Cochrane Handbook for Systematic Reviews of Interventions 4.2.6

Updated September 2006

Table Of Contents

About the handbook

Editors

Julian PT Higgins and Sally Green.

How to cite this version of the Handbook

Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]. In: The Cochrane Library, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd.

or

Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]. http://www.cochrane.org/resources/handbook/hbook.htm (accessed 6th October 2006).

When referring to a specific section or subsection refer to it by the title and section number, NOT page numbers. For example:

Higgins JPT, Green S, editors. Formulating the problem. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]; Section 4. In: The Cochrane Library, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd.

or

Higgins JPT, Green S, editors. Formulating the problem. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]; Section 4. http://www.cochrane.org/resources/handbook/hbook.htm (accessed 6th October 2006).

When referring to a section that has section editors listed, use:

Deeks JJ, Higgins, JPT, Altman DG, editors. Analysing and presenting results. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]; Section 8. In: The Cochrane Library, Issue 4, 2006. Chichester, UK: John Wiley & Sons, Ltd.

or

Deeks JJ, Higgins, JPT, Altman DG, editors. Analysing and presenting results. In: Higgins JPT, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions 4.2.6 [updated September 2006]; Section 8. http://www.cochrane.org/resources/handbook/hbook.htm (accessed 6th October 2006).

Contact addresses

Sally Green
Australasian Cochrane Centre
Monash Institute of Health Services Research
Monash Medical Centre
Locked Bag 29
Clayton, Victoria 3168
Australia
Phone: +61 3 9594 7530
Fax: +61 3 9594 7554
Email: sally.green@med.monash.edu.au

Julian Higgins
MRC Biostatistics Unit
Institute of Public Health
Robinson Way
Cambridge CB2 2SR
United Kingdom
Phone: +44 1223 330 396
Fax: +44 1223 330 388
Email: julian.higgins@mrc-bsu.cam.ac.uk

Jane Lane (Administration)
Public Health Genetics Unit

Strangeways Laboratory

Worts Causeway

Cambridge CB1 8RN

United Kingdom

Phone: +44 1223 742003

Email: jane.lane@srl.cam.ac.uk

Updates and corrections

See http://www.cochrane.org/resources/handbook/

Sources of support

Present sources of support

Department of Health and Ageing, Australia

Monash University, Australia

Medical Research Council, United Kingdom

Public Health Genetics Unit, United Kingdom

Previous sources of support

National Health Service Research & Development Programme, England

Health Research Board, Ireland

National Institute of Public Health, Norway

Copenhagen Hospital Corporation, Denmark

Health Services Research and Development Service and the University of Texas Health Science Center, San Antonio, USA

US Veterans Health Administration, USA

Oxford Regional Health Authority, UK

Nuffield Provincial Hospitals Trust, UK

LW Frohlich Fund, USA

Norwegian Ministry of Health and Social Affairs, Norway

Norwegian Research Council, Norway

Glaxo Wellcome, Norway

Pfizer, Norway

Acknowledgements

Many people have contributed to this and preceding editions of the Cochrane Collaboration's Handbook, both directly and indirectly (through workshops, helpful suggestions and inspiration). Grateful acknowledgement is offered to everyone who has made its preparation worthwhile, by enthusiastically receiving the Cochrane Collaboration and committing his or her own time to making it a reality.

The first edition of the Handbook (1994) was developed by Andy Oxman, Iain Chalmers, Mike Clarke, Murray Enkin, Ken Schulz, Mark Starr, Kay Dickersin, Andrew Herxheimer and Chris Silagy with administrative support from Sally Hunt.

The second and third editions of the Handbook (1995 and 1997) were edited by Andy Oxman and Cynthia Mulrow.

Subsequent editions of the Handbook were edited by Mike Clarke and Andy Oxman (1999 onwards), and then by Phil Alderson, Sally Green and Julian Higgins (2003 onwards). Julian Higgins and Sally Green are now the editors of the Handbook (since 2005).

The Handbook editors are supported by advice from the Handbook Advisory Group, with administrative support provided byJane Lane and technical support by Jacob Riis. In addition to the editors, the current membership of this Group is: Lisa Askie, Chris Cates, Jon Deeks, Matthias Egger, Davina Ghersi, Donna Gilles, Paul Glasziou, Andrew Herxheimer, Carol Lefebvre, Harriet McLehose, Philippa Middleton, Ruth Mitchell, David Moher, Miranda Mugford, Jane Noyes, Donald Patrick, Jennie Popay, Barney Reeves, Jacob Riis, Jonathan Sterne, Lesley Stewart, , Jessica ThomasJayne Tierney, Danielle Wheeler.

We would also like to thank the following people for their past or continuing contributions to the Handbook: Christina Aguilar, Doug Altman, Bob Badgett, Hilda Bastian, Lisa Bero, Michael Brand, Joe Cavellero, Mildred Cho, Lelia Duley, Frances Fairman, Jeremy Grimshaw, Gord Guyatt, Peter Gøtzsche, Jeph Herrin, Nicki Jackson, Monica Kjeldstrøm, Jos Kleijnen, Valerie Lawrence, Eric Mannheimer, Rasmus Moustgaard, Melissa Ober, Drummond Rennie, Dave Sackett, Mark Starr, Nicola Thornton, Luke Vale, Veronica Yank.

We are moving towards a system where each section of the Handbook will have its own editors who will be listed at the start of the section, and contributors who will be listed at the end of the section. Over time, we hope this will make it clearer who has been responsible for contributing to the Handbook, and give credit appropriately.

Section 5 of the Handbook was prepared first, in 1994, by Kay Dickersin and Carol Lefebvre as a document entitled Establishing and Maintaining Registers of RCTs. Many others contributed to it and Kay Dickersin was the editor. This document was updated by Kay Dickersin and Kristen Larson in 1995. In 1997, a major revision was undertaken by Mike Clarke for inclusion in version 3 of the Cochrane Reviewer's Handbook. The current version included was largely rewritten in 2002 by Eric Manheimer; with input from Kay Dickersin and members of the Handbook Advisory Group.

What's new

Corrections and changes in Version 4.2.6 (September 2006) of the Handbook

1. Section 3.1 (Cover sheet): New format for entering list of authors for citation in RevMan
2. Section 3.3 (Abstract): New format for citing CENTRAL
3. Section 3.4 (Text of a review): New guidance on length of a review (word limits)
4. Appendix 5b.1: Corrections to search strategies ('Human' to 'Humans' in MEDLINE; correction of line number)

Corrections and changes in Version 4.2.5 (May 2005) of the Handbook

1. In Section 2, information on conflict of interest and commercial sponsorship has been updated to reflect Collaboration's policy on commercial sponsorship, and replaces the previous Appendix 2b.

2. A new Section 3 ('Guide to the contents of a protocol and review') replaces the previous Appendix 2a ('Guide to the format of a Cochrane review'). This guide has been extensively revised and updated. These revisions include:

   (i) A new list of recommended subheadings for Cochrane reviews

   (ii) New guidance for Plain Language Summaries

   (iii) Recommendations on writing the 'Implications for Research' section from the Database of Uncertainty about the Effect of Treatments (DUET) project

   (iv) Clarification of what may appear under 'Published notes'

   (v) Clarification of guidance that only contributions of co-authors should be listed under 'Contributions'. Others should be listed, with permission, under 'Acknowledgements'.

3. Two new sections have been added to Section 8.11 (Special topics). 8.11.2 Cluster-randomized trials and 8.11.3 Cross-over trials.
4. A new Appendix of the Handbook on including adverse effects in Cochrane reviews has been added as Appendix 6b. Some new terms have been added to the glossary: adverse effect (replacing adverse reaction), adverse event, safety, side effect and tolerability. The previous Appendix 6 (Reviews including non-randomised studies) has been renumbered as Appendix 6a, accordingly.

Corrections and changes in Version 4.2.4 (March 2005) of the Handbook

1. Julian Higgins and Sally Green are now the editors of the Handbook. Phil Alderson stepped down at the end of 2004, and we would like to thank him for all of his invaluable contributions.
2. We have begun to replace the term ‘reviewer’ with the term ‘author’, in line with the decision of Steering Group in Feb-Mar 2004. This has so far been implemented in Sections 1, 2 and 8.
3. The title of the Handbook has changed from Cochrane Reviewers’ Handbook to Cochrane Handbook for Systematic Reviews of Interventions. For details of how to cite the Handbook, please refer to How to cite this version of the Handbook.
4. Sections 1-3 and Section 10 have been restructured and extensive revisions have been made to Section 2. Section 2 is now titled ‘Preparing a Cochrane review’. It incorporates the previous content of Section 3, new discussions of review teams (including advisory groups), sections on training and software from the previous Section 10 (both updated), and a reduction in the length of the section on open learning materials.
5. A new Section 3 will be available in a future issue, and will comprise an updated and extended version of the existing Appendix 2a. Guide to the format of a Cochrane review.

Corrections and changes in Version 4.2.3 (November 2004) of the Handbook

The Glossary has been extensively revised.

Corrections and changes in Version 4.2.2 (March 2004) of the Handbook

Minor corrections have been made to Section 8.

The search strategies for MEDLINE in Appendix 5b have been checked and updated to reflect changes in MeSH terms. The only substantive change is the replacement of ANIMAL with ANIMALS.

Corrections and changes in version 4.2.1 (December 2003) of the Handbook

‘About the Handbook’ has been updated to reflect changes to the editorial team, and changes in the Cochrane Collaboration’s publishers.

Section 8 has been substantially revised and updated. This is the first half of what will be the new section 8 and covers core material. The second half is planned for late 2004. In the meantime, two sections from the old version of section 8 have been retained.

Appendix 8a Effect measures for dichotomous data has been removed as it is now covered in Section 8.

A new Appendix 8a has been added, covering advice on graphical presentation of results.

Appendix 2a.1 has been updated to clarify whose contact details should be entered if none of the reviewers will be a contact reviewer.

Appendix 2a.3 has been updated to make clear that links to additional figures should not be included in an abstract.

Text in section 2.0 and Appendix 2a.2 has been revised to reflect current arrangements for preparing synopses.

In appendix 5b.1 and 5b.2, search strategies for SilverPlatter-MEDLINE and Ovid-MEDLINE have been corrected.

The URL for the Cochrane Collaboration site has been corrected in the Glossary.

Corrections and changes in version 4.2.0 (March 2003) of the Handbook

Major corrections and changes:

Section 5 and Appendices 5: these have been revised and updated.

Section 8: this has been amended slightly to mention the addition of the generic inverse variance method to RevMan 4.2 and the ability to include additional figures.

Minor corrections:

Acknowledgements: the help of additional people in the preparation of this version of the Handbook has been acknowledged. Additional figures: relevant sections have been amended to note the ability to include additional figures in Cochrane reviews (using RevMan 4.2).

Appendix 2a, section 2a.1: the categories for sources of support have been changed to

‘internal’ and ‘external’.

Appendix 2a, section 2a.3: the way to describe a search of the ‘Cochrane Central Register of Controlled Trials (CENTRAL)’ in the abstract for a Cochrane review has been clarified.

Appendix 2c: the permission to publish form has been removed while it is being revised (this has also led to a change in section 2.2.3).

Some typographical mistakes have been corrected.

Corrections and changes in version 4.1.6 (January 2003) of the Handbook

Major corrections and changes:

Section 1: this has been updated and information has been added on The Cochrane Collaboration Open Learning Material for Cochrane Reviewers.

Section 3.2: the policy on the withdrawal of protocols has been updated.

Section 10.10: the policy on the withdrawal of reviews has been added.

Appendix 1: this has been added to provide information on The Cochrane Collaboration Open Learning Material for Cochrane Reviewers.

Minor corrections:

Some typographical corrections have been made.

The name for the ‘Cochrane Controlled Trials Register (CENTRAL/CCTR)’ has been changed to the ‘Cochrane Central Register of Controlled Trials (CENTRAL)’.

Sources of support: this has been updated to reflect the support from the National Health Service Research & Development Programme, UK and the Health Research Board, Ireland.

Acknowledgements: the help of additional people in the preparation of this version of the Handbook has been acknowledged.

Section 4.5: an additional example (children versus adults) has been added of why separate reviews might be done.

Section 9.7: this has been amended to clarify the distinction between ‘no evidence of an effect’ and ‘evidence of no effect’.

Appendix 2a, 2a.4 Text, Results: this has been amended to clarify the distinction between ‘no evidence of an effect’ and ‘no evidence of effect’.

Appendix 2a, 2a.5 Conflict of interest: suggested wording if there no known conflicts of interest has been changed to ‘None known’.

Corrections and changes in version 4.1.5 (April 2002) of the Handbook

Internet addresses: the list of Internet addresses has been reduced to the three official Cochrane Collaboration sites that are mirrors of each other (i.e. www.cochrane.de, www.cochrane.org and www.update-software.com/ccweb).

Appendix 2a, Section 2a.2: it has been clarified that help with synopses should be sought directly from the Cochrane Consumer Network, rather than the Australasian Cochrane Centre.

Corrections and changes in version 4.1.4 (October 2001) of the Handbook

Major corrections:

Section 2.3: the suggested wording to use when versions of Cochrane reviews are published in paper journals has been revised.

Minor corrections:

Section 9.7: advice has been added on the balanced interpretation of analyses when the confidence interval for the effect estimate overlaps the null value.

Section 10.11: the address of the Comments and Criticisms web page has been updated.

Corrections and changes in version 4.1.3 (June 2001) of the Handbook

Minor corrections:

Section 9.7: this has been expanded to include more discussion of the interpretation of results that are not statistically significant.

Appendix 5a: a contact address has been added for the International Register of Clinical Trials Registers.

Corrections and changes in version 4.1.2 (March 2001) of the Handbook

Major corrections:

Appendix 6: this has been replaced with an updated version.

Minor corrections:

Section 1.0: the new name (Cochrane Methodology Register) has replaced "Cochrane Review Methodology Database".

Glossary: Three terms have been added: inter-rater reliability, intra-rater reliability and N of 1 randomised trial.

Corrections and changes in version 4.1.1 (December 2000) of the Handbook

Major corrections:

Section 10.10: the revised Cochrane Collaboration policy that reviews should be updated at least every 2 years (instead of every year) has been added. This policy was agreed by the Steering Group in October 2000.

Minor corrections:

Section 10.10: The Collaboration policy that protocols that have not been converted into full reviews within two years should generally be withdrawn from the CDSR (stated in section 3.2) has been restated here.

Section 10.11: the mention that software is being developed to help Criticism Editors to coordinate the reviewers' responses to comments and criticisms has been deleted.

Appendix 5a: The list of registers has been replaced by the URLs for online registers of registers.

Corrections and changes in version 4.1.0 (June 2000) of the Handbook

Major corrections and changes:

Chapter 2: additional guidance has been added on the publication of Cochrane Reviews in journals.

Chapter 5: this has been updated.

Chapter 6: this has been updated.

Chapter 11: a new section (11.6) has been added on the conversion of reviews that used individual patient data into Cochrane Reviews

Appendix 2a, synopses: The guidance on preparing synopses has been changed to reflect the new policy that responsibility for the approval of the synopsis to be included in a Cochrane review rests solely with the relevant review group.

Appendix 2a: a section has been added to show the elements of Cochrane protocols and reviews that should be published.

Minor corrections:

Acknowledgements: the help of additional people in the preparation of this version of the Handbook has been acknowledged.

Appendix 2a, Text: The importance of keeping searches up-to-date has been added to the guidance on the content of the Search strategy section of the text of a Cochrane Review.

Appendix 2a, references: the title of the Flanagin 1998 reference has been corrected.

Corrections and changes in Version 4.0.0 (July 1999) of the Handbook

The Handbook has been thoroughly revised to take account of the changes in RevMan. We have also taken the opportunity to update several other sections of the Handbook.

Corrections and changes in Version 3.0.2 (September 1997) of the Handbook

Major corrections and changes

1. In appendix 2c, 'Conditions of publication', it has now been specified that a new 'Conditions of publication' form should be filled out with each substantive revision of a review.

2. In order to keep version numbers of the Handbook consistent with version numbers of RevMan, the Handbook will now make use of three digits:

· the first digit indicates a new release of RevMan and the Handbook,

· the second digit indicates an interim release of RevMan and the Handbook,

· the third digit indicates changes to the Handbook only.

Minor corrections and changes

1.Section 5.5 on handsearching has been updated to take account of the development of the control register on studies that might be relevant for inclusion in Cochrane Reviews (CENTRAL).

2. The glossary has three additions; *CENTRAL*, *trend*, and *peer review*. The terms *Handsearching* and *Cochrane Controlled Trials Register (CCTR)* have been updated.

3. Synapse Publishing Inc. have put a version of the Handbook on the WWW at the following address: hhtp://www.medlib.com/cochranehandbook/.

4. Corrections have been made to the references in appendix 2a.6.

5. The list of handbook versions and related resources has been updated.

6. About the Handbook and What's New have been updated.

Corrections and changes in Version 3.0.1 (December 1996) of the Handbook

Major corrections and changes

1. Appendix 11a, 'Practical methodology of meta-analyses (overviews) using updated individual patient data', was added to the Handbook.

2. Appendix 5a, 'Registers of clinical trials', was updated.

Minor corrections and changes

1. All references to publications included in the Cochrane Library were updated ('How to cite the Handbook'; references: section 1; references: section 3; references: section 4; references: section 6; references: section 8; Appendix 5b; Appendix 5c.).

Corrections and changes in Version 3.0.0 (October 1996) of the Handbook

1. Editorial responsibility

Responsibility for maintaining material formerly contained in Sections I to V of The Cochrane Collaboration Handbook was devolved as described below. The Handbook now consists solely of what was formerly Section VI: Preparing and Maintaining Systematic Reviews (Oxman, 1995). Cynthia Mulrow, director of the San Antonio Cochrane Center, joined Andy Oxman as co-editor. The entire Handbook was revised in response to suggestions we have received regarding the previous edition of the Handbook and the Training Manual prepared by the San Antonio Cochrane Center.

Editorial responsibilities for written material prepared on behalf of the Cochrane Collaboration has been evolving and it became clear in 1995 that new arrangements were required to deal with new circumstances. At its meeting 27 February 1996 in San Francisco the Steering Group established an Editorial Board to oversee the preparation of written material prepared on behalf of the Collaboration. This is one of five groups responsible for core functions that report directly to the Steering Group. The other groups responsible for core functions are the Software Development Group, the Trials Registers Development Group, a group responsible for forthcoming Colloquia, and the editorial team for the Handbook.

Further changes in editorial responsibility were proposed by Iain Chalmers and Andy Oxman to accommodate several developments, including:

- potential duplication of effort, and confusion regarding the roles of the Editorial Board and the Handbook editorial team
- the availability of CDSR and the development of modules in CDSR for Cochrane Centres, Fields, MGs and the Consumer Network as well as for CRGs
- the establishment of an elected Steering Group with representatives for each type of entity and the formation of groups responsible for core functions, which are directly responsible to the Steering Group

The proposed changes were circulated to all registered groups and approved by the Steering Group at its meeting 19 August 1996. The new arrangements are as follows:

Material about	**Responsible group**	**Current co-ordinator**
The Collaboration	Editorial Board	Jos Kleijnen
Core Functions:		
Handbook	Handbook Advisory Group	Andy Oxman &

		Cynthia Mulrow
Software	Software Development Group	Monica Fischer
Trials registers	Trials Registers Development Group	Kay Dickersin & Jean-Pierre Boissel
Registered groups:		
CRGs	CRG reps on Steering Group	CRG reps to decide
Cochrane Centres	Centre directors on Steering Group	Peter Gøtzsche
Fields	Field rep on Steering Group	Field rep to decide
Consumer Network	Consumer reps on Steering Group	Consumer reps to decide
MGs	MG rep on Steering Group	Andy Oxman

2. Abstracts

Abstracts are no longer optional and the subheadings used in abstracts have been changed to:

- Objectives
- Search strategy
- Selection criteria
- Data collection & analysis
- Main results

(see section 2a.2 in appendix 2a .)

3. Descriptions of methods used by Collaborative Review Groups

All reviews should state specifically when the register of trials maintained by the CRG responsible for the review was last searched for relevant studies. Descriptions of the methods used to develop and maintain CRG registers of trials are included in CRG modules published in the *Cochrane Database of Systematic Reviews* (*CDSR*). Other standardised methods used by a CRG should also be described in the group's module. Reviewers should state explicitly that they have used these methods and when they have used methods that differ from the standard methods used by a group.

4. Reviews of non-experimental evidence

Some CRGs, Fields and Methods Groups (MGs) have begun to explore ways of incorporating non-experimental evidence in reviews when this is appropriate. These developments are reflected in changing the terminology from 'trials' to 'studies' and adding 'Types of studies' as a new subheading under 'Selection criteria'.

5.Links between the Handbook and related resources

The Handbook is being linked to several related resources (see 'About the Handbook'). These include: the *Cochrane Review Methodology Database*, the San Antonio Cochrane Center's Training Manual, Review Manager, a glossary, a frequently asked questions (FAQ) list, a library of examples, a library of slides, a register of empirical methodological studies, systematic reviews of those studies, and modules prepared by MGs for inclusion in *CDSR*.

6. Conflict of interest

A conflict of interest statement will be included in all Cochrane Reviews beginning with the second issue of the Cochrane Library in 1997 (see section 2a.2 and section 2a.4 in appendix 2a).

Corrections and changes to the Glossary

The following changes and corrections have been made to the March 2001 version of the Glossary:

The following terms have been added to the Glossary:

- Inter-rater reliability
- Intra-rater reliability
- N of 1 randomised trial

The following changes and corrections have been made to the February 2000 version of the Glossary:

The following terms in the Glossary have been updated:

- Bayesian approach
- Case study
- CENTRAL
- Cochrane Database of Systematic Reviews
- Cochrane Library
- Cochrane Review
- Cochrane Reviewers' Handbook
- Cochrane Review Methodology Database (CRMD)
- Cohort study
- Confounding
- Coordinator
- CRMD
- Economic analysis
- Editorial team
- Expected date (of a Cochrane Review)
- Funnel plot
- Handsearching
- Meta-analysis
- Methodological quality
- Methods Group (MG)
- Minimisation
- Negative study
- Observational study
- Parent Database
- Peer review
- Phase II studies
- Primary study
- P-value
- Quality score
- Random selection
- Randomised controlled trial (RCT)
- Referee process
- Relative Risk (RR)
- Risk difference (RD)
- Run-in period
- Sequential trial
- Trials register
- Unit of allocation
- World Wide Web

The following terms have been added to the Glossary:

- Cochrane Methodology Register
- Meerkat
- Review Group Coordinator
- RGC
- SMD
- WMD

1 Introduction

Edited by Julian PT Higgins and Sally Green

1.1 Systematic reviews and the Cochrane Handbook

Healthcare providers, consumers, researchers, and policy makers are inundated with unmanageable amounts of information. We need systematic reviews to efficiently integrate valid information and provide a basis for rational decision making (Mulrow 1994). Systematic reviews establish where the effects of healthcare are consistent and where they may vary significantly. Consistent research results can be applied across populations, settings, and small differences in treatment (e.g. dose). The use of explicit, systematic methods in reviews limits bias (systematic errors) and reduces chance effects, thus providing more reliable results upon which to draw conclusions and make decisions (Antman 1992, Oxman 1993b). Meta-analysis, the use of statistical methods to summarise the results of independent studies, can provide more precise estimates of the effects of healthcare than those derived from the individual studies included in a review (Oxman 1993a, Sacks 1987, L'Abbe 1987, Thacker 1988) and allows decisions that are based on the totality of the available evidence.

Wider recognition of the key role of reviews in synthesising and disseminating the results of research has prompted people to consider the validity of reviews. In the 1970s and early 1980s, psychologists and social scientists drew attention to the systematic steps needed to minimise bias and random errors in reviews of research (Light 1971, Glass 1976, Rosenthal 1978, Jackson 1980, Cooper 1982). It was not until the late 1980s that people drew attention to the poor scientific quality of healthcare review articles (Mulrow 1987, Yusuf 1987, Oxman 1988). However, recognition of the need for systematic reviews of healthcare has grown rapidly and continues to grow, as reflected by the number of articles about review methods and empirical studies of the methods used in reviews, published in the Cochrane Methodology Register; the number of systematic reviews included in the Database of Abstracts of Reviews of Effects; and the rapid growth in the number of reviews published within The Cochrane Collaboration in the Cochrane Database of Systematic Reviews. All the above databases are published and updated quarterly in The Cochrane Library (The Cochrane Library).

This Handbook builds on the work of a large number of people, including those involved in Cochrane Methods Groups, practical experience and feedback from Collaborative Review Groups (which have taken on the daunting task of systematically reviewing the effects of healthcare within their areas of interest), Cochrane Centres (which provide training and support for authors (reviewers) of Cochrane reviews and others involved in the review process) and Cochrane Fields (which represent broad areas of health care).Whenever possible recommendations made here are based on empirical evidence and advice from Cochrane Methods Groups.

Our aim is to help review authors make good decisions about the methods they use, rather than dictate arbitrary standards. However, where The Cochrane Collaboration has laid down policy, which must be followed by Cochrane authors, this is made clear. The guidelines provided here are intended to help review authors to be systematic and explicit (not mechanistic!) about the questions they pose and how they derive answers to those questions. These guidelines are not a substitute for good judgement.

The Cochrane Collaboration and the Cochrane Handbook for Systematic Reviews of Interventions focus particularly on systematic reviews of randomised controlled trials (RCTs) because they are likely to provide more reliable information than other sources of evidence on the differential effects of alternative forms of healthcare (Kunz 2003). Systematic reviews of other types of evidence can also help those wanting to make better decisions about healthcare, particularly forms of care where RCTs have not been done and may not be possible or

appropriate. Furthermore, RCTs are particularly suited to questions of effectiveness, but may be less suitable for considerations of safety or adverse effects. The basic principles of reviewing research are the same, whatever type of evidence is being reviewed. Although we focus mainly on systematic reviews of RCTs, we address issues specific to reviewing other types of evidence when this is relevant. Fuller guidance on such reviews is being developed. In 2003, The Cochrane Collaboration decided to develop a database of systematic reviews of diagnostic test accuracy that, in time, will complement the Cochrane Database of Systematic Reviews (CDSR) on The Cochrane Library. A separate Handbook for these reviews is in development.

Cochrane reviews have a standard format that we outline in Section 2.2. Those preparing a review should begin by developing a protocol (Section 2.1), which contains the background and objectives along with an outline of the proposed search methods and plans for collecting and analysing data. Editorial base staff of Collaborative Review Groups appraise and give feedback on these protocols before reviews are conducted. The protocol will also be published in the CDSR and may be subject to comments and criticisms. A detailed description of what should appear in each section of a protocol and each section of a review is provided in Section 3 of this Handbook.

The main body of the Handbook is organised in seven sections according to the steps of preparing and maintaining a systematic review:

- Formulating the problem
- Locating and selecting studies
- Quality assessment of studies
- Collecting data
- Analysing and presenting results
- Interpreting results
- Improving and updating reviews

In the final section we take up specific issues about using individual patient data in reviews.

This Handbook is continually updated to reflect advances in systematic review methodology and in response to feedback from users. If you have any feedback about the Handbook please email Jane Lane (jane.lane@srl.cam.ac.uk).

1.2 Contributions

This section builds on earlier versions of the Handbook. For details of previous authors and editors of the Handbook, please refer to the Acknowledgements section.

Contributing authors (March 2005): Sally Green, Julian Higgins

Editors: Julian Higgins and Sally Green

1.3 References

Antman 1992. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA 1992; 268:240-8.

CDSR 2003. Cochrane Database of Systematic Reviews. In: The Cochrane Library, Issue 1, 2003. Oxford: Update Software. Updated quarterly.

CMR 2003. Cochrane Methodology Register. In The Cochrane Library, Issue 1, 2003. Oxford: Update Software; Updated quarterly.

Cooper 1982. Cooper HM. Scientific guidelines for conducting integrative research reviews. Rev Educ Res 1982; 52:291-302.

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

Jackson 1980. Jackson GB. Methods for integrative reviews. Rev Educ Res 1980; 50:438-60.

Kunz 2003. Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials (Cochrane Methodology Review). In: The Cochrane Library, Issue 1, 2003. Oxford: Update Software.

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

Light 1971. Light RJ, Smith PV. Accumulating evidence: procedures for resolving contradictions among different research studies. Harv Educ Rev 1971; 41:429-71.

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

Mulrow 1994. Mulrow CD. Rationale for systematic reviews. BMJ 1994; 309:597-9.

NHS CRD 2003. NHS Centre for Reviews and Dissemination. Database of Abstracts of Reviews of Effectiveness. In: The Cochrane Library, Issue 1, 2003. Oxford: Update Software. Updated quarterly.

Oxman 1988. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. Can Med Assoc J 1988; 138:697-703.

Oxman 1993a. Oxman AD. Meta-statistics: Help or hindrance? ACP J Club 1993; 118:A-13.

Oxman 1993b. Oxman AD, Guyatt GH. The science of reviewing research. Ann NY Acad Sci 1993; 703:125-33.

Rosenthal 1978. Rosenthal R. Combining results of independent studies. Psychol Bull 1978; 85:185-93.

Sacks 1987. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987; 316:450-5.

Thacker 1988. Thacker SB. Meta-analysis: a quantitative approach to research integration. JAMA 1988; 259:1685-9.

Yusuf 1987. Yusuf S, Simon R, Ellenberg S (eds). Proceedings of 'Methodologic issues in overviews of randomized clinical trials'. Stat Med 1987; 6:217-409.

2 Format of a Cochrane review

Edited by Julian PT Higgins and Sally Green

2.1 Rationale for protocols

Preparing a review is a complex process that comprises many judgements, as well as decisions about the process and the resources needed (Section 2.3). As in any scientific endeavour, the methods to be used should be established beforehand. However, reviews are by their nature retrospective, since the studies included are usually identified after they have been completed and reported. Therefore, it is important to make the process as rigorous and well defined as possible (Light 1984b) while maintaining a practical perspective. The author's (reviewer's) knowledge of the results of an included study may influence:

- the definition of a systematic review question
- the criteria for study selection
- the comparisons for analyses
- the outcomes to be reported in the review

While the intention should be that a review will adhere to the published protocol, just as protocols for randomised trials 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 recognised that this is not always possible or appropriate. Changes in the protocol should not be made on the basis of how they affect the results of the review. 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, changes in the protocol should be documented and reported in the methods section of the completed review, and 'sensitivity analyses' (see section 8.10) of the impact of such decisions on the results of the review should be made when possible.

2.2 Format of a Cochrane review

All Cochrane reviews of interventions have the same format. There are several reasons for this. It helps readers to find the results of research quickly and to assess the validity, applicability and implications of those results. It guides authors to report their work explicitly and concisely, and minimises 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 original 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 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 of Systematic Reviews (CDSR), the format must be well defined and fixed. Some minor changes have been made from the format described in the first (1994) edition of the Handbook. These changes have been made based on the experience of Collaborative Review Groups, feedback from users of Cochrane reviews and suggestions brought forward through the Handbook Advisory Group and the RevMan Advisory Group, which has developed specifications for the software that is used to prepare Cochrane reviews. The Review Manager (RevMan) software is

designed to help authors construct reviews in the appropriate format and to prepare files required to transfer reviews electronically.

Each review consists of:

- a cover sheet - giving the title, citation details and contact addresses
- a plain language summary
- an abstract - using a structured format
- the text of the review - consisting of an introduction (background and objective), methods (selection criteria, search methods, data collection and data analysis), results (description of studies, methodological quality, and results of analyses), discussion, authors' conclusions, acknowledgements and conflicts of interest
- tables and figures - showing characteristics of the included studies, specification of the interventions that were compared, the results of the included studies, a log of the studies that were excluded, and additional tables and figures relevant to the review
- references

Each protocol consists of:

- a cover sheet – giving the title, citation details and contact addresses
- the text of the protocol – consisting of an introduction (background and objective), methods (selection criteria, search methods, data collection and data analysis), acknowledgements and conflicts of interest
- tables and figures - relevant to the background or methods
- references

Standard headings and tables embedded in RevMan guide review authors when preparing their report and make it easier for readers to identify information that is of particular interest to them. The headings are listed below. The content that should follow each heading is described in Section 3 (Guide to the contents of a protocol and review).

2.2.1 Detailed outline of a protocol for a Cochrane review

The following elements define a complete protocol for a Cochrane review, and indicate how the protocol appears in the Cochrane Database of Systematic Reviews. If any of the sections marked below with a * are empty, the protocol should not be published until something has been added to the section.

Cover sheet:

*Title

*Name of contact author

*List of authors for citation

Contributions

Sources of support

Internal

External

What's New

Text

Issue protocol first published

Published notes

***Text of review:**

Background

Objectives

Criteria for selecting studies for this review

Types of studies

Types of participants

Types of interventions

Types of outcome measures

Search strategy for identification of studies

Methods of the review

Acknowledgements

Conflicts of interest

References:

Other references

Additional references

Tables and figures:

Additional tables

Additional figures

Comments and criticisms:

Title

Summary

Reply

Contributors

2.2.2 Detailed outline of a Cochrane review

The following elements define a complete Cochrane review, and indicate how the review appears in the Cochrane Database of Systematic Reviews. If any of the sections marked below with a * are empty, the review should not be published until something has been added to the section.

Cover sheet:

*Title

*Name of contact author

*List of authors for citation

Contributions

- Sources of support
  - Internal
  - External
- What's New
  - Text
  - Issue protocol first published
  - Issue review first published
  - *Date of last substantive update
  - Date new studies sought but none found
  - Date new studies found but not yet included/excluded
  - Date new studies found and included or excluded
  - Date authors' conclusions section amended
- Published notes

Plain Language Summary

***Abstract:**

- Background
- Objectives
- Search strategy
- Selection criteria
- Data collection & analysis
- Main results
- Authors' conclusions

***Text of review:**

- Background
- Objectives
- Criteria for selecting studies for this review
  - Types of studies
  - Types of participants
  - Types of interventions
  - Types of outcome measures
- Search strategy for identification of studies
- Methods of the review
- Description of studies
- Methodological quality of included studies
- Results
- Discussion
- Authors' conclusions
  - Implications for practice

Implications for research

Acknowledgements

Conflicts of interest

References:

References to studies

Included studies

Excluded studies

Studies awaiting assessment

Ongoing studies

Other references

Additional references

Other published versions of this review

Tables and figures:

Characteristics of included studies

Characteristics of excluded studies

Characteristics of ongoing studies

Comparisons, data and graphs

Additional tables

Additional figures

Comments and criticisms:

Title

Summary

Reply

Contributors

2.3 Logistics of doing a review

2.3.1 Motivation for undertaking a review

Preparation of a systematic review can be motivated by a number of factors. For example, reviews can be conducted in an effort to resolve conflicting evidence, to answer questions where the answer is uncertain, to explain variations in practice or simply to confirm the appropriateness of current practice. The primary aim of Cochrane reviews should be to summarise and help people to understand the evidence. Review authors must be careful not to impose their own values and preferences on others when addressing the questions they pose. They should help people make practical decisions about healthcare. This has important implications for deciding whether or not to undertake a Cochrane review, how to formulate the problem that a review will address, how to develop the protocol and how to present the results of the review.

- Questions should address the choices (practical options) people face when deciding about healthcare.
- Reviews should address outcomes that are meaningful to people making decisions about healthcare.
- Review authors should describe how they will address adverse effects as well as benefits.
- The methods used in a review should be selected to optimise the likelihood that the results will provide the best current evidence upon which to base decisions, and should be described in sufficient detail in the protocol for the readers to fully understand the planned steps.
- It is important to let people know when there is no reliable evidence, or no evidence about particular outcomes that are likely to be important to decision makers.
- It is not helpful to include evidence for which there is a high risk of bias in a review, even if there is no better evidence. (See Section 6 for a more detailed discussion of bias).
- Similarly, it is not helpful to focus on trivial outcomes simply because those are what researchers have chosen to measure in the individual studies.
- So far as is possible, it is important to take an international perspective. The evidence collected should not be restricted by nationality or language without good reason, background information such as prevalence and morbidity should where possible take a global view, and some attempt should be made to put the results of the review in a broad context.

2.3.2 Registering a protocol

The first step in the review process is to agree a review topic with the relevant Collaborative Review Group (CRG). A title will be registered, possibly after discussion among the CRG editors, and the review authors will be invited to submit a protocol. Once a protocol has been completed it will be sent to the CRG for editors and staff at the editorial base to consider. When they are satisfied with the protocol they will include it in the CRG's module for publication and dissemination in CDSR. Editors and authors should not include a protocol in a module unless there is a firm commitment to complete the review within a reasonable time frame and to keep it up-to-date once it is completed.

It is Collaboration policy that protocols that have not been converted into full reviews within two years should generally be withdrawn from the CDSR. If a protocol is withdrawn for any reason other than it being superseded by a review, a withdrawal notice should be published in CDSR for one issue. Thereafter, information on the withdrawal of the protocol should be noted in the CRG's module.

