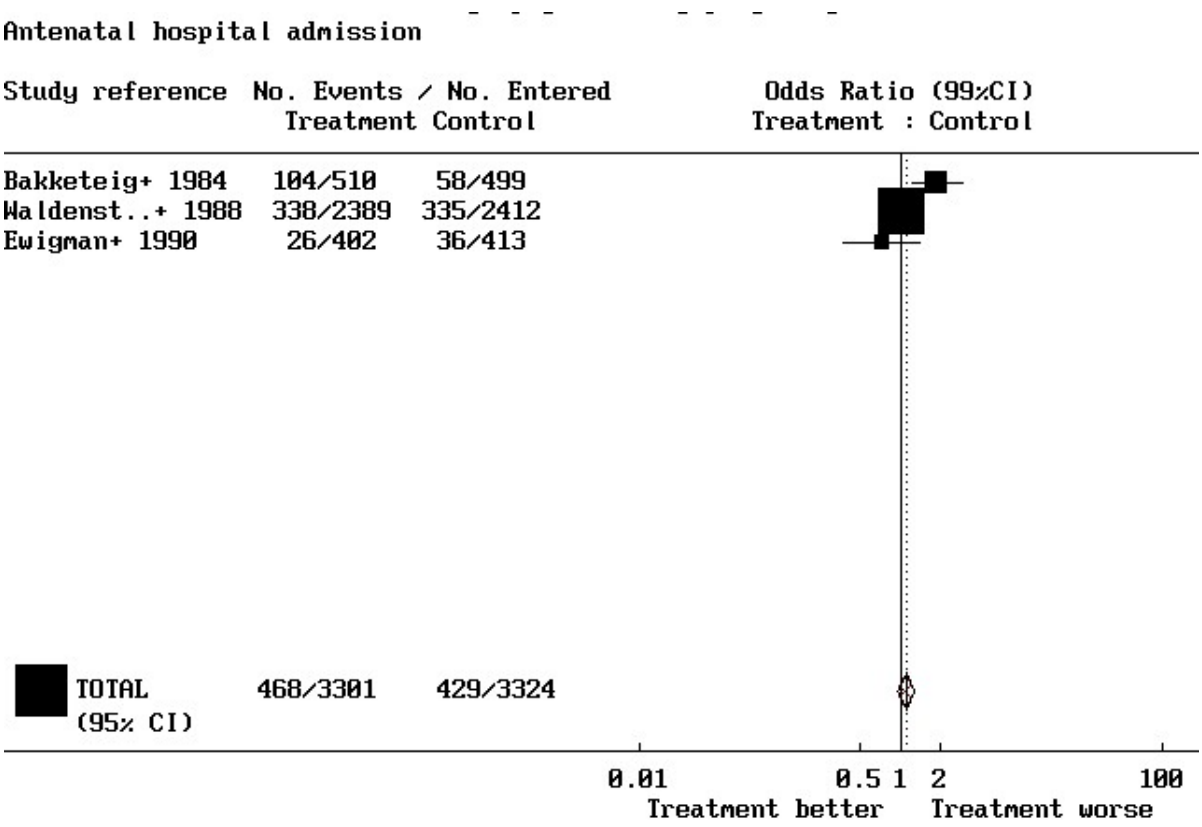


Figure. Results for each trial for a single outcome



11. KEEPING A REVIEW UP TO DATE

Reviewers are responsible for revising a review when new evidence becomes available or in light of valid criticisms. Each collaborative review group is responsible for developing appropriate mechanisms for ensuring that new trials are identified and brought to the attention of the reviewer and for handling criticisms that are received in response to the review.

12. REVIEWS USING INDIVIDUAL PATIENT DATA

For systematic reviews using individual patient data, the first step may be the formation of a collaborative trialists' group, such as the Early Breast Cancer Trialists' Collaborative Group, involving most or all of the research groups who have conducted RCTs in the relevant area. The organization and coordination of such a group, and the design and conduct of reviews using individual patient data will be the subject of a Cochrane Workshop being organized by Lesley Stewart, Mike Clarke, Max Parmar and Doug Altman for April 1994.