2.3.3 The review team

It is recommended that Cochrane reviews be undertaken by more than one person. This ensures that tasks such as selection of studies for inclusion and data extraction can be performed by at least two people independently, increasing the chance that errors are detected. If more than one group or individual expresses an interest in undertaking a review on the same topic, it is likely that a CRG will encourage them to work together.

Review teams must include expertise in the topic area being reviewed and expertise in systematic review methodology (including epidemiological and statistical expertise). First-time authors are encouraged to work with others who are experienced in the process of systematic reviews and to attend training events organised by the Collaboration (see Section 2.3.5 Training). The Collaboration is committed to user-involvement in principle (the tenth

principle of the Collaboration is enabling wide participation), and encourages review authors to seek and incorporate the views of users, including consumers, clinicians and those from varying regions and settings in the development of protocols and reviews. Where a review topic is of particular relevance in a region or setting (for example reviews of malaria in the developing world), involvement of people from that setting is encouraged.

2.3.3.1 Consumer involvement

The Cochrane Collaboration encourages healthcare consumer involvement, either as part of the review team or in the editorial process, in developing Cochrane reviews. Consumer involvement helps ensure that reviews:

- address problems that are important to people
- take account of outcomes that are important to those affected
- are accessible to people making decisions
- adequately reflect variability in the values and conditions of people, and the circumstances of healthcare in different countries

Relatively little is known about the effectiveness of various means of involving consumers in the review process or, more generally, in the spectrum of healthcare research. The Collaboration is dedicated to consumer involvement in principle. This is based on our values, good logic, and evidence that the views and perspectives of consumers often differ greatly from those of healthcare providers and researchers (Bastian 1998). Researchers and research funders have generally failed to ensure that healthcare research adequately meets the needs of those ultimately affected. Because of conflicting values and interests, it is unlikely that this situation will improve substantially without appropriate mechanisms for involving consumers in decisions about research. However, to ensure the effectiveness of consumer involvement, creativity and a critical approach must be used to develop and evaluate the mechanisms that are used.

This exploration of effective methods of involving consumers is being done in a variety of ways by CRGs, through the activities of the Cochrane Consumer Network and by other entities within the Collaboration. The Consumers and Communication CRG is currently reviewing evidence on the effects of consumer participation in systematic reviews, as well as in research more generally (a current protocol on The Cochrane Library, 'Interventions for promoting consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material' (Nilsen 2004) will inform this effort. This practical experience and formal evaluations will provide a basis for guidelines on how to ensure that consumer involvement effectively contributes to ensuring the quality and accessibility of Cochrane reviews.

Consumers are participating in the development of protocols and reviews in the following ways:

- helping CRGs to determine topics and issues for reviews and establish priority lists for reviews
- as co-authors
- as part of a consumer consultation during protocol and review development (including by questionnaire, direct dialogue or interview, in focus groups, and email discussion groups or teleconferences)
- as referees during the editorial process

Whenever consumers (or others) are consulted during the development of a protocol or review, their contribution should be acknowledged in the acknowledgement section of the protocol or review. Where input to the review is more substantive, formal inclusion in the list of review authors for citation may also be appropriate, as it is for other contributors.

Literature by consumers, or surveys and studies exploring consumers' views, can also be discussed within the review to ensure that issues of importance to consumers are addressed.

2.3.3.2 Advisory groups

Systematic reviews are likely to be more relevant to the end user and of higher quality if they are informed by advice from people with a range of experiences, in terms of both the topic and the methodology (Thomas 2004, NHS CRD 2001, Rees 2004). Decisions made in the early stages of the review process influence the content of the protocol and the subsequent review. As the priorities of decision-makers and consumers may be different from those of the review's authors, it is important that authors address the questions of importance to stakeholders and include relevant interventions, outcomes and populations. It may be useful to form an advisory group of people, including consumers, with relevant interests, skills and commitment. Their input will need to coordinated to inform key review decisions.

An advisory group might consist of four or more people, covering perspectives such as:

- practitioners
- potential recipients/consumers
- methodologists
- policy makers
- funders

The Effective Public Health Practice Project, Canada, has found that six members can cover all areas and is manageable for public health reviews.

The broader the review, the broader the experience required of advisory group members.

It is important to consider the needs of resource-poor countries in the review process. To increase the relevance of systematic reviews authors could also consult health professionals in developing countries to identify priority topics on which reviews should be conducted (Richards 2004). It may also be important to include vulnerable and marginalised people in the advisory group (Steel 2001) in order to ensure that the conclusions regarding the value of the interventions are well informed and applicable to all groups in society. A range of consultative processes is available that authors can use to engage the relevant groups, or to gain advice. A commitment from authors to involving a broad spectrum of lay and scientific advisors is evident in how they recruit members of their advisory group.

Terms of reference, job descriptions or person specifications for an advisory group may be developed to ensure there is clarity about the task(s) required. Examples are provided in briefing notes for researchers (Hanley 2000) or at the INVOLVE website (www.invo.org.uk). Advisory group members may be involved in one or more of the following tasks:

- making and refining decisions about the interventions of interest, the populations to be included, priorities for outcomes and, possibly, sub-group analyses
- providing or suggesting important background material that elucidates the issues from different perspectives
- helping to interpret the findings of the review
- designing a dissemination plan and assisting with dissemination to relevant groups

An example of the benefits of using an advisory group in the planning process

A review of HIV prevention for men who have sex with men (Rees, 2004) employed explicit consensus methods to shape the review with the help of practitioners, commissioners and researchers. An advisory group was convened of people from research/academic, policy and service organisations and representatives from charities and organisations that have emerged from and speak on behalf of people living with, or affected by, HIV/AIDS. The group met three times over the course of the review.

The group was presented with background information about the proposed review: its scope, conceptual basis, aims, research questions, stages and methods. Discussion focused on the policy relevance and political background/context to the review; the inclusion criteria for literature (interventions, outcomes, sub-groups of men); dissemination strategies; and timescales. Two rounds of voting identified and prioritised outcomes for analysis. Open discussion identified sub-groups of vulnerable men. A framework for characterising interventions of interest was refined through advisory group discussions.

The review followed this guidance by adopting the identified interventions, populations and outcomes to refine the inclusion criteria, performing a meta-analysis as well as sub-group analyses. The subsequent product included synthesised evidence directly related to health inequalities.

2.3.4 Software and the Information Management System

Since The Cochrane Collaboration was established in 1993, several tools and systems have been developed to help facilitate the electronic production and publication of Cochrane reviews and other material. These tools and systems make up the Cochrane Information Management System (IMS). A key component of the IMS is the review-authoring tool, Review Manager (RevMan), which is used by authors to prepare Cochrane protocols and reviews in the format described in Section 2.2. RevMan is currently distributed as copyrighted freeware and as such is available to all. However, technical support is only provided to authors who have registered their reviews with a CRG. RevMan will continue to be developed to support standards and guidelines for Cochrane reviews, and provide improved analytic methods, 'online' help and error checking mechanisms, as these evolve.

The ongoing development of the IMS is overseen by the Information Management Advisory Group with guidance from the relevant advisory groups.

More information about The Cochrane Collaboration’s software, such as the latest versions and planned developments, is available at the IMS website: www.cc-ims.net.

2.3.5 Training

It is important to ensure that those contributing to the work of the Collaboration have the knowledge, skills and support that they need to do a good job. Training may be needed by review authors, editors, criticism editors, referees, CRG Co-ordinators and Trials Search Co-ordinators, hand-searchers, trainers and users of Cochrane reviews. We focus here on the training needs of review authors and editors to help them to prepare and maintain high quality reviews.

While some review authors who join a CRG have training and experience in conducting a systematic review, many do not. Cochrane Centres are responsible for developing training materials and organising training workshops for members of CRGs. Each CRG is responsible for ensuring that the members of the group, including review authors, 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.

Training for review authors is delivered in many countries by Cochrane Centres and CRGs. Training timetables are listed on The Cochrane Collaboration’s training website (www.cochrane.org/resources/training.htm), along with various training resources, including The Cochrane Collaboration’s open learning material (see Section 2.6).

2.3.6 Editorial procedures of a Review Group

The editorial team of each CRG is responsible for maintaining a module, which includes information about the Group. Any specific methods used by the CRG, beyond the standard methods specified in the Handbook, should be documented in their module, including:

- methods used to review protocols
- any standard methodological criteria for including studies in reviews
- the search methods and specific search strategies used to develop and maintain the Specialised Register used by the CRG, and method of distributing potentially relevant citations or full-text reports to authors
- any additional search methods that authors are instructed to use routinely
- any standard methods used to select studies for reviews and any templates for inclusion assessment forms
- any standard criteria or methods used to assess the methodological quality of included studies
- any standard methods used for data collection and any templates for data extraction forms
- any standard methods used for synthesising 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 categorise interventions (see section 9.6)
- any specific rules used for preparing the standard tables and figures
- the methods used to keep reviews up-to-date and respond to criticisms

Descriptions of specific methods used by each CRG are published as part of the group's module in The Cochrane Library. Authors are recommended to familiarise themselves with the contents of their Group’s module.

2.3.7 Resources for a systematic review

Individual Cochrane reviews are prepared by authors working in CRGs. Each CRG has an editorial team responsible for producing a module of edited reviews for dissemination through the Cochrane Database of Systematic Reviews in The Cochrane Library.

Because The Cochrane Collaboration is built around CRGs, it is important that each author is linked with one from the beginning of the process. Besides ensuring that Cochrane reviews are carried out appropriately, this structure reduces the burden placed on individual authors 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 authors
- establishing specific standards and procedures for the CRG
- ensuring that authors receive the methodological support they need

The main resource required by authors is their own time. The majority of authors will contribute their time free of charge because it will be viewed as part of their existing efforts to keep up-to-date in their areas of interest. In some cases, authors 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 studies, the methods used (e.g., the extent of efforts to obtain unpublished information), the experience of the authors, 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 authors to estimate the amount of time that will be required. These tasks include:

- training
- meetings
- protocol development
- searching for studies
- assessing citations and full-text reports of studies for inclusion in the review
- assessing the quality of included studies and obtaining data
- pursuing missing data and unpublished studies
- analysing the data
- interpreting the results and preparing a report
- keeping the review up-to-date

Resources that might be required for these tasks, in addition to the authors' time, include:

- searching (identifying studies is primarily the responsibility of the editorial team of the CRG. However, authors may share this responsibility and it may be appropriate to search additional databases for a specific review.)
- help for library work, interlibrary loans and photocopying
- a second author, to assess studies for inclusion, assess the quality of included studies, obtain data and check analyses
- statistical support for synthesising (if appropriate) the results of the included studies
- equipment (e.g. computing hardware and software)
- supplies and services (long distance telephone charges, facsimiles, paper, printing, photocopying, audio-visual and computer supplies)
- office space for support staff
- travel funds

2.3.8 Seeking funding

Many organisations currently provide funding for systematic reviews and additional agencies are likely to recognise the importance of supporting this type of work in the future. These include research funding agencies, those organisations that provide or fund healthcare 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 organisation 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 studies
- Selection of studies for inclusion

- Assessments of the validity of included studies
- Obtaining data for the included studies
- Analysis
- Inferences and presentation of results
- Time-chart for major activities
- Budget

A time chart with target dates for accomplishing key tasks can help with scheduling the time needed to complete a review. Such targets may vary widely from review to review. Authors, together with the editorial team for the CRG, must determine an appropriate time frame for a specific review. An example of a time chart with target dates is:

Month

1 – 2 Preparation of protocol

3 – 8 Searches for published and unpublished studies

2 – 3 Pilot test of inclusion criteria

3 – 8 Inclusion assessments

3 Pilot test of validity criteria

3 – 10 Validity assessments

3 Pilot test of data collection

3 – 10 Data collection

3 – 10 Data entry

5 – 11 Missing information

8 – 10 Analysis

1 – 11 Preparation of report

12 - Keeping the review up-to-date

2.4 Publication of Cochrane reviews in print journals and books

Authors may wish to seek co-publication of Cochrane reviews in peer-reviewed healthcare journals, particularly in those journals that have expressed enthusiasm for co-publication of Cochrane reviews. For The Cochrane Collaboration, there is one essential condition of co-publication: Cochrane reviews must remain free for dissemination in any and all media, without restriction from any of them. To ensure this, Cochrane authors grant the Collaboration world-wide licences for these activities, and do not sign over exclusive copyright to any journal or other publisher. A journal is free to request a non-exclusive copyright that permits it to publish and re-publish a review, but this cannot restrict the publication of the review by The Cochrane Collaboration in whatever form the Collaboration feels appropriate. To republish material published in the Cochrane Database of Systematic Reviews elsewhere, most particularly in print journals, authors must complete a ‘permission to publish’ form available in the Cochrane Manual (www.cochrane.org/admin/manual.htm), along with an explanation of the procedures to follow.

Authors are strongly discouraged from publishing Cochrane reviews in journals before they are ready for publication in CDSR. This applies particularly to Centre directors and editors of CRGs. However, journals will sometimes insist that the publication of the review in CDSR should not precede publication in print. When this is the case, authors should submit a review

for publication in the journal after agreement from their CRG editor and before publication in CDSR. Publication in print should not be subject to lengthy production times, and authors should not unduly delay publication of a Cochrane review either because of delays from a journal or in order to resubmit their review to another journal.

Journals can also request revision of a review for editorial or content reasons. External peer review provided by journals may enhance the value of the review and should be welcomed.

Journals generally may require shorter reviews than those published in CDSR. Selective shortening of reviews may be appropriate, but there should not be any substantive differences between the review as published in the journal and CDSR. If a review is published in a journal, it should be noted that a fuller and maintained version of the review is available in CDSR. Typically, this should be done by including a statement such as the following in the introduction: 'A more detailed review will be published and updated in the Cochrane Database of Systematic Reviews' The reference should be to the protocol for the review published in CDSR. A similar statement should be included in the introduction if a review is published in CDSR prior to publishing a version of the review in a journal. After a version of a Cochrane review has been published in a journal, a reference to the journal publication must be added under the heading 'Other published versions of this review'. Authors are also encouraged to add the following statement to versions of Cochrane reviews that are published in journals:

'This paper is based on a Cochrane review first published [or most recently substantively amended, as appropriate] in The Cochrane Library YYYY, Issue X (see www.thecochranelibrary.com for information). Cochrane reviews are regularly updated as new evidence emerges and in response to comments and criticisms, and The Cochrane Library should be consulted for the most recent version of the review.'

The following modification of the disclaimer published in The Cochrane Library should be added to Cochrane reviews published in journals.

'The results of a Cochrane review can be interpreted differently, depending on people's perspectives and circumstances. Please consider the conclusions presented carefully. They are the opinions of review authors, and are not necessarily shared by The Cochrane Collaboration.'

The passage below can be provided to journal editors upon submission of a review for publication, and the letter of submission should be copied to the CRG editorial base for information. This policy and procedure may be new to some journal editors and may require direct discussion with the journal editor. The CRG editorial base should be informed of any problems encountered in this process. The following passage is suggested for inclusion in letters of submission to journal editors:

'This systematic review has been prepared under the aegis of The Cochrane Collaboration, an international organisation that aims to help people make well-informed decisions about healthcare by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions. The Collaboration's publication policy permits journals to publish reviews, with priority if required, but permits The Cochrane Collaboration also to publish and disseminate such reviews. Cochrane reviews cannot be subject to the exclusive copyright requested by some journals.'

2.5 Publication of previously published reviews as Cochrane reviews

Most reviews that have been conducted by authors outside of The Cochrane Collaboration (referred to as 'previously published reviews' here) require substantial additional work before they can be published as a Cochrane review in CDSR. In light of this additional work and

substantial differences from the previously published review, the Cochrane review can be considered a new publication. The previously published version of the review must be referenced in the Cochrane review under the heading 'Other published versions of this review'. However, it is generally not necessary to seek permission from the publisher of the previously published review.

Occasionally a Cochrane review will be similar enough to a previously published review that the only change is in the formatting of the review. In these cases authors should obtain permission from the publisher of the previously published review prior to publishing the review in CDSR. If authors are in doubt about whether they should request permission, they are encouraged to do so. This is unlikely to present a problem, provided it is done well in advance of the planned submission to CDSR. If it is known in advance that there is interest in publishing in CDSR a version of a review already published in a journal, authors should not assign exclusive copyright to the journal (see Section 2.4). The Cochrane Collaboration does not require exclusive copyright. It is therefore not a problem to publish a version of a Cochrane review in a journal after it has been published in CDSR, provided it is not called a Cochrane review and that it is acknowledged that it is based on a Cochrane review (see Section 2.4).

The conversion of individual patient data reviews into Cochrane reviews is discussed in section 11.6.

2.6 Conflict of interest and commercial sponsorship

Cochrane reviews should be free of any real or perceived bias introduced by the receipt of any benefit in cash or kind, any hospitality, or any subsidy derived from any source that may have or be perceived to have an interest in the outcome of the review. There should be a clear barrier between the production of Cochrane reviews and any funding from commercial sources with financial interests in the conclusions of Cochrane reviews. Thus, sponsorship of a Cochrane review by any commercial source or sources (as defined above) is prohibited. Other sponsorship is allowed, but a sponsor should not be allowed to delay or prevent publication of a Cochrane review and a sponsor should not be able to interfere with the independence of the authors of reviews in regard to the conduct of their reviews. The protocol for a Cochrane review should specifically mention that a sponsor cannot prevent certain outcome measures being assessed in the review.

These rules also apply to ‘derivative products’ (containing Cochrane reviews) so that commercial sponsors cannot prevent or influence what would be included in such products. Receipt of benefits from any source of sponsored research must be acknowledged and conflicts of interest must be disclosed in CDSR and other publications that emanate from the Collaboration.

The Cochrane Collaboration code of conduct for avoiding potential financial conflicts of interest appears in Box 2.6. If a proposal for undertaking a review raises a question of serious conflict of interest, this should be forwarded to the Collaboration’s funding arbiter (fundingarbiter@cochrane.org) for review. It is not mandatory to send funding proposals to the local Cochrane Centre or Steering Group prior to accepting them. However, this would be desirable in the cases of restricted donations, or any donation that appears to conflict with the general principle noted above.

It is impossible to abolish conflict of interest, since the only person who does not have some vested interest in a subject is somebody who knows nothing about it (Smith 1994). Financial conflicts of interest cause the most concern, can and should be avoided, but must be disclosed if there are any. Any secondary interest (such as personal conflicts) that might unduly influence judgements made in a review (concerning, for example, the inclusion or exclusion of studies, assessments of the validity of included studies or the interpretation of results) should be disclosed.

Disclosing a conflict of interest does not necessarily reduce the worth of a review and it does not imply dishonesty. However, conflicts of interest can influence judgements in subtle ways. Authors should let the editors of their Collaborative Review Group know of potential conflicts even when they are confident that their judgements were not or will not be influenced. Editors may decide that disclosure is not warranted or they may decide that readers should know about such a conflict of interest so that they can make up their own minds about how important it is. Decisions about whether or not to publish such information should be made jointly by authors and editors.

To help ensure the integrity and perceived integrity of Cochrane reviews, all authors must sign the relevant statements in the form giving the Cochrane Collaboration permission to publish their review in addition to disclosing conflicts of interest, and the editorial team of each Collaborative Review Group (CRG) must also disclose any potential conflict of interest that they might have, both on their module and within relevant reviews.

Box 2.6 The Cochrane Collaboration Code of Conduct for Avoiding Potential Financial Conflicts of Interest

General Principle

The essential activity of The Cochrane Collaboration is co-ordinating the preparation and maintenance of systematic reviews of the effects of health care interventions performed by individual reviewers/authors according to procedures specified by The Cochrane Collaboration. The performance of the review must be free of any real or perceived bias introduced by receipt of any benefit in cash or kind, any hospitality, or any subsidy derived from any source that may have or be perceived to have an interest in the outcome of the review. All entities that constitute The Cochrane Collaboration must accept this General Principle as a condition of participation in the organisation.

Policy

(i) Receipt of benefits from any source of sponsored research must be acknowledged and conflicts of interest must be disclosed in the Cochrane Database of Systematic Reviews and other publications that emanate from The Cochrane Collaboration.

(ii) If a reviewer/author is involved in a trial included in his/her review, this must be acknowledged, as it could be perceived as a potential conflict of interest.

(iii) If a proposal raises a question of serious conflict of interest, this should be forwarded to the local Cochrane Centre for review (and the Steering Group notified accordingly). If the issue involves a Cochrane Centre, the issue should be referred to the Steering Group.

(iv) It is not mandatory to send funding proposals to the local Cochrane Centre or Steering Group prior to accepting them. However, such reviews would be desirable in cases of restricted donations, or any donation that appears to conflict with the General Principle.

(v) The Steering Group should receive (and review at least annually) information about all external funds accepted by Cochrane entities. The Steering Group will use this information to prepare and distribute an annual report on the potential conflicts of interest attendant on The Cochrane Collaboration's solicitation and use of external funds.

(vi) The Steering Group is considering constituting an Ethics Sub-Group to view potential conflicts of interest, to offer recommendations for their resolution, and to consider appropriate sanctions to redress violations of the General Principle.

2.7 The Cochrane Collaboration Open Learning Material

In 2002, *The Cochrane Collaboration Open Learning Material for Reviewers* was prepared to accompany the Handbook in helping people who are working on a Cochrane review. It does not replace the Handbook, instead it provides a framework for progressing through the Handbook, supplementing it with examples and activities along the way. The first version of the open learning material (Version 1.1) was made available on the internet in November 2002. It can be accessed at http://www.cochrane-net.org/openlearning/.

Along with the Handbook, this material will stand alone, offering an alternative to face-to-face training, especially for those authors living and working away from easy access to the training offered by Cochrane Centres and Collaborative Review Groups. For those able to access this face-to-face training, this material will serve as a useful resource to remind them of what they learned.

The open learning material takes a step-by-step approach to Cochrane reviews, exploring each step individually, signposting appropriate links and references and providing examples and activities to help make sense of the information. The material is organised in modules, with modules relating to consecutive sections of a review. There are also some additional modules relating to issues of reviewing that do not occur in all Cochrane reviews.

2.8 Contributions

This section builds on earlier versions of the Handbook. For details of previous authors and editors of the Handbook, please refer to the Acknowledgements section.

Contributing authors (March 2005): Ginny Brunton, Sally Green, Julian Higgins, Monica Kjeldstrøm, Nicki Jackson, Sandy Oliver

Comments on drafts (March 2005): Chris Cates, Carol Lefebvre, Philippa Middleton, Lesley Stewart

Editors: Julian Higgins and Sally Green

2.9 References

Bastian 1998. Bastian H. Speaking up for ourselves: the evolution of consumer advocacy in health care. International Journal of Technology Assessment in Health Care 1998;14:3-23.

Hanley 2000. Hanley B, Bradburn J, Gorin S et al. Involving Consumers in Research and Development in the NHS: briefing notes for researchers. Winchester: Help for Health Trust, 2000. Available at

http://www.hfht.org/ConsumersinNHSResearch/pdf/involving_consumers_in_rd.pdf.

Light 1984b. Light RJ, Pillemer DB. Organizing a reviewing strategy. In: Summing Up: The Science of Reviewing Research. Cambridge, Massachusetts: Harvard University Press, 1984; 13-31.

NHS CRD 2001. Centre for Reviews and Dissemination (Undertaking Systematic Reviews of Research on Effectiveness. CRD's Guidance for those Carrying Out or Commissioning Reviews. CRD Report Number 4 (2nd Edition) March 2001), at http://www.york.ac.uk/inst/crd/report4.htm

Nilsen 2004. Nilsen ES, Myrhaug HT, Johansen M, Oliver S, Oxman AD. Interventions for promoting consumer involvement in developing healthcare policy and research, clinical practice guidelines and patient information material. The Cochrane Database of Systematic Reviews 2004, Issue 1. Art. No.: CD004563. DOI: 10.1002/14651858.CD004563.

Rees 2004. Rees R, Kavanagh J, Burchett H, Shepherd J, Brunton G, Harden A, Thomas J, Oliver S, Oakley A (2004) HIV Health Promotion and Men who have Sex with Men (MSM): A systematic review of research relevant to the development and implementation of effective and appropriate interventions. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Richards 2004. Richards T. Poor countries lack relevant health information, says Cochrane editor. BMJ 2004; 328:310.

Smith 1994. Smith R. Conflict of interest and the BMJ. BMJ 1994; 308:4-5

Steel 2001. Steel R. Involving marginalised and vulnerable groups in research: a discussion document. Consumers in NHS research. 2001.

http://www.invo.org.uk/pdf/Involving_Marginalised_Groups_in_Research.pdf

Thomas 2004. Thomas BH, Ciliska D, Dobbins M, Micucci S. A Process for Systematically Reviewing the Literature: Providing the Research Evidence for Public Health Nursing Interventions. Worldviews on Evidenced-Based Nursing 2004;1(3):176-184(9).

3 Guide to the contents of a protocol and review

Edited by Julian PT Higgins and Sally Green

3.1 Cover sheet

The cover sheet includes the following information. Note that the dates fields are not all published in the *Cochrane Database of Systematic Reviews (CDSR)*. They should all be completed by the author (reviewer) or Collaborative Review Group (CRG) in RevMan, although the list of fields is currently under review.

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 reading the title on its own should be able to decide quickly whether the review addresses a question of interest. At its most basic, a title should take the structure 'Intervention for condition'. Other structures are included in the Style Guidelines for Cochrane reviews (http://www.liv.ac.uk/lstm/ehcap/CSR/home.html). Specific outcomes should be mentioned only rarely within the title. If so, this should usually be done as a subtitle separated by a colon from the main title.

Version: One version of each review must be marked as the primary version and this is the one that should be submitted for publication in the *CDSR*.

Status: This specifies what stage the review is at: title, protocol or full review. Titles are only used internally, within Collaborative Review Groups, and are not included in the *CDSR*.

Date edited: This date is updated automatically any time the review is amended.

Date of last substantive update: See under list of dates, below.

Date next stage expected: This must be completed in RevMan for protocols so that users of the *CDSR* can be informed when they can expect the completed review to be available. It can also be completed for full reviews to inform users of the *CDSR* when an updated review is likely to be available.

Contact author: This should provide the contact details for the person to whom correspondence about the review should be addressed, and who has agreed to take responsibility for maintaining and developing the review. This usually is the person who takes responsibility for developing and organizing the review team, communicates with the editorial base, ensures that the review is prepared within agreed timescales, submits it to the editorial base, communicates feedback to co-authors and ensures that the updates are prepared.

The contact author need not be the first listed author, and the choice of contact author will not affect the citation for the review. If the contact author no longer wishes to be responsible for a published review and another member of the review team does not wish to take responsibility

for it, then the Review Group Co-ordinator (RGC) should be listed as the contact author, and the former contact author listed as a co-author. The RGC need not be listed as a co-author.

Co-authors: This should be a list of co-authors on the review. Authorship of all scientific papers (including Cochrane protocols and reviews) establishes accountability, responsibility and credit (Rennie 1997, Flanagin 1998, Rennie 1998). When deciding who should go in the byline for Cochrane reviews, it is important to distinguish individuals who have made a substantial contribution to the review (and who should be listed) and those who have made other contributions, which should be noted in the Acknowledgements section. Authorship should be based on substantial contributions to all of the following three steps, based on ICMJE 1997:

- conception and design of study, or analysis and interpretation of data
- drafting the review or revising it critically for important intellectual content
- final approval of the version to be published.

Brief contact details of co-authors may be published within the completed protocol or review, so authors should ensure that these fields are completed and up-to-date in RevMan. The fields that must be completed are the First name(s) and Last name of the co-author, Organisation and Country. If a co-author does not have a publishable address, but should still appear in the byline for the citation, then the Organisation and Country should be those of the Review Group (for example, 'Smith J. c/o Cochrane Pregnancy and Childbirth Group, UK').

Contributions: The names and contribution of the present co-authors should be described in this section. One author, usually the contact author, should be identified as the guarantor of the review. All authors should discuss and agree on their respective descriptions of contribution before the review is submitted for publication on the *CDSR*. When the review is updated, this section should be checked and revised as necessary to ensure that it is accurate and up-to-date.

The following potential contributions have been adapted from Yank 1999. This a suggested scheme and the section should describe what people did, rather than attempt to identify which of these categories someone's contribution falls within. Ideally, the contributors should describe their contribution in their own words:

Conceiving the review

Designing the review

Coordinating the review

Data collection for the review

Designing search strategies

Undertaking searches

Screening search results

Organising retrieval of papers

Screening retrieved papers against inclusion criteria

Appraising quality of papers

Extracting data from papers

Writing to authors of papers for additional information

Providing additional data about papers

Obtaining and screening data on unpublished studies

Data management for the review

Entering data into RevMan

Analysis of data

Interpretation of data

Providing a methodological perspective

Providing a clinical perspective

Providing a policy perspective

Providing a consumer perspective

Writing the review

Providing general advice on the review

Securing funding for the review

Performing previous work that was the foundation of the current study

List of authors for citation: This will be used to generate the by-line for the published review. A strict format is necessary in order for a computer script to provide the correct components of names. The format is Last-name Initial(s), without personal title (such as Dr) or internal punctuation but with a comma between names (for example, 'Jepson RG, Mihaljevic L, Craig JC'). Multiple initials should not be separated by a space. Surname prefixes should precede the surname and surname suffixes follow it (for example, 'Hayden JA, van Tulder MW, Malmivaara Jr A'). It is preferable to use the abbreviated forms of surname suffixes, for example 'Jr' not 'Junior'. Hyphens are permitted in surname and initials, so Marie-Claire Gene Lautrec would become 'Lautrec M-CG' and Deborah Pentesco-Gilbert would become 'Pentesco-Gilbert D'.

The list of authors for citations can be the name of an individual, several individuals, a collaborative group (for example, 'Early Breast Cancer Trialists' Collaborative Group') or a combination of one or more authors and a collaborative group. Where group authors are included in list they should be separated by a comma from other authors ('Jones BA, Smith PJ, Early Breast Cancer Trialists' Collaborative Group') and extraneous text such as 'on behalf of' should not be included. Ideally, the order of authors should relate to their relative contributions to the review. The person who contributed most should be listed first.

Sources of support to the review: Authors 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 are divided into 'internal' (provided by the institutions at which the review was produced) and 'external' (provided by other institutions or funding agencies).

What's new: This should describe the changes to the protocol or review since it was last published in the *CDSR*. At each update of a review, substantive or not, the 'What's new' field should contain the calendar date of the change and a description of what was changed. This might be, for example, a brief summary of how much new information has been added to the review (for example, number of studies, participants or extra analyses) and any important changes to the conclusions, results or methods of the review.

Issue protocol first published: The issue of *The Cochrane Library* where the protocol was first published (for example, Issue 2, 2004).

Issue review first published: The issue of *The Cochrane Library* where the full review was first published (for example, Issue 1, 2005).

Date of last substantive update: The author(s) and/or editors of a CRG should decide whether an amendment is substantive or not. Substantive amendments are ones that are sufficient to recommend that previous readers of the review should look at the updated version. For example, important changes in the conclusions of the review or the list of studies that are included or excluded may qualify as substantive amendments. New protocols, reviews and substantive updates should have a date that is within the three months leading up to the submission deadline date for inclusion of the review in the *CDSR*.

Date of last minor update: The most recent date on which the review was updated, but this update is not sufficient to recommend that previous readers of the review should look at the new version (in which case the review should be classified as a substantive update).

Date review re-formatted: The most recent date on which structural changes were made to the review (for example, the addition of a new fixed heading), usually because of a new version of RevMan. This field is not applicable to most reviews.

Date new studies sought but none found: The most recent date on which a search was done for new studies but none were found.

Date new studies found but not yet included or excluded: The most recent date on which a search was done for new studies and some were found and added to the list of studies awaiting assessment or ongoing studies.

Date new studies found and included or excluded: The most recent date on which a search was done for new studies and some were found and added to the list of included or excluded studies.

Date authors' conclusions section amended: The most recent date on which the Authors' Conclusions section was amended in such a way that it is recommended that previous readers of the review should look at the new version. Details of the change should be reported under 'What's new', above.

Date comment / criticism added: The most recent date on which a comment or criticism was added to the review.

Date response to comment / criticism added: The most recent date on which a reply to a comment or criticism was added to the review.

Unpublished CRG notes: These will not be published in the *CDSR* but can be used as a space for temporary notes.

Published notes: These will be published in the *CDSR*. They may include

- editorial notes and comments from the CRG, for example where issues highlighted by editors or referees are believed worthy of publication alongside the review;
- a summary of previous changes to the review. Changes since the previous published version must be stated under 'What's new'.

The published notes must be completed for all withdrawn publications to give the reason for withdrawal. Only the cover sheet and published notes are published for withdrawn protocols and reviews.

Amended sections: These boxes can be checked to make it easier for co-authors or the CRG's editorial team to locate changes in the review. This information is not published in the *CDSR*.

3.2 Plain language summary

The plain language summary (formerly called the 'synopsis') aims to summarise the review in an easily understood style which would be understandable by consumers of healthcare. Plain language summaries are made freely accessible on the internet, so will often be read as stand-alone documents. Plain language summaries have two parts. The first part is a restatement of the review's title using plain language terms. This does not need to be declarative but does need to include participants, intervention and outcome when included in the title of the review. The heading should be no more than 256 characters in length, should be written in sentence case (ie with a capital at the beginning of the title and for names, but the remainder in lower case- see example plain language summary), but should not end with a full stop. The title of the plain language summary should, where the review title is easily understood, simply restate the review's title.

The second part or body of the summary should be no more than 400 words in length and should include:

- A statement about why the review is important: for example definition of and background to the health care problem, signs and symptoms, prevalence, description of the intervention and the rationale for its use.
- The main findings of the review: this could include numerical summaries when the review has reported results in numerical form, but these should be given in general and easily understood forms. Results in the plain language summary should not be presented any differently from in the review (ie no new results should appear in the summary.Where possible an indication of the number of trials and participants on which the findings are based should be stated.
- A comment on any adverse effects.
- A brief comment on any limitations of the review (for example trials in very specific populations or poor methods of included trials).

At the end of the plain language summary authors may give web links (for example to other information or decision aids on CRG websites, providing that these comply with the Cochrane Collaboration policy on web links. There should not be graphs or pictures in the plain language summary. As with other components of a Cochrane review, plain language summaries should follow the format of the Cochrane Style Guide.

3.2.1 Process of finalising a plain language summary

The first draft of the plain language summary should usually be written by the review authors and submitted with the review to the relevant CRG. This draft may be subject to alteration,

and authors should anticipate one or more iterations. Many CRGs have plain language summary writing skills within their editorial team. Where this is not available, a central support service is available to assist CRGs in their writing and editing . This service is co-ordinated by the Cochrane Consumer Network (ccnet-contact@cochrane.de), but should be accessed through the CRG (i.e. review authors needing assistance with writing a plain language summary should contact their CRG).

Further information on the process of finalising plain language summaries is available in the Cochrane Manual.

3.3 Abstract

All full reviews must include an abstract of not more than 400 words. It should be kept as brief as possible without sacrificing important content. Abstracts to Cochrane reviews are published on MEDLINE and made freely accessible on the internet, so will often be read as stand-alone documents. They should, therefore, summarise the key methods and content of the review and not contain any material that is not in the review. The content must be consistent with the text, data and conclusions of the review and not include references to any information outside the review. Links to other parts of the review (such as references, studies, additional tables and additional figures) may not be inserted in the abstract. A hypothetical example is included in Box 3.3.

Abstracts should be made as readable as possible without compromising scientific integrity. They should primarily be targeted to healthcare decision makers (clinicians, consumers and policy makers) rather than just researchers. Terminology should be reasonably comprehensible to a general rather than a specialist heathcare audience. Abbreviations should be avoided, except where they are widely understood (for example, HIV). Where essential, other abbreviations should be spelt out (with the abbreviations in brackets) on first use. Names of drugs and interventions that can be understood internationally should be used wherever possible.

The content under each heading in the abstract should be as follows:

Background: This should be one or two sentences to explain the context or elaborate on the purpose and rationale of the review.

Objectives: This should be a precise statement of the primary objective of the review, ideally in a single sentence. Where possible the style should be of the form ‘To assess the effects of *[intervention or comparison] for [health problem] for/in [types of people, disease or problem and setting if specified]’.*

Search strategy: This should list the sources and the dates of the last search, for each source, using the active form ‘We searched….’ or, if there is only one author, the passive form can be used, for example, ‘Database X, Y, Z were searched’. Search terms should not be listed here. If the CRG’s Specialised Register was used, this should be listed first in the form ‘Cochrane X Group Specialised Register’. The order for listing other databases should be the Cochrane Central Register of Controlled Trials, MEDLINE, EMBASE, other databases. The date range of the search for each database should be given. For the Cochrane Central Register of Controlled Trials this should be in the form ‘Cochrane Central Register of Controlled Trials (*The Cochrane Library* 2005, *Issue* 1)’. For most other databases such as MEDLINE, it should be in the form ‘MEDLINE (January 1966 to December 2004)’. Searching of bibliographies for relevant citations can be covered in a generic phrase ‘reference lists of

articles'. If there were any constraints based on language or publication status, these should be listed. If individuals or organisations were contacted to locate studies this should be noted and it is preferable to use 'We contacted pharmaceutical companies' rather than a listing of all the pharmaceutical companies contacted. If journals were specifically handsearched for the review, this should be noted but handsearching to help build the Specialised Register of the CRG should not be listed.