13. MAINTAINING AND RAISING THE STANDARDS OF COCHRANE REVIEWS

If Cochrane Reviews are to be useful to those who want to take more informed decisions in health care and research, then they must be trustworthy, and transparently so. As made clear in this Section of the Handbook, the Collaboration will be using explicit methods to produce reviews, and this feature alone will make them more useful to users than the vast majority of reviews that are currently available. Textbooks and review articles with 'Materials and Methods' sections remain rare.

Above a certain guaranteed minimum standard, the reviews contributed to the Cochrane Database of Systematic Reviews will vary in the level of methodological quality that it has been possible for the reviewers to achieve. The 'gold standard' will continue to be represented by systematic reviews, conducted under the aegis of collaborative trialists' groups, that are based on individual patient data for all patients entered into all of the trials meeting the entry criteria for the review. Such reviews require not only substantial resources (including time), they also depend on the success of negotiations among the trialists. These factors should not be underestimated. Furthermore, because 'the best can be the enemy of the good', it will be important to do empirical research to learn more than is currently known about which methodological standards are essential, and which desirable, in attempts to avoid bias.

Mechanisms for maintaining and raising the standards of Cochrane Reviews are summarized in the table below and described in the text that follows. The people who are responsible, or able to contribute in each of the ways listed, have been indicated with a check mark.

Table: Mechanisms for maintaining and raising the standards of Cochrane Reviews

	Field coordinators	Editors	Reviewers	Cochrane centres	Users
ATTRACTING DEDICATED PARTICIPANTS:					
Editors	U			U	
Reviewers	U	U			
Field coordinators	U			U	
Cochrane centres				U	
Correction of misjudgments	U	U		U	
ESTABLISHING AND DEVELOPING STANDARDS AND GUIDELINES:					
Clear standards and guidelines		U		U	
USING RIGOROUS REVIEW METHODS					
Thorough searches for relevant studies		U	U		
Explicit criteria for selecting and assessing studies		U	U		
Multiple reviewers, where appropriate		U	U		
Ongoing collection of missing data		U	U		
Collection of individual patient data, if possible		U	U		
Use of appropriate statistical techniques		U	U	U	
Sensitivity analyses		U	U		
Cautious use of subgroup analyses		U	U		
Careful interpretation of results		U	U		U
Full reporting of materials and methods		U	U		
SOFTWARE SUPPORT					
Software for reviewers				U	
Software for editors				U	
TRAINING					
Training for reviewers		U	U	U	
Training for editors		U		U	
ONGOING OPEN PEER REVIEW					
Comments on draft reviews solicited from peers	U	U	U		U
Review of drafts by editors		U			
Publication of substantive criticisms	U	U			U
Amendment of reviews in light of valid criticisms	U	U	U		U
Publication of dissenting opinions	U	U	U		U
Amendment of reviews in light of new evidence	U	U	U		U

13.1 Attracting dedicated participants

The quality of the Cochrane Collaboration, viewed from any of a variety of

perspectives, will reflect the characteristics of the individuals contributing to it. In general, these people have selected themselves during the early phases of the Collaboration's development. The community of people who have experience preparing systematic reviews of RCTs remains small, and among those who have this experience, only some will wish to commit themselves to taking on the substantial commitment that is expected of anyone who joins the Collaboration. In other words, the disincentives that confront those who are wondering whether they should become involved are a useful screening test in themselves.

13.2 Establishing and developing standards and guidelines

The Handbook is the Cochrane Collaboration's most tangible manifestation of the development of standards and guidelines. As experience accumulates among people trying to apply these, and in the light of relevant research findings, the standards and guidelines promulgated throughout the Collaboration will be modified.