Selection criteria: These should be given as '*[type of study]* of *[type of intervention or comparison]* in *[disease, problem or type of people]*'. Outcomes should only be listed here if the review was restricted to specific outcomes.

Data collection and analysis: This should be restricted to how data were extracted and assessed, and not include details of what data were extracted. This section should cover whether extraction and quality assessment of studies were done by more than one person. If the authors contacted investigators to obtain missing information, this should be noted here. What steps, if any, were taken to identify adverse effects should be noted.

Main results: This section should begin with the total number of trials and participants included in the review, and brief details pertinent to the interpretation of the results (for example, the quality of the studies overall or a comment on the comparability of the studies, if appropriate). It should address the primary objective and be restricted to the main qualitative and quantitative results (generally including not more than six key results). The outcomes included should be selected on the basis of which are most likely to help someone making a decision about whether or not to use a particular intervention. Adverse effects should be included if these are covered in the review. If necessary, the number of studies and participants contributing to the separate outcomes should be noted, along with concerns over quality of evidence specific to these outcomes. The results should be expressed narratively as well as quantitatively if the numerical results are not clear or intuitive (such as those from a standardised mean differences analysis). The summary statistics in the abstract should be the same as those selected as the defaults for the review, and should be presented in a standard way, such as 'odds ratio 2.31 (95% confidence interval 1.13 to 3.45)'. Ideally, risks of events (percentage) or averages (for continuous data) should be reported for both comparison groups. If overall results are not calculated in the review, a qualitative assessment or a description of the range and pattern of the results can be given. However, 'vote counts' in which the numbers of 'positive' and 'negative' studies are reported should be avoided.

Authors' conclusions: The primary purpose of the review should be to present information, rather than to offer advice. The Authors' conclusions should be succinct and drawn directly from the findings of the review so that they directly and obviously reflect the main results. Assumptions should not be made about practice circumstances, values, preferences, tradeoffs; and the giving of advice or recommendations should generally be avoided. Any important limitations of data and analyses should be noted. Important conclusions about the implications for research should be included if these are not obvious.

Box 3.3 Hypothetical example of an abstract

Almonds and raisins in the treatment of influenza in adults

Peach A, Apricot D, Plum P

Background

Almonds and raisins both have antiviral properties, but they are not widely used due to incomplete knowledge of their properties and concerns about possible adverse effects.

Objectives

To assess the effects of almonds and raisins in adults with influenza.

Search strategy

We searched the Cochrane Acute Respiratory Infections Group trials Specialised Register (15 February 2005), the Cochrane Central Register of Controlled Trials (The Cochrane Library Issue 1, 2005), MEDLINE (January 1966 to January 2005), EMBASE (January 1985 to December 2002) and reference lists of articles. We also contacted manufacturers and researchers in the field.

Selection criteria

Randomised and quasi-randomised studies comparing almonds and/or raisins with placebo, or comparing doses or schedules of almonds and /or raisins in adults with influenza.

Data collection

Two authors independently assessed trial quality and extracted data. We contacted study authors for additional information. We collected adverse effects information from the trials.

Main results

Seventeen trials involving 689 people were included. Five trials involving 234 people compared almonds with placebo. Compared to placebo, almonds significantly shortened duration of fever by 23% (by 1.00 days, 95% confidence interval 0.73 to 1.29). Six trials involving 256 people compared raisins with placebo. Raisins significantly shortened duration of fever by 33% compared to placebo (by 1.27 days, 95% confidence interval 0.77 to 1.77). The small amount of information available directly comparing almonds and raisins (two trials involving 53 people) indicated that the efficacy of the two drugs was comparable, although the confidence intervals were very wide. Based on four trials of 73 people, central nervous system effects were significantly more common with almonds than raisins (relative risk 2.58, 95% confidence interval 1.54 to 4.33).

Authors' conclusions

Almonds and raisins appear to be equally effective in the treatment of influenza. Both drugs appear to be relatively well tolerated, although raisins may be safer.

3.4 Text of a review

The text of the review should be as succinct and readable as possible. Although there is no formal word limit on Cochrane reviews, review authors should consider 10,000 words an absolute maximum unless there is special reason to write a longer review. The majority of

reviews should be substantially shorter than this. A review should be written so that someone who is not an expert in the area can understand it, in light of the following policy statement, reported in Cochrane News 1999; 15: 14):

"The target audience for Cochrane reviews is people making decisions about healthcare. This includes healthcare professionals, consumers and policy makers with a basic understanding of the underlying disease or problem.

It is a part of the mission and a basic principle of The Cochrane Collaboration to promote the accessibility of systematic reviews of the effects of healthcare interventions to anyone wanting to make a decision about healthcare. However, this does not mean that Cochrane reviews must be understandable to anyone, regardless of their background. This is not possible, any more than it would be possible for Cochrane reviews to be written in a single language that is understandable to everyone in the world. It is important to translate the content, or elements of the content, of reviews into different languages and formats targeted at different audiences including healthcare professionals, consumers and policy makers in a variety of circumstances.

Cochrane reviews should be written so that they are easy to read and understand by someone with a basic sense of the topic who may not necessarily be an expert in the area. Some explanation of terms and concepts is likely to be helpful, and perhaps even essential. However, too much explanation can detract from the readability of a review. Simplicity and clarity are also vital to readability.

The readability of Cochrane reviews should be comparable to that of a well-written article in a general medical journal."

The text of a Cochrane review contains a number of fixed headings that are embedded in RevMan. Subheadings may be added by the author at any point. Certain specific headings are recommended for use by all authors, but are not mandatory and should be avoided if they make individual sections needlessly short. Wording for further subheadings that may or may not be relevant to a particular review is also provided. In the rest of this section, the relevant category from these (fixed, recommended, optional) is noted for each of the headings described.

Background [fixed, level 1 heading]

Well-formulated review questions usually do not appear out of thin air. They occur in the context of an already formed body of knowledge. This context should be addressed in the background section of the review. This background helps set the rationale for the review, and should explain why the questions being asked are important. It should be presented in a fashion that is understandable to the users of the health care under investigation, and should be concise (generally around one page when printed).

Description of the condition [recommended, level 2 heading]

The review should begin with a brief description of the condition being addressed and its significance. It may include information about the biology, diagnosis, prognosis and public health importance (including prevalence or incidence).

Description of the intervention [recommended, level 2 heading]

A description of the experimental intervention(s) should place it in the context of any standard, or alternative interventions. It should be made clear what role the comparator intervention(s) have in standard practice.

How the intervention might work [recommended, level 2 heading]

Systematic reviews gather evidence to assess whether the expected effect of an intervention does indeed occur. This section might describe the theoretical reasoning why the interventions under review might have an impact on potential recipients, for example, by relating a drug intervention to the biology of the condition. Authors may refer to a body of empirical evidence such as similar interventions having an impact, or identical interventions having an impact on other populations. Authors may also refer to a body of literature that justifies the possibility of effectiveness.

Although every review, just like every intervention, is based on a theory, this may not be explicit or well explored. Controversy remains about whether or not theory makes a difference to intervention effectiveness, but as Oakley (1999) points out "the importance or unimportance of theory is unlikely to emerge unless review activity is structured to cross problem/outcome areas, and allow for the classification of interventions according to their theoretical base."

Why it is important to do this review [recommended, level 2 heading]

The background helps set the rationale for the review, and should explain why the questions being asked are important. It might also mention why this review was undertaken and how it might relate to a wider review of a general problem.

Objectives [fixed, level 1 heading]

This should begin with a precise statement of the primary aim of the review, including the intervention(s) reviewed and the targeted problem. This might be followed by a series of specific objectives relating to different participant groups, different comparisons of interventions or different outcome measures.

Methods sections

The Methods section in a protocol should be written in the future tense. Because Cochrane reviews are updated as new evidence accumulates, methods outlined in the protocol should generally anticipate a sufficiently large number of studies to address the review's objectives (even if it is known this is not the case).

The Methods section in a review should be written in the past tense, and should describe what was done to obtain the results and conclusions of the current version of the review. Often a review is unable to implement all of the methods outlined in the protocol, usually because there is insufficient evidence. In such circumstances, it is recommended that the methods that were not implemented still be outlined in the review, so that it serves as a protocol for future updates of the review. Some CRGs have policies on this issue, and these should be available from the Review Group Co-ordinator. Examples include adding an additional subsection at the end of 'Methods of the review', or including the methods for future updates in an additional table.

Criteria for considering studies for this review [fixed, level 1 heading]

The criteria used to select studies for inclusion in the review must be clearly stated.

Types of studies [fixed, level 2 heading]

Eligible study designs should be stated here, along with any thresholds for inclusion based on the conduct or quality of the studies. For example, 'All randomised controlled comparisons'

or ‘All randomised controlled trials with blind assessment of outcome’. Exclusion of particular types of randomised studies (for example, cross-over trials) should be justified.

Types of participants [fixed, level 2 heading]

The diseases or conditions of interest should be described here, including any restrictions on diagnoses, age groups and settings. Subgroup analyses should not be listed here.

Types of interventions [fixed, level 2 heading]

Experimental and control interventions should be defined here, making it clear which comparisons are of interest. Restrictions on dose, frequency, intensity or duration should be stated. Subgroup analyses should not be listed here.

Types of outcome measures [fixed, level 2 heading]

Note that outcome measures do not always form part of the criteria for including studies in a review. If they do not, then this should be made clear. Outcome measures of interest should be listed in this section whether or not they form part of the inclusion criteria.

Primary outcomes [recommended, level 3 heading]

Primary outcomes should normally reflect at least one potential benefit and at least one potential area of harm, and should be as few as possible.

Secondary outcomes [recommended, level 3 heading]

Non-primary outcomes should be listed here.

It may be helpful to use the following optional (level 3) headings:

Adverse outcomes

Economic data

Timing of outcome assessment

Search strategy for identification of studies [fixed, level 1 heading]

The data sources used to identify studies should be summarised. The following headings are recommended. Further details of the contents of these sections are discussed in Section 5.2.2 Documenting a search strategy. Some CRGs have a standard paragraph they ask their authors to use which refers to the Group’s generic searching activities as detailed in the editorial information for the CRG. Before starting to develop this section, authors should contact their CRG for guidance.

Electronic searches [recommended, level 2 heading]

The bibliographic databases searched, the dates and periods searched and any constraints, such as language should be stated. The full search strategies for each database should be listed here or in an Additional table. If a CRG has developed a Specialised Register of studies and this is searched for the review, a standard description of this register can be referred to but information should be included on when and how the Specialised Register was most recently searched for the current version of the review and the search terms used should be listed.

Other sources [recommended, level 2 heading]

List grey literature sources, such as reports and conference proceedings. If journals are specifically handsearched for the review, this should be noted but handsearching done by the authors to help build the Specialised Register of the CRG should not be listed. List people (for example, trialists, experts) and/or organisations that were contacted. List any other sources, which may include, for example, reference lists, the World Wide Web or personal collections of articles.

The following *optional* headings may be used, either in place of 'Other sources' (level 2) or as subheadings (level 3).

Grey literature

Handsearching

Reference lists

Correspondence

Methods of the review [fixed, level 1 heading]

This should describe the methods for data collection and analysis. In the future this will be renamed 'Data collection and analysis'.

Selection of studies [recommended, level 2 heading]

The method used to apply the selection criteria. Whether they are applied independently by more than one author should be stated, along with how any disagreements are resolved.

Data extraction and management [recommended, level 2 heading]

The method used to extract or obtain data from published reports or from the trialists (for example, using a data extraction/data collection form). Whether data are extracted independently by more than one author should be stated, along with how any disagreements are resolved. If relevant, methods for processing data in preparation for analysis should be described.

Assessment of methodological quality of included studies [recommended, level 2 heading]

The method used to assess methodological quality. Whether methods are applied independently by more than one author should be stated, along with how any disagreements are resolved. The tool(s) used should be described or referenced, with an indication of how the results are incorporated into the interpretation of the results.

Measures of treatment effect [recommended, level 2 heading]

The effect measures of choice should be stated. For example, odds ratio (OR), risk ratio (RR) or risk difference (RD) for dichotomous data; difference in means (MD) or standardised difference in means (SMD) for continuous data. Alternatively, *optional* headings specific to the type of data may be used, such as:

Dichotomous data

Continuous data

Time-to-event data

Unit of analysis issues [recommended, level 2 heading]

Special issues in the analysis of studies with non-standard designs, such as cross-over trials, cluster-randomised trials and non-randomised studies, should be addressed (see Section 8.3. Study designs and identifying the unit of analysis). Alternatively, *optional* (level 2) headings specific to the types of studies may be used, such as:

Studies with multiple treatment groups

Cross-over trials

Cluster randomised trials

Dealing with missing data [recommended, level 2 heading]

Strategies for dealing with missing data should be described. This will principally include missing participants due to drop-out (whether an intention-to-treat analysis will be conducted), and missing statistics (such as standard deviations or correlation coefficients).

Assessment of heterogeneity [recommended, level 2 heading]

Approaches to addressing clinical heterogeneity should be described, along with how the authors will determine whether a meta-analysis is considered appropriate. Methods for identifying statistical heterogeneity should be stated (for example, visually, using a chi-squared test, or using I). See Section 8.7 Heterogeneity.

Assessment of reporting biases [recommended, level 2 heading]

How publication bias, and other reporting biases are addressed (for example, funnel plots, statistical tests, imputation). Authors should remember that asymmetric funnel plots are not necessarily caused by publication bias (and that publication bias does not necessarily cause asymmetry in a funnel plot). See Section 8.11.1 Publication bias and funnel plots.

Data synthesis (meta-analysis) [recommended, level 2 heading]

The choice of meta-analysis method should be stated, including whether a fixed effect or a random effects model is used. If meta-analyses are not undertaken, systematic approaches to synthesising the findings of multiple studies should be described.

Subgroup analysis and investigation of heterogeneity [recommended, level 2 heading]

All planned subgroup analyses should be listed (or independent variables for meta-regression). Any other methods for investigating heterogeneity of effects should be described.

Sensitivity analysis [recommended, level 2 heading]

This should describe analyses aimed at determining whether conclusions are robust to decisions made during the review process, such as inclusion/exclusion of particular studies from a meta-analysis, imputing missing data or choice of a method for analysis.

The following *optional* (level 2) headings may be helpful:

Economic issues

Methods for future updates

Results sections

The text of a protocol ends just before the results sections. The results sections begin with a description of the studies identified by the review, which should start with a summary of the inclusion/exclusion of studies.

Description of studies [fixed, level 1 heading]

Results of the search [recommended, level 2 heading]

The results sections should start with a summary of the results of the search (for example, how many references were retrieved by the electronic searches).

Included studies [recommended, level 2 heading]

It is essential that the number of included studies is clearly stated. This section should comprise a succinct summary of the information contained in the 'Characteristics of Included Studies' table. Key characteristics of the included studies should be described, including the study participants, interventions and outcome measures in the included studies and any important differences among the studies. The sex and age range of participants should be stated here except where their nature is obvious (for example, if all the participants are pregnant). Authors should note any other characteristics of the studies that they regard as important for readers of the review to know. The following *optional* (level 3) subheadings may be helpful:

Design

Sample sizes

Setting

Participants

Interventions

Outcomes

Excluded studies [recommended, level 2 heading]

This should refer to the information contained in the 'Characteristics of Excluded Studies' tables, providing a succinct summary of why studies were excluded from the review.

The following *optional* (level 2) headings may be used:

Ongoing studies

Studies awaiting assessment

New studies found at this update

Methodological quality of included studies [fixed, level 1 heading]

This should summarise the general quality of the included studies, its variability across studies and any important flaws in individual studies. The criteria that were used to assess the risk of bias should be described or referenced under 'Methods' and not here. How each study was rated on each criterion should be reported in an additional table and not described in detail in the text, which should be a concise summary.

For large reviews, aspects of the quality assessment may be summarised for the primary outcomes under the following headings.

Allocation [recommended, level 2 heading]

Attempts to conceal allocation of intervention assignment and methods for generation of the sequence of allocations should be summarised here, along with any judgements concerning the risk of bias that may arise from the methods used.

Blinding [recommended, level 2 heading]

A summary of who was blinded during the conduct and analysis of the trial should be reported here. Blinding of outcome assessment should be summarised for each main outcome. Judgements concerning the risk of bias associated with blinding should be summarised.

Follow-up and exclusions [recommended, level 2 heading]

The completeness of data should be summarised here for each of the main outcomes. Concerns over exclusion of participants and excessive (or differential) drop-out should be reported.

Selective reporting [recommended, level 2 heading]

Concerns over the selective availability of data should be summarised here, including evidence of selective reporting of outcomes, timepoints, subgroups or analyses.

Other potential sources of bias [recommended, level 2 heading]

Any other potential concerns should be summarised here.

Results [fixed, level 1 heading]

This should be a summary of the main findings on the effects of the interventions studied in the review. The section should directly address the objectives of the review rather than list the findings of the included studies in turn. The results of individual studies, and any statistical summary of these, should be included in Data tables. Subheadings are encouraged if they make reading easier (for example, for each different participant group, comparison or outcome measure if a review addresses more than one). Any sensitivity analyses that were undertaken should be reported.

Authors 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'. When there is inconclusive evidence, it is wrong to claim that it shows that an intervention has ‘no effect’ or is ‘no different’ from the control intervention. In this situation, it is safer to report the data, with a confidence interval, as being compatible with either a reduction or an increase in the outcome.

Discussion [fixed, level 1 heading]

A structured discussion can aid the systematic consideration of the implications of the review (Docherty 1999).

Summary of main results (benefits and harms) [recommended, level 2 heading]

Summarise the main findings and outstanding uncertainties, balancing important benefits against important harms.

Overall completeness and applicability of evidence [recommended, level 2 heading]

Are the studies identified sufficient to address all of the objectives of the review? Have all relevant types of participants, interventions and outcomes been investigated? Describe the relevance of the evidence to the review question. This should lead to an overall judgement of the external validity of the review. Comments on how the results of the review fit into the context of current practice might be included here, although authors should bear in mind that current practice might vary internationally.

Quality of the evidence [recommended, level 2 heading]

Do the studies identified allow a robust conclusion regarding the objective(s) that they address? Summarise the amount of evidence that has been included (numbers of studies, numbers of participants), review the general methodological quality of the studies, and reiterate the consistency of their results. This should lead to an overall judgement of the internal validity of the results of the review.

Potential biases in the review process [recommended, level 2 heading]

State the strengths and limitations of the review with regard to preventing bias. These may be factors within, or outside, the control of the review authors. The discussion might include whether all relevant studies were identified, whether all relevant data could be obtained, or whether the methods used (for example, searching, study selection, data extraction, analysis) could have introduced bias.

Agreements and disagreements with other studies or reviews [recommended, level 2 heading]

Comments on how the included studies fit into the context of other evidence might be included here, stating clearly whether the other evidence was systematically reviewed.

Authors' conclusions / Reviewers' conclusions [fixed, level 1 heading]

The primary purpose of the review should be to present information, rather than to offer advice. Conclusions of the authors are divided into two sections:

Implications for practice [fixed, level 2 heading]

The implications for practice should be as practical and unambiguous as possible. They should not go beyond the evidence that was reviewed and be justifiable by the data presented in the review. 'No evidence of effect' should not be confused with 'evidence of no effect'.

Implications for research [fixed, level 2 heading]

This section of Cochrane reviews is used increasingly often by people making decisions about future research, and authors should try to write something that will be useful for this purpose. As with the 'Implications for Practice', the content should be based on the available evidence and should avoid the use of information that was not included or discussed within the review.

In preparing this section, authors should consider the different aspects of research, perhaps using types of study, participant, intervention and outcome as a framework. Implications for *how* research might be done and reported should be distinguished from *what* future research should be done. For example, the need for randomised trials rather than other types of study, for better descriptions of studies in the particular topic of the review, or for the routine collection of specific outcomes, should be distinguished from the lack of a continuing need for a comparison with placebo if there is an effective and appropriate active treatment, or for the need for comparisons of specific named interventions, or for research in specific types of people.

It is important that this section is as clear and explicit as possible. General statements that contain little or no specific information, such as "Future research should be better conducted" or "More research is needed" are of little use to people making decisions, and should be avoided.

Acknowledgements [fixed, level 1 heading]

This section should be used to acknowledge any individuals or organisations that the authors wish to acknowledge including individuals who are not listed among the authors. This would include any previous authors of the Cochrane review and might include the contributions of the editorial team of the CRG. Permission should be obtained from persons acknowledged.

Potential conflict of interest [fixed, level 1 heading]

Authors should report any conflict of interest that might be perceived by others as being capable of influencing their judgements, including personal, political, academic and other possible conflicts, as well as financial conflicts. Authors must state if they have been involved in a study included in the review. Details of the Collaboration's policy on conflicts of interest appear in 2.6 Conflict of interest and commercial sponsorship.

Financial conflicts of interest cause the most concern, and should be avoided, but must be reported if there are any. Any secondary interest (such as personal conflicts) that might unduly influence judgements made in a review (concerning, for example, the inclusion or exclusion of studies, assessments of the validity of included studies or the interpretation of results) should be reported.

If there are no conflicts of interest, this should be stated explicitly, for example, by writing 'None known'.

3.5 References

Authors should check all references for accuracy (Dickersin 1986, Eichorn 1987).

3.5.1 References to studies

Studies are organised under four fixed headings:

Included studies: Studies that specifically meet the inclusion criteria and are included in the review should be listed here.

Excluded studies: Studies that specifically do not meet the inclusion criteria and are not included in the review should be listed here.

Studies awaiting assessment: Relevant studies that have been identified, but cannot be assessed for inclusion until additional data or information are obtained, should be listed here. These need not be cited in the text of the review.

Ongoing studies: Studies that are ongoing but meet (or appear to meet) the inclusion criteria should be listed here.

Each of these headings can include multiple studies (or no studies). A study is identified by a 'Study ID'. A year can be associated with each study (usually the year of completion, or the publication year of the primary reference to that study). In addition, each study should be assigned a category of 'Data source' from among the following.

- Published data only
- Published and unpublished data
- Unpublished data only
- Unpublished data sought but not used

Each study can have multiple references. Each reference has its own 'Reference ID'. A single reference for each study should be awarded the status of 'Primary reference'.

3.5.2 Other references

References other than those to studies are divided among two categories:

Additional references: Other references cited in the text should be listed here, including those cited in the background and methods sections. If a report of a study is cited in the text for some reason other than referring to the study (for example, because of some background or methodological information in the report), it should be listed here as well as under the relevant study.

Other published versions of this review: References to other published versions of the review in a journal, textbook or the CDSR should be listed here.

Note: RevMan also includes a 'Classification pending' category to facilitate organisation of references while preparing a review. Any references remaining in this category when the review is submitted are not published.

3.6 Tables

3.6.1 Characteristics of included studies

This is a standard table with seven columns: study ID, methods, participants, interventions, outcomes, notes and allocation concealment. Authors must decide what characteristics of the included studies are likely to interest users of the review. It is possible to use codes so that each column can include several subcategories of information; for example, an author could include country, setting, age and sex under 'participants'. Information on the funding of a study could be included under 'notes'. Footnotes should be used for explanations of any abbreviations used (these will be published in the *CDSR*).

3.6.2 Characteristics of excluded studies

Studies meeting the inclusion criteria, or appearing to meet the inclusion criteria, that were excluded should be identified and the reason for exclusion should be given (for example, inappropriate control group). This should be kept brief, and a single reason for exclusion is usually sufficient.

3.6.3 Characteristics of ongoing studies

This is a standard table with seven columns: Study ID, Trial name or title, Participants, Interventions, Outcomes, Starting date, Contact information and Notes. Footnotes should be used for explanations of any abbreviations used in the table (these will be published in the *CDSR*).

3.6.4 Comparisons and data

Results of studies included in a review are organised in a hierarchy: studies are nested within (optional) sub-categories, which are nested within outcomes, which are nested within comparisons. A study can be included several times among the analyses, but no more than once within any specific sub-category (or within each outcome if there are no sub-categories).

Authors should avoid listing many comparisons or outcomes for which there are no data in the review since each comparison generates a graph even if it contains no data and analysis. Instead, authors should note these comparisons in the text of their review.

Comparison: The comparisons should correspond to the questions or hypotheses under 'Objectives'.

Outcome: Five types of outcomes are possible: dichotomous data, continuous data, individual patient data ('O – E' and 'V' statistics), generic inverse variance (estimate and standard error) and other data (text only). Detailed discussion of data analysis appears in Section 8.

Sub-category: These are sub-categories of studies so that studies can be displayed separately within the given outcome. Sub-categories may relate to subgroup analyses (for example, trials using different doses of a drug) or to a sub-division of the outcome (for example, short-term, medium-term, long-term).

Study: Data for each study must be entered in a standardised format specific to the outcome under which they are appearing.

3.6.5 Additional tables

Additional tables may be used for information that cannot be conveniently placed in the text or in fixed tables. Examples include:

- information to support the background
- details of search methods
- details of quality assessments of included studies
- results that do not fit into 'Comparisons and data' tables

Table number

A number for the table, which must be unique within the current review. This is used for linking to the table from the text and for ordering the tables in the RevMan *Tree view*.

Title

A brief and informative title for the table, which will appear with it.

3.7 Figures

3.7.1 Analyses

Forest plots illustrating data, effect estimates and results of meta-analyses are generated automatically by RevMan Analyses from the 'Comparisons and data', and included in the published review. The author is able to control whether, and how, meta-analyses are performed.

3.7.2 Additional figures

Additional figures can be used to include graphs and other images that are not generated automatically when a review is published in the *CDSR*. Additional figures should never be used for content that can be included in other ways in RevMan, for example as standard graphs in the Table of comparisons or as Additional tables. Funnel plots can be generated by RevMan Analyses for inclusion as additional figures. Other graphs and images may come from other sources.

The images included in RevMan will not be edited or otherwise improved by others, but will be published 'as is'. It is therefore important that images are fully fit for publication. Figures showing statistical analyses should follow the relevant guidance prepared by the Statistical Methods Group (Appendix 8a).

Figure ID

An identifier (maximum 20 characters) for the figure, which must be unique within the current review. This is used for linking to the figure from the text and for ordering the figures in the RevMan *Tree view*. It is not possible to use an ID for a figure that has already been used for a reference because RevMan stores additional figures as a special type of reference. When using several figures, the IDs should be consistent and consecutive, i.e. Figure 01, Figure 02, etc.

There should always be at last one link to a figure in the text; otherwise the figure will not be displayed in the published version.

Caption

A description of the figure, which will appear next to it. If permission to publish a copyrighted figure is granted, the final phrase of the figure caption must be: "Copyright © [Year] [Name of copyright holder, or other required wording]: reproduced with permission.".

- Warning! Large images take up lots of disk space. A single large image can easily take up ten times the total space used for the text and tables of the review. This leads to very large export files. Scanned images can be especially space-consuming because the resolution may be much higher than needed. Always use images with a good balance between resolution and detail, and include as few images as possible.

3.8 Comments & Criticisms

Summary, Reply and **Contributors** are subheadings in this section. The summary should be prepared by the criticisms editor for the CRG in consultation, if necessary, with the person submitting the comment. A reply to this should then be prepared by the author(s) of the

review. Details of the people who contributed to this process should be given. Further information on the comments and criticisms and the updating of reviews is given in Section 10.7.

3.9 Contributions

This section builds on earlier versions of the Handbook. For details of previous authors and editors of the Handbook, please refer to the Acknowledgements section. The list of recommended headings was developed with valuable input from Mike Clarke, Sally Hopewell, Jacqueline Birks, numerous Review Group Co-ordinators, participants at a workshop on bias susceptibility (May 2005) and members of the Handbook Advisory Group.

Contributing authors (May 2005): Ginny Brunton, Mike Clarke, Mark Davies, Frances Fairman, Sally Green, Julian Higgins, Nicki Jackson, Harriet MacLehose, Sandy Oliver, Peter Tugwell, Janet Wale.

Comments on drafts (May 2005): Lisa Askie, Sonja Henderson, Carol Lefebvre, Philippa Middleton, Rasmus Moustgaard, Rebecca Smyth.

Editors: Julian Higgins and Sally Green.

3.10 References

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

Docherty 1999. Docherty M, Smith R. The case for structuring the discussion of scientific papers. BMJ 1999; 318: 1224-5

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

Flanagin 1998. Flanagin A, Carey LA, Fontarosa PB, Philips SG, Pace BP, Lundberg GD, Rennie D. Prevalence of articles with honorary articles and ghost authors in peer-reviewed medical journals. JAMA 1998; 280: 222-4.

ICMJE 1997. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. Canadian Medical Association Journal 1997; 156: 270-85.

Oakley 1999. Oakley A. An infrastructure for assessing social and educational interventions: the same or different? Background paper for the meeting at The School of Public Policy, University College London, 15-16 July 1999, 10pp. http://www.ucl.ac.uk/spp/download/publications/Annexe4.pdf

Rennie 1997. Rennie D, Emanuel L, Yank V. When authorship fails: a proposal to make contributors accountable. JAMA 1997;278:579-85.

Rennie 1998. Rennie D, Yank V. If authors become contributors, everyone would gain, especially the reader. American Journal of Public Health 1998;88:828-30.

Yank 1999. Yank V, Rennie D. Disclosure of researcher contributions: a study of original research articles in the Lancet. Annals of Internal Medicine 1999; 130: 661-70.

4 Formulating the problem

4.1 Rationale for well-formulated questions

Poorly focused questions lead to unclear decisions about what research to include and how to summarise it.

As with any research, the first and most important decision in preparing a review is to determine its focus (Light 1984b). This is best done by asking clearly framed questions. Such questions are essential for determining the structure of a review (Jackson 1980, Cooper 1984, Hedges 1994). Specifically, they will guide much of the review process including strategies for locating and selecting studies or data, for critically appraising their relevance and validity, and for analysing variation among their results.

In addition to guiding the review process, a review's questions and objectives are used by readers in their initial assessments of relevance. The readers use the stated questions and objectives to judge whether the review is likely to be interesting and directly relevant to the issues they face.

4.2 Key components of a question

There are several key components to a well-formulated question (Richardson 1995, Counsell 1997) and these should be set in the Criteria for selecting studies section of the review. A clearly defined question should specify the types of people (participants), types of interventions or exposures, and the types of outcomes that are of interest. In addition, the types of studies that are relevant to answering the question should be specified. In general the more precise one is in defining components, the more focused the review. Equal precision in addressing each component is not necessary. For example, one might wish to concentrate on various treatments for a particular stage of breast cancer, or alternately to focus on a particular drug for any stage of breast cancer. In the former example the stage and severity of the disease would be defined very precisely within the Types of participants. Whereas, in the latter example, the treatment formulation would be defined very precisely within the Types of intervention.

An overview of the key components follows with examples of useful issues to consider for each component. Authors need to ensure that they understand the terminology used to describe these components in different places and settings.

4.2.1 What types of people (participants)?

It is often helpful to consider the types of people that are of interest in two steps. First, define the diseases or conditions that are of interest. Explicit criteria sufficient for establishing the presence of the disease or condition should be developed. Second, identify the population and setting of interest. This involves deciding whether one is interested in a special population group determined on the basis of factors such as age, sex, race, educational status, or the presence of a particular condition such as angina or shortness of breath. One might also be interested in a particular setting on the basis of factors such as whether people are living in the community; are hospitalised, in nursing homes or chronic care institutions; or are outpatients.

Any restrictions with respect to specific population characteristics or settings should be based on sound evidence. For example, focusing a review of the effectiveness of mammographic screening on women between 40 and 50 years old can be justified on the basis of biological plausibility, previously published systematic reviews and existing controversy. On the other hand, focusing a review on a particular subgroup of people on the basis of their age, sex or

astrological birth-sign simply because of personal interests when there is no underlying biologic or sociological justification for doing so should be avoided. When there is uncertainty about whether there are important differences in effects among various subgroups of people, it is probably best to include all of the relevant subgroups and then test for important and plausible differences in effect in the analysis (see section 4.5 below and section 8).

4.2.2 What types of comparisons (interventions)?

The next key component of a well-formulated question is to specify the interventions that are of interest. It is also important to define the interventions against which these will be compared, such as the types of control groups that are acceptable for the review. Give thought to whether persons in a control group might receive interventions other than a placebo, and whether those interventions overlap in any way with the active intervention being tested. This issue is discussed further in the section on assessing the quality of studies (section 6).

4.2.3 What types of outcomes?

The third key component of a well-formulated question is the delineation of particular outcomes that are of interest. While all important outcomes should be included in Cochrane reviews, trivial outcomes should not be included. Authors need to avoid overwhelming readers with data that is of little or no importance. At the same time that they must be careful not to leave out important data. If explicit criteria are necessary for establishing the presence of those outcomes these should be specified. Likewise if combinations of outcomes will be considered these need to be specified. For example, if a study only has data on nonfatal and fatal strokes combined, will this be included if the question specifically relates to stroke death?

In general, Cochrane reviews should include all reported outcomes that are likely to be meaningful to people making a decision about the healthcare problem the review addresses. Beyond this, it may be important to specify outcomes that are important to decision makers, even when it is unlikely that data will be found. For example, quality of life is an important outcome, perhaps the most important outcome, for people considering whether or not to use chemotherapy for advanced cancer, even if the available studies only report survival data. In addition, authors (reviewers) should indicate how they will try to include data on adverse effects in their review. In regard to this, rather than including an exhaustive list of adverse outcomes it may be more informative to summarise 'severe' (e.g. severe enough to require withdrawal of treatment) and minor adverse outcomes and include appropriate description of these.

It is sometimes possible to acquire unpublished data from investigators in order to disentangle combined outcomes, as well as for other purposes (see section 7). Before excluding a study that seems to meet criteria for relevance, but has not reported results in a way that is adequate for the review, it is worth considering trying to obtain the necessary information from the investigators.

4.2.4 What types of study designs?

Certain study designs are superior to others when answering particular questions. Randomised controlled trials (RCTs) are considered by many the *sine qua non* when addressing questions regarding therapeutic efficacy, whereas other study designs are appropriate for addressing other types of questions. For example, questions relating to aetiology or risk factors may be addressed by case-control and cohort studies. Authors should consider up-front what study designs are likely to provide reliable data with which to answer their questions.

Other aspects relevant to study design that are worth initial consideration are whether to review studies that: have a placebo comparison group, evaluate outcomes in an unbiased manner, or have a certain length of follow-up. The more restrictive authors are in matching questions to particular aspects of design, the less likely they are to find data specific to the restricted question. However, reviewing studies that are unlikely to provide reliable data with which to answer the question is a poor use of time and can result in misleading conclusions. If, for example, one is interested in whether a therapy improves survival in patients with a chronic condition, it might be inappropriate to look at studies of very short duration, except to make explicit the fact that they cannot address the question of interest.

Because Cochrane reviews address questions about the effects of healthcare, they focus primarily on RCTs. There are two reasons why one should be cautious about including non-randomised studies in a review of the effects of healthcare, both relating to bias. First, although it is possible to control for confounders that are known and measured using other study designs, randomisation is the only way to control for confounders that are not known or not measured. For clinical interventions, deciding who receives an intervention and who does not is influenced by many factors, including prognostic factors. Empirical evidence suggests that, on average, non-randomised studies tend to overestimate the effects of healthcare (Sacks 1982, Chalmers 1983, Schulz 1995). However, a systematic methodology review has shown that the extent and even the direction of bias in non-randomised studies is often impossible to predict (Kunz 1998).

Second, although it is often difficult to locate RCTs (Dickersin 1994) and reviews that fail to include unpublished trials may be biased towards overestimating the effectiveness of an intervention (Dickersin 1993). The efforts of the Cochrane Collaboration to identify RCTs have not been matched for the identification of other types of studies. Consequently, including studies other than controlled trials in a review may require additional efforts to identify studies and to keep the review up-to-date, and might increase the risk that the result of the review will be influenced by publication bias.

Despite the above concerns, it may sometimes be appropriate to conduct a systematic review of non-randomised studies of the effects of healthcare. For example, occasionally the course of a disease is so uniform or the effects of an intervention are so dramatic that it is unnecessary and unethical to conduct RCTs. Under such circumstances it would be senseless to restrict a review to RCTs. While attention to the risk of bias should guide decisions about what types of study designs to include in a review, individual authors and Collaborative Review Groups must decide what types of studies are best suited to specific questions.

4.3 Using the key components of a question to locate and select studies

Once one has a well-formulated question, one should determine which key components to focus on in initial searching strategies. For Cochrane reviews searching for studies is greatly facilitated by the availability of specialised registers compiled by CRGs. However, the extent to which these registers are developed varies and it may be necessary for authors to conduct supplemental searches.

Searches that demand the simultaneous presence of several components or very specific formulations of certain components are likely to be too specific and miss important information. For example, if one searches for studies addressing long-term effects of insulin therapy on renal function in type II diabetics by demanding that they be indexed as 'type II diabetes', 'insulin', 'renal function' and 'long-term', relevant studies are likely to be missed. On the other hand if 'insulin' or 'type II diabetes' is used alone as a search term, hundreds of irrelevant reports are likely to be identified.

In general, useful key components to use when searching include the condition or disease of interest and the intervention or exposure being evaluated. Although one may be specifically interested in a particular setting, studies are often not indexed by the type of setting in electronic databases. Also, multiple outcomes may be evaluated in studies, some of which may be relevant to the review, but not part of the indexing of the article. This issue is discussed further in the next section on locating and selecting studies (section 5).

Whatever search strategies are used, it will be necessary to go through a number of reports and decide which ones are relevant and which ones are not relevant. Formulating a question in terms of the types of participants, interventions, outcomes and study designs of interest will lead naturally to specifying the criteria that will be used to select studies. However, some additional effort is often needed to clarify the selection criteria and develop decision rules that are sensible and reproducible. If, for example, you are reviewing studies of therapies for constipation, you must decide if you will review studies addressing acute and/or chronic constipation as well as acceptable criteria for acute and chronic. Are you interested in the entire spectrum of severity of constipation or only in severe constipation and how will you define 'severe'? Do you want to review studies that define constipation on the basis of a certain frequency of bowel movements per week or limit yourself to studies that define constipation on the basis of symptoms such as straining and hard stools? Will you only review studies that have determined the underlying pathophysiologic mechanism of constipation or limit your review to certain specific pathophysiologic disorders? Will you consider studies that merely state that participants were 'constipated'.

4.4 Using the key components of a question to guide data collection

Details relevant to key components of questions are what authors will be collecting from individual studies. Thus well-formulated questions are directly linked to the data collection process because they guide: determination of final criteria that will be used to select appropriate studies for review, and what data should be abstracted from studies meeting those selection criteria. Components of questions may also be directly related to how one chooses to present and analyse data. These issues are discussed further in section 6, section 7 and section 8.

4.5 Broad versus narrow questions

The questions that a review addresses may be broad or narrow in scope. For example, a review might address a broad question regarding whether antiplatelet agents in general are effective in preventing thrombotic events in humans. Alternatively, a review might address whether a particular antiplatelet agent, such as aspirin, is effective in decreasing the risks of a particular thrombotic event, stroke, in elderly persons with a previous history of stroke. As another example, separate reviews might be done to investigate the effectiveness of antibiotics to treat respiratory tract infections in young children and adults.

Determining the scope of a review question is a decision dependent upon multiple factors including perspectives regarding a question's relevance and potential impact; supporting theoretical, biologic and epidemiological information; the potential generalisability and validity of answers to the questions; and available resources.

There are several advantages and disadvantages to initially asking broad or narrow questions. Narrowly focused reviews may not be generalisable to multiple settings, populations and formulations of an intervention. They can also result in spurious or biased conclusions in the same way that subgroup analyses sometimes do (see section 8.7). For example, a review of

the effectiveness of aspirin for preventing strokes in women could lead to a false conclusion that aspirin was not effective in women when in truth there were not enough data to detect any significant difference in effect between men and women. A narrow focus is at high risk of resulting in biased conclusions when the author is familiar with the literature in an area and narrows the inclusion criteria in such a way that one or more studies with results that are in conflict with the author's beliefs are excluded. There is also a danger that the known results of a series of studies of a class of interventions might influence the choice of a specific intervention from this class for a narrow review.

The validity of very broadly defined reviews may be criticised for mixing apples and oranges, particularly when there is good biologic or sociological evidence to suggest that various formulations of an intervention behave very differently or that various definitions of the condition of interest are associated with markedly different effects of the intervention. It is fine to mix apples and oranges, if your question is about fruit, but not if your question is about vitamin C and you know that apples and oranges are different with respect to vitamin C.

Searches for data relevant to broad questions may be more time-consuming and more expensive than searches relevant to narrowly defined questions. As broad questions may be addressed by large sets of heterogeneous studies, the synthesis and interpretation of data may be particularly challenging. Broadly focused reviews can also become unwieldy to present, maintain and understand.

One option that has been found useful is to build a broadly focused review on the basis of a series of more narrowly focused reviews. For example, healthcare providers and pregnant women who want to quit smoking are likely to want to know which smoking cessation strategy to use - a broad question. A review that helps them to answer this question could be built upon a series of more focused reviews that ask what the effectiveness of a specific strategy, such as behaviour modification, is. Whether it makes most sense to start with narrower questions and build up to a broader question, or to start with a broad question and then divide it into a number of smaller questions depends on the nature of the problem (e.g. how complex it is, how well understood it is, how much research is available) and the particular circumstances of the authors and their CRG (e.g. how well developed their specialised register is, the availability of resources, time and interest).

4.6 Changing questions

While questions should be posed in the protocol before initiating the full review, these questions should not become a straightjacket that prevents exploration of unexpected issues (NHS CRD 1996). Reviews are analyses of existing data that are constrained by previously chosen study populations, settings, intervention formulations, outcome measures and study designs. It is generally not possible to formulate an answerable question for a review without knowing some of the studies relevant to the question, and it may become clear that the questions a review addresses need to be modified in light of evidence accumulated in the process of conducting the review.

Although a certain fluidity and refinement of questions is to be expected in reviews as one gains a fuller understanding of the problem, it is important to guard against bias in modifying questions. *Post-hoc* questions are more susceptible to bias than those asked *a priori*, and data-driven questions can generate false conclusions based on spurious results. Any changes to the protocol that result from revising the question for the review should be documented. When refining questions it is useful to ask the following questions:

- What is the motivation for the refinement?
- Was it made after you had seen and been influenced by results from a particular study or was it simply that you had not initially considered alternate but acceptable ways of defining the participants, interventions or outcomes of interest?

- Are your search strategies appropriate for the refined question (especially any that have already been undertaken)?
- Is your data collection tailored to the refined question?

4.7 References

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

Cooper 1984. Cooper HM. The problem formulation stage. In: Cooper HM, editor. Integrating Research. A Guide for Literature Reviews. Newbury Park: Sage Publications, 1984; 19-37.

Counsell 1997. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997; 127: 380-7.

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

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

Hedges 1994. Hedges LV. Statistical considerations. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York: Russell Sage Foundation, 1994; 30-3.

Jackson 1980. Jackson GB. Methods for integrative reviews. Rev Educ Res 1980; 50:438-60.

Kunz 1998. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised trials. BMJ 1998; 317: 1185-90.

Light 1984. Light RJ, Pillemer DB. Organizing a reviewing strategy. In: Summing Up: The Science of Reviewing Research. Cambridge, Massachusetts: Harvard University Press, 1984; 13-31.

NHS CRD 1996. NHS Centre for Reviews and Dissemination. Undertaking Systematic Reviews of Research on Effectiveness (CRD Report 4). York: The University of York, 1996.

Richardson 1995. Richardson WS, Wilson MS, Nishikawa J, Hayward RSA. The Well-built Clinical Question: A Key to Evidence Based Decisions. ACP J Club 1995; A12-3.

Sacks 1982. Sacks HS, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. Am J Med 1982; 72:233-40.

Schulz 1995. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995; 273:408-12.

5 Locating and selecting studies

Systematic reviews of the effects of health care interventions generally focus on reports from randomized controlled trials (RCTs), when such data are available, because of the general acceptance that this study design will lead to the most reliable estimates of effects. A comprehensive search for relevant RCTs, which seeks to minimize bias, is one of the essential steps in doing a systematic review, and one of the factors that distinguishes a systematic review from a traditional review.

A 'quick and dirty' search of, for example MEDLINE, is generally not considered adequate. Studies have shown that only 30 - 80% of all known published RCTs were identifiable using MEDLINE (depending on the area or specific question) (Dickersin 1994). Even if relevant records are in MEDLINE it can be difficult to retrieve them easily. A comprehensive search is important not only for ensuring that as many studies as possible are identified but also to minimize selection bias for those that are found. Relying exclusively on a MEDLINE search may retrieve a set of reports unrepresentative of all reports that would have been identified through a comprehensive search of several sources. For example, the majority of the journals indexed in MEDLINE are published in English. If studies showing an intervention to be effective are more likely to be published in English, then any summary of only the English language reports retrieved through a MEDLINE search may result in an overestimate of effectiveness due to a language bias (Gregoire 1995; Moher 1996; Egger 1997; Juni 2002). In addition, the results of many studies are never published, and most of these probably remain unknown. If studies showing an intervention to be effective are more likely to be published, then any summary of only the published reports may result in an overestimate of effectiveness due to a publication bias (Simes 1986; Dickersin 1987; Simes 1987; Begg 1988; Hetherington 1989; Easterbrook 1991; Dickersin 1993; Song 2000).

This section contains information about locating and selecting studies for systematic reviews. The first section describes some of the sources and approaches that can be used. The second section provides guidance on developing and documenting search strategies and organizing the records retrieved.

5.1 Searching for studies

5.1.1 Electronic databases

A search for relevant studies generally begins with health-related electronic bibliographic databases. Searches of electronic databases are generally the easiest and least time-consuming way to identify an initial set of relevant reports. Some electronic bibliographic databases, such as MEDLINE and EMBASE, include abstracts for the majority of recent records. Often a searcher can determine an article's relevance to a review based on the abstract, and can thereby avoid retrieving the full journal article, if the reported study is clearly not eligible for inclusion. Another advantage of these databases is that they can be searched electronically, for either words in the title and abstract, or using standardized subject related indexing terms that have been assigned to the record. For example, the MEDLINE indexing term RANDOMIZED-CONTROLLED-TRIAL (Publication Type) was introduced in 1991 and allows a user to search for articles describing individual randomized trials.

Hundreds of electronic bibliographic databases exist. Some databases, such as MEDLINE/PubMed and EMBASE, cover all areas of health care and index journals published from around the world. Other databases, such as the Australasian Medical Index, the Chinese Biomedical Literature Database, the Latin American Caribbean Health Sciences Literature (LILACS), and the Japan Information Centre of Science and Technology File on Science, Technology and Medicine (JICST-E) index journals published in specific regions of

the world. Others, such as the Cumulative Index of Nursing and Allied Health (CINAHL) and AIDSLINE, focus on specific areas of health. The Cochrane Collaboration has been developing an electronic database of reports of controlled trials ("CENTRAL") that is now the best single source of information about records that relate to studies, which might be eligible for inclusion in Cochrane reviews (Dickersin 2002). Details of other databases that might contain eligible records are available in the Gale Directory of Online, Portable and Internet databases (http://www.dialog.com). The three electronic bibliographic databases generally considered as the richest sources of trials - MEDLINE, EMBASE, and CENTRAL - are described in more detail below.

5.1.1.1 MEDLINE and EMBASE

Index Medicus (published by the US National Library of Medicine (NLM)) and Excerpta Medica (published by Elsevier) are indexes of healthcare journals that are available in electronic form as MEDLINE and EMBASE, respectively. MEDLINE indexes about 4600 journals from the United States and 70 other countries, and in February 2002 contained over 11 million records from 1966 forward. (Some pre-1966 records have been added recently.) PubMed is a free, online MEDLINE database that also includes up-to-date citations not yet indexed (http://www.ncbi.nlm.nih.gov). EMBASE, which is often considered the European counterpart to MEDLINE, indexes nearly 4000 journals from over 70 countries and, in May 2002, contained approximately 9 million citations.

The overlap in journals covered by MEDLINE and EMBASE has been estimated to be approximately 34% (Smith 1992). The actual degree of reference overlap depends on the topic, with reported overlap values in particular areas ranging from 10% to 75% (Kleijnen 1992; Odaka 1992; Smith 1992; Rovers 1993; Ramos-Remus 1994). Studies comparing searches of the two databases have generally concluded that a comprehensive search requires that both databases be searched. Although MEDLINE and EMBASE searches tend not to identify the same sets of references, they have been found to return similar numbers of relevant references.

MEDLINE and EMBASE can be searched using standardized subject terms assigned by indexers employed by the publishing organization. Standardized subject terms (as part of a "controlled vocabulary") are useful because they provide a way of retrieving articles that may use different words to describe the same concept and because they provide information beyond what is simply contained in the words of the title and abstract. Using the appropriate standardized subject terms, a simple search strategy can quickly identify articles pertinent to the topic of interest. This approach works well if the goal is to identify a few good articles on a topic or to identify one particular article. However, when searching for studies for a systematic review the precision with which subject terms are applied to references should be viewed with healthy skepticism. Authors may not describe their methods or objectives well, indexers are not always expert in the subject area of the article that they are indexing, and indexers make mistakes, like all people. In addition, the available indexing terms might not correspond to the terms the searcher wishes to use. The controlled vocabulary search terms for MEDLINE and EMBASE are not identical. Search strategies need to be customized for each database. One way to begin to identify controlled vocabulary terms for a particular database is to retrieve articles from that database, which meet the inclusion criteria for the review and to note common text words and the terms the indexers had applied to the articles, which could then be used for a full search.

Assuming that search results from each database are of approximately equal value, the choice of which to search first may often be a matter of cost, with MEDLINE typically being the less costly option. As noted earlier, PubMed provides free online access to MEDLINE. Other NLM databases, including AIDSLINE, and HealthSTAR are being phased out and their unique journal citations are migrating to PubMed. PubMed also provides links to full-text versions of articles on other publishers' web sites. A particularly useful feature of PubMed is that a list of 'Related articles' can be obtained for each relevant record identified. The NLM is

developing a new database, called the Gateway, which allows users to search PubMed and multiple other NLM retrieval systems simultaneously. The current Gateway (http://gateway.nlm.nih.gov/gw/Cmd) searches PubMED, OLDMEDLINE, LOCATORplus, MEDLINEplus, DIRLINE, AIDS Meetings, Health Services Research Meetings, Space Life Sciences Meetings, and HSRProj.

5.1.1.2 The Cochrane Central Register of Controlled Trials (CENTRAL)

The Cochrane Central Register of Controlled Trials (CENTRAL) serves as the most comprehensive source of records related to controlled trials. As of January 2003, CENTRAL contained just over 350,000 citations to reports of trials and other studies potentially relevant to Cochrane reviews. CENTRAL includes citations to reports of controlled trials that might not indexed in MEDLINE, EMBASE or other bibliographic databases; citations published in many languages; and citations that are available only in conference proceedings or other sources that are difficult to access (Dickersin 2002). Guidance on searching CENTRAL has been prepared as part of the CENTRAL Management Plan (http://www.cochrane.us/manage.htm). Many of the records in CENTRAL have been identified through systematic searches of MEDLINE and EMBASE, as described in the paragraph below.

The US Cochrane Center (as the former New England Cochrane Center, Providence Office) and the UK Cochrane Centre have searched MEDLINE for publication years 1966-2000 using phases 1 and 2 of the Cochrane highly sensitive search strategy (Appendix 5b) (Dickersin 1994). Each year, the US Cochrane Center updates this searching of MEDLINE. Hundreds of thousands of records have been retrieved and reviewed to date. If, on the basis of their title and abstract, the retrieved citations were judged to meet the Cochrane definitions for reports of randomized controlled trials (RCTs) and controlled clinical trials (CCTs), they have been assigned the Publication Type RANDOMIZED CONTROLLED TRIAL or CONTROLLED CLINCIAL TRIAL in MEDLINE and also included in CENTRAL (with the permission of the NLM) (see Appendix 5a.1 for Cochrane and Appendix 5a.2 for NLM definitions of RCT and CCT).

Similarly, in an ongoing project, the UK Cochrane Centre is retrieving records from EMBASE, checking their titles and abstracts and submitting these for inclusion in CENTRAL when appropriate (with the permission of Elsevier). A search of EMBASE using five free text terms (ie, random*, crossover*, cross-over*, factorial*, and placebo*), and covering the years 1974-1999, was run in 1999 to identify reports of trials. The results of this search are published in each quarterly release of CENTRAL. Additional searching of EMBASE began in December 2000, and this stage of the project includes searching using additional free text terms and EMBASE (EMTREE) thesaurus terms (Dickersin 2002).

Other general healthcare databases published in Australia, China, and Brazil are undergoing similar systematic searches to identify reports of trials for CENTRAL. The Australasian Cochrane Centre is coordinating the search of the National Library of Australia's Australasian Medical Index; the Chinese Cochrane Centre is coordinating the search of the Chinese Biomedical Literature Database; and the Brazilian Cochrane Centre is coordinating the search of the Pan American Health Organization's database LILACS (Latin American Caribbean Health Sciences Literature).

Each Collaborative Review Group (CRG) is responsible for the development of a subject specific specialized register of trials, which serves to ensure that individual authors (reviewers) within the CRG have easy and reliable access to the maximum possible number of studies relevant to their review topic. Typically, the editorial team will assume at least some, if not all, responsibility for examining new studies and forwarding them to appropriate authors. CRGs use all the methods described in this chapter to identify trials for their specialized registers, with the exception of generalized searches of MEDLINE and EMBASE, which, as described above, are performed by the US Cochrane Center and the United

Kingdom Cochrane Centre. Many CRGs also have systems to ensure that reports identified by authors for their review(s) are contributed to the CRG's specialized register. The registers should, in turn, be submitted for inclusion in CENTRAL. Thus, records included in the specialized register of one CRG become accessible to all other CRGs through CENTRAL.

More detailed information about the development and contents of CENTRAL is included in a recent article (Dickersin 2002) and *The Cochrane Library* help file for CENTRAL.

5.1.1.3 SciSearch

SciSearch is an electronic database that lists published "source" articles from 4500 major scientific and technical journals and the articles that cite them. SciSearch can be used to identify studies for a review by identifying in the database a known relevant source article, and checking each of the articles citing the source article, to see if it is also relevant to the review. It is a way of searching forward in time from the publication of an important article. SciSearch also includes reference lists for records it indexes.

5.1.2 Handsearching

Handsearching involves a manual page-by-page examination of the entire contents of a journal issue to identify all eligible reports of trials, whether they appear in articles, abstracts, news columns, editorials, letters or other text. Handsearching health care journals is a necessary adjunct to searching electronic databases for at least two reasons: 1) not all trial reports are included on electronic bibliographic databases, and 2) even when they are included, they may not be indexed with terms that allow them to be easily identified as trials. Each journal year should be handsearched thoroughly and competently by a well-trained handsearcher for all reports of trials so that once a journal year has been handsearched, it will not need to be searched again. A recent study has found that a combination of handsearching and electronic searching is necessary for full identification of relevant reports published in journals that are indexed in MEDLINE, especially for articles published before 1991 when the NLM system for indexing trial reports was not as well developed as it is today and for those articles that are in parts of journals (such as supplements and correspondence) which are not indexed in MEDLINE (Hopewell 2002).

To facilitate the identification of all published trials the Cochrane Collaboration has organized extensive handsearching efforts. Overall coordination of the Collaboration's handsearch of the world's medical literature is managed by the US Cochrane Center, which oversees prospective registration of all potential handsearching on the Master List of Journals being Searched (http://www.cochrane.us/cochranemainpage.asp). Almost 2200 journals have been, or are being, searched within the Collaboration, and are included in the Master List. "Stand-alone" conference proceedings being searched are also included. The Master List enables search progress to be recorded and monitored for each title and also serves to prevent the duplication of effort that might otherwise arise if journals or conference proceedings in overlapping specialties were to be searched by more than one group or individual.

Cochrane entities and authors can prioritize handseaching based on where they expect to identify the most trial reports. This prioritization can be informed by searching CENTRAL, MEDLINE, and EMBASE in a topic area and identifying which journals appear to be associated with the most retrieved citations. Preliminary evidence suggests that most of the journals with a high yield of trial reports are indexed in MEDLIINE (Dickersin 2002), but this may reflect the fact that Cochrane contributors have concentrated early efforts on searching these journals.

Conference proceedings are important to handsearch because individual conference. These abstracts are not included on MEDLINE and are not usually included in other databases.

Abstracts and other grey literature have been shown to be sources of approximately 10% of the studies referenced in Cochrane reviews (Mallett 2002). Over one-half of trials reported in conference abstract never reach full publication, and those that are eventually published in full have been shown to be systematically different than those that are never published in full (Scherer 2003). In addition, grey literature in general has been found to be more likely than health care journals to contain 'negative' reports (McAuley 2000). Thus, failure to identify trials reported in conference proceedings might affect the results or threaten the validity of a systematic review.

Authors who wish to handsearch journals or conference proceedings to identify reports of studies for their review should first consult with the editorial based of their CRG. The CRG's Trials Search Coordinator/Review Group Coordinator can determine whether the journal or conference proceedings has already been searched, and, if it has not, the Coordinator can register the search on the Master List and provide training in handsearching. Training material is available on the US Cochrane Center web site (http://www.cochrane.us/hsmain.htm) All correspondence regarding the initiation of a journal search, progress of a journal search, status of a search etc needs to be between staff at the US Cochrane Center and the Trials Search Coordinator/Review Group Coordinator.

5.1.3 Checking reference lists

Authors should check the reference lists of articles obtained (including those from previously published systematic reviews) to identify relevant reports. The process of following up references from one article to another is generally an efficient means of identifying studies for possible inclusion in a review. Because investigators may selectively cite studies with positive results (Gotzsche 1987; Ravnskov 1992), reference lists should never be used as a sole approach to identifying reports for a review, but rather as an adjunct to other approaches.

5.1.4 Checking other reviews

Some of the most convenient and obvious sources of references to potentially relevant studies are existing reviews. Copies of previously published reviews on the topic of interest should be obtained and checked for references to the original studies. As well as *the Cochrane Database of Systematic Reviews, The Cochrane Library* includes the Database of Abstracts of Reviews of Effects (DARE) a database produced by the NHS Centre for Reviews and Dissemination in York, UK, that provides information on previously published reviews of the effects of healthcare. MEDLINE, EMBASE and other bibliographic databases can also be used to identify review articles. In MEDLINE, the most appropriate review articles would be indexed under the Publication Type terms META-ANALYSIS and REVIEW, ACADEMIC. Search strategies have been developed to enhance identification of these types of publication (Boynton 1998).

5.1.5 Print versions of electronic databases

While MEDLINE and EMBASE include citations from 1966 and 1974 to the present, respectively, Index Medicus and Excerpta Medica, the print versions of these databases, include citations from 1879 and 1948, respectively. Searching the earlier printed subject indexes may be worthwhile, especially if there is reason to believe that there were early studies of the intervention being reviewed.

Science Citation Index is the print version of SciSearch (see Section 5.1.1.3) and is used for the same general purpose, i.e. for listings of where a published article was subsequently cited. Science Citation Index is more comprehensive than SciSearch, which began in 1974.

5.1.6 Identifying unpublished studies

Some completed studies are never published. If it could be assumed that unpublished studies of a given intervention were comparable to published studies on the same intervention, the failure to identify unpublished results would not be an important threat to the validity of a systematic review. However, an association between significant results and publication has been documented across a number of studies (Dickersin 1997). Finding out about unpublished studies, and including them in a systematic review, when eligible, may be important to minimizing bias. Unfortunately, there is no easy way to obtain information about studies that have been completed but never published.

Colleagues can be an important source of information about unpublished studies, and informal channels of communication can sometimes be the only means of identifying unpublished data. Formal letters of request for information can also be used to identify completed but unpublished studies. One way of doing this is to send a comprehensive list of relevant articles along with the inclusion criteria for the review to the first author of reports for included studies, asking if they know of any additional studies (published or unpublished) that might be relevant. It may also be desirable to send the same letter to other experts and pharmaceutical companies or others with an interest in the area. However, it should be borne in mind that asking researchers for information about completed but never published studies has not typically been fruitful (Hetherington 1989; Horton 1997).

Identifying ongoing studies may also be important so that when a review is later updated, these can be assessed for possible inclusion. Unfortunately no single, central register of ongoing randomized trials currently exists and instead there are hundreds of distinct, predominantly online registers that vary widely in content, quality, and accessibility. These may have limited use as a means of identifying studies relevant to systematic reviews. Various efforts have been made by independent groups to begin to provide central access to ongoing trials, mostly through web sites that provide links to hundreds of registers of ongoing clinical trials. Two such examples are TrialsCentral™ (www.trialscentral.org) and Current Controlled Trials (www.controlled-trials.com). Current Controlled Trials also has a searchable database of information about thousands of ongoing and completed trials, including those registered on ClinicalTrials.gov.

5.1.7 Evidence on adverse effects

The first sources to investigate for information on adverse effects are reports from trials or other studies included in the systematic review. Excluded reports might also provide some useful information.

There are a number of sources of information on adverse effects of drugs, including Current Problems produced by the UK Medicines Control Agency (http://www.open.gov.uk/mca), MedWatch produced by the US Food and Drug Administration, and the Australian Adverse Drug Reactions Bulletin (http://www.health.gov.au/). Other regulatory authorities and the drug manufacturer may also be able to provide some information. Information on adverse effects might also be sought from other types of studies than those considered appropriate for the systematic review (e.g. cohort and case-control studies, uncontrolled trials, case series and case reports). However, all such studies and reports are subject to bias to a greater extent than randomized trials, and findings must be interpreted with caution.

5.2 Developing and documenting a search strategy for studies and organizing search results

5.2.1 Developing a search strategy

The ultimate goal in developing a specialized register for a CRG is that it can serve as an all-inclusive source of reports relevant to the CRG's scope and topic area, such that a relatively simple search using some key words related to the intervention could be run against the specialized register to identify all relevant studies. Most CRG specialized registers have not yet reached this point of comprehensiveness. Nevertheless, for many CRGs, the specialized register is still the best available source of studies for a given review. Different CRGs have different systems of ensuring authors have access to reports included in their specialized registers. Many Trials Search Coordinators/Review Group Coordinators search their CRG's specialized register for authors on request. Specialized registers can also be searched through CENTRAL, which contains a recent version of the registers for most CRGs.

It is always necessary to strike a balance between comprehensiveness and precision when developing a search strategy. Increasing the comprehensiveness of a search entails reducing its precision and retrieving more non-relevant articles. Developing a search strategy is an iterative process in which the terms that are used are modified, based on what has already been retrieved. There are diminishing returns for search efforts; after a certain stage, each additional unit of time invested in searching returns fewer references that are relevant to the review. Consequently there comes a point where the rewards of further searching may not be worth the effort required to identify the additional references. The decision as to how much to invest in the search process depends on the question a review addresses, the extent to which the CRG's specialised register is developed, and the resources that are available.

It is a good idea to search other electronic bibliographic databases regardless of whether CENTRAL or a CRG's specialized register is searched. If authors wish to conduct their own additional searches, information specialists with expertise in electronic searching should be sought to design and run the search strategy. The assistance of an information specialist should help to avoid many errors, and ensure that database-specific search term syntax will be appropriate and that advanced searching techniques (e.g. 'exploding' controlled vocabulary terms) can be employed where available. If information specialists are involved in developing the search strategy, they should be made aware of the greater importance of high recall (i.e. sensitivity) as compared to precision in searching for studies for systematic reviews. Ideally, authors should be present when the search is done. There are often costs associated with searching each database and with each record that is downloaded. Therefore, judgments about what to download often need to be made while the search is being done. The exact search performed and material retrieved for each search should be recorded in the Search Strategies for Identification of Studies section of the Cochrane review.

An electronic search strategy should generally have three sets of terms: 1) terms to search for the health condition of interest; 2) terms to search for the intervention(s) evaluated; and 3) terms to search for the types of study design to be included (typically randomized trials). The exception to this is CENTRAL, which aims to contain only reports with study designs possibly relevant for inclusion in Cochrane reviews, so searches of CENTRAL should be based on health condition and intervention only. A good approach to developing an electronic search strategy is to begin with multiple terms that describe the health condition of interest and join these together with the Boolean 'OR' operator. This means you will retrieve articles

containing at least one of these search terms. You can do likewise for a second set of terms related to the intervention(s) and for a third set of terms related to the appropriate study design. These three sets of terms can then be joined together with the 'AND' operator. This final step of joining the three sets with the 'AND' operator limits the retrieved set to articles of the appropriate study design that address both the health condition of interest and the intervention(s) to be evaluated. A note of caution about this approach is warranted however: if an article does not contain at least one term from each of the three sets, it will not be identified. For example, if an index term has not been added to the record for the intervention or the intervention is not mentioned in the title and abstract, the article would be missed. A possible remedy is to omit one of the three sets of terms and decide which records to check on the basis of the number retrieved and the time available to check them.

No language restrictions should be included in the search strategy. Date restrictions should be applied only if it is known for certain that relevant studies could only have been reported during a specific time period.

A Trials Search Coordinators or information specialist can often be helpful in suggesting terms for the health condition and intervention. In general, both controlled vocabulary terms and text words (i.e. those found in the title or abstract) should be used. You should assume that earlier articles are harder to identify. For example, abstracts are not included in MEDLINE for most articles published before 1976 and, so, text word searches will only apply to titles. In addition, few MEDLINE indexing terms relating to study design were available before the 1990s. In designing a search strategy, it may be helpful to look at published papers on the same topic and check the controlled vocabulary terms and text words. Although a research question may address particular populations, settings or outcomes, these concepts are often not well indexed with controlled vocabulary terms and generally do not lend themselves well to searching.

The Cochrane highly sensitive search strategy for MEDLINE (Dickersin 1994; Robinson 2002) was developed specifically with the needs of Cochrane reviews in mind. The earliest version of this search strategy was developed in 1994 and subsequent versions have been developed, each with a different syntax, specific to the version of MEDLINE being searched (e.g. Silver Platter MEDLINE, OVID MEDLINE, PubMed) (Appendix 5b).

As noted in Section 5.1.1.2, the first two phases of the strategy have already been applied to search MEDLINE for all years from 1966 to 2000. Records resulting from the search were downloaded, printed out, and classified as definite or possible randomized or quasi-randomized trials, or not using the information in the title and abstract. If no abstract was available, the decision was based on the title alone. Because identification relied solely on the titles and, where available, the abstracts, some relevant articles may not have been identified. Therefore, it may still be worthwhile for authors to search MEDLINE using the Cochrane highly sensitive search strategy and to obtain and check the full reports of possibly relevant citations.

None of the terms from phase 3 of the Cochrane highly sensitive search strategy were used for generalized searching for controlled trial reports on MEDLINE noted above because of a pilot assessment which showed an unfavorable ratio of effort and expense to results (Clarke 1999).

CRGs typically use phases 1-3 of the Cochrane highly sensitive search strategy plus subject matter terms (using the Boolean "AND") for searching MEDLINE. In developing a search strategy for other electronic bibliographic databases, the terms used to identify trials would generally be similar or the same as terms from the Cochrane highly sensitive search strategy. If an information specialist is assisting with developing a search strategy, she should be made aware of the Cochrane highly sensitive search strategy and how it is used.

5.2.2 Documenting a search strategy

5.2.2.1 Electronic databases

The search strategy for electronic databases should be described in sufficient detail in a review that the process could be replicated. The following information should be included for each electronic bibliographic database each time it is searched, including CENTRAL and specialized registers:

- Title of database searched (e.g. MEDLINE)
- Name of the host (e.g. Silver Platter version 2.0)
- Date search was run (month, day, year)
- Years covered by the search
- Complete search strategy used, including all search terms (preferably cut and pasted rather than retyped)
- One or two sentence summary of the search strategy indicating which lines of the search strategy were used to identify records related to the health condition and intervention, and which lines were used to identify studies of the appropriate design
- The absence of any language restrictions

A description of a search strategy for electronic databases is included as Appendix 5c.

5.2.2.2 Journal Handsearching

Any journal years searched specifically for the review should be listed in the Search Strategies for Identification of Studies section of the review, by journal title, in alphabetical order. Ideally the full titles should be used for the journals. The months and years searched should be stated.

Example: British Journal of Surgery January 1948 December 1998

5.2.2.3 Conference Proceedings

Details of the conference proceedings searched for the review should be provided as follows:

Proceedings with a title in addition to the conference name:

> Child abuse and neglect: a medical community response. 1st AMA National Conference on Child Abuse and Neglect; 1984 Mar 30 31; Chicago.

Proceedings without a separate title:

> Symposium on Nasal Polyp; 1984 Oct 5 6; Tokyo.

Proceedings in a language other than English:

> Patologia de cancer de higado. Primera Reunion Germano Espanola de Anatomia Patologica [Pathology of liver cancer. 1st German Spanish Meeting on Pathological Anatomy]; 1988 Sep 23 25; Granada, Spain.

Proceedings also published as part of a journal:

> Symposium on Vaccination against Hepatitis B; 1990 Sep 9; Goteburg, Sweden. (Scandinavian Journal of Infectious Diseases.1991 Supplement; 38).

Note whether the printed proceedings were handsearched or an electronic database was searched.

5.2.2.4 Efforts to identify unpublished studies

Provide a brief summary including databases searched (e.g. SIGLE, National Research Register, HSRProj), giving database details as described in 5.2.2.1. Include also efforts to contact investigators for information about unpublished studies.

5.2.2.5 Other sources

Provide a brief summary of other sources searched (e.g. bibliographies, reference lists and web sites) specifically for the review, giving details of date searched, search terms used, and web sites if relevant.

The search strategies used to develop the specialized register of a CRG are described in their module and should not be reported in the text of Cochrane reviews, but it is helpful to include details of the strategy used to search the specialized register.

5.2.3 Selecting studies

It is generally for authors to decide which study design(s) to include in their review. Most Cochrane reviews include only randomized or quasi-randomized trials (Appendix 5a). Some reviews are more restrictive, and include only randomized trials, while others are less restrictive, and include other study designs as well, particularly when few randomized trials addressing the topic of the review are identified. For example, many of the reviews from the Cochrane Effective Practice and Organization of Care (EPOC) Collaborative Review Group include before-and-after studies and interrupted time series in addition to randomized and quasi-randomized trials.

The process by which studies will be selected for inclusion in a review should be described in the review protocol. The selection of studies for consideration for inclusion in a review is a process that involves several stages. The first stage of checking the results of an electronic search involves assessing titles and abstracts to determine whether each article might meet predetermined eligibility criteria. Authors must decide if more than one of them will assess the records retrieved by electronic databases. There is evidence that using at least two authors has an important effect on reducing the possibility that relevant reports will be discarded (Edwards 2002). If, given the information available, it can be determined that an article definitely does not meet inclusion criteria, it can be rejected. If the title or abstract leave room for doubt that the article cannot definitely be rejected, the full text of the article should be obtained. Reading the full text may lead the authors to exclude the study because it does not meet inclusion criteria. If the article is not rejected, information from it may then be formally extracted as described in Section 7. At all but the last stage of the selection process it is important to err on the side of over-inclusion because once a study has been excluded from the selection process it is unlikely to be reconsidered. Articles about which there is some doubt which are included at one stage can be excluded at a latter stage when more information becomes available.

All reports of studies that are identified as potentially eligible must be assessed to see whether they meet the inclusion criteria for the review. Authors must decide:

- whether more than one author will assess the relevance of each report
- whether the decisions concerning relevance will be made by content area experts, non-experts, or both

- whether the people assessing the relevance of studies will know the names of the authors, institutions, journal of publication and results when they apply the inclusion criteria
- how disagreements will be handled if more than one author applies the criteria to each article

Decisions about which studies to include in a review often involve judgment. To help ensure that these judgments are reproducible, it is desirable for more than one author to apply the inclusion criteria to all the potentially relevant reports that are retrieved. However, the approach used varies from review to review. Whatever the case, the number of people assessing the relevance of each report 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 authors in a particular CRG).

Experts in a particular area frequently have pre-formed opinions that can bias their assessments of both the relevance and validity of articles (Cooper 1989; Oxman 1993b). Thus, while it is important that at least one author is knowledgeable in the area under review, it may be an advantage to have a second author who is not an expert in the area.

Some authors may decide that assessments of relevance should be made by people who are blind or masked 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 (Berlin 1997a; Berlin 1997b). However, this takes much time, and may not be warranted given the resources required and the uncertain benefit in terms of protecting against bias (Berlin 1997b).

Disagreements about whether a study 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 authors. 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 study without additional information. In these cases, authors may choose to categorize the study in their review 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.

One approach to determining which studies to identify in the review as ‘excluded’ is to list any studies about which it is plausible to expect that a reader would question why the study was not included. This covers all studies that apparently meet the selection criteria but have had to be excluded and also any that do not meet all of the criteria but are well known, in the same general area as the review and likely to be thought relevant by some readers. By listing such studies as excluded and giving the reason for exclusion, the author can show that consideration has been given to these studies.