13.3 Using rigorous review methods

It is neither feasible nor desirable to dictate the decisions that a reviewer should take. These will vary from review to review depending on the topic, the nature of the available evidence and the resources available to the reviewer. However, in general, the validity of Cochrane Reviews will be ensured by:

- searching as thoroughly as possible for trials meeting the inclusion criteria of a review, relying as much as possible on centralized efforts to assist with this and ensure the thoroughness and efficiency with which RCTs are identified
- use of explicit criteria for selecting trials for inclusion in a review and for assessing the quality of included studies
- application of these criteria by more than one reviewer where appropriate and feasible, to ensure the reproducibility of the judgements that are made
- ongoing efforts to collect missing information that might contribute importantly to a review, to the extent possible depending on the availability of resources and data
- collection of individual patient data from trialists where appropriate and feasible, to the extent possible depending on the availability of resources and data
- use of appropriate statistical techniques, where appropriate, to synthesize results
- use of sensitivity analyses to test the robustness of the results of a review relative to any judgements or assumptions
- cautious use of subgroup analyses and avoidance of over-interpretation of any subgroup analyses that are undertaken**Error! Bookmark not defined.Error! Bookmark not defined.**
- carefully drawn conclusions, including practical implications and implications for

future research, based on cautious interpretation of results - taking into account the limitations of the review, possible tradeoffs between the expected benefits, possible harms and costs of the interventions compared, and variability in the context in which the interventions might be applied

- full reporting of the materials and methods used in undertaking the review

Just as it is possible to update Cochrane Reviews in the light of new evidence, it is possible to improve upon the methods. Moreover, because the methods are explicitly reported in Cochrane Reviews, users can judge for themselves the validity of the results of a review.

13.4 Software support

The Cochrane Review Manager software is designed to assist reviewers in constructing reviews in the structured format described in Chapter 0. The software will continue to be developed and will incorporate standards and guidelines for Cochrane Reviews, 'online' help and error checking mechanisms, as these evolve. For example, future editions of the software might incorporate double entry for the results of each trial and enhanced support to guide reviewers through the analysis according to guidelines developed on the basis of the Workshop on statistical methods for synthesizing the results of RCTs.**Error! Bookmark not defined.** One way in which the software currently contributes to ensuring the validity of Cochrane Reviews is by facilitating registration of *a priori* protocols for planned reviews, as described in Chapter 0.

13.5 Training for reviewers

While some reviewers who join a collaborative review group will have training and experience in conducting a systematic review, many will not. The Cochrane centres are responsible for developing training materials and organizing training workshops for registered review groups. Each collaborative review group is responsible for ensuring that the members of the group have adequate training and methodological support. Training materials and opportunities for training will continue to be developed and will evolve to reflect the needs of the Collaboration and its standards and guidelines.

13.6 Training for editors

Editors in a collaborative review group need skills related to the field covered by the review group, skill and experience in assessing clinical trials, and an ability to edit scientific material for publication. The first two are not discussed further here, except to say that the first requires some first-hand experience of work in the relevant field, and a critical understanding of the main research methods used in it. The second can be acquired by taking part in the design, execution and analysis of RCTs.

An editor in a collaborative review group may have to:

- identify and define topics to be reviewed;
- identify potential reviewers who are capable of preparing a systematic review, and brief them;
- support and help reviewers to overcome difficulties encountered in preparing their reviews;
- check and ensure that reviews make scientific sense and address relevant issues, using referees if necessary;
- check and ensure that reviews are in the standard format agreed by the Cochrane Collaboration; and
- arrange that existing reviews are updated, by organising the retrieval of newly registered RCTs within the scope of the collaborative review group, sending them to reviewers, and helping them to update their reviews.

Some of these tasks are common to many types of medical editorial work, others are specific to the work of the Cochrane Collaboration as a whole, and a few to particular fields or collaborative review groups. The training of editors is therefore made up of general editorial training, training within the Cochrane Collaboration, and training on the job within collaborative review groups.

General training is probably most efficiently provided in courses, workshops and seminars organised for science editors, for example by the European Association of Science

Editors (EASE) or the North American Council of Biology Editors (CBE). The UK Cochrane Centre is a member of EASE; it may be useful for other Cochrane Centres to join EASE or CBE.