5.2.4 Keeping track of identified studies

Specially designed reference management systems such as ProCite, Reference Manager, and EndNote are useful and relatively easy to use to keep track of reports of studies. ProCite is the most widely used package and the one for which support to editorial bases is most widely available. It is also the preferred database for submitting controlled trials and specialized registers to CENTRAL. ProCite eases the work of identifying duplicate references. In addition, it facilitates storage of information about the methods and process of a search. For example, separate unused fields in ProCite can be used to store 1) when and from whom an article was ordered, and the date of article receipt; 2) reasons for article exclusion; and 3) name of electronic bibliographic database source from which an article was identified.

General database packages such as Access and FoxPro include powerful query capabilities and lend themselves well to customisation, but require some programming and database design skills to set up. An Access-based software (called 'MeerKat') has been developed by the UK Cochrane Centre, in association with Update Software, to address the specific needs of CRGs in managing their specialised registers (http://www.update-software.com/meerkat/). MeerKat allows for a specialized register to be organized around studies, instead of the publications or reports generated from these studies. Each study may have several associated reports. For example, a single randomized trial may have reports that relate to plans for the trial, baseline characteristics of the trial participants, initial results from the trial, and final results from the trial. In MeerKat, each of these reports can be associated with the corresponding study. MeerKat has also been designed specifically to facilitate the work of the Review Group Coordinator/Trials Search Coordinator. For example, MeerKat can produce tables to indicate which records have been assigned to a particular author or topic, and which records have been submitted to CENTRAL. MeerKat also allows complex database searches, including wildcard searches, Boolean searches, and searches of only specific fields. If adopted, MeerKat may ease the task of managing references within a CRG.

5.3 Summary

Conducting a comprehensive, objective, and reproducible search for studies can be the most time-consuming and challenging task in preparing a systematic review. Yet it is also one of the most important. Identifying all relevant studies, and documenting the search for studies with sufficient detail so that it can be reproduced is, after all, largely what distinguishes a systematic review from a traditional narrative review. Although currently it is necessary to search multiple sources to identify relevant published studies, it is envisioned that CENTRAL will eventually become a comprehensive source for published studies, thus reducing the searching burden for authors. Identifying ongoing studies, however, will continue to remain a challenge until a comprehensive, searchable, ongoing trials register is produced to track, organize, and disseminate reports for ongoing studies, as CENTRAL doing for reports of studies that have been published (Lefebvre 2001).

5.4 References

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

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

Berlin 1997a. Berlin JA, Miles CG, Crigliano MD. Does blinding of readers affect the results of meta-analyses? Results of a randomized trial. Online J Curr Clin Trials 1997

Berlin 1997b. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Bliding Study Group [letter]. Lancet 1997; 350: 185-6.

Boynton 1998. Boynton J, Glanville J, McDaid D, Lefebvre. Identifying systematic reviews in MEDLINE: developing an objective approach to search strategy design. Journal of Information Science 1998;24:137-57

Clarke 1999. Clarke M. Feasibility study of the search terms in phase 3 of the Highly Sensitive Search Strategy for the MEDLINE recode project. The Cochrane Collaboration Methods Groups Newsletter 1999;3:20-1. (http://www.cochrane.de/newslett/1999mg.pdf, accessed 28 February 2003)

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

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

Dickersin 1992. 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.

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

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

Dickersin 1997. Dickersin K. How important is publication bias? A synthesis of available data. AIDS Education and Prevention 1997;9 Suppl A:15-21.

Dickersin 2002. Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, McDonald S, and the CENTRAL Development Group. Development of the Cochrane Collaboration's CENTRAL Register of Controlled Clinical Trials. Evaluation and the Health Professions 2002;25:38-64.

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

Edwards 2002. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Statistics in Medicine 2002; 21:1635-40.

Egger 1997. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997;350:326-329.

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

Gregoire 1995. Gregoire G, Derderian F, LeLorier J. Selecting the language of the publications included in a meta-analysis: is there a tower of Babel bias? J Clin Epidemiol 1995; 48.159-63.

Hetherington 1989. 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.

Hopewell 2002. Hopewell S, Clarke M, Lusher A, Lefebvre C, Westby M. A comparison of handsearching versus MEDLINE searching to identify reports of randomized controlled trials. Statistics in Medicine 2002;21:1625-34.

Horton 1997. Horton R. Medical editors trial amnesty. Lancet 1997;350:756.

Juni 2002. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. International Journal of Epidemiology 2002;31:115-23.

Kleijnen 1992. Kleijnen J, Knipschild P. The comprehensiveness of Medline and Embase computer searches. Searches for controlled trials of homeopathy, ascorbic acid for common cold and ginkgo biloba for cerebral insufficiency and intermittent claudication. Pharm Weekbl Sci 1992; 14:316-20.

Lefebvre 2001. Lefebvre C, Clarke M. Identifying randomized trials. In: Egger M, Davey Smith G, Altman D, editors. Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group, 2001:69-86.

Mallett 2002. Mallett S, Hopewell S, Clarke M. Grey literature in systematic reviews: The first 1000 Cochrane systematic reviews. 4th Symposium on Systematic Reviews: Pushing the Boundaries, Oxford, UK, July 2-4, 2002.

McAuley 2000. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet 2000;356:1228-31.

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

Odaka 1992. Odaka T, Nakayama A, Akazawa K, Sakamoto M, Kinukawa N, Kamakura T, et al. The effect of a multiple literature database search a numerical evaluation in the domain of Japanese life science. J Med Syst 1992; 16:77-81.

Oxman 1993b. Oxman AD, Guyatt GH. The science of reviewing research. Ann NY Acad Sci 1993; 703:125-33.

Petrosino 2000. Petrosino A, Boruch RF, Rounding C, McDonald S, Chalmers I. The Campbell Collaboration Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR) to

facilitate the preparation and maintenance of systematic reviews of social and educational interventions. Evaluation and Research in Education 2000;14:206-19.

Ramos-Remus 1994. Ramos-Remus C, Suarez-Almazor M, Dorgan M, Gomez-Vargas A, Russell AS. Performance of online biomedical databases in rheumatology. J Rheumatol 1994; 21(10):1912-21.

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

Robinson 2002. Robinson KA. Dickersin K. Development of a highly sensitive search strategy for the retrieval of controlled trials using PubMed. International Journal of Epidemiology 2002;31:150-3.

Rovers 1993. Rovers JP, Janosik JE, Souney PF. Crossover comparison of drug information online database vendors: Dialog and MEDLARS. Annals of Pharmacotherapy 1993; 27(5):634-9.

Scherer 2003. Scherer RW, Langenberg P. Full publication of results initially presented in abstracts (Cochrane Methodology Review). In: The Cochrane Library, Issue 1, 2003. Oxford: Update Software.

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

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

Smith 1992. Smith BJ, Darzins PJ, Quinn M, Heller RF. Modern methods of searching the medical literature. Med J Aust 1992; 157:603-11.

Song 2000. Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. Publication and related biases. Health Technology Assessment 2000;4(10).

6 Assessment of study quality

Quality assessment of individual studies that are summarised in systematic reviews is necessary to limit bias in conducting the systematic review, gain insight into potential comparisons, and guide interpretation of findings. Factors that warrant assessment are those related to applicability of findings, validity of individual studies, and certain design characteristics that affect interpretation of results. Applicability, which is also called external validity or generalisability by some, is related to the definition of the key components of well-formulated questions outlined in section 4. Specifically, whether a review's findings are applicable to a particular population, intervention strategy or outcome is dependent upon the studies selected for review, and on how the people, interventions and outcomes of interest were defined by these studies and the authors (reviewers).

Interpretation of results is dependent upon the validity of the included studies and other characteristics. For example, a review may summarise twenty valid trials that evaluate the effects of antiischemic agents on symptoms of chest pain in adults with prior myocardial infarction. However, the trials may examine different preparations and doses of antiischemic agents and may have varying durations. These latter issues would affect interpretation though they may not be directly relevant to the internal validity of the trials. Examples of how to abstract data related to applicability and design factors likely to affect the interpretation are in section 7. The remainder of this section will focus on assessing the validity of individual studies included in a systematic review. As most Cochrane reviews focus on randomised trials, it concentrates on how to appraise the validity from these studies.

6.1 Validity

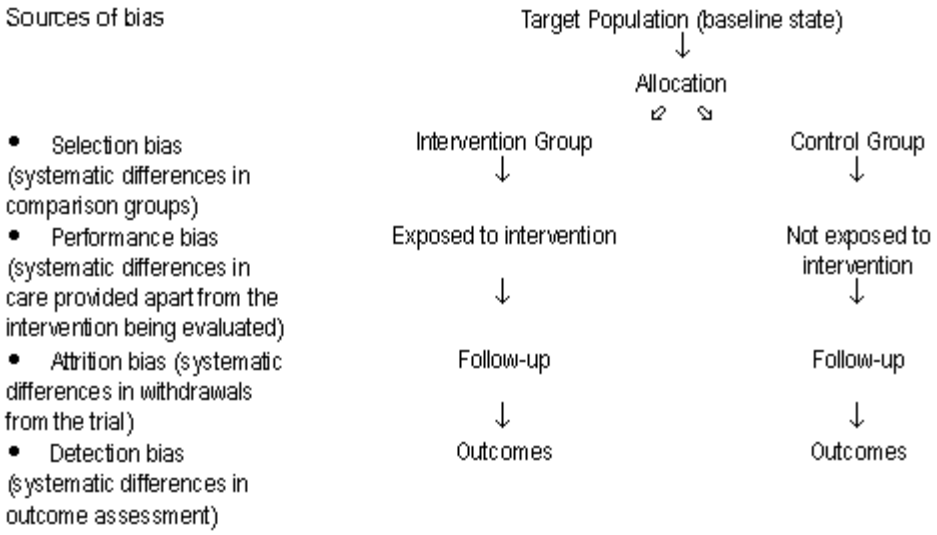
In the context of a systematic review, the validity of a study is the extent to which its design and conduct are likely to prevent systematic errors, or bias (Moher 1995). An important issue that should not be confused with validity is precision. Precision is a measure of the likelihood of chance effects leading to random errors. It is reflected in the confidence interval around the estimate of effect from each study and the weight given to the results of each study when an overall estimate of effect or weighted average is derived. More precise results are given more weight.

Variation in validity can explain variation in the results of the studies included in a systematic review. More rigorous studies may be more likely to yield results that are closer to the 'truth'. Quantitative analysis of results from studies of variable validity can result in 'false positive' conclusions (erroneously concluding an intervention is effective) if the less rigorous studies are biased toward overestimating an intervention's effectiveness. They might also come to 'false negative' conclusions (erroneously concluding no effect) if the less rigorous studies are biased towards underestimating an intervention's effect (Detsky 1992).

It is important to systematically complete critical appraisal of all studies in a review even if there is no variability in either the validity or results of the included studies. For instance, the results may be consistent among studies but all the studies may be flawed. In this case, the review's conclusions would not be as strong as if a series of rigorous studies yielded consistent results about an intervention's effect.

6.2 Sources of bias in trials of healthcare interventions

There are four sources of systematic bias in trials of the effects of healthcare: selection bias, performance bias, attrition bias and detection bias (see figure below). Unfortunately, we do not have strong empirical evidence of a relationship between trial outcomes and specific criteria or sets of criteria used to assess the risk of these biases (Moher 1995, Moher 1996b). There is, however, a logical basis for suspecting such relationships and good reason to consider these four potential biases when assessing studies for a review (Feinstein 1985).



6.3 Selection bias

One of the most important factors that may lead to bias and distort treatment comparisons is that which can result from the way that comparison groups are assembled (Kunz 1998). Using an appropriate method for 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 and the participants themselves should remain unaware of the next assignment in the sequence until after the decision about eligibility has been made. Then, after 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 individuals making the allocation. This will be most securely achieved if an assignment schedule generated using true randomisation is administered by someone who is not responsible for recruiting subjects, such as someone based in a central trial office or pharmacy. If such central randomisation cannot be organised, then other precautions are required to prevent manipulation of the allocation process by those involved in recruitment.

The process of concealing assignment until treatment has been allocated has sometimes been referred to as 'randomisation blinding' (Chalmers 1983). This term does not clearly distinguish concealed allocation from blinding of patients, providers, outcome evaluators and analysts and is unsatisfactory for three reasons. First, the reason for concealing the assignment schedule is to eliminate selection bias. In contrast, blinding (used after the allocation of the intervention) reduces performance and detection biases. Second, from a practical standpoint, concealing allocation up to the point of assignment is always possible, regardless of the study question, but blinding after allocation may be impossible, as in trials comparing surgical with

medical treatment. Third, control of selection bias is relevant to the trial as a whole, and thus to all outcomes being compared. In contrast, control of detection bias is often outcome-specific and may be accomplished successfully for some outcomes in a study but not others. Thus, blinding up to allocation and blinding after allocation are addressing different sources of bias, are inherently different in their practicability and may apply to different components of a study. To clearly distinguish these different forms and purposes of 'blinding', we will refer to the process of concealing assignments as allocation concealment and reserve blinding for measures taken to reduce bias after the intervention has been assigned.

Empirical research has shown that lack of adequate allocation concealment is associated with bias (Chalmers 1983, Schulz 1995, Moher 1998). Indeed, concealment has been found to be more important in preventing bias than other components of allocation, such as the generation of the allocation sequence (e.g., computer, random number table, alternation). Thus, studies can be judged on the method of allocation concealment. Information should be presented that provides some assurance that allocations were not known until, at least, the point of allocation. The method for assigning participants to interventions should be robust against patient and clinician bias and its description should be clear. The following are some approaches that can be used to ensure adequate concealment schemes.

- centralised (e.g. allocation by a central office unaware of subject characteristics) orpharmacy-controlled randomisation
- pre-numbered or coded identical containers which are administered serially to participants
- on-site computer system combined with allocations kept in a locked unreadable computer file that can be accessed only after the characteristics of an enrolled participant have been entered
    - sequentially numbered, sealed, opaque envelopes

Other approaches may include approaches similar to ones listed above, along with reassurance that the person who generated the allocation scheme did not administer it. Some schemes may be innovative and not fit any of the approaches above, but still provide adequate concealment.

Approaches to allocation concealment that should be considered clearly inadequate include: alternation; the use of case record numbers, dates of birth or day of the week, and any procedure that is entirely transparent before allocation, such as an open list of random numbers. When studies do not report any concealment approach, adequacy should be considered unclear. Examples include merely stating that a list or table was used, only specifying that sealed envelopes were used and reporting an apparently adequate concealment scheme in combination with other information that leads the author to be suspicious. When authors enter studies into RevMan they are required to indicate whether allocation concealment was adequate (A), unclear (B), inadequate (C), or that allocation concealment was not used (D) as a criterion to assess validity.

6.4 Performance bias

Performance bias refers to systematic differences in the care provided to the participants in the comparison groups other than the intervention under investigation. To protect against unintended differences in care and placebo effects, those providing and receiving care can be 'blinded' so that they do not know the group to which the recipients of care have been allocated. Some research suggests that such blinding is important in protecting against bias (Karlowski 1975, Colditz 1989, Schulz 1995). Studies have shown that contamination (provision of the intervention to the control group) and cointervention (provision of unintended additional care to either comparison group) can affect study results (CCSG 1978,

Sackett 1979b). Furthermore, there is evidence that participants who are aware of their assignment status report more symptoms, leading to biased results (Karlowski 1975). For these reasons, authors may want to consider the use of ' blinding' as a criterion for validity. This can be done with the following questions: Were the recipients of care unaware of their assigned intervention? Were those providing care unaware of the assigned intervention?

A third question addressing blinding and detection bias is often added: Were persons responsible for assessing outcomes unaware of the assigned intervention? This addresses detection bias, as noted below.

Authors working on topics where blinding is likely to be important may want to develop specific criteria for judging the appropriateness of the method that was used for blinding. In some areas it may be desirable to use the same criterion across reviews, in which case a Collaborative Review Group (CRG) might want to agree to a standard approach for assessing blinding (Chalmers 1989, Schulz 1995, Jadad 1996, Moher 1996b).

6.5 Attrition bias

Attrition bias refers to systematic differences between the comparison groups in the loss of participants from the study. It has been called exclusion bias. It is called attrition bias here to prevent confusion with pre-allocation exclusion and inclusion criteria for enrolling participants. Because of inadequacies in reporting how losses of participants (e.g. withdrawals, dropouts, protocol deviations) are handled, authors should be cautious about implicit accounts of follow-up. The approach to handling losses has great potential for biasing the results and reporting inadequacies cloud this problem. What is reported, or more frequently implied, in study reports on attrition after allocation has not been found to be consistently related to bias (Schulz 1995). Thus authors should be cautious about using reported follow-up as a validity criterion, particularly when it is implied rather than explicitly reported. This is a general recommendation, however, and may not apply to certain topic areas that have higher quality reporting or where it is possible to obtain missing information from investigators.

6.6 Detection bias

Detection bias refers to systematic differences between the comparison groups in outcome assessment. Trials that blind the people who will assess outcomes to the intervention allocation should logically be less likely to be biased than trials that do not. Blinding is likely to be particularly important in research with subjective outcome measures such as pain (Karlowski 1975, Colditz 1989, Schulz 1995). However, at least two empirical studies have failed to demonstrate a relationship between blinding of outcome assessment and study results. This may be due to inadequacies in the reporting of studies (Reitman 1988).

Bias due to the selective reporting of results is somewhat different from bias in outcome assessment. This source of bias may be important in areas where multiple outcome measures are used, such as evaluations of treatments for rheumatoid arthritis (Gotzsche 1989). Therefore, authors may want to consider specification of predefined primary outcomes and analyses by the investigators as indicators of validity. Alternatively, selective reporting of particular outcomes could be taken to suggest the need for better reporting and efforts by authors to obtain missing data.

6.7 Approaches to summarising the validity of studies

6.7.1 Simple approaches

There are several ways to rate validity. One is to rate individual criteria as 'met', 'unmet', or 'unclear' and to use individual criteria, such as adequacy of allocation concealment, in sensitivity analyses (see section 8.10). However, if several explicit criteria are used to assess validity, it is desirable to summarise these so as to derive an overall assessment of how valid the results of each study are. A simple approach to doing this is to use three categories such as the following:

Risk of bias	Interpretation	Relationship to individual criteria
A. Low risk of bias	Plausible 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 partly met
C. High risk of bias	Plausible bias that seriously weakens confidence in the results	One or more criteria not met

The relationships suggested above will most likely be appropriate if only a few assessment criteria are used and if all the criteria address only substantive, important threats to the validity of study results. In general and when possible, authors should obtain further information from the authors of a report when it is unclear whether a criterion was met.

6.7.2 'Quality' scales and checklists

David Moher and his colleagues identified 25 scales and 9 checklists that have been used to assess the validity and 'quality' of randomised controlled trials (Moher 1995, Moher 1996b). These scales and checklists include anywhere from 3 to 57 items and take from 10 to 45 minutes to complete. Almost all of the items in the instruments are based on suggested or 'generally accepted' criteria that are mentioned in clinical trial textbooks. Many of the instruments are liable to confuse the quality of reporting with the validity of the design and conduct of a trial. Moreover, scoring is based on whether something was reported (such as how participants were allocated) rather than whether it was done appropriately in the study. Many also contain items that are not directly related to 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 applicability than validity).

Because there is no 'gold standard' for the 'true' validity 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 scale for assessing the validity of randomised trials, the relationship between such a score and the degree to which a study is free from bias is not obvious. None of the currently available scales for measuring the validity or 'quality' of trials can be recommended without reservation. If authors or CRGs choose to use such a scale, it must be with caution.

Most of the available scales for assessing the validity of randomised controlled trials derive a summary score by adding the scores (with or without differential weights) for each item.

While this approach offers appealing simplicity, it is not supported by empirical evidence (Emerson 1990, Schulz 1995). Notably, scales with multiple items and complex scoring systems take more time to complete than simple approaches. They have not been shown to provide more reliable assessments of validity (Jüni 1999). They may carry a greater risk of confusing the quality of reporting with the validity of the study. They are more likely to include criteria that do not directly measure internal validity, and they are less likely to be transparent to users of the review. For these reasons, it is preferable to use simple approaches for assessing validity that can be fully reported (i.e. how each trial scored on each criterion).

6.8 Bias in non-experimental studies

The Non-randomised Studies Methods Group are preparing guidance on the use of non-randomised studies in Cochrane reviews (Appendix 6a). In the meantime, this section describes some issues that should be considered in assessing the validity of non-randomised studies. The logical reason for focusing on randomised controlled trials in Cochrane reviews is that randomisation is the only means of allocation that controls for unknown and unmeasured confounders as well as those that are known and measured. Differences between comparison groups in prognosis, responsiveness to treatment or exposure to other factors that affect outcomes can distort the apparent magnitude of effects of the intervention of interest. It is possible to control or adjust for confounders that are known and measured in observational studies, such as case-control and cohort studies. However, it is not possible to adjust for those factors that are not known to be confounders or that were not measured. Unfortunately it can rarely, if ever, be assumed that all important factors relevant to prognosis and responsiveness to treatment are known, and for those that are known difficulties can arise in measuring and accounting for them in analyses. Empirical evidence supports these logical concerns (Kunz 1995). Selection bias can distort effects in either direction, causing them to appear either larger or smaller than they are. It is generally not possible to predict the magnitude, and often not even the direction of this bias in specific studies. However, on average, selection bias tends to make treatment effects appear larger than they are and the size of these distortions can be as large or larger than the size of the effects that are being measured (Kunz 1995).

Despite these concerns, there is sometimes good reason to rely on observational studies for information about the effects of healthcare interventions, and to include such studies in Cochrane reviews. For example, well designed observational studies have provided useful data regarding the effects of interventions such as mandatory use of helmets by motorcyclists, screening for cervical cancer, dissemination of clinical practice guidelines to change professional practice and rare adverse effects of medication.

Various criteria have been suggested to critically appraise the validity of observational studies (Horwitz 1979, Feinstein 1982, Levine 1994, Bero 1999). In general, the same four sources of bias noted above can be applied to other types of comparative studies, as illustrated below:

Source of bias	Cohort studies	Case-control studies
Selection bias	Control for confounders	Matching
Performance bias	Measurement of exposure	Measurement of exposure
Attrition bias	Completeness of follow-up	Completeness of follow-up
Detection bias	Blinding	Case definition

Concerns about attrition bias are similar in randomised trials, cohort studies and case-control studies and relate to the extent that all participants in a study are appropriately accounted for in its results. Concerns about detection bias are also similar for cohort studies, and are related to the case definition that is used in case-control studies (since people are entered into such studies based on knowledge of the outcome of interest). The major difference between randomised trials and observational studies has to do with selection bias and the need to identify and account for potential confounders in observational studies. To do this authors

must make judgements about what confounders are important and the extent to which these were appropriately measured and controlled for. Assessing 'performance bias' is also more difficult in observational studies since it is necessary to measure exposure to the intervention of interest and ensure that there were not differences in the exposure of the comparison groups to other factors that could affect outcomes. In addition to considerations of blinding, which are similar to those in randomised trials, it is important to consider whether exposure was measured in a similar and unbiased way in the groups being compared. So, for example, in addition to concerns about bias due to confounders in cohort and case control studies of the effects of post-menopausal hormone replacement therapy, investigators and authors must ensure that use of hormones was measured in an unbiased way.

In summary, a great deal of judgement is necessary in assessing the validity of observational studies. Judgement is also needed when the validity of randomised trials is assessed, but the nature of observational studies makes them even more difficult to critically appraise. This requires a thorough understanding of both the problem that is the focus of the review and methodological considerations. Caution is needed.

6.9 Application of quality assessment criteria

Several basic decisions must be made regarding the assessment of studies, similar to those made regarding the process of selecting studies (section 5.2.3). A prime consideration is the number of authors. Should there be one or more than one? How many are necessary and how many are too many? Will authors review the same articles to maximise reliability or mutually exclusive sets of reports to minimise workload? A concomitant consideration is the backgrounds of the different authors and whether previous training and experience in study design or critical appraisal will be required.

Conducting systematic reviews with multiple authors is a two-sided coin. On the one hand it may limit bias, minimise errors and improve reliability of findings, but having more than one creates the potential for disagreement among authors. When multiple authors will be involved, there should be an explicit procedure or decision rule identified *a priori* for identifying and resolving disagreement. As a general rule, we recommend that at least two authors assess information that involves subjective interpretation and information that is critical to the interpretation of results (e.g., outcome data). Section 7 describes methods for reaching and monitoring consensus when more than one author is used.

Regardless of the number of authors, it is important to first test any assessment criteria that are planned using a pilot sample of articles to ensure that the appraisal criteria can be applied consistently. Three to six papers that span a range of low to high risk bias might provide a suitable sample for this.

Should authors be especially trained in research methods, the content area of a review or both? Although experts in content areas may have pre-formed opinions that can bias their assessments (Oxman 1993b), they may nonetheless give more consistent assessments of the validity of studies than persons without content expertise (Jadad 1996). They may also have valuable insights that are different than those that someone with methodological expertise alone would have. It would seem intuitively desirable to use both content experts and non-experts and to ensure that both have an adequate understanding of the relevant methodological issues.

Authors must also decide whether those assessing study validity will be blinded to the names of the authors, institutions, journal and results of a study when they assess its methods. Some empirical evidence suggests that blind assessment of reports might produce lower and more consistent scores than open assessments (Jadad 1996). Other empirical evidence suggests little benefits from blind assessments (Berlin 1997). However, blinded assessments are very time consuming. Authors must weigh their potential benefits against the costs involved when

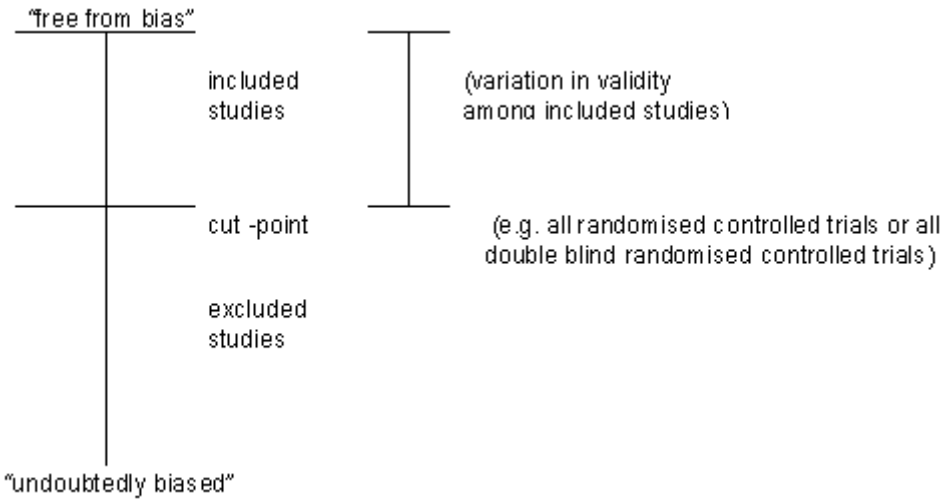
deciding whether or not to blind the authors. Further research is underway comparing blind and open assessments of study validity and these results may help guide this decision.

6.10 Incorporating assessments of study validity in reviews

There are several ways in which validity assessments can be used in a review:

- as a threshold for inclusion of studies
- as a possible explanation for differences in results between studies
- in sensitivity analyses
- as weights in statistical analysis of the study results

Failure to meet one or more validity criteria may indicate such a high risk of bias in some reviews that it constitutes grounds for exclusion of those studies. For example, for highly subjective outcomes such as pain, authors may decide to include only studies that prevent 'performance bias' by blinding participants. The decision about where to set the cut point for inclusion can be conceptualised as existing on a continuum between 'free from bias' and 'undoubtedly biased' as illustrated below:



If authors raise the methodological cut-point for including studies, there will be less variation in validity among the included reports. Assessments of validity would then categorise studies by the risk of bias within the range above the inclusion cut-point. With a sufficiently high cut-point, any variation in validity among included studies may be too small to be important.

There are several methods to examine whether validity may explain differences among study results (Detsky 1992). Visual plots of the results arranged in order of their validity can be used. A second approach is to analyse subgroups of studies above a methodological cut-point, which should, preferably, be specified *a priori*, in the protocol of the review. This approach can be used whether or not the study results are heterogeneous, by doing 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. A third approach is to combine the results of each study sequentially in order of their assessed validity ('cumulative meta-analysis'), examining the impact on the overall results as trials of decreasing validity are included (see section 8.11.6).

A fourth approach is to use statistical methods to weight studies according to their assessed validity or to use 'meta-regression' to explore the relationship between validity and the magnitude of effect across studies (see section 8.8.1). Statistical methods for combining the results of studies generally weight the influence of each study by the inverse of the variance for the estimated measure of effect. In other words, studies with more precise results (narrower confidence intervals) are given more weight. It is also possible to weight studies according to validity so that more valid studies have more influence on the summary result. The main objection to this approach is that there is no empirical basis for determining how much weight to assign to different validity criteria or for quantitatively relating differences on 'quality' scales to differences in the risk of bias between studies.

It is possible using RevMan 4.0 to order studies according to either adequacy of concealment of allocation or 'user defined' assessments of validity. Subgroup analyses based on assessments of validity can be done, although a test of statistical significance of differences between subgroups of studies has not been implemented. RevMan does not include an option for weighting studies by methodological validity and meta-regression is not possible using RevMan 4.0.

6.11 Limitations of quality assessment

There are two major difficulties with assessing the validity of studies. The first is inadequate reporting of trials (SORT 1994, Schulz 1994, WGRR 1994, Begg 1996). It is possible to assume if something was not reported it was not done. However, this is not necessarily correct. Authors should attempt to obtain additional data from investigators as necessary, but this may be difficult. The application of standards for reporting trials (SORT 1994, WGRR 1994, Begg 1996) should facilitate the assessment of study validity in the future.

The second limitation, which in part is a consequence of the first, is limited empirical evidence of a relationship between parameters thought to measure validity and actual study outcomes. As noted above, there is empirical evidence suggesting that, on average, both inadequate concealment of allocation and lack of double blinding result in over-estimates of the effects of treatment. Clearly much more research needs to be done to establish which criteria for assessing validity are indeed important determinants of study results and when. Improved reporting of methods will facilitate such research. Meanwhile, authors should avoid the use of 'quality scores' and undue reliance on detailed quality assessments. It is not supported by empirical evidence, it can be time-consuming and it is potentially misleading.

6.12 References

Begg 1996. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-9.

Berlin 1997. Berlin JA, on behalf of the University of Pennsylvania meta-analysis blinding study group. Does blinding of readers affect the results of meta-analyses? Lancet 1997; 350: 185-6.

Bero 1999. Bero L, Grilli R, Grimshaw J, Mowatt G, Oxman A, Zwarenstein M (editors). Effective Practice and Organisation of Care Module. In: The Cochrane Library, Issue 2, 1999. Oxford: Update Software.

CCSG 1978. The Canadian Cooperative Study Group. The Canadian trial of aspirin and sulfinpyrazone in threatened stroke. N Engl J Med 1978; 299:53-9.

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

Chalmers 1989. Chalmers I, Hetherington J, Elbourne,D, Keirse MJNC, Enkin M. Materials and methods used in synthesizing evidence to evaluate the effects of care during pregnancy and childbirth. In: Chalmers I, Enkin M, Keirse MJNC, editors. Effective Care in Pregnancy and Childbirth. Oxford: Oxford University Press, 1989; 39-65.

Colditz 1989. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: medical. Stat Med 1989; 8:441-54.

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

Emerson 1990. 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.

Feinstein 1982. Feinstein AR, Horwitz RI. Double standards, scientific methods, and epidemiological research. N Engl J Med 1982; 307:1611-7.

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

Gotzsche 1989. Gotzsche 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.

Horwitz 1979. Horwitz RI, Feinstein AR. Methodological standards and contradictory results in case-control research. Am J Med. 1979; 66:556-64.

Jadad 1996. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clin Trials 1996; 17:1-12.

Jüni 1999. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999; 282: 1054-60.

Karlowski 1975. Karlowski TR, Chalmers TC, Frenkel LD, Kapikian AZ, Lewis TL, Lynch JM. Ascorbic acid for the common cold: a prophylactic and therapeutic trial. JAMA. 1975; 231:1038-42.

Kleijnen 1997. Kleijnen J, Gotzsche P, Kunz RA, Oxman AD, Chalmers I. So what's so special about randomization? In: Chalmers, I, Maynard, A, editors. Non-Random Reflections on Health Services Research. London: BMJ, 1997; 93-106.

Kunz 1995. Kunz RA, Oxman AD. Empirical evidence of selection bias in studies of the effects of health care: a systematic review. Presented at the Cochrane Colloquium, Oslo, 5-8 October, 1995.

Kunz 1998. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998; 317:1185-90.

Levine 1994. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, IV: how to use an article about harm. JAMA 1994; 271:1615-9.

Moher 1995. 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. Controlled Clin Trials 1995; 16:62-73.

Moher 1996b. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. Int J Tech Assess in Health Care 1996; 12:195-208.

Moher 1998a. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-13.

Oxman 1993b. Oxman AD, Guyatt GH. The science of reviewing research. Ann NY Acad Sci. 1993;703:125-33.

Reitman 1988. Reitman D, Chalmers TC, Nagalingam R, Sacks H. Can efficacy of blinding be documented by meta-analysis? Presented to the Society for Clinical Trials, San Diego, 23-26 May, 1988.

Sackett 1979b. Sackett DL. Bias in analytic research. J Chronic Dis 1979; 32:51-63.

Schulz 1994. Schulz KF, Chalmers I, Grimes DA, Altman DG. Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynecology journals. JAMA 1994;272:125-8

Schulz 1995. Schulz KF, Chalmers I, Hayes RJ, Altman D. Empirical evidence of bias. JAMA 1995; 273:408-12.

SORT 1994. The Standards of Reporting Trials Group. A proposal for structured reporting of randomized controlled trials. JAMA 1994; 272:1926-31.

WGRR 1994. Working Group on Recommendations for Reporting of Clinical Trials in the Biomedical Literature. Call for comments on a proposal to improve reporting of clinical trials in the biomedical literature. Ann Intern Med 1994; 121:894-5.

7 Collecting data

7.1 Rationale for data collection forms

The data collection form is a bridge between what has been reported by primary investigators (e.g journal articles, project reports, personal communications) and what is ultimately reported by an author (reviewer). The data collection form serves at least three important functions. First, the data collection form is directly linked to the formulated review question and planned assessment of included studies and, therefore, provides a visual representation of these. Second, the data collection form is the historical record of the multitude of decisions (and changes to decisions) that occur throughout the review process. Third, the data collection form is the data repository from which the analysis will emerge.

Given the important functions of data collection forms, ample time and thought should be invested in their designs. Because each review is different, data collection forms will vary across reviews. However, there are similarities regarding types of information that are important, and forms can be adapted from one review to the next.

7.2 Electronic versus paper data collection forms

Should authors design paper data collection forms or automate the review process with electronic data collection forms? Paper forms can be easier to design because electronic forms require computer programming knowledge. On the other hand, large amounts of data from reviews involving large numbers of studies are more easily stored and retrieved with electronic than paper forms. Electronic forms eliminate the need for data entry separate from data abstraction. They also can be used to calculate simple variables or conversions (e.g. pounds to kilograms) for data that is presented in various formats in different studies. Both electronic and paper forms can be designed to provide an historical record of decisions and refinements that occur throughout the review process.

Many authors use a double-abstraction process whereby two independent assessments of each study can be compared and reconciled if necessary. When using a paper data collection form, the comparison process is simple: one form is used to mark and correct errors and disagreements. Comparing double-abstractions using electronic forms is fast but requires the writing and testing of programs within the structure of the database being used. Identifying and addressing errors and disagreements among authors may be more difficult with electronic than paper forms. This is because fields or areas of data collection forms that allow open-ended responses are not easily compared electronically. Amendments or changes to original forms may be more difficult with electronic than paper forms because of programming issues. A final potential drawback to electronic data collection forms is whether they will be compatible with Review Manager (RevMan) which is used to generate and store the final review. Although there are ways to transfer data from electronic data collection forms to RevMan, this might not be straightforward and should, ideally, be planned in advance.

If an electronic form is used, consider the following guides. First, do not program the electronic form until you have designed, piloted, and refined a paper copy of the form. Such pilot tetsting ideally involves more than one author and several articles. Second, when designing the data collection form, consider the needs of the data entry person, structure the form in a logical manner and make coding of responses as consistent and straightforward as possible. Third, when choosing an electronic database or spreadsheet, check whether it can create an electronic file that will be transferable to RevMan. Fourth, don't forget to develop quality control mechanisms for assessing and correcting data entry errors.

7.3 Data management and software

A variety of software and data management programs may be helpful in the systematic review process. Spreadsheet software such as QuatroPro, Excel and Lotus or database programs such as FoxPro or DataEase can be used for electronic data collection forms. Software such as DBMSCOPY may be useful for converting such database files into files compatible with data analysis, if analyses not available in RevMan are planned (see section 8.8).

7.4 Key components of a data collection form

There is no single correct way to design a data collection form. The following suggestions are based on experience. When adapting or designing a data collection form authors first should consider how much information they want to collect. Overly detailed collection can result in forms that are longer than original study reports, tedious and boring to complete, and wasteful of author time. On the other hand, if forms are not sufficiently detailed and omit key data, authors may have to re-abstract studies using supplemental data collection forms. Having to review a study a second time can be frustrating and time-consuming.

7.4.1 Information about study references and authors

Because data collection forms are adaptable across reviews and some authors participate in multiple reviews, a clear title of the review is needed and the name of the author who is abstracting data should be recorded. It is useful to leave space after the title so authors can write notes specific to the study being abstracted. This avoids placing notes, questions or reminders on the last page of the form where they are least likely to be noticed. Important notes may be entered into RevMan in the 'notes' column of the Characteristics of Included Studies table, or in the text of the review. Every Cochrane review is assigned a unique identifier. This should be included next to the title on the data collection form. Forms occasionally have to be revised. Coding the form with a revision date or version number reduces the chances of erroneously using an outdated form by mistake.

Each included study must be given a study identifier that is used in RevMan. Authors may need to collect data from multiple reports of the same trial. It is a good idea to record the source of key information, including where it was found in a report or if information was obtained from unpublished sources and personal communications. Any unpublished information that is used should be written and coded in the same way as published information.

7.4.2 Verification of study eligibility

Although the search and selection process should have weeded out most ineligible studies, it is good to verify study eligibility at the time of data abstraction or collection. Verification information should occur early because the remainder of the form pertains to studies which meet inclusion criteria and the extraction of data from studies that will be excluded is a waste of resources.

Cochrane reviews include an excluded studies table for studies that appear to meet the inclusion criteria and which others might believe to be relevant, but upon closer inspection were excluded. The verification information on the data collection form can be a mechanism for coding reasons why such studies were excluded. For example, an author may only include truly randomised trials in a review. A verification query on the data collection form might be: Randomised? Yes, No, Unclear. If the study used alternate allocation, the answer to the query is no, and this information would be entered in RevMan as the reason for exclusion.

7.5 Study characteristics

When assessing each study, it is necessary to code specific study characteristics. These can be categorized into groups that match information that will be entered into RevMan: methods, participants, interventions, and outcomes. Information under participants might include details relevant to the study setting and diagnostic criteria for the condition of interest. The development of this part of the data collection form deserves careful thought and pilot testing. Data that is collected should be directly linked to the review question(s) and planned analysis strategies. It should be collected in a format conducive to logical entry into RevMan.

7.5.1 Methods

Different research methods can influence study outcomes by introducing bias and artefacts in study results. For example, whether allocation was adequately concealed is important, as discussed in section 6. When entering information about particular studies in RevMan, it will be necessary to code allocation concealment as adequate (A), unclear (B), inadequate (C) or not used (D). Data collection forms should reflect these assessments. Other methods features that may be relevant include study duration; type of trial such as parallel or cross-over design; patient, provider and outcome assessor blinding; amount of drop-outs and cross-overs; cointerventions and other potential confounders. The methods part of the data collection form should include any validity criteria that are used.

7.5.2 Participants

Characteristics of participants may vary substantially across studies and some Collaborative Review Groups (CRGs) have developed standards regarding which characteristics should be collected. Typically, items that should be collected are those that could affect study results or help users assess applicability. For example, if the author has reason to suspect important treatment effect differences between various ethnic populations, this information should be collected. If treatment effects are thought constant over ethnic groups, and if such information would not be useful to help apply results, it should not be collected. Items that are often useful for assessing applicability include age and sex. Occasionally, other sociodemographic items such as education level are important as well as items addressing the presence of important comorbid conditions.

If the settings of studies are likely to influence treatment effects or applicability, they should be assessed. Typical settings that are involved in healthcare intervention studies are: acute care hospitals, emergency facilities, offices or clinics, extended care facilities such as boarding and nursing homes, and communities. Sometimes studies are conducted in different geographical regions that have important differences in cultural characteristics that could affect delivery of an intervention and its outcomes. Sometimes temporal settings indicate important technology differences. If such items are important for the interpretation of the review, they should be assessed.

Diagnostic criteria that were used to define the condition of interest can be a particularly important source of clinical heterogeneity and should be described. For example, in a review of drug therapy for congestive heart failure, it is important to know how the definition and severity of heart failure was determined in each study (e.g. systolic or diastolic dysfunction, severe systolic dysfunction with ejection fractions of < 20%, etc.). Similarly, in a review of antihypertensive therapy, it is important to describe baseline levels of blood pressure of participants.

7.5.3 Interventions

The intervention and how it was delivered should be described. For trials of pharmaceutical agents, routes of delivery (e.g., oral, intravenous), doses, and timing (e.g. within 24 hours of diagnosis) may be assessed. Treatment length also may be recorded here, particularly if it was different than study follow-up length and was not recorded under methods. For complex interventions such as those that evaluate psychotherapy, behavioural and educational approaches or healthcare delivery strategies, it is important to collect information that will help to disentangle the underlying relationships. This includes information about who delivered the intervention, its contents, format, timing, etc.

For trials that do not utilise placebos and those that evaluate complicated interventions, it is also important to collect information regarding what was given to the control group. This will help guide later decisions about whether it is reasonable to combine data across studies; since marked heterogeneity in what is received by control groups may be a reason for not combining studies, or for doing sensitivity analyses.

7.5.4 Outcome measures and results

What may appear to be obvious and simple may in fact be one of the more difficult sections of the data collection form to design. Reports of studies often include more than one outcome (mortality, morbidity, quality of life, etc.), may report the same outcome using different measures, may include outcomes for subgroups and may report outcomes measured at different points in time. The author needs to integrate what type of outcome information is needed to answer the review's question(s) with what is likely to be in the reports of studies. To avoid hidden mistakes outcomes should be collected in the format they were reported and transformed in a subsequent step. For cross-over trials and trials with outcome assessments at various periods of follow-up, decisions will need to be made about which outcomes to assess (see section 8.11.5 and section 8.9.1 respectively).

Authors should consider formatting the forms to match RevMan data tables. For example, if the author plans to use continuous data, the following information is required for each comparison group: the number of participants, the mean and the standard deviation. However, these data fields may be insufficient because there is great variation in what researchers report and fail to report. In this example, investigators may have reported a confidence interval for the mean difference and not reported any standard deviations, or they may simply have reported the value of a test statistic (t test, F test, chi-square test, etc.) or a p-value. Data collection forms should incorporate flexibility for addressing this type of variability in outcome assessments. For more detail, regarding what outcome information is necessary for specific types of analyses, see section 8.

7.6 Coding format and instructions for coders

Accurate coding is extremely important. The coding should not be so complicated that the abstractor is easily confused or likely to make poor decisions. Authors need instructions and decision rules on the data collection form. There are varying preferences regarding where instructions should be included. One approach is to insert the instruction adjacent or near to the data field that is to be coded. In some cases, instructions can be lengthy and may have to be placed on a separate page. Regardless of the approach used (most likely it will be a mixture), it is crucial for authors to practice using the form and receive, or give, training if the form was designed by someone other than the person using it.

7.7 Pilot testing and form revisions

All forms should be pilot tested using a representative sample of the studies to be reviewed. This test is likely to identify data that are not needed or are missing. Abstractors may provide feedback that certain coding instructions are confusing or incomplete (e.g. all of the types of responses might not be described). When multiple authors are participating on a project, there may need to be a consensus among them before the form is modified to avoid any misunderstandings or later disagreements. Depending on the complexity of the review and the experiences from piloting, additional pilot tests may be necessary.

Problems with the data collection form will occasionally surface after pilot testing has been completed and the form has been revised. In fact, it is rare for a data collection form to not require any modifications after it has been piloted. When changes have to be made to the form or coding instructions, be sure to correct the forms of those studies that have already been reviewed. In some situations, it may only be necessary to clarify coding instructions without modifying the actual data collection form.

7.8 Reliability of data collection

Reliability refers here to the degree to which different people review a study in the same way. For example, did each author agree on the presence of comorbidity among subjects in a specific trial? Did authors agree on the outcome data in each comparison group? When more than one person is reviewing data, there will inevitably be disagreements. Multiple authors need to develop a plan for comparing information in their data collection forms and for reaching consensus when there are disagreements. Consensus can be achieved by discussion among authors or by using an additional independent arbitrator. It is also important to plan how the 'consensus' agreement will be recorded. There are at least three possibilities: 1) use one author's form and record changes after consensus in red ink; 2) use a separate printed form; or 3) enter only the consensus data onto an electronic form. Keeping the 'consensus' information separate is essential for assessing the reliability of coding.

It may not be important to formally examine reliability for all of the collected data; for example, an author may elect to limit the evaluation of reliability to the coding of outcomes and for validity assessments. There is no fixed standard for the degree of reliability that is adequate or how to assess reliability. However, it is important to examine reliability throughout the data collection process. For example, if after reaching consensus on the first few studies, the authors note a frequent disagreement for specific data, then coding instructions may need modification. Authors may display 'coder drift' (a change over time in how information is coded), indicating a possible need for re-training or re-coding. This can only be identified when reliability is examined throughout the project.

7.9 Blinded data extraction

UNDER CONSTRUCTION – A section is being prepared on the issue of whether data extraction should be done blinded; for example to the authors and journal and to the results when assessing quality. Although there is some evidence that blinded assessments of the quality of trials may be more reliable and different from assessments that are not blinded (Jadad 1996, Moher 1998b), blinding is difficult to achieve, time consuming and may not substantially alter the results of a review (Berlin 1997a, Berlin 1997b).

7.10 Collection of data from investigators

Authors will often find that they are unable to extract all of the information they are interested in from published reports, both with regard to the details of the study and its numerical results. In such circumstances, the authors need to determine how to collect the missing information. They might wish to contact the original investigators and should, for example, consider whether they will contact them with an open-ended request, send them their standard data collection form, request individual patient data (see section 11) or do something else.

7.11 References

Jadad 1996. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clin Trials 1996; 17:1-12.

Moher 1998b. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses?. Lancet 1998; 352:609–13.

Berlin 1997a. Berlin JA, Miles CG, Crigliano MD. Does blinding of readers affect the results of meta-analyses? Results of a randomized trial. Online J Curr Clin Trials 1997.

Berlin 1997b. Berlin JA. Does blinding of readers affect the results of meta-analyses? University of Pennsylvania Meta-analysis Blinding Study Group. Lancet 1997; 350: 185-6.

8 Analysing and presenting results

Edited by Jonathan J Deeks, Julian PT Higgins and Doug G Altman on behalf of the Cochrane Statistical Methods Group.

Do not start here! Please consult Sections 2 to 6 before reading this Section. It can be tempting to jump prematurely into a statistical analysis when undertaking a systematic review. The production of a diamond at the bottom of a plot is an exciting moment for many authors (reviewers), but results of meta-analyses can be very misleading if suitable attention has not been given to formulating the review question; specifying inclusion criteria; identifying, selecting and critically appraising studies; collecting appropriate data; and deciding what would be meaningful to analyse.

Within this section 'RevMan' is used to refer to the Cochrane Collaboration's Review Manager software including its statistical component, which is now called RevMan Analyses. Previous versions of RevMan used a statistical program called MetaView, which is currently one option for viewing graphs in The Cochrane Library. Thus people reading a review may see a slightly different output to that the reviewer sees in RevMan.

8.1 Planning the analysis

While in primary studies the investigators select and collect data from individual patients, in systematic reviews the investigators select and collect data from primary studies. While primary studies include analyses of their patients, Cochrane reviews contain analyses of the primary studies. Analyses may be narrative, such as a structured summary and discussion of the studies' characteristics and findings, or quantitative, that is involving statistical analysis. **Meta-analysis** – the statistical combination of results from two or more separate studies – is the most commonly used statistical technique. Cochrane review writing software (RevMan) can perform a variety of meta-analyses, but it must be stressed that meta-analysis is not appropriate in all Cochrane reviews. Issues to consider when deciding whether a meta-analysis is appropriate in your review are discussed in this section and in 8.1.2 When not to use meta-analysis in a review.

Studies comparing health care interventions, notably randomised trials, use the outcomes of participants to compare the effects of different interventions. Meta-analyses focus on pair wise comparisons of interventions, such as an experimental intervention versus a control intervention, or the comparison of two experimental interventions. The terminology used throughout this section of the Handbook (experimental versus control interventions) implies the former, but is intended to include the latter.

The contrast between the outcomes of two groups treated differently is known as the effect or the treatment effect. Whether analysis of included studies is narrative or quantitative, a general framework for synthesis may be provided by considering four questions:

1. What is the direction of effect?
2. What is the size of effect?
3. Is the effect consistent across studies?
4. What is the strength of evidence for the effect?

Meta-analysis provides a statistical method for (1)-(3). Assessment of (4) relies additionally on judgements based on assessments of study design and study quality, as well as statistical measures of uncertainty.

Narrative synthesis uses subjective (rather than statistical) methods to follow through stages (1)-(4) for reviews where meta-analysis is either not feasible or not sensible. In a narrative

synthesis the method used for each stage should be pre-specified, justified and followed systematically. Bias may be introduced if the results of one study are inappropriately stressed over those of another.

The analysis plan follows from the scientific aim of the review. Reviews have different types of aims, and may therefore contain different approaches to analysis.

1. The most straightforward Cochrane review assembles studies that make one particular comparison between two treatment options, for example, comparing inhaled steroids with placebo for bronchiectasis. Meta-analysis and related techniques can be used if there is a consistent outcome measure to:
    i. establish whether there is evidence of an effect;
    ii. estimate the size of the effect and the uncertainty surrounding that size; and
    iii. investigate whether the effect is consistent across studies.
2. Some reviews may have a broader focus than a single comparison. The first is where the intention is to identify and collate all studies in a particular field. An example of such a review is that of topical treatments for fungal infections of the skin and nails of the foot, which included studies of any topical treatment. The second, related aim is that of identifying a 'best' intervention. A review of interventions for emergency contraception sought that which was most effective (while also considering potential adverse effects). Such reviews may include multiple comparisons and meta-analyses between all possible pairs of treatments, and require care when it comes to planning analyses – see 8.1.4 Which comparisons should be made?
3. Occasionally review comparisons have particularly wide scopes that make the use of meta-analysis problematic. For example, a review of media-based behavioural treatments for behavioural disorders in children covers diverse media-based treatments (including written material and film) and diverse behavioural problems (including Attention Deficit/Hyperactivity Disorder and enuresis). When reviews contain very diverse studies a meta-analysis might be useful to answer the overall question of whether there is evidence that, for example, media-based treatments can work (but see 8.1.2 When not to use meta-analysis in a review). But use of meta-analysis to describe the size of effect may not be meaningful if the implementations are so diverse that an effect estimate cannot be interpreted in any specific context.
4. An aim of some reviews is to investigate the relationship between the size of an effect and some characteristic(s) of the studies. This is uncommon as a primary aim in Cochrane reviews, but may be a secondary aim. For example, in the review of inhaled steroids for bronchiectasis, there was interest in whether the administered dose of steroid affected its efficacy. Such investigations of heterogeneity need to be undertaken with care: see 8.8 Investigating heterogeneity.

8.1.1 Why perform a meta-analysis in a review?

The value a meta-analysis can add to a review depends on the context in which it is used, as described in 8.1 Planning the analysis. Reasons for considering including a meta-analysis in a review are:

1. To increase power. Power is the chance of detecting a real effect as statistically significant if it exists. Many individual studies are too small to detect small effects, but when several are combined there is a higher chance of detecting an effect.
2. To improve precision. The estimation of a treatment effect can be improved when it is based on more information.
3. To answer questions not posed by the individual studies. Primary studies often involve a specific type of patient and explicitly defined interventions. A selection of

studies in which these characteristics differ can allow investigation of the consistency of effect and, if relevant, allow reasons for differences in effect estimates to be investigated.

4. To settle controversies arising from apparently conflicting studies or to generate new hypotheses. Statistical analysis of findings allows the degree of conflict to be formally assessed, and reasons for different results to be explored and quantified.

Of course, the use of statistical methods does not guarantee that the results of a review are valid, any more than it does for a primary study. Moreover, like any tool, statistical methods can be misused.

8.1.2 When not to use meta-analysis in a review

If used appropriately, meta-analysis is a powerful tool for deriving meaningful conclusions from data and can help prevent errors in interpretation. However, there are situations in which a meta-analysis can be more of a hindrance than a help. A common criticism of meta-analyses is that they 'combine apples with oranges'. If studies are clinically diverse then a meta-analysis may be meaningless, and genuine differences in effects may be obscured. A particularly important type of diversity is in the comparisons being made by the primary studies. Often it is nonsensical to combine all included studies in a single meta-analysis: sometimes there is a mix of comparisons of different treatments with different comparators, each combination of which may need to be considered separately. Further, it is important not to combine outcomes that are too diverse.

Decisions concerning what should and should not be combined are inevitably subjective, and are not amenable to statistical solutions but require discussion and clinical judgement. In some cases consensus may be hard to reach.

Meta-analyses of poor quality studies may be seriously misleading. If bias is present in each (or some) of the individual studies, meta-analysis will simply compound the errors, and produce a 'wrong' result that may be interpreted as having more credibility.

Finally, meta-analyses in the presence of serious publication and/or reporting biases may produce an inappropriate summary.

8.1.3 What does a meta-analysis entail?

While the use of statistical methods in reviews can be extremely helpful, the most essential element of an analysis is a thoughtful approach, to both its narrative and quantitative elements. This entails consideration of the following questions:

1. Which comparisons should be made?
2. Which study results should be used in each comparison?
3. What is the best summary of effect for each comparison?
4. Are the results of studies similar within each comparison?
5. How reliable are those summaries?

The first step in addressing these questions is to decide which comparisons to make (8.1.4 Which comparisons should be made?). The next step is to prepare tabular summaries of the characteristics and results of the studies that are included in each comparison (8.2 Types of data and effect measures, 8.5 Extraction of study results). It is then possible to derive estimates of effect across studies in a systematic way (8.6 Summarising effects across studies), to measure and investigate differences among studies (8.7 Heterogeneity) and to interpret the findings and conclude how much confidence should be placed in them (8.X Issues in interpretation).

8.1.4 Which comparisons should be made?

The first and most important step in planning the analysis is to specify the pair wise comparisons that will be made. The comparisons addressed in the review should relate clearly and directly to the questions or hypotheses that are posed when the review is formulated (see Section 4). It should be possible to specify in the protocol of a review the main comparisons that will be made. However, it will often be necessary to modify comparisons and add new ones in light of the data that are collected. For example, important variations in the intervention may only be discovered after data are collected.

Decisions about which studies are similar enough for their results to be grouped together require an understanding of the problem that the review addresses, and judgement by the author and the user. The formulation of the questions that a review addresses is discussed in Section 4. Essentially the same considerations apply to deciding which comparisons to make, which outcomes to combine and which key characteristics (of study design, participants, interventions and outcomes) to consider when investigating variation in effects (heterogeneity). These considerations must be addressed when setting up the Table of Comparisons in RevMan and in deciding what information to put in the table of Characteristics of Included Studies.

8.1.5 Writing the analysis section of the protocol

The analysis section of a Cochrane review protocol may be more susceptible to change than other protocol sections (such as criteria for including studies and how methodological quality will be assessed). It is rarely possible to anticipate all the statistical issues that may arise, for example, finding outcomes that are similar but not the same as each other; outcomes measured at multiple or varying time-points; and use of concomitant treatments

However the protocol should provide a strong indication as to how the author will approach the statistical evaluation of studies' findings. At least one member of the review team should be familiar with the majority of the contents of Section 8 when the protocol is written. As a guideline we recommend that the following be addressed (more details of all the issues may be found in the rest of Section 8):

1. ensure that the analysis strategy firmly addresses the stated objectives of the review (8.1 Planning the analysis);
2. consider which types of study design would be appropriate for the review. Parallel group trials are the norm, but other randomized designs may be appropriate to the topic (e.g. cross-over trials, cluster randomized trials, factorial trials). Decide how such studies will be addressed in the analysis (See 8.11.1 Publication bias and funnel plots)
3. decide whether a meta-analysis is intended and consider how the decision as to whether a meta-analysis is appropriate will be made (8.1.1 Why perform a meta-analysis in a review? 8.1.2 When not to use meta-analysis in a review);
4. determine the likely nature of outcome data (e.g. dichotomous, continuous etc) (8.2 Types of data and effect measures);
5. consider whether it is possible to specify in advance what treatment effect measures will be used (e.g. risk ratio, odds ratio or risk difference for dichotomous outcomes, mean difference or standardised mean difference for continuous outcomes) (8.6.3.4 Which measure for dichotomous outcomes? 8.6.4.1 Which measure for continuous outcomes?);
6. decide how statistical heterogeneity will be identified (8.7.2 Identifying and measuring heterogeneity);

7. decide whether random effects meta-analyses, fixed effect meta-analyses or both methods will be used for each planned meta-analysis (8.7.4 Incorporating heterogeneity into random effects models);
8. consider how clinical and methodological diversity (heterogeneity) will be assessed and whether (and how) these will be incorporated into the analysis strategy (8.7 Heterogeneity and 8.8 Investigating heterogeneity);
9. decide how quality of included studies will be assessed and addressed in the analysis (Section 6, Assessing trial quality);
10. pre-specify characteristics of the studies that may be examined as potential causes of heterogeneity. (8.8.4 Selection of study characteristics for subgroup analyses and meta-regression);
11. consider how missing data will be handled (e.g. imputing data for intention-to-treat analyses) (8.X Missing data);
12. decide whether (and how) evidence of possible publication and/or reporting biases will be sought (8.11.1 Publication bias and funnel plots).
13. It may become apparent when writing the protocol that additional expertise is likely to be required: see 8.X Where to go for help.

8.2 Types of data and effect measures

The starting point of all meta-analyses of studies of effectiveness involves the identification of the data type for the outcome measurements.

Through Section 8 we consider outcome data to be of five different types:

1. Dichotomous (or binary) data, where each individual's outcome is one of only two possible categorical responses;
2. Continuous data, where each individual's outcome is a measurement of a numerical quantity;
3. Ordinal data (including measurement scales), where the outcome is one of several ordered categories, or generated by scoring and summing categorical responses;
4. Counts and rates calculated from counting the number of events that each individual experiences;
5. Time-to-event (typically survival) data that analyse the time until an event occurs, but where not all individuals in the study experience the event (censored data).

The ways in which the effect of a treatment can be measured depends on the nature of the data being collected. In this section we briefly examine the types of outcome data that might be encountered in systematic reviews of clinical trials, and review definitions, properties and interpretation of standard measures for treatment effect. In Section 8.6.3.4 Which measure for dichotomous outcomes? and Section 8.6.4.1 Which measure for continuous outcomes? we discuss issues in the selection of one of these measures for a particular meta-analysis.

8.2.1 Effect measures for dichotomous outcomes

Dichotomous outcome data arise when the outcome for every participant is one of two possibilities, for example, dead or alive, or clinical improvement or no clinical improvement. This section considers the possible summary statistics when the outcome of interest has such a binary form. The most commonly encountered effect measures used in clinical trials with dichotomous data are:

- the risk ratio (RR) (also called the relative risk);

- the odds ratio (OR);
- the risk difference (RD) (also called the absolute risk reduction, ARR);
- the number needed to treat (NNT).

Details of the calculations of the first three of these measures are given in Box 8.2.1. Numbers needed to treat are discussed in detail in 8.X Re-expressing meta-analysis results as NNTs.

Aside: As events may occasionally be desirable rather than undesirable, it would be preferable to use a more neutral term than risk (such as probability), but for the sake of convention we use the terms risk ratio and risk difference throughout. We also use the term 'risk ratio' in preference to 'relative risk' for consistency with other terminology. The two are interchangeable and both conveniently abbreviate to 'RR'. Note also that we have been careful with the use of the words 'risk' and 'rates'. These words are often treated synonymously. However, we have tried to reserve use of the word 'rate' for the data type 'counts and rates' where it describes the frequency of events in a measured period of time.

Box 8.2.1 Calculation of RR, OR and RD from a 2´2 Table

The results of a clinical trial can be displayed as a 2´2 table:

	Event	No event	Total
Intervention	a	b	a+b
Control	c	d	c+d

where a, b, c and d are the numbers of participants with each outcome in each group. The following summary statistics can be calculated:

$$\text{risk ratio} = \frac{\text{risk of event in intervention group}}{\text{risk of event in control group}} = \frac{a/(a+b)}{c/(c+d)}$$

$$\text{odds ratio} = \frac{\text{odds of event in intervention group}}{\text{odds of event in control group}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

$$\text{risk difference} = \text{risk of event in intervention group - risk of event in control group}$$

$$= \frac{a}{a+b} - \frac{c}{c+d}$$

8.2.1.1 Risk and odds

In general conversation the terms 'risk' and 'odds' are used interchangeably (as are the terms 'chance', 'probability' and 'likelihood') as if they describe the same quantity. In statistics,

however, risk and odds have particular meanings and are calculated in different ways. When the difference between them is ignored the results of a systematic review may be misinterpreted.

Risk is the concept more familiar to patients and health professionals. Risk describes the probability with which a health outcome (usually an adverse event) will occur. In research, risk is commonly expressed as a decimal number between 0 and 1, although it is occasionally converted into a percentage. It is simple to grasp the relationship between a risk and the likely occurrence of events: in a sample of 100 people the number of events observed will on average be the risk multiplied by 100. For example, when the risk is 0.1, about ten people out of every 100 will have the event, when the risk is 0.5, about 50 people out of every 100 will have the event.

Odds is a concept that is more familiar to gamblers. The odds is the ratio of the probability that a particular event will occur to the probability that it will not occur, and can be any number between zero and infinity. In gambling, the odds describes the ratio of the size of the potential winnings to the gambling stake; in health care it is the ratio of the number of people with the event to the number without. It is commonly expressed as a ratio of two integers. For example, an odds of 0.01 is often written as 1:100, odds of 0.33 as 1:3, and odds of 3 as 3:1. Odds can be converted to risks, and risks to odds, using the formulae:

$$\text{risk} = \frac{\text{odds}}{1 + \text{odds}} \quad ; \quad \text{odds} = \frac{\text{risk}}{1 - \text{risk}}$$

The interpretation of an odds is more complicated than for a risk. The simplest way to ensure that the interpretation is correct is to first convert the odds into a risk. For example, when the odds are 1:10, or 0.1, one person will have the event for every 10 who do not, and, using the above formula, the risk of the event is 0.1/(1+0.1) = 0.091. In a sample of one hundred, about nine individuals will have the event and 91 will not. When the odds is equal to one, one person will have the event for every one who does not, so in a sample of 100, 100 ´ 1/(1+1) = 50 will have the event and 50 will not.

The difference between odds and risk is small when the event is rare (as illustrated in the first example above where a risk of 0.091 was seen to be similar to an odds of 0.1). When events are common, as is often the case in clinical trials, the differences between odds and risks are large. For example, a risk of 0.5 is equivalent to an odds of 1; and a risk of 0.95 is equivalent to odds of 19.

Measures of effect for clinical trials with dichotomous outcomes involve comparing either risks or odds from two treatment groups. To compare them we can look at their ratio (risk ratio or odds ratio) or their difference in risk (risk difference).

8.2.1.2 Measures of relative effect: the risk ratio and odds ratio

Measures of relative effect express the outcome in one group relative to that in the other. The risk ratio (relative risk) is the ratio of the risk of an event in the two groups whereas the odds ratio is the ratio of the odds of an event (Box 8.2.1). For both measures a value of one indicates that the estimated effects are the same for both treatments.

Neither the risk ratio nor the odds ratio can be calculated for a trial if there are no events in the control group. This is because, as can be seen from the formulae in box 8.2.1, we would be trying to divide by zero. The odds ratio also cannot be calculated if everybody in the intervention group experiences an event. In these situations, and others where standard errors cannot be computed, it is customary to add ½ to each cell of the 2x2 table (RevMan automatically makes this correction when necessary). In the case where no events (or all

events) are observed in both groups the trial provides no information about relative probability of the event and is automatically omitted from the meta-analysis. This is entirely appropriate. Zeros arise particularly when the event of interest is rare – such events are often unintended adverse outcomes. For further discussion of choice of effect measures for such sparse data (often with lots of zeros) see 8.X Rare events (including zero frequencies).

Risk ratios describe the multiplication of the risk that occurs with use of the intervention. For example, a risk ratio of 3 implies that events with treatment are three times more likely than events without treatment. Alternatively we can say that treatment increases the risk of events by 100 × (RR – 1)% = 200%. Similarly a risk ratio of 0.25 is interpreted as the probability of an event with treatment being one-quarter of that without treatment. This may be expressed alternatively by saying that treatment decreases the risk of events by 100 × (1 –RR)% = 75%. This is known as the relative risk reduction. The interpretation of the clinical importance of a given risk ratio cannot be made without knowledge of the typical risk of events without treatment: a risk ratio of 0.75 could correspond to a clinically important reduction in events from 80% to 60%, or a small, less clinically important reduction from 4% to 3%.

The numerical value of the observed risk ratio must always be between 0 and 1/ CGR, where CGR (abbreviation of 'control group risk', sometimes referred to as the CER or control event rate) is the observed risk of the event in the control group (expressed as a number between 0 and 1). This means that for common events large values of risk ratio are impossible. For example, when the observed risk of events in the control group is 0.66 (or 66%) then the observed risk ratio cannot exceed 1.5. This problem applies only for increases in risk, and causes problems only when the results are extrapolated to risks above those observed in the trial

Odds ratios, like odds, are more difficult to interpret (Sackett 1996, Sinclair 1994). Odds ratios describe the multiplication of the odds of the outcome that occur with use of the intervention. To understand what an odds ratio means in terms of changes in numbers of events it is simplest to first convert it into a risk ratio, and then interpret the risk ratio in the context of a typical baseline risk (BR) without treatment, as outlined above. Formulae for converting an odds ratio to a risk ratio, and vice versa, are:

$$RR = \frac{OR}{1-(BR \times (1 - OR))} \quad ; \quad OR = \frac{RR\ (1 - BR)}{1 - (BR \times RR)} \quad ,$$

where BR is the typical risk of an event without treatment (as a number between 0 and 1). Please note that this conversion requires specification of a value of BR. Often the value of CGR is used, but use of different values of baseline risk will give different answers when the conversion is made. Sometimes it may be sensible to calculate the RR for more than one value of the BR.

8.2.1.3 Warning: OR and RR are not the same

Because risk and odds are different when events are common, the risk ratio and the odds ratio also differ when events are common. The non-equivalence of the risk ratio and odds ratio does not indicate that either is wrong: both are entirely valid ways of describing a treatment effect. Problems may arise, however, if the odds ratio is misinterpreted as a risk ratio. For treatments that increase the chances of events, the odds ratio will be larger than the risk ratio, so the misinterpretation will tend to overestimate the treatment effect, especially when events are common (with, say, risks of events more than 20%). For treatments that reduce the chances of events, the odds ratio will be smaller than the risk ratio, so that again misinterpretation overestimates the effect of treatment. This error in interpretation is

unfortunately quite common in published reports of individual studies and systematic reviews.

8.2.1.4 Measure of absolute effect: the risk difference

The risk difference is the difference between the observed risks (proportions of individuals with the outcome of interest) in the two groups (Box 8.2.1). The risk difference can be calculated for any trial, even when there are no events in either group. The risk difference is straightforward to interpret: it describes the actual difference in the risk of events that was observed with treatment and with control; for an individual it describes the estimated difference in the probability of experiencing the event. However, the clinical importance of a risk difference may depend on the underlying risk of events. For example, a risk difference of 0.02 (or 2%) may represent a small, clinically insignificant change from a risk of 58% to 60% or a proportionally much larger and potentially important change from 1% to 3%. Although there are arguments that the risk difference provides more complete information than relative measures (Sackett 1997, Laupacis 1988) it is still important to be aware of the underlying risk of events and consequences of the events when interpreting a risk difference.

The risk difference is naturally constrained (like the risk ratio), which may create difficulties when applying results to other patient groups and settings. For example, if a trial or meta-analysis estimates a risk difference of –0.1 (or –10%), then for a group with an initial risk of, say, 7% the outcome will have an impossible estimated negative probability of –3%. Similar scenarios for increases in risk occur at the other end of the scale. Such problems can arise only when the results are applied to patients with different risks from those observed in the trial(s).

The number needed to treat is obtained from the risk difference. Although it is often used to summarise results of clinical trials, NNTs cannot be combined in a meta-analysis (see 8.6.3.4 Which measure for dichotomous outcomes?).

8.2.1.5 What is the event?

In the context of dichotomous outcomes, health care interventions are intended either to reduce the risk of occurrence of an adverse outcome or increase the chance of a good outcome. All of the effect measures described above apply equally to both scenarios.

In many situations it is natural to talk about one of the outcome states as being an event. For example, when participants have particular symptoms at the start of the trial the event of interest is usually recovery or cure. If participants are well or alternatively at risk of some adverse outcome at the beginning of the trial, then the event is the onset of disease or occurrence of the adverse outcome. Because the focus is usually on the experimental intervention group, a trial in which the experimental intervention reduces the occurrence of an adverse outcome will have an odds ratio and risk ratio less than one, and a negative risk difference. A trial in which the experimental intervention increases the occurrence of a good outcome will have an odds ratio and risk ratio greater than one, and a positive risk difference (see Box 8.2.1).

However, it is possible to switch events and non-events and consider instead the proportion of patients not recovering or not experiencing the event. For meta-analyses using risk differences or odds ratios the impact of this switch is of no great consequence: the switch simply changes the sign of a risk difference, whilst for odds ratios the new odds ratio is the reciprocal (1/x) of the original odds ratio.

By contrast, switching the outcome can make a substantial difference for risk ratios, affecting the effect estimate, its significance, and the consistency of treatment effects across studies. This is because the precision of a risk ratio estimate differs markedly between situations with low risks of events and situations with high risks of events. In a meta-analysis the effect of this reversal cannot easily be predicted. The identification, before data analysis, of which risk

ratio is more likely to be the most relevant summary statistic is therefore important and discussed further in 8.6.3.4 Which measure for dichotomous outcomes?.

8.2.2 Effect measures for continuous outcomes

The term ‘continuous’ in statistics conventionally refers to data that can take any value in a specified range. When dealing with numerical data, this means that any number may be measured and reported to arbitrarily many decimal places. Examples of truly continuous data are weight, area, volume and blood concentrations. In practice, in Cochrane reviews we can use the same statistical methods for other types of data, most commonly measurement scales and counts of large numbers of events (see 8.2.3 Effect measures for ordinal outcomes (including measurement scales).

Two summary statistics are commonly used for meta-analysis of continuous data: the mean difference and the standardised mean difference. These can be calculated whether the data from each individual are single assessments or change from baseline measures. It is also possible to measure effects by taking ratios of means, or by comparing statistics other than means (e.g. medians). However, methods for these are under development and are not addressed here.

8.2.2.1 The mean difference (and ‘WMD’)

The ‘difference in means’ is a standard statistic that measures the absolute difference between the mean value in the two groups in a clinical trial. It estimates the amount by which the treatment changes the outcome on average. It can be used as a summary statistic in meta-analysis when outcome measurements in all trials are made on the same scale. Analyses based on this effect measure are termed weighted mean difference (WMD) analyses in RevMan and the Cochrane Database of Systematic Reviews (CDSR). This name is potentially confusing. This is for three reasons. First, the measure is a difference in means and not a mean of differences. Second, although the meta-analysis computes a weighted average of these differences in means, no weighting is involved in calculation of a statistical summary of a single trial. Third, all meta-analyses involve a weighted combination of estimates, yet we don’t use the word ‘weighted’ when referring to other methods.

8.2.2.2 The standardised mean difference

The **standardised mean difference** is used as a summary statistic in meta-analysis when the trials all assess the same outcome, but measure it in a variety of ways (for example, all trials measure depression but they use different psychometric scales). In this circumstance it is necessary to standardise the results of the trials to a uniform scale before they can be combined. The standardised mean difference expresses the size of the treatment effect in each trial relative to the variability observed in that trial. (Again in reality the treatment effect is a difference in means and not a mean of differences.):

$$\text{SMD} = \frac{\text{Difference in mean outcome between groups}}{\text{Standard deviation of outcome among participants}}$$

Thus trials for which the difference in means is the same proportion of the standard deviation will have the same SMD, regardless of the actual scales used to make the measurements.

However, the method assumes that the differences in standard deviations among trials reflect differences in measurement scales and not real differences in variability among trial populations. This assumption may be problematic in some circumstances where we expect real differences in variability between the participants in different trials. For example, where pragmatic and explanatory trials are combined in the same review, pragmatic trials may

include a wider range of participants and may consequently have higher standard deviations. The overall treatment effect can also be difficult to interpret as it is reported in units of standard deviation rather than in units of any of the measurement scales used in the review, but in some circumstances it is possible to transform the effect back to the units used in a specific trial (see Section 8.X Re-expressing standardised mean differences).