Training within the Cochrane Collaboration will take the form of workshops and seminars organised by the Cochrane centres, with participation of experienced editors from established collaborative review groups and beginners. Some of these workshops will be a general sharing of editorial experience, others may focus on particular specialised aspects of editing - such as the correct use of meta-analysis, how to plan the work of a collaborative review group and how to help and support the reviewers.

Within each collaborative review group, the senior editor or editors should take responsibility for the systematic development of the editorial skills of the other editors. Since this task will be new for most senior editors, it may be useful for the Cochrane Collaboration to hold introductory workshops.

When a new editor joins a group, the senior editor(s) should assess her or his experience and skills. The new editor should begin by editing one or more reviews together with one of the established editors, who can act as supervisor and tutor. They can together make an outline plan for the new editor's training, revising it as the training proceeds. It is desirable for a collaborative review group's editorial office to keep a formal or informal record of the work experience, courses attended, etc, of each editor. Such a record should incidentally also be of value for the individuals, as a part of their curriculum vitae that might otherwise remain inaccessible.

13.7 Open peer review and up-dating of reviews

It is important to have efficient arrangements for criticising the reviews prepared by contributors to the Cochrane Collaboration, and for amending reviews in the light of valid criticisms. Developing these arrangements is facilitated if the potential of electronic publication is exploited imaginatively. At present, opportunities for criticising reviews before they are published in print are restricted by the number and competence of the referees selected by editors. After a review has been printed, opportunities for published criticism are usually limited to the few letters that editors can accept for publication, or to book reviews, that are often unhelpfully brief and non-specific. It is also frustrating that there is no straightforward way in which the authors of printed reviews can amend their reports after taking account of valid criticisms.

The Cochrane Collaboration aims to create an iterative system through which successive versions of each review will reflect not only the emergence of new data, but also valid criticisms, solicited or unsolicited, from whatever source. This process is facilitated by the contact details both for reviewers and for editorial offices on each review. Successive versions of a particular review, together with any intervening criticisms, will be archived electronically.

REFERENCES

1. Mulrow CD. The medical review article: state of the science. Ann Int Med 1987; 106:485-8.

2. 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. JAMA 1992; 268:240-8.

3. Oxman AD, Guyatt GH, Singer J, Goldsmith CH, Hutchison BG, Milner RA, Streiner DL. Agreement among reviewers of review articles. J Clin Epidemiol 1991; 44:91-8.

4. Oxman AD, Guyatt, GH. The science of reviewing research. Ann NY Acad Sci. In press.

5. Glass GV, McGaw B, Smith M. Meta-Analysis in Social Research. Newbury Park: Sage Publications, 1981.

6. Light RJ, Pillemer DB. Summing Up: The science of Reviewing Research. Cambridge: Harvard University Press, 1984.

7. Yusuf S, Simon R, Ellenberg S, eds. Proceedings of 'Methodologic Issues in Overviews of Randomized Clinical Trials'. Stat Med 1987; 6:217-409.

8. Cooper H, Hedges LV, eds. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1993.

9. Crowley P. Antenatal corticosteroids for the prevention of respiratory distress syndrome in preterm babies. In: Enkin MW, Keirse MJNC, Renfrew MJ, Neilson JP, eds.The Cochrane Pregnancy and Childbirth Database. [Derived from The Cochrane Database of Systematic Reviews]. Published through Cochrane Updates on Disk. Oxford: Update Software, 1993, Disk Issue 2.

10. Haynes RB, Mulrow CD, Huth EJ, Altman DG, Gardner MJ. More informative abstracts revisited. Ann Int Med 1990; 113:69-76.

11. Mulrow CD, Thacker SB, Pugh JA. A proposal for more informative abstracts of review articles. Ann Int Med 1988; 108:613-5.

12. Silagy CA, Mant DC, Fowler GH, Lodge M. The effectiveness of nicotine replacement therapy in smoking cessation. Online J Curr Clin Trials [serial online]. In press.

13. Chalmers I, Enkin M, Keirse MJNC, eds. Effective Care in Pregnancy and Childbirth. Oxford: Oxford University Press, 1989.