The term ‘effect size’ is frequently used in the social sciences, particularly in the context of meta-analysis. Effect sizes typically, though not always, refer to versions of the standardised mean difference. It is recommended that the term ‘standardised mean difference’ be used in Cochrane reviews in preference to ‘effect size’ to avoid confusion with the more general medical use of the latter term as a synonym for ‘treatment effect’ or ‘effect estimate’. The particular definition of standardised mean difference used in Cochrane reviews is the effect size known in social science as Hedges’ (adjusted) g.

It should be noted that the SMD method does not correct for differences in the direction of the scale. If some scales increase with disease severity whilst others decrease it is essential to multiply the mean values from one set of trials by –1 (or alternatively to subtract the mean from the maximum possible value for the scale) to ensure that all the scales point in the same direction. Any such adjustment should be described in the statistical methods section of the review. The standard deviation does not need to be modified.

8.2.3 Effect measures for ordinal outcomes (including measurement scales)

Ordinal outcome data arise when each participant is classified in a category and when the categories have a natural order. For example, a ‘trichotomous’ outcome with an ordering to the categories, such as the classification of disease severity into ‘mild’, ‘moderate’ or ‘severe’ is of ordinal type. As the number of categories increases, ordinal outcomes acquire properties similar to continuous outcomes, and probably will have been analysed as such in a clinical trial.

Measurement scales are one particular type of ordinal outcome frequently used to measure conditions that are difficult to quantify, such as behaviour, depression, and cognitive abilities. Measurement scales typically involve a series of questions or tasks, each of which is scored, and the scores then summed to yield a total ‘score’. If the items are not considered of equal importance a weighted sum may be used. See Box 8.2.3 for an example.

It is important to know whether scales have been validated: that is, that they have been proven to measure the conditions that they claim to measure. When a scale is used to assess an outcome in a clinical trial the cited reference to the scale should be studied in order to understand the objective, the target population and the assessment questionnaire. As investigators often adapt scales to suit their own purpose by adding, changing or dropping questions, check whether an original or adapted questionnaire is being used. This is particularly important when pooling outcomes for a meta-analysis. Clinical trials may appear to use the same rating scale, but closer examination may reveal differences that must be taken into account. It is possible that modifications to a scale were made in the light of the results of a trial, in order to highlight components that appear to benefit from an experimental intervention.

Specialist methods are available for analysing ordinal outcome data that describe effects in terms of **proportional odds ratios**, but they are not available in RevMan, and become unwieldy (and unnecessary) when the number of categories is large. In practice longer ordinal scales are often analysed in meta-analyses as continuous data, whilst shorter ordinal scales are often made into binary data by combining adjacent categories together. Scales may sometimes be analysed as dichotomous data if an established defensible cut-point is available. Inappropriate choice of a cut-point can induce bias, particularly if it is chosen to maximise the difference between two intervention arms in a clinical trial.

Where ordinal scales are summarised using methods for binary data, one of the two sets of grouped categories is defined to be the event and treatment effects are described using risk ratios, odds ratios or risk differences (see 8.2.1 Effect measures for dichotomous outcomes). When ordinal scales are summarised using methods for continuous data, the treatment effect is expressed as a difference in means or standardised difference in means (see 8.2.2 Effect measures for continuous outcomes). Difficulties will be encountered if trials have summarised their results using medians (see 8. 5.2 Data extraction for continuous data).

Unless individual patient data are available, the analyses reported by the investigators in the clinical trials typically determine the approach that is used in the meta-analysis.

Box 8.2.3

An example of a scale is the Clinical Dementia Rating (CDR) (Berg 1988). The CDR is a quantitative global assessment of the severity of dementia. The clinician rates the patient's cognitive function in each of six categories: memory, orientation, judgement and problem solving, function in community affairs, function in home and hobbies, and function in personal care. Impairment is rated in each category on a five point scale (none=0, questionable=0.5, mild=1, moderate=2, severe=3). From these six ratings the CDR is established from a simple algorithm that is slightly more complex than an average. The result is a rating of no dementia (CDR=0), questionable (CDR=0.5), mild (CDR=1), moderate (CDR=2) and severe dementia (CDR=3). A second scale is formed by summing the category scores with equal weights. This is called the CDR sum of boxes and it has a range of 0 - 18.

8.2.4 Effect measures for counts and rates

Some types of event can happen to a person more than once, for example, a myocardial infarction, fracture, an adverse reaction or a hospitalisation. It may be preferable, or necessary, to address the number of times these events occur rather than simply whether each person experienced any event (that is, rather than treating them as dichotomous data). We refer to this type of data as count data. For practical purposes, **count data** may be conveniently divided into counts of rare events and counts of common events.

Counts of rare events are often referred to as ‘Poisson data’ in statistics. Analyses of rare events often focus on rates. Rates relate the counts to the amount of time during which they could have happened. For example, the result of one arm of a clinical trial could be that 18 myocardial infarctions (MIs) were experienced, across all participants in that arm, during a period of 314 person-years of follow-up, the rate is 0.057 per person year or 5.7 per 100 person years. The summary statistic used in meta-analysis is the rate ratio (also abbreviated to RR), which compares the rate of events in the two groups by dividing one by the other. It is also possible to use a difference in rates as a summary statistic, although this is much less common.

Counts of more common events, such as counts of decayed, missing or filled teeth, may often be treated in the same way as continuous outcome data. The treatment effect used will be the mean difference which will compare the difference in the mean number of events (possibly standardised to a unit time period) experienced by participants in the intervention group compared to participants in the control group.

8.2.4.1 Warning: counting events or counting participants?

A common error is to attempt to treat count data as dichotomous data. Suppose that in the example just presented, the 314 person-years arose from 157 patients observed on average for 2 years. One may be tempted to quote the results as 18/157. This is inappropriate if multiple MIs from the same patient could have contributed to the total of 18 (say if the 18 arose through 12 patients having single MIs and 3 patients each having 2 MIs). It is also possible

that the total number of events could theoretically exceed the number of patients, making the results nonsensical. For example, over the course of one year, 35 epileptic participants in a trial may experience 63 seizures among them.

8.2.5 Effect measures for time-to-event (survival) outcomes

Time-to-event data arise when interest is focused on the time elapsing before an event is experienced. They are known generically as **survival data** in statistics, since death is often the event of interest, particularly in cancer and heart disease. Time-to-event data consist of pairs of observations for each individual: (i) a length of time during which no event was observed, and (ii) an indicator of whether the end of that time period corresponds to an event or just the end of observation. Participants who contribute some period of time that does not end in an event are said to be 'censored'. Their event-free time contributes information and they are included in the analysis. Time-to-event data may be based on events other than death, such as recurrence of a disease event (for example, time to the end of a period free of epileptic fits) or discharge from hospital.

Time-to-event data can sometimes be analysed as dichotomous data. This requires the status of all patients in a trial to be known at a fixed time-point. For example, if all patients have been followed for at least 12 months, and the proportion who have incurred the event before 12 months is known for both groups, then a 2x2 table can be constructed (see Box 8.2.1) and treatment effects expressed as risk ratios, odds ratios or risk differences.

It is not appropriate to analyse time-to-event data using methods for continuous outcomes (e.g. using mean times-to-event) as the relevant times are only known for the subset of participants who have had the event. Censored participants must be excluded, which may well introduce bias.

The most appropriate way of summarising time-to-event data is to use methods of survival analysis and express the treatment effect as a **hazard ratio**. Hazard is similar in notion to risk, but is subtly different in that it measures instantaneous risk and may change continuously (for example, your hazard of death changes as you cross a busy road). A hazard ratio is interpreted in a similar way to a risk ratio, as it describes how many times more (or less) likely a participant is to suffer the event at a particular point in time if they receive the experimental rather than the control intervention. When comparing treatments in a trial or meta-analysis a simplifying assumption is often made that the hazard ratio is constant across the follow-up period, even though hazards themselves may vary continuously. This is known as the proportional hazards assumption.

8.2.6 Expressing treatment effects on log scales

The values of ratio treatment effects (such as the odds ratio, risk ratio, rate ratio and hazard ratio) undergo log transformations before being analysed, and they may occasionally be referred to in terms of their log transformed values. Typically the natural log (log base e) transformation is used.

Ratio summary statistics all have the common feature that the lowest value that they can take is 0, that the value 1 corresponds with no treatment effect, and the highest value that an odds ratio can ever take is infinity. This number scale is not symmetric. For example, whilst an odds ratio of 0.5 (a halving) and an OR of 2 (a doubling) are opposites such that they should average to no effect, the average of 0.5 and 2 is not an OR of 1 but an OR of 1.25. The log transformation makes the scale symmetric: the log of zero is minus infinity, the log of one is zero, and the log of infinity is infinity. In the example, the log of the OR of 0.5 is -0.69 and the log of the OR of 2 is 0.69. The average of -0.69 and 0.69 is 0 which is the log transformed value of an OR of 1, correctly implying no average treatment effect.

Graphics for ratio scale meta-analysis usually use a log scale. This has the effect of making the confidence intervals appear symmetric for the same reasons.

8.3 Study designs and identifying the unit of analysis

An important principle in clinical trials is that the analysis must take into account the level at which randomization occurred. In most circumstances the number of observations in the analysis should match the number of 'units' that were randomized. In a simple parallel group design for a clinical trial, participants are individually randomized to one of two intervention groups, and a single measurement for each outcome from each participant is collected and analysed. However, there are numerous variations on this design. Authors should consider whether in each trial

- groups of individuals were randomized together to the same intervention (i.e. cluster randomized trials);
- individuals undergo more than one intervention (e.g. in a cross-over trial, or simultaneous treatment of multiple sites on each individual);
- there are multiple observations for the same outcome (e.g. repeated measurements, recurring events, measurements on different body parts).

There follows a more detailed list of situations in which unit-of-analysis issues commonly arise, together with directions to relevant discussions elsewhere in the Handbook.

8.3.1 Cluster randomized trials

In cluster randomized trials, groups of participants are randomized to different interventions. For example, the groups may be schools, villages, medical practices, patients of a single doctor or families. See 8.11.2 Cluster-randomized trials.

8.3.2 Cross-over trials

In a cross-over trial all participants receive all interventions in sequence – they are randomized to an ordering of interventions, and participants act as their own control. See 8.11.3 Cross-over trials.

8.3.3 Repeated observations on participants

In studies of long duration, results may be presented for several periods of follow-up (for example, at 6 months, 1 year and 2 years). Results from more than one time point for each trial cannot be combined in a standard meta-analysis without a unit of analysis error. Some options are:

- to obtain individual patient data and perform an analysis (such as time-to-event analysis) that uses the whole follow up for each participant. Alternatively, compute an effect measure for each individual participant which incorporates all time points, such as total number of events, an overall mean, or a trend over time. Occasionally, such analyses are available in published reports;
- to define several different outcomes, based on different periods of follow-up, and to perform separate analyses. For example, time frames might be defined to reflect short-term, medium-term and long-term follow-up;
- to select a single time point and analyse only data at this time for trials in which it is presented. Ideally this should be a clinically important time point. Sometimes it might be

chosen to maximise the data available, although authors should be aware of the possibility of reporting biases;

- to select the longest follow-up from each trial. This may induce a lack of consistency across studies that gives rise to heterogeneity.

8.3.4 Events that may re-occur

If the outcome of interest is an event that can occur more than once, then care must be taken to avoid a unit-of-analysis error. Count data should not be treated as if they are dichotomous data. See 8.2.4 Effect measures for counts and rates.

8.3.5 Multiple treatment attempts

Similarly, multiple treatment attempts per participant can cause a unit of analysis error. Care must be taken to ensure that the number of participants randomized, and not the number of treatment attempts, is used to calculate confidence intervals. For example, in subfertility studies, women may undergo multiple cycles, and authors might erroneously use cycles as the denominator rather than women. This is similar to the situation in cluster randomized trials, except that each participant is the 'cluster'. See methods described in 8.11.2 Cluster randomized trials.

8.3.6 Multiple body parts I: body parts receive the same treatment

In some trials, whole people are randomized, but multiple parts of the body receive the same treatment and the number of body parts is used as the denominator in the analysis. For example, eyes may be mistakenly used as the denominator without adjustment for the non-independence between eyes. This is similar to the situation in cluster randomized trials, except that participants are the 'clusters'. See methods described in 8.11.2 Cluster randomized trials.

8.3.7 Multiple body parts II: body parts receive different treatments

A different situation is that in which different parts of the body are randomized to *different* treatments. 'Split-mouth' designs in oral health are of this sort, in which different areas of the mouth are assigned different interventions. These are similar to cross-over trials. See methods described in Section 8.11.3 Cross-over trials. It is important to distinguish these studies from those in which participants receive multiple versions of the *same* treatment.

8.3.8 Multiple intervention groups

Trials that compare more than two intervention groups need to be treated with care. A serious unit of analysis problem arises if the same group of participants is included twice in the same meta-analysis (for example, if 'Dose 1 vs Placebo' and 'Dose 2 vs Placebo' are both included in the same meta-analysis, with the same placebo patients in both comparisons). See 8.X Trials with more than two treatment groups.

8.4 Intention to treat issues

From the emphasis given to proper randomisation it follows that analysis of a randomised trial should ideally compare the groups exactly as randomised. Often some participants are excluded, either because they were lost to follow up and no outcome was obtained, or for some deviation from the protocol, such as receiving the wrong treatment or no treatment, lack

of compliance, or ineligibility. Alternatively, it may be impossible to measure certain outcomes for all participants because their availability depends on another outcome (see 8.4.4 Identifying conditional outcomes only available for subsets of participants).

8.4.1 What are intention-to-treat analyses?

An estimated treatment effect may be biased if some randomised participants are excluded from the analysis. Imbalances in such omissions between groups may be especially indicative of bias. Intention-to-treat (ITT) analysis aims to include all participants randomized into a trial irrespective of what happened subsequently (Lewis 1993, Newell 1992). ITT analyses are generally preferred as they are unbiased, and also because they address a more pragmatic and clinically relevant question.

The simple idea of an ITT analysis, to include all randomised patients, is not always easy to implement, and there are confusions about terminology. There are two criteria for an ITT analysis:

1. Trial participants should be analysed in the groups to which they were randomised regardless of which (or how much) treatment they actually received, and regardless of other protocol irregularities, such as ineligibility.
2. All participants should be included regardless of whether their outcomes were actually collected.

There is no clear consensus on whether both criteria should be applied (Hollis 1999). While the first is widely agreed, the second is contentious, since to include participants whose outcomes are unknown (mainly through loss to follow up) involves ‘filling-in’ (‘imputing’) missing data.

Many trials report having undertaken ITT analyses when they have met only the first of the two criteria, the second being impossible to achieve when contact is lost with the trial participants. An analysis in which data are analysed for every participant for whom the outcome was obtained is more properly called an **available case analysis**. Some trial reports present analyses of the results of only those participants who completed the trial *and* who complied with (or received some of) their allocated treatment. Some authors incorrectly call these ITT analyses, but they are in fact **per-protocol** or **treatment-received** analyses. Here we interpret the term ITT to mean that both of the above criteria are fulfilled. Authors should critically consider and report which type of analysis each trial has presented. Authors should avoid using the terms ‘intention-to-treat’ and ‘ITT’ without explicitly defining them.

8.4.1.1 Available case analyses

In most situations authors should attempt to extract from papers the data to enable at least an **available case analysis**. Avoidable exclusions should be reinstated if possible. The proportion of participants in each study arm who do not provide outcome data should be noted in the Study Characteristics table.

Three types of exclusions deserve specific mention. First, some trial participants may legitimately be excluded (i.e. without introducing bias) if their reason for exclusion was specified in the protocol and relates only to information collected before randomisation. For example, a condition may be defined by delayed blood tests on samples taken before randomization. Such exclusions are generally unwise, however, as the results do not then relate to the real clinical situation.

Second, and by contrast, exclusions immediately post-randomisation (and perhaps before treatment) may introduce bias, as they could be related to the treatment allocation.

Third, if dropout is very high or is different across treatment groups then the systematic review's protocol may dictate that a study be given a low quality rating and perhaps excluded from a meta-analysis (though usually not from the systematic review).

Many (but not all) people consider that available case and ITT analyses are not appropriate when assessing unintended (adverse) effects, as it is wrong to attribute these to a treatment that somebody did not receive. As ITT analyses tend to bias the results towards no difference they may not be the most appropriate when attempting to establish equivalence or non-inferiority of a treatment.

8.4.1.2 Full intention-to-treat analyses

In some rare situations it is possible to create a genuine ITT analysis from information presented in the text and tables of the paper, or by obtaining extra information from the author about participants who were followed up but excluded from the trial report. If this is possible without imputing study results it should be done.

Otherwise an intention to treat analysis can only be produced by using imputation. This involves making assumptions about the outcomes of participants for whom no outcome was recorded, and making up data for these participants. Some statistical techniques exist for imputing data but, ultimately, assessing the results of trials in the presence of more than minimal amounts of missing data is a matter of judgement. Statistical analysis cannot reliably compensate for missing data (Unnebrink 2001). No assumption is likely adequately to reflect the truth, and the impact of any assumption should be assessed by trying more than one method as a sensitivity analysis (see 8.10 Sensitivity analyses).

In the next two sections we consider some ways to take account of missing observations for dichotomous or continuous outcomes. Although imputation is possible, at present a sensible decision in most cases is to include data for only those participants whose results are known, and discuss the potential impact of the missing data. Where imputation is used the methods and assumptions for imputing data for dropouts should be described in the Methods section of the protocol and review.

8.4.2 ITT issues for dichotomous data

Percentages of participants for whom no outcome data were obtained should always be collected and reported in the Characteristics of Included Studies table; note that the percentages may vary by outcome. However, there is no consensus on the best way to handle these participants in an analysis. There are two basic options, and it may be wise to plan to undertake both and compare their results in a sensitivity analysis (see 8.10 Sensitivity analyses).

- Available case analysis: Include data on only those whose results are known, using as a denominator the total number of people who completed the trial for the particular outcome in question. The potential impact of the missing data on the results should be considered in the interpretation of the results of the review. This will depend on the degree of 'missingness', the frequency of the events and the size of the pooled effect estimate. Variation in the degree of missing data across studies may also be considered as a potential source of heterogeneity.
- ITT analysis using imputation: Base an analysis on the total number of randomized participants, irrespective of how the original trialists analysed the data. This will involve 'imputing' (a formal term for 'making up') outcomes for the missing patients. Studies with imputed data will be given more weight than they warrant if entered as dichotomous data into RevMan. It is possible to determine more appropriate weights; consultation with a statistician is recommended.

There are several approaches to imputing dichotomous outcome data. One common approach is to assume either that all missing participants experienced the event, or that all missing participants did not experience the event. The choice among these assumptions should be based on clinical judgement as to what would be the most likely outcome. An alternative approach is to impute data according to the event rate observed in the control group, or according to event rates among completers in the separate groups. None of these assumptions is likely to reflect the truth, and the latter achieves little other than an unwarranted inflation of the precision of effect estimates. Thus this approach is generally not recommended. The impact of any assumptions can be tested by undertaking sensitivity analyses where first it is assumed that all missing participants in the first group incurred the event and those in the second group did not, and then assuming the opposite. When missing data are common, these worst-case/best-case scenarios will cover a very wide range of possible treatment effects and thus the analysis will not be very informative. However, when missing data are not common and this procedure is done across all trials in the review with little impact on the results, it can be concluded that the missing data could not affect the outcome of the review.

8.4.3 ITT issues for continuous data

In full ITT analyses, all participants who did not receive the assigned intervention according to the protocol as well as those who were lost to follow-up are included in the analysis. Inclusion of these in an analysis requires that means and standard deviations for all randomized participants are available. As for dichotomous data, dropout rates should always be collected and reported in the Characteristics of Included Studies table. There are two basic options, and it may be wise to plan to undertake both and formally compare their results in a sensitivity analysis (see 8.10 Sensitivity analyses).

- Available case analysis: Include data only on those whose results are known. The potential impact of the missing data on the results should be considered in the interpretation of the results of the review. This will depend on the degree of ‘missingness’, the pooled estimate of the treatment effect and the variability of the outcomes. Variation in the degree of missing data may also be considered as a potential source of heterogeneity.
- ITT analysis using imputation: Base an analysis on the total number of randomized participants, irrespective of how the original trialists analysed the data. This will involve imputing outcomes for the missing patients. Approaches to imputing missing continuous data in the context of a meta-analysis have received little attention in the methodological literature. In some situations it may be possible to exploit standard (although often questionable) approaches such as ‘last observation carried forward’, or, for change from baseline outcomes, to assume that no change took place, but such approaches generally require access to the raw patient data. Inflating the sample size of the available data up to the total numbers of randomized participants is based on an assumption that those dropping out from the study were a random sample of all those included, and is not recommended as it will artificially inflate the precision of the effect estimate.

8.4.4 Identifying conditional outcomes only available for subsets of participants

Some trial outcomes may only be applicable to a proportion of participants. For example, in subfertility trials the proportion of clinical pregnancies that miscarry following treatment is often reported. By definition this outcome excludes participants who do not achieve an interim state (clinical pregnancy), so the comparison is not of all participants randomized. As a general rule it is better to re-define such outcomes so that the analysis includes all randomized participants. In this example, the outcome could be whether the woman has a ‘successful pregnancy’ (becoming pregnant and reaching, say, 24 weeks or term).

Another example is a morbidity outcome measured in the medium or long term (e.g. development of chronic lung disease), when there is a distinct possibility of a death preventing assessment of the morbidity. A convenient way to deal with such situations is to combine the outcomes, for example as 'death or chronic lung disease'.

Some intractable problems arise when a continuous outcome (say a measure of functional ability or quality of life following stroke) is measured only on those who survive to the end of follow-up. Two unsatisfactory alternatives exist: (a) imputing zero functional ability scores for those who die (which may not appropriately represent the death state and will make the outcome severely skewed), and (b) analysing the available data (which must be interpreted as a non-randomised comparison applicable only to survivors).

8.5 Extraction of study results

This section outlines the data that need to be extracted from trial reports for analyses of each of the data types described in 8.2 Types of data and effect measures. For many studies the required data will be presented clearly. However, sometimes the required data may be obtained only indirectly, and the relevant results may not be obvious. This section provides some useful tips and techniques to deal with these situations.

The section concludes with some important considerations that despite being mentioned last must be considered before starting the data extraction process. First, a common error when extracting data is to fail to recognise what the unit of analysis should be. A unit of analysis error may arise when results entered into an analysis do not suitably reflect the design of the study. It is important to recognise such situations. Second, intention-to-treat analyses may require collection of data from different parts of a paper.

8.5.1 Data extraction for dichotomous outcomes

Dichotomous data are described in 8.2.1 Effect measures for dichotomous outcomes. The only data required for a dichotomous outcome are the numbers in each of the two categories in each of the intervention groups the numbers needed to fill in the four boxes *a*, *b*, *c* and *d* in Box 8.2.1. The data are often available as the number assessed and the number incurring the event of interest in each group. Difficulties may be experienced in clearly identifying the numbers actually assessed for each outcome due to poor reporting, and occasionally the numbers incurring the event need to be derived from percentages (although it is not always clear which denominator to use, and rounded percentages may be compatible with more than one numerator).

See also 8.6.3 Meta-analysis of dichotomous outcomes.

8.5.1.1 Extracting effect estimates calculated from dichotomous outcomes

Sometimes the numbers of participants and numbers of events are not available, but results calculated from them are. For example, an estimate of an odds ratios or a risk ratio may be present in an abstract, while the full text of the paper cannot be obtained so further data are unavailable. Such data may be included in meta-analyses only if they are accompanied by measures of uncertainty such as a 95% confidence interval or an exact *P*-value. The numbers then must be analysed using the generic inverse variance method in RevMan (see 8.6.2 A generic inverse variance approach to meta-analysis). This requires the author to enter an estimate and a standard error for each study. The process of obtaining a suitable estimate and standard error from a confidence interval or *P*-value is described in 8.5.6 Obtaining standard errors from confidence intervals and P-values.

A limitation of this approach is that estimates and standard errors of the same effect measure must be calculated for all the other studies in the same meta-analysis, even if they provide the

original numbers of participants and events. If the numbers of events and participants are known the necessary summary statistics may be obtained from RevMan (entering the data as dichotomous data), and copied manually into the data entry window for the generic inverse variance outcome. The confidence intervals estimated in RevMan will need to be converted into standard errors.

When extracting data from non-randomized studies, and from some randomized studies, adjusted odds ratios may be available from logistic regression analyses. The process of data extraction, and analysis using the generic inverse variance method, is the same as for unadjusted estimates.

8.5.2 Data extraction for continuous outcomes

Continuous data are described in 8.2.2 Effect measures for continuous outcomes. To perform a meta-analysis of continuous data using either mean differences or standardised mean differences one needs to extract the mean values of the outcomes, the standard deviations of the outcomes, and the number of participants on whom the outcome was assessed in each of the two groups.

In many cases the relevant information can be extracted directly from trial reports in a straightforward way. However, due to poor and variable reporting occasionally it is difficult or impossible to obtain the necessary information from the data summaries presented. Trials vary in the statistics they use to summarise average (sometimes using medians rather than means) and variation (sometimes using standard errors, confidence intervals, interquartile ranges and ranges rather than standard deviations).

When needed, missing information and clarification about the statistics presented should always be sought from the authors. However, for several of the measures of variation there is an approximate or direct algebraic relationship with standard deviations, so it may be possible to obtain the required statistic even if it is not published directly in the paper as is explained in the subsections that follow. For more details and examples see (Deeks 1997a, Deeks 1997b).

A particularly misleading error is to misinterpret a standard error as a standard deviation. Unfortunately it is not always clear what is being reported and some intelligent reasoning may be required. Standard deviations and standard errors are occasionally confused by authors of trial reports, and the terminology is used inconsistently.

See also 8.6.4 Meta-analysis of continuous outcomes.

8.5.2.1 Medians

The median is very similar to the mean when the distribution of the data is symmetrical, and so occasionally can be used directly in meta-analyses. However, means and medians can be very different from each other if the data are skewed, and medians are often the summary statistic of choice when data are skewed (see 8.5.2.11 Skewed data).

8.5.2.2 Standard errors of group means

Standard deviations are obtained by multiplying standard errors of means by the square-root of the sample size:

$$SD = SE \times \sqrt{N}$$

When making this transformation ensure that standard errors are standard errors of means calculated from within a treatment group and not standard errors of the difference in means computed between treatment groups.

8.5.2.3 Confidence intervals for group means

Confidence intervals for means can also be used to calculate standard deviations via calculation of the standard error of the mean. The following applies to confidence intervals for mean values calculated within treatment group results and not from comparisons of treatments. Most confidence intervals are 95% confidence intervals. If the sample size is large (say bigger than 100), the 95% confidence interval is 3.92 (2 x 1.96) standard errors wide. The standard deviation for each group is obtained by dividing the length of the confidence interval by 3.92, and then multiplying by the square root of the sample size:

$$SD = \sqrt{N} \times (\text{upper limit} - \text{lower limit}) / 3.92$$

For 90% confidence intervals divide by 3.29 rather than 3.92, for 99% confidence intervals divide by 5.15.

If the sample size is smaller than 60 then confidence intervals should have been calculated using a value from a *t*-distribution. The numbers 3.92, 3.29 and 5.15 need to be replaced with slightly larger numbers specific to both the *t*-distribution and the sample size which can be obtained from tables of the *t*-distribution with degrees of freedom equal to the group sample size minus 1. (Relevant details of the *t*-distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example the *t*-value for a 95% confidence interval from a sample size of 27 can be obtained by typing =**tinv(1-0.95,27-1)** in a cell in a Microsoft Excel spreadsheet.)

As an example, consider data presented as follows:

Group	*Sample size*	*Mean*	*95% CI*
Experimental intervention	25	32.1	(30.0, 34.2)
Control intervention	22	28.3	(26.5, 30.1)

The confidence intervals should have been based on *t*-distributions with 24 and 21 degrees of freedom respectively. The relevant numbers for the divisor are then 2 x 2.06 = 4.12 and 2 x 2.08 = 4.16. The standard deviations for the two groups are √25 x (34.2 30.0)/4.12 = 5.10 and √22 x (30.1 26.5)/4.16= 4.06.

It is important to check that the confidence interval is symmetrical about the mean (the distance between the lower limit and the mean is the same as the distance between the mean and the upper limit). If this is not the case the confidence interval may have been calculated on transformed values (see Section 8.5.2.11 Skewed data below).

8.5.2.4 *t*-values, standard errors and confidence intervals for differences in means

The same ingredients of means, standard deviations and sample sizes are involved in *t*-tests used to compute the statistical significance of differences in means. The methods do not actually estimate the two standard deviations observed in the two groups but estimate the average of their values. This simplification does not matter for the purpose of meta-analysis.

The *t*-value is the ratio of the difference in means to the standard error of the difference in means. Computing the standard deviation first involves computing the standard error of the difference in means by dividing the difference in means (MD) by the *t*-value:

$$\text{standard error of difference in means} = \frac{\text{MD}}{t}$$

If a 95% confidence interval is available for the difference in means, then the same standard error can be calculated as:

$$\text{SE} = (\text{upper limit lower limit})/3.92$$

as long as the trial is large. For 90% confidence intervals divide by 3.29 rather than 3.92, for 99% confidence intervals divide by 5.15. If the sample size is small then confidence intervals should have been calculated using a *t*-distribution. The numbers 3.92, 3.29 and 5.15 need to be replaced with larger numbers specific to both the *t*-distribution and the sample size, and can be obtained from tables of the *t*-distribution with degrees of freedom equal to NE + NC 2, where NE and NC are the sample sizes in the two groups. (Relevant details of the *t*-distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example the *t*-value for a 95% confidence interval from a comparison of a sample size of 27 with a sample size of 24 can be obtained by typing =**tinv(1-0.95,27+24-2)** in a cell in a Microsoft Excel spreadsheet).

The standard deviation can then be obtained from the standard error of the difference in means using the following formula:

$$\text{standard deviation} = \frac{\text{standard error of difference in means}}{\sqrt{\left(\frac{1}{\text{N}_\text{E}} + \frac{1}{\text{N}_\text{C}}\right)}}.$$

See below (Section 8.5.2.5 *P*-values) for an example. This standard deviation must be entered into RevMan for both intervention groups.

Related methods can be used to derive standard deviations from certain F-statistics, although methods are somewhat complex and advice of a knowledgeable statistician is recommended.

8.5.2.5 P-values

Where actual *P*-values obtained from *t*-tests are quoted, it is possible to extract standard deviations by first obtaining the corresponding *t*-value from a table of the *t*-distribution (noting that the degrees of freedom are given by $\text{N}_\text{E} + \text{N}_\text{C}$ 2), and then transforming the *t*-value into a standard deviation as described in 8.5.2.4 *t*-values, standard errors and confidence intervals for differences in means.

As an example, consider a trial of an experimental intervention (NE = 25) versus a control intervention (N_c = 22), where the difference in means was MD = 3.8. It is noted that the *P*-value for the comparison was *P* = 0.008 obtained using a two-sample t-test.

The *t*-statistic that corresponds with a *P*-value of 0.008 and 25+22-2=45 degrees of freedom is *t* = 2.78. This can be obtained from a table of the *t*-distribution with 45 degrees of freedom or a computer (for example, by entering =**tinv(0.008, 45)** into any cell in a Microsoft Excel spreadsheet).

The standard error of the difference in means is obtained by dividing the MD (3.8) by the *t*-value (2.78), which gives 1.37. To calculate the standard deviation from the *t*-statistic we use

$$\text{standard deviation} = \frac{1.37}{\sqrt{\left(\frac{1}{25} + \frac{1}{22}\right)}} = 4.69.$$

Note that this standard deviation is the average of the standard deviations of the experimental and control arms, and must be entered into RevMan for both groups.

Difficulties are encountered when levels of significance are reported (such as $P<0.05$ or even P=NS which usually implies $P>0.05$) rather than exact P-values. A conservative approach would be to take the P-value at the upper limit (e.g. for $P<0.05$ take $P=0.05$, for $P<0.01$ take $P=0.01$ and for $P<0.001$ take $P=0.001$). However, this is not a solution for results which are reported as P=NS. It may be preferable to impute a value for the standard deviation for studies that report P=NS from those observed in other studies rather than inevitably introducing bias by excluding them from the meta-analysis (see 8.X Missing Data).

8.5.2.6 Interquartile ranges

Interquartile ranges describe where the central 50% of participants outcomes lie. When sample sizes are reasonably large and the distribution of the outcome is similar to the normal distribution, the width of the interquartile range will be approximately 1.35 standard deviations. In other situations, and especially when the outcomes distribution is skewed, it is not possible to estimate a standard deviation from an interquartile range. Note that the use of interquartile ranges rather than standard deviations can often be taken as an indicator that the outcomes distribution is skewed.

8.5.2.7 Ranges

Ranges are very unstable and, unlike other measures of variation, increase when the sample size increases. They describe the extremes of observed outcomes rather than the average variation. It is not possible to reliably estimate a standard deviation from a range. One common approach has been to make use of the fact that, with normally distributed data, 95% of values will lie within 2xSD either side of the mean. The SD may therefore be estimated to be approximately one quarter of the typical range of data values. This method is not robust and is discouraged.

8.5.2.8 No information on variability

If none of the above methods allow calculation of the standard deviation(s) from the trial report (and the information is not available directly from the trialists) then, in order to perform a meta-analysis, an author is forced either to exclude the study and risk introducing bias, or to impute missing data (see 8.X Missing data) and risk making a different type of error. Alternatively a narrative approach to synthesis may be used. It is valuable to tabulate available results for all studies included in the systematic review, even if they cannot be included in a formal meta-analysis.

8.5.2.9 Change from baseline

A common feature of continuous data (and also possible with ordinal data) is that a measurement used to assess the outcome of each participant is also measured at baseline, that is at or before randomization into the trial. This gives rise to the possibility of using differences in **changes from baseline** (also called a **change score**) as the primary outcome. Authors are advised not to focus on change from baseline unless this method of analysis was used in some of the trial reports.

When addressing change from baseline, a single measurement is created for each participant, obtained either by subtracting the final measurement from the baseline measurement or by subtracting the baseline measurement from the final measurement. Analyses then proceed as

for any other type of continuous outcome variable using the changes rather than the final measurements.

The principal difficulty associated with change from baseline analyses is the availability of data from published reports. It is very common for standard deviations of the changes to be unavailable. A common situation is that the following data are available:

	Baseline	*Final*	*Change*
Experimental intervention (sample size n_1)	mean, SD	mean, SD	mean
Control intervention (sample size n_2)	mean, SD	mean, SD	mean

Note that the mean change in each group can always be obtained by subtracting the final mean from the baseline mean even if it is not presented explicitly. However, the information in this table does *not* allow us to calculate the standard deviation of the changes. We cannot know whether the changes were very similar or very variable. Some other information in a paper may help us determine the standard deviation of the changes. If statistical analyses comparing the changes themselves are presented (e.g. confidence intervals, *t*-values or *P*-values) then the techniques described above (see Sections 8.5.2.3 to 8.5.2.5) may be used.

In other situations it is possible to impute standard deviations for the changes. Follmann (Follmann 1992) discusses techniques for imputing missing standard deviations, some of which are described in Section 8.5.2.10 Imputing standard deviations for changes from baseline. However, all imputation techniques involve making assumptions about unknown statistics, and it is best to avoid using them wherever possible If they are used the impact of the imputations should be tested in planned sensitivity analyses (see 8.10 Sensitivity analyses). Imputed standard deviations should not be used for a majority of studies in a meta-analysis, but may be reasonable for a small proportion of studies comprising a small proportion of the data if it enables them to be combined with other studies for which full data are available.

Authors are advised to extract data on both change from baseline and final value outcomes if the required means and standard deviations are available. Commonly an author will find that they end up with a mixture of changes from baseline and final values for trials included in a review. Some trials will report both; others will report only change scores or only final values. As explained in Section 8.6.4.2 Meta-analysis of change scores, both final values and change scores can often be combined in the same analysis so this is not necessarily a problem.

A final problem with using change from baseline measures is that often baseline and final measurements will be reported for different numbers of participants due to missed visits and study withdrawals. It may be difficult to identify the subset of participants who report both baseline and final value measurements for whom change scores can be computed.

8.5.2.10 Imputing standard deviations for changes from baseline

A hidden number known as the correlation coefficient describes how similar the baseline and final measurements were across participants. Here we describe (1) how to estimate the correlation coefficient from a study that is reported in considerable detail and (2) how to impute a change from baseline standard deviation in another study, making use of an imputed correlation coefficient. Note that the methods in (2) are applicable both to correlation coefficients obtained using (1) and to correlation coefficients obtained in other ways (for example, by reasoned argument). These methods should be used sparingly, if at all. This is

partly because one can never be sure that an imputed correlation is appropriate (correlations between baseline and final values will, for example, decrease with increasing time between baseline and final measurements, as well as depending on the outcomes and characteristics of the participants). A further reason is that a comparison of final measurements in a randomised trial in theory estimates the same quantity as the comparison of changes from baseline, so imputation is often not necessary to enable trials to be included in the analysis.