14. Dickersin K, Hewitt P. Look before you quote. BMJ 1986; 293:1000-2.

15. Eichorn P, Yankauer A. Do authors check their references? A survey of accuracy of references in three public health journals. Am J Public Health 1987; 77:1011-2.

16. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted

to biomedical journals. Can Med Assoc J 1992; 146:861-8.

17. List of Journals Indexed in Index Medicus. Bethesda: National Library of Medicine, 1993.

18. Moher D, Jadad A, Nichol G, Penman M, Tugwell T, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Ottawa: University of Ottawa, 1993. Unpublished manuscript.

19. Sacks HS, Berrier J, Reitman D, et al. Meta-analyses of randomized controlled trials. N Engl J Med 1987; 316:450-5.

20. L'AbbJ KA, Detsky AS, O'Rourke K. Meta-analysis in clinical research. Ann Int Med 1987; 107:224-33.

21. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991; 11:1271-8

22. Antiplatelet Trialists' Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. BMJ 1988; 296:320-31.

23. Grtzsche PC. Reference bias in reports of drug trials. BMJ 1987; 295:654-6.

24. Online Databases in the Medical and Life Sciences. New York: Cuadra/Elsevier, 1987.

25. Dickersin K, Min YI. NIH clinical trials and publication bias. Online J Curr Clin Trials [serial online] 1993, Doc No 50.

26. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results: follow-up of applications submitted to two institutional review boards. JAMA 1992; 263:374-8.

27. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet 1991; 337:867-72.

28. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. J Natl Cancer Inst 1989; 81:107-15.

29. Hetherington J, Dickersin K, Chalmers I, Meinert CL. Retrospective and prospective identification of unpublished controlled trials: lessons from a survey of obstetricians and pediatricians. Pediatrics 1989; 84:374-80.

30. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. J R Statist Soc A 1988; 151:445-63.

31. Simes RJ. Confronting publication bias: a cohort design for meta-analysis. Stat Med 1987; 6:11-29.

32. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H. Publication bias and clinical trials. Controlled Clin Trials 1987; 8:343-53.

33. Simes RJ. Publication bias: the case for an international registry of clinical trials. J Clin Oncol 1986, 4:1529-41.

34. Pro-Cite [computer program]. DOS version 2.1. Ann Arbor: Personal Bibliographic Software Inc, 1993.

35. Reference Manager [computer program]. Windows version 5.5. Carlsbad: Research Information Systems Inc., 1993.

36. Cooper H, Ribble RG. Influences on the outcome of literature searches for integrative research reviews. Knowledge 1989; 10:179-201.

37. Cooper H. 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 Press, 1986, pp 341-63.

38. Detsky AS, Naylor CD, O'Rourke K, McGeer AJ, L'AbbJ KA. Incorporating variations in the quality of individual randomized trials into meta-analysis. J Clin Epidemiol 1992; 45:255-65.

39. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds JM, Gavaghan DJ, McQuay HJ. Development of an instrument to assess the scientific quality of pain research reports. Oxford: Oxford University, 1993. Unpublished manuscript.

40. 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 & Prevention, 1993. Unpublished manuscript.

41. Feinstein AR. Clinical Epidemiology: The Architecture of Clinical Research. Philadelphia: Saunders, 1985: 39-52.

42. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. N Engl J Med 1983; 309:1358-61.

43. 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.

44. Altman DG, DorJ CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990; 335:149-53.

45. Pocock SJ. Clinical Trials: A Practical Approach. Chichester: Wiley, 1983.

46. Mosteller F, Gilbert JP, McPeek B. Reporting standards and research strategies for controlled trials: agenda for the editor. Controlled Clin Trials 1980; 1:37-58.

47. Grtzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. Controlled Clin Trials 1989; 10:3-56.

48. Reitman D, Chalmers TC, Nagalingam R, Sacks H. Can efficacy of blinding be documented by meta-

analysis? Paper presented to the Society for Clinical Trials, San Diego, 23-26 May, 1988.

49. Chalmers TC, Berrier J, Hewitt P, Reitman D, Nagalingam R, Sacks H. Logistics of determining rates of rare side effects by meta-analysis. Paper presented at the 9th Annual Meeting of the Society for Clinical Trials, 22-25 May, San Diego, 1988.

50. Enkin MW, Keirse MJNC, Renfrew MJ, Neilson JP, eds.The Cochrane Pregnancy and Childbirth Database. [Derived from The Cochrane Database of Systematic Reviews]. Published through Cochrane Updates on Disk. Oxford: Update Software, 1993, Disk Issue 2.

51. Chalmers I, ed. Oxford Database of Perinatal Trials. Version 1.2. Oxford: Oxford University Press, 1992, Disk Issue 7.

52. Yusuf 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.

53. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116:78-84.

54. Laird NM, Mosteller F. Some statistical methods for combining experimental results. Int J of Technology Assessment in Health Care 1990; 6:5-30.

55. Cooper HM, Rosenthal R. Statistical versus traditional procedures for summarizing research findings. Psychol Bull 1980; 87:442-9.

56. Davis DA, Thomson MA, Oxman AD, Haynes RB. Evidence for the effectiveness of CME: a review of 50 randomized controlled trials. JAMA 1992; 268:1111-7.

57. 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; 27:335-71.

58. Schulz KF, Altman DG. Statistical methods for data synthesis. Cochrane Workshop Report. Oxford: UK Cochrane Centre, 1993.

59. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. N Engl J Med 1988; 318:1728-33.

60. Walker AM, Martin-Moreno JM, Artalejo FR. Odd man out: a graphical approach to meta-analysis. Am J Public Health 1988; 78:961-6.

61. Laird NM, Mosteller F. Some statistical methods for combining experimental results. Int J Tech Asses Health Care 1990; 6:5-30.

62. Altman DG. Confidence intervals in research evaluation [Editorial]. ACP J Club. 1992 Mar-Apr:A28 (Ann Intern Med. vol 116, suppl 2).

63. Thompson SG, Pocock SJ. Can meta-analyses be trusted? Lancet 1991; 338:1127-30.

64. Berlin JA, Begg CB, Louis T. An assessment of publication bias using a sample of published clinical trials. J Am Stat Assoc 1989; 84:381-92.

65. Making clinical trialists register. Lancet 1991, 338:244-5.

66. Schneider B. Analysis of clinical trial outcomes: alternative approaches to subgroup analysis. Controlled Clin Trials 1989; 10:176S-186S.

67. Breslow NE, Day NE. Analysis of Case-Control Studies, Vol 1 of Statistical Methods in Cancer Research. Lyon: IARC Sci Publ, 1980: 122-59.

68. Bracken MB. Statistical methods for analysis of effects of treatment in overviews of randomized trials. In: Effective Care of the Newborn Infant, eds JC Sinclair, MB Bracken. Oxford: Oxford University Press, 1992: 13-8.

69. 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; 102:305S-311S.

70. Hill AB. Principles of Medical Statistics. 9th ed. London: Lancet, 1971: 312-20.

71. Eddy DM: Anatomy of a decision. JAMA 1990; 263:441-3.

72. Mugford M, Kingston J, Chalmers I. Reducing the incidence of infection after caesarean section: implications of prophylaxis with antibiotics for hospital resources. BMJ 1989; 299:1003-6.

73. Mugford M, Piercy J, Chalmers I. Cost implications of different approaches to the prevention of respiratory distress syndrome. Arch Dis Child 1991; 66:757-64.

74. Enkin M, Keirse MJNC, Renfrew MJ, Neilson JP. A Guide to Effective Care in Pregnancy and Childbirth. 2nd ed. Oxford: Oxford University Press, 1994.

75. Neilson JP. Routine ultrasonography in early pregnancy. In: Enkin MW, Keirse MJNC, Renfrew MJ, Neilson JP, eds.The Cochrane Pregnancy and Childbirth Database. [Derived from The Cochrane Database of Systematic Reviews]. Published through Cochrane Updates on Disk. Oxford: Update Software, 1993, Disk Issue 2.