1. Suppose a study is available that presents the following information:

	Baseline	*Final*	*Change*
Experimental intervention (sample size n_1)	$mean_1(B)$, $SD_1(B)$	$mean_1(F)$,$SD_1(F)$	$mean_1(C)$,$SD_1(C)$
Control intervention (sample size n_2)	$mean_2(B)$, $SD_2(B)$	$mean_2(F)$,$SD_2(F)$	$mean_2(C)$,$SD_2(C)$

An analysis of change from baseline is available from this study, using only the data in the final column. We can use the other data from the study to estimate the correlation coefficient in the experimental intervention, r_1, as follows:

$$r_1 = \frac{SD_1(B)^2 + SD_1(F)^2 - SD_1(C)^2}{2 \times SD_1(B) \times SD_1(F)},$$

and similarly for the control intervention, r_2. Where either *SD(F)* or *SD(B)* are unavailable, then it may be substituted by the other if it is reasonable to assume that the intervention does not alter the variability of the outcome measure. Correlation coefficients lie between 1 and 1. If zero or a negative number is obtained, then there is no value in using change from baseline and an analysis of final values should be performed. Assuming the correlation coefficients from the two intervention groups are similar, a simple average will provide a reasonable measure of the similarity of baseline and final measurements across individuals. If the correlation coefficients differ, then either the sample sizes are too small for reliable estimation, or the intervention is affecting the variability in outcome measures, and the use of imputation is best avoided. Before imputation is undertaken it is recommended that correlation coefficients are computed for many (if not all) studies in the meta-analysis and it is noted whether or not they are consistent. Imputation should be done only as a very tentative analysis if correlations are inconsistent.

2. To impute the standard deviation of a change from baseline, when baseline and final standard deviations are known, we use an imputed value R_1 for the correlation coefficient. The value R_1 might be imputed from another study in the meta-analysis (using the method in (1) above), it might be imputed from elsewhere, or it might be hypothesised based on reasoned argument. In all of these situations, a sensitivity analysis should be undertaken, trying different values of R_1, to determine whether the overall result of the analysis is robust to the use of imputed correlation coefficients.

   To obtain a standard deviation of the change from baseline for the experimental intervention, use

$$SD_1(C) = \sqrt{SD_1(B)^2 + SD_1(F)^2 - (2 \times R_1 \times SD_1(B) \times SD_1(F))},$$

and similarly for the control intervention. Again, if either *SD(F)* or *SD(B)* are unavailable, then one may be substituted by the other if it is reasonable to assume that the intervention does not alter the variability of the outcome measure.

As an example, given the following data:

	Baseline	*Final*	*Change*
Experimental intervention (sample size 35)	mean=12.4 SD=4.2	mean=15.2 SD=3.8	mean=2.8
Control intervention (sample size 38)	mean=10.7 SD=4.0	mean=13.8 SD=4.4	mean=3.1

and using an imputed correlation coefficient of 0.5, we can impute the standard deviation for the change score in the control group as:

$$SD_2(C) = \sqrt{4.0^2 + 4.4^2 - (2 \times 0.5 \times 4.0 \times 4.0)} = 4.21.$$

8.5.2.11 Skewed data

Analyses based on means or standardised means are appropriate for data that are at least approximately normally distributed, and for data from very large trials. If the true distribution of outcomes is asymmetrical then the data are said to be skewed. Methods for meta-analysing skewed data are lacking at present, though they are the subject of current research.

Transformation of the original outcome data may substantially reduce skewness. Reports of trials may present results on a transformed scale, usually a log scale. More often they do not. Collection of appropriate data summaries from the trialists, or acquisition of individual patient data, is currently the approach of choice. Appropriate data summaries and analysis strategies for the individual patient data will depend on the situation. Consultation with a knowledgeable statistician is advised.

With the more common positive skewness, presentation of a geometric mean with its 95% confidence interval is equivalent to an analysis of a log transformation of the data. The difference in means of the log transformed data may be obtained from a ratio of geometric means (geometric mean ratio, GMR) as log(GMR), and the standard error of this difference as [log(lower confidence limit for GMR) log(upper confidence limit for GMR)]/3.92. The standard deviation of the log transformed data may be determined from the standard error as described above (see Sections 8.5.2.2 to 8.5.2.5). This approach depends on being able to obtain transformed data for all trials. Log-transformed and untransformed data can not be mixed in a meta-analysis.

Skewness can sometimes be diagnosed from the means and standard deviations of the outcomes. A rough check is available, but it is only valid if a lowest or highest possible value for an outcome is known to exist. Thus the check may be used for outcomes such as weight, volume and blood concentrations, which have lowest possible values of 0, or for scale outcomes that may have lowest and highest possible values. The check is not appropriate for change from baseline measures. The check involves calculating the observed mean minus the lowest possible value (or the highest possible value minus the observed mean), and dividing

this by the standard deviation. A ratio less than 2 suggests skewness. If the ratio is less than 1 there is strong evidence of a skewed distribution (Altman 1996).

It should be noted that skewness is not necessarily a problem for meta-analyses in RevMan if the sample sizes in the individual studies are large.

8.5.2.12 Extracting effect estimates calculated from continuous data

Sometimes only effect estimates (estimates of a mean difference or standardized mean difference) are available with a standard error or confidence interval. If this is the case, the analysis should be performed using the generic inverse variance method in RevMan (8.6.2 A generic inverse variance approach to meta-analysis). This requires the author to enter the estimate and standard error for each study. The process of obtaining a suitable standard error from a confidence interval for a mean difference is described in 8.5.2.4 *t*-values, standard errors and confidence intervals for differences in means. For standardized mean differences, see 8.5.6 Obtaining standard errors from confidence intervals and *P*-values.

A limitation of this approach is that all other studies in the same meta-analysis must provide estimates and standard errors of the same effect measure, even if they provide the six numbers usually required to analyse continuous data. However, the necessary numbers may be obtained from RevMan (entering the data as continuous data), and copied manually into the data entry window for a generic inverse variance outcome, converting the confidence interval into a standard error.

When extracting data from non-randomized studies, and from some randomized studies, adjusted estimates of mean differences may be available from multiple regression analyses and analyses of covariance. The process of data extraction and analysis using the generic inverse variance method is the same as for unadjusted estimates.

8.5.3 Data extraction for ordinal outcomes and measurement scales

Ordinal data and measurement scales are described in 8.2.3 Effect measures for ordinal outcomes (including measurement scales). The data that need to be extracted for ordinal outcomes depend on whether the ordinal scale will be dichotomised for analysis (see 8.5.1 Data extraction for dichotomous data), treated as a continuous outcome (see 8.5.2 Data extraction for continuous data) or analysed directly as ordinal data. This decision, in turn, will be influenced by the way in which authors of the trials analysed their data. Thus it may be impossible to pre-specify whether data extraction will involve calculation of numbers of participants above and below a defined threshold, or mean values and standard deviations. In practice, it is wise to extract data in all forms in which they are given as it will not be clear which is the most common until all trials have been reviewed, and in some circumstances more than one form of analysis may justifiably be included in a review.

Where ordinal data are being dichotomised and there are several options for selecting a cutpoint (or the choice of cutpoint is arbitrary) it is sensible to plan from the outset to investigate the impact of choice of cutpoint in a sensitivity analysis (see 8.10 Sensitivity analyses). To do this it is necessary to collect the data that would be used for each alternative dichotomisation. Hence it is preferable to record the numbers in each category of short ordinal scales to avoid having to extract data from a paper multiple times. This approach of recording all categorisations is also sensible when trials use slightly different short ordinal scales, and it is not clear whether there will be a cutpoint that is common across all the trials which can be used for dichotomisation.

It is also necessary to record the numbers in each category of the ordinal scale for each treatment group if the proportional odds ratio method (see 8.2.3 Effect measures for ordinal outcomes (including measurement scales) will be used.

8.5.4 Data extraction for counts and rates

Counts and rates are described in 8.2.4 Effect measures for counts and rates. Data that are inherently counts may be analysed in several ways. The essential decision is whether to make the outcome of interest dichotomous, continuous, time-to-an-event or a rate. A common error is to treat counts directly as dichotomous data, using as sample sizes either the total number of participants or the total number of, say, person-years of follow-up. Neither of these approaches is appropriate for an event that may occur more than once for each participant. This becomes obvious when the total number of events exceeds the sample size, leading to nonsensical results. Although it is preferable to decide how count data will be analysed in advance, the choice is often determined by the format of the available data, and thus cannot be decided until the majority of studies have been reviewed.

8.5.4.1 Extracting counts as dichotomous data

To consider the outcome as a dichotomous outcome, the author must determine the number of participants in each intervention group, and the number of participants in each intervention group who experience *at least one event* (or some other appropriate criterion which classified all participants into one of two possible groups). Any time element in the data is lost through this approach, though it may be possible to create a series of dichotomous outcomes, for example at least one stroke during the first year of follow-up, at least one stroke during the first two years of follow-up, and so on. Such data may be hard to derive from published reports. See also 8.6.3 Meta-analysis of dichotomous outcomes.

8.5.4.2 Extracting counts as continuous data

To extract counts as continuous data, guidance in 8.5.2 Data extraction for continuous outcomes should be followed, although particular attention should be paid to the likelihood that the data will be highly skewed. See also 8.6.4 Meta-analysis of continuous outcomes.

8.5.4.3 Extracting counts as time-to-event data

For rare events that can happen more than once, an author may be faced with studies that treat the data as time-to-*first*-event. To extract counts as time-to-event data, guidance in 8.5.5 Data extraction for time-to-event outcomes should be followed. See also 8.6.8 Meta-analysis of time-to-event outcomes.

8.5.4.4 Extracting counts as rate data

To analyse rate data an author should extract the total number of events in each group, and the total amount of person-time at risk in each group. Unlike for dichotomous data, the total number of events may include multiple events for some participants, and may even exceed the total number of participants. Note that the total number of participants is not required for an analysis of rate data but you will probably wish to record it as part of the trial description. See also 8.6.7 Meta-analysis of counts and rates.

8.5.4.5 Extracting effect estimates calculated from rate data

Sometimes detailed data on events and person-years at risk are not available, but results calculated from them are. For example, an estimate of a rate ratio or rate difference may be present in an abstract, while the full text of the paper unavailable. Such data may be included in meta-analyses only if they are accompanied by measures of uncertainty such as a 95% confidence interval. See 8.5.6 Obtaining standard errors from confidence intervals and *P*-

values. When extracting data from non-randomized studies, and from some randomized studies, adjusted rate ratios may be available from Poisson regression analyses. Data extraction is the same as for unadjusted rate ratios.

8.5.5 Data extraction for time-to-event outcomes

Meta-analysis of time-to-event data commonly involves obtaining individual patient data from the trialists, re-analysing the data to obtain estimates of the log hazard ratio and its standard error, and then performing a meta-analysis. Conducting a meta-analysis using summary information from published papers or trial reports is often problematic as the most appropriate summary statistics are typically not explicitly presented.

Two approaches can be used to obtain estimates of log hazard ratios regardless of whether individual patient data or aggregate data are being used.

In the first approach an estimate of the log hazard ratio can be obtained from statistics computed during a logrank analysis. Collaboration with a knowledgeable statistician is advised if this approach is followed. The log hazard ratio (experimental relative to control) is estimated by (O - E)/V, which has standard error $1/\sqrt{V}$, where O is the observed number of events on the experimental intervention, E is the logrank expected number of events on the experimental intervention, (O - E) is the logrank statistic and V is the variance of the logrank statistic. It is therefore necessary to obtain values of O - E and V for each study.

These statistics are easily computed if individual patient data are available, and can sometimes be extracted from quoted statistics and survival curves as discussed by Parmar, Torri and Stewart (Parmar 1998). Alternatively, use can sometimes be made of aggregated data for each treatment group in each trial. For example, suppose that the data comprise the number of participants who have the event during the first year, second year, etc., and the number of participants who are event free and still being followed up at the end of each year. A logrank analysis can be performed on these data, to provide the (O - E) and V values, although careful thought needs to be given to the handling of censored times. Because of the coarse grouping the log hazard ratio is estimated only approximately, and in some reviews has been referred to as a log odds ratio (Early Breast Cancer Trialists' Collaborative Group 1990). If the time intervals are large, a more appropriate approach is one based on interval-censored survival (Collett 1994).

The second approach can be used if trialists have analysed the data using a Cox proportional hazards model, or if a Cox model is fitted to individual patient data. Cox models produce direct estimates of the log hazard ratio and its standard error. If the hazard ratio is quoted in a report together with a confidence interval or *P*-value, estimates of standard error can be obtained as described in 8.5.6 Obtaining standard errors from confidence intervals and *P*-values.

8.5.6 Obtaining standard errors from confidence intervals and P-values

Estimates of an effect measure of interest are typically presented along with a confidence interval or a *P*-value. On occasion, the data contributing to the estimate (for example, numbers of events and participants, or means and standard deviations) cannot be extracted. In such situations it may still be possible to include the data in a meta-analysis using the generic inverse variance method, which requires only an estimate and a standard error from each study (See 8.6.2 A generic inverse variance approach to meta-analysis). This section describes how to obtain a standard error from a confidence interval or a *P*-value. If extracting data concerning a mean from one treatment arm, or the difference between two means, then section 8.5.2 Data extraction for continuous data should be followed instead.

The procedure for obtaining a standard error depends on whether the effect measure is a ratio measure (e.g. odds ratio, risk ratio, hazard ratio, rate ratio) or an absolute measure (e.g. mean difference, standardized mean difference, risk difference).

8.5.6.1 Standard error for absolute (difference) measures

If a 95% confidence interval is available for an absolute measure of treatment effect, then the standard error can be calculated as

SE = (upper limit lower limit)/3.92

For 90% confidence intervals divide by 3.29 rather than 3.92; for 99% confidence intervals divide by 5.15.

Where exact *P*-values are quoted alongside estimates of treatment effect, it is possible to estimate standard errors. While all tests of statistical significance produce *P*-values, different tests use different mathematical approaches to obtain a *P*-value. The method here assumes *P*-values have been obtained through a particular simple approach known as a Wald test. Where significance tests have used other mathematical approaches the estimated standard errors may not coincide exactly with the true standard errors.

The first step is to obtain the *Z*-value corresponding to the reported *P*-value from a table of the standard normal distribution. A standard error may then be calculated as

SE = treatment effect estimate / Z

As an example, suppose a conference abstract presents an estimate of a risk difference of 0.03 ($P = 0.008$). The *Z*-statistic that corresponds with a *P*-value of 0.008 is $Z = 2.652$. This can be obtained from a table of the standard normal distribution or a computer (for example, by entering **=abs(normsinv(0.008/2)** into any cell in a Microsoft Excel spreadsheet). The standard error of the risk difference is obtained by dividing the risk difference (0.03) by the *Z*-value (2.652), which gives 0.011.

8.5.6.2 Standard error for ratio measures

The process of obtaining standard errors for ratio measures is similar to that for absolute measures, but with an additional first step. Analyses of ratio measures are performed on the log scale (see 8.2.6 Expressing treatment effects on log scales). For a ratio measure *R*, such as an odds ratio or hazard ratio, first calculate

lower limit = log(lower confidence limit given for *R*)

upper limit = log(upper confidence limit given for *R*)

treatment effect estimate = log(*R*)

Then the formulae in Section 8.5.6.1 Standard error for absolute (difference) measures can be used. Note that the standard error refers to the log of the ratio measure. When using the generic inverse variance method in RevMan, the data should be entered on the log scale, that is as log(*R*) and the standard error of log(*R*), as calculated here (see 8.6.2 A generic inverse variance approach to meta-analysis).

8.6 Summarising effects across studies

An important step in a systematic review is the thoughtful consideration of whether it is appropriate to combine the numerical results of all, or perhaps some, of the studies. Such a ‘meta-analysis’ yields an overall statistic (together with its confidence interval) that summarises the effectiveness of the experimental intervention compared with a control

intervention (see 8.1 Planning the analysis). This section describes the principles and methods used to carry out a meta-analysis for the main types of data encountered.

Formulae for all the methods described and a much longer discussion of the issues discussed in this section appears in Deeks et al (Deeks 2001a) and Deeks and Altman (Deeks 2001b).

8.6.1 Principles of meta-analysis

All commonly used methods for meta-analysis follow the following basic principles.

1. Meta-analysis is typically a two-stage process. In the first stage, a summary statistic is calculated for each study. For controlled trials, these values describe the treatment effects observed in each individual trial. For example, the summary statistic may be a risk ratio if the data are dichotomous or a difference between means if the data are continuous.
2. In the second stage, a summary (pooled) treatment effect estimate is calculated as a weighted average of the treatment effects estimated in the individual studies. A weighted average is defined as

$$\text{weighted average} = \frac{\text{sum of}\left(\text{estimate} \times \text{weight}\right)}{\text{sum of weights}} = \frac{\sum T_i W_i}{\sum W_i}$$

   where *Ti* is the treatment effect estimated in study *i*, W_i is the weight given to study *i* and the summation is across all studies. Note that if all the weights are the same then the weighted average is equal to the mean treatment effect. The bigger the weight given to study *i* the more it will contribute to the weighted average. The weights are therefore chosen to reflect the amount of information that each trial contains. For ratio measures (OR, RR, etc.) *Ti* is the logarithm of the measure.
3. The combination of treatment effect estimates across studies may optionally incorporate an assumption that the studies are not all estimating the same treatment effect, but estimate treatment effects that follow a distribution across studies. This is the basis of a **random effects meta-analysis** (see Section 8.7.4 Incorporating heterogeneity in random effects models). Alternatively, if it is assumed that each study is estimating exactly the same quantity a **fixed effect meta-analysis** is performed.
4. The standard error of the summary (pooled) treatment effect can be used to derive a confidence interval which communicates the precision (or uncertainty) of the summary estimate, and to derive a *P*-value (significance level) which communicates the strength of the evidence against the null hypothesis of no treatment effect.
5. As well as yielding a summary quantification of the pooled effect, all methods of meta-analysis can incorporate an assessment of whether the variation among the results of the separate studies is compatible with random variation, or whether it is large enough to indicate inconsistency of treatment effects across studies (see 8.7 Heterogeneity).

8.6.2 A generic inverse variance approach to meta-analysis

A very common and simple version of the meta-analysis procedure is commonly referred to as the **inverse variance method**. This approach was implemented in its most basic form in RevMan version 4.2, although it has been used behind the scenes in certain meta-analyses of both dichotomous and continuous data.

The inverse variance method is so named because the weight given to each study is chosen to be the inverse of the variance of the effect estimate (i.e. one over the square of its standard error). Thus larger studies, which have smaller standard errors, are given more weight than

smaller studies, which have larger standard errors. This choice of weight minimises the imprecision (uncertainty) of the pooled effect estimate.

A fixed effect meta-analysis using the inverse variance method calculates a weighted average as

$$\text{generic inverse variance weighted average} = \frac{\sum\left(T_i / S_i^2\right)}{\sum\left(1/S_i^2\right)}$$

where T_i is the treatment effect estimated in study i, S_i is the standard error of that estimate and the summation is across all studies. The basic data required for the analysis are therefore an estimate of the treatment effect and its standard error from each study.

8.6.2.1 Random effects (DerSimonian and Laird) method for meta-analysis

A variation on the inverse variance method is to incorporate an assumption that the different studies are estimating different, yet related, treatment effects. This produces a random effects meta-analysis, and the simplest version is known as the DerSimonian and Laird method (DerSimonian 1986). Random effects meta-analysis is discussed in 8.7.4 Incorporating heterogeneity into random effects models. To undertake a random effects meta-analysis, the standard errors of the study-specific estimates (S_i above) are adjusted to incorporate a measure of the extent of variation, or heterogeneity, among the treatment effects observed in different studies. The size of this adjustment can be estimated from the treatment effects and standard errors of the studies included in the meta-analysis.

8.6.2.2 The generic inverse variance outcome type in RevMan 4.2

Estimates and standard errors may be entered directly into RevMan 4.2 (and subsequent versions) under the 'Generic inverse variance' outcome. The software will undertake fixed effect meta-analyses and random effects (DerSimonian and Laird) meta-analyses, along with assessments of heterogeneity. For ratio measures of treatment effect, the data should be entered as logarithms (for example as a log odds ratio and the standard error of the log odds ratio). However, it is straightforward to instruct the software to display results on the original (e.g. odds ratio) scale. Rather than displaying summary data separately for the treatment groups, the forest plot will display the estimates and standard errors as they were entered beside the study identifiers. It is possible to supplement or replace this with a column providing the sample sizes in the two groups.

Note that the ability to enter estimates and standard errors directly into RevMan creates a high degree of flexibility in meta-analysis. For example, it facilitates the analysis of properly analysed cross-over trials, cluster randomised trials and non-randomized studies, as well as outcome data that are ordinal, time-to-event or rates. However, in most situations for analyses of continuous and dichotomous outcome data it is still preferable to enter more detailed data into RevMan (i.e. specifically as simple summaries of dichotomous or continuous data for each group). This avoids the need for the author to calculate effect estimates, and allows the use of methods targeted specifically at different types of data (see 8.6.3 Meta-analysis of dichotomous outcomes and 8.6.4 Meta-analysis of continuous outcomes). Also, it is helpful for the readers of the review to see the summary statistics for each treatment group in each trial.

8.6.3 Meta-analysis of dichotomous outcomes

There are four widely used methods of meta-analysis for dichotomous outcomes, three fixed effect methods (Mantel-Haenszel, Peto and Inverse Variance) and one random effects method (DerSimonian and Laird). The Mantel-Haenszel, Peto and DerSimonian and Laird methods

are available as options in RevMan analyses for dichotomous data, and the inverse variance analysis can be performed by using the generic inverse variance outcome data method (see 8.6.2.2 The generic inverse variance outcome type in RevMan 4.2). The Peto method can only pool odds ratios whilst the other three methods can pool odds ratios, risk ratios and risk differences. Formulae for all of the meta-analysis methods are given by Deeks et al (Deeks 2001a).

Note that zero cells (e.g. no events in one group) cause problems with computation of estimates and standard errors with some methods. The RevMan software automatically adds 0.5 to each cell of the 2x2 table for any such study.

8.6.3.1 Mantel-Haenszel methods

The Mantel-Haenszel methods (Mantel 1959, Greenland 1985) are the default fixed effect methods of meta-analysis programmed in RevMan. When data are sparse, either in terms of event rates being low or trial size being small, the estimates of the standard errors of the effect estimates that are used in the inverse variance methods may be poor. Mantel-Haenszel methods use a different weighting scheme that depends upon which effect measure (e.g. risk ratio, odds ratio, risk difference) is being used. They have been shown to have better statistical properties when there are few events. As this is a common situation in Cochrane reviews, the Mantel-Haenszel method is generally preferable to the inverse variance method. In other situations the two methods give similar estimates.

8.6.3.2 Peto odds ratio method

Peto's method (Yusuf 1985) can only be used to pool odds ratios. It uses an inverse variance approach but utilises an approximate method of estimating the log odds ratio, and uses different weights. An alternative way of viewing the Peto method is as a sum of 'O – E' statistics. Here, O is the observed number of events and E is an expected number of events in the experimental intervention group of each trial.

The approximation used in the computation of the log odds ratio works well when treatment effects are small (odds ratios are close to one), events are not particularly common and the trials have similar numbers in experimental and control groups. In other situations it has been shown to give biased answers. As these criteria are not always fulfilled, Peto's method is not recommended as a default approach for meta-analysis.

Corrections for zero cell counts are not necessary when using Peto's method. Perhaps for this reason, this method performs well when events are very rare (Deeks 1998a) (see 8.X Rare events (including zero frequencies)). Also, Peto's method can be used to combine dichotomous outcome data with data from time-to-event analyses where log-rank tests have been used (see 8.6.8 Meta-analysis of time-to-event outcomes).

8.6.3.3 DerSimonian and Laird random effects method

The DerSimonian and Laird random effects method (DerSimonian 1986) incorporates an assumption that the different studies are estimating different, yet related, treatment effects. As described in 8.6.2.1 Random effects (DerSimonian and Laird) method for meta-analysis the method is based on the inverse variance approach, making an adjustment to the study weights according to the extent of variation, or heterogeneity, among the varying treatment effects. The DerSimonian and Laird method and the inverse variance method will give identical results when there is no heterogeneity among the studies (and thus also gives results similar to the Mantel-Haenszel method in many situations). Where there is heterogeneity, confidence intervals for the average treatment effect will be wider if the DerSimonian and Laird method is used rather than a fixed effect method, and corresponding claims of statistical significance will be more conservative. It is also possible that the central estimate of the treatment effect

will change if there are relationships between observed treatment effects and sample sizes. See 8.7.4 Incorporating heterogeneity into random effects models for further discussion of these issues.

8.6.3.4 Which measure for dichotomous outcomes?

Summary statistics for dichotomous data are described in 8.2.1 Effect measures for dichotomous outcomes. The effect of treatment can be expressed as either a relative or an absolute effect. The risk ratio (relative risk) and odds ratio are relative measures, while the risk difference and number needed to treat are absolute measures. A further complication is that there are in fact two risk ratios. We can calculate the risk ratio of an event occurring or the risk ratio of no event occurring. These give different pooled results in a meta-analysis, sometimes dramatically so.

The selection of a summary statistic for use in meta-analysis depends on balancing three criteria (Deeks 2002). First, we desire a summary statistic that gives values that are similar for all the trials in the meta-analysis and subdivisions of the population to which the treatment will be applied. The more consistent the summary statistic the greater is the justification for expressing the effect of treatment as a single summary number. Second, the summary statistic must have the mathematical properties required for performing a valid meta-analysis. Third, the summary statistic should be easily understood and applied by those using the review. It should present a summary of the effect of the intervention in a way that helps readers to interpret and apply the results appropriately. Among effect measures for dichotomous data, no single measure is uniformly best, so the choice inevitably involves a compromise.

Consistency: Empirical evidence suggests that relative effect measures are, on average, more consistent than absolute measures. For this reason it is wise to avoid performing meta-analyses of risk differences, unless there is a clear reason to suspect that risk differences will be consistent in a particular clinical situation. On average there is little difference between the odds ratio and risk ratio in this regard (Deeks 2002). When the trial aims to reduce the incidence of an adverse outcome (see 8.2.1.5 What is the event?) there is empirical evidence that risk ratios of the adverse outcome are more consistent than risk ratios of the non-event (Deeks 2002). Selecting an effect measure on the basis of what is the most consistent in a *particular* situation is not a generally recommended strategy, since it may lead to a selection that spuriously maximises the precision of a meta-analysis estimate.

Mathematical properties: The most important mathematical criterion is the availability of a reliable variance estimate. The number needed to treat does not have a simple variance estimator and cannot easily be used directly in meta-analysis, although it can be computed from the other summary statistics (see 8.X Re-expressing meta-analysis results as NNTs). There is no consensus as to the importance of two other often cited mathematical properties: the fact that the behaviour of the odds ratio and the risk difference do not rely on which of the two outcome states is coded as the event, and the odds ratio being the only statistic which is unbounded (see 8.2.1 Effect measures for dichotomous outcomes).

Ease of interpretation: The odds ratio is the hardest summary statistic to understand and to apply in practice, and many practising clinicians report difficulties in using them. There are many published examples where authors have misinterpreted odds ratios from meta-analyses as if they were risk ratios. There must be some concern that routine presentation of the results of systematic reviews as odds ratios will lead to frequent overestimation of the benefits and harms of treatments when the results are applied in clinical practice. Absolute measures of effect are also thought to be more easily interpreted by clinicians than relative effects (Sinclair 1994), although they are less likely to be generalisable.

It seems important to avoid using summary statistics for which there is empirical evidence that they are unlikely to give consistent estimates of treatment effects (the risk difference) and it is impossible to use statistics for which meta-analysis cannot be performed (the number needed to treat). Thus it is generally recommended that analysis proceeds using risk ratios

(taking care to make a sensible choice over which category of outcome is classified as the event) or odds ratios. It may be wise to plan to undertake a sensitivity analysis to investigate whether choice of summary statistic (and selection of the event category) is critical to the conclusions of the meta-analysis (see 8.10 Sensitivity analyses).

It is often sensible to use one statistic for meta-analysis and re-express the results using a second, more easily interpretable statistic. For example, meta-analysis may often be best performed using relative effect measures (risk ratios or odds ratio) and the results re-expressed using absolute effect measures (risk differences or numbers needed to treat – see 8.X Re-expressing meta-analysis results as NNTs). If odds ratios are used for meta-analysis they can also be re-expressed as risk ratios (see 8.2.1 Effect measures for dichotomous outcomes). In all cases the same formulae can be used to convert upper and lower confidence limits. However, it is important to note that all of these transformations require specification of a value of baseline risk indicating the likely risk of the outcome in the population to which the results will be applied. Where the chosen value for baseline risk is close to the average of the control group event rates across the trials the same estimates of NNT will be obtained regardless of whether odds ratios or risk ratios are used for meta-analysis. Where the chosen baseline risk differs from the average control group event rate, the predictions of absolute benefit will differ according to which summary statistic was used for meta-analysis.

8.6.4 Meta-analysis of continuous outcomes

Two methods of analysis are available in RevMan for meta-analysis of continuous data, one fixed effect method and one random effects method. The default fixed effect method uses the inverse variance approach whilst the random effects method uses the DerSimonian and Laird random effects approach. The methods will give exactly the same answers when there is no heterogeneity. Where there is heterogeneity, confidence intervals for the average treatment effect will be wider if the DerSimonian and Laird method is used rather than a fixed effect method, and corresponding *P*-values will be less significant. It is also possible that the central estimate of the treatment effect will change if there are relationships between observed treatment effects and sample sizes. See 8.7.4 Incorporating heterogeneity into random effects models for further discussion of these issues.

Authors should be aware that an assumption underlying methods for meta-analysis of continuous data is that the outcomes have a normal distribution in each treatment arm in each study. This assumption may not always be met, although it is unimportant in very large studies. It is useful to consider the possibility of skewed data (see 8. 5.2.11 Skewed data).

8.6.4.1 Which measure for continuous outcomes?

There are two summary statistics used for meta-analysis of continuous data, the mean difference (MD) and the standardised mean difference (SMD) (see 8.2.2 Effect measures for continuous outcomes). Selection of summary statistics for continuous data is principally determined by whether trials all report the outcome using the same scale (when the mean difference can be used) or using different scales (when the standardised mean difference has to be used).

It is important to note the different roles played in the two approaches by the standard deviations of outcomes observed in the two groups.

For the mean difference method the standard deviations are used together with the sample sizes to compute the weight given to each study. Studies with small standard deviations are given relatively higher weight whilst studies with larger standard deviations are given relatively smaller weights. This is appropriate if variation in standard deviations between studies reflects differences in the reliability of outcome measurements, but is probably not appropriate if the differences in standard deviation reflect real differences in the variability of outcomes in the study populations.

For the standardised mean difference approach the standard deviation is used to standardise the mean differences to a single scale (see 8.2.2.2 The standardised mean difference), as well as in the computation of study weights. It is assumed that variation between standard deviations reflects only differences in measurement scales and not differences in the reliability of outcome measures or variability among trial populations.

These limitations of the methods should be borne in mind where unexpected variation of standard deviations across studies is observed.

8.6.4.2 Meta-analysis of change scores

In some circumstances an analysis based on changes from baseline will be more efficient and powerful than comparison of final values as it removes a component of between person variability from the analysis. However, calculation of a change score requires measurement of the outcome twice and in practice may be less efficient for outcomes which are unstable or difficult to measure precisely, where the measurement error may be larger than true between person baseline variability. Change from baseline outcomes may also be preferred if they have a less skewed distribution than final measurement outcomes. Although sometimes used as a device to ‘correct’ for unlucky randomization, this practice is not recommended.

In practice an author is likely to discover that the trials included in a review may include a mixture of change from baseline and final value scores. However, mixing of outcomes is not a problem when it comes to meta-analysis. There is no statistical reason why trials with change from baseline outcomes should not be combined in a meta-analysis with trials with final measurement outcomes when using the weighted mean difference method in RevMan. In a randomized trial, mean differences based on changes from baseline can usually be assumed to be addressing exactly the same underlying treatment effects as analyses based on final measurements. That is to say, the difference in mean final values will on average be the same as the difference in mean change scores. If the use of change scores does increase precision, the studies presenting change scores will appropriately be given higher weights in the analysis than they would have received if final values had been used, as they will have smaller standard deviations.

When combining the data authors must be careful to use the appropriate means and standard deviations (either of final measurements or of changes from baseline) for each trial. Since the mean values and standard deviations for the two types of outcome may differ substantially it may be advisable to place them in separate subgroups to avoid confusion for the reader, but the results of the subgroups can legitimately be pooled together.

However, final value and change scores should not be combined together as standardised mean differences, since the difference in standard deviation reflects not differences in measurement scale, but differences in the reliability of the measurements.

8.6.5 Combining dichotomous and continuous outcomes

Occasionally authors encounter a situation where data for the same outcome are presented in some studies as dichotomous data and in other studies as continuous data. For example, scores on depression scales can be reported as means or as the percentage of patients who were depressed at some point after an intervention (i.e. with a score above a specified cut-point). This type of information is often easier to understand and more helpful when it is dichotomised. However, deciding on a cut-point may be arbitrary and information is lost when continuous data are transformed to dichotomous data.

There are several options for handling combinations of dichotomous and continuous data. Generally, it is useful to summarise results from all the relevant, valid studies in a similar way, but this is not always possible. It may be possible to collect missing data from investigators so that this can be done. If not, it may be useful to summarise the data in three

ways: by placing the continuous data in a Continuous Data Table, dichotomous data in a Dichotomous Data Table and all of the data in an Other Data Table.

There are statistical approaches available which will re-express odds ratios as standardised mean differences (and vice versa) which allow dichotomous and continuous data to be pooled together, subject to making particular distributional assumptions. Based on an assumption that the underlying distribution of the continuous measurement in each treatment group follows a logistic distribution (which is a symmetrical distribution similar in shape to the normal distribution but with more data in the distributional tails), and that the variability of the outcomes is the same in both treated and control participants, the odds ratios can be re-expressed as a standardised mean difference according to the following simple formula (Chinn 2000):

$$\mathrm{SMD} = \frac{\sqrt{3}}{\pi} \log \mathrm{OR} \, .$$

The standard error of the log odds ratio can be converted to the standard error of a standardised mean difference by multiplying by the same constant (0.5513). Alternatively standardised mean differences can be re-expressed as log odds ratios by multiplying by $\pi/\sqrt{} = 1.8140$.

Once standardised mean differences and standard errors have been computed for all trials in the meta-analysis they can be combined using the generic inverse variance method in RevMan (version 4.2 or later). Standard errors will first need to be computed for all trials by entering the data in RevMan as dichotomous and continuous outcome type data as appropriate, and converting the confidence intervals for the resulting log odds ratios and standardised mean differences into standard errors (see 8.5.6 Obtaining standard errors from confidence intervals and *P*-values).

8.6.6 Meta-analysis of ordinal and measurement scale outcomes

Ordinal and measurement scale outcomes are most commonly meta-analysed as dichotomous data (if so see Section 8.6.3) or continuous data (if so see Section 8.6.4) depending on the way that the trialists performed the original analyses.

Occasionally it is possible to analyse the data using proportional odds models where ordinal scales have a small number of categories, the numbers falling into each category for each treatment group can be obtained, and the same ordinal scale has been used in all trials. This approach may make more efficient use of all available data than dichotomisation, but requires access to advanced statistical software and results in a summary statistic for which it is challenging to find a clinical meaning.

The proportional odds model uses the proportional odds ratio as the measure of treatment difference (Agresti 1996). Suppose that there are 3 categories, which are ordered in terms of desirability such that 1 is the best and 3 the worst. The data could be dichotomised in 2 ways. That is, category 1 constitutes a success and categories 2-3 a failure, or categories 1-2 constitute a success and category 3 a failure. A proportional odds model would assume that there is an equal odds ratio for both dichotomies of the data. Therefore, the odds ratio calculated from the proportional odds model can be interpreted as the odds of success on the experimental intervention relative to control, irrespective of how the ordered categories might be divided into success or failure. Methods (specifically polychotomous logistic regression models) are available for calculating trial estimates of the log odds ratio and its standard error and for conducting a meta-analysis in advanced statistical software packages (Whitehead 1994).

Estimates of log odds ratios and their standard errors from a proportional odds model may be meta-analysed using the generic inverse variance method in RevMan version 4.2 or later (see 8.6.2.2 The generic inverse variance outcome type in RevMan 4.2). Both fixed effect and

random effects methods of analysis are available. If the same ordinal scale has been used in all studies, but has in some reports been presented as a dichotomous outcome, it may still be possible to include all studies in the meta-analysis. In the context of the 3 category model, this might mean that for some studies category 1 constitutes a success, while for others both categories 1 and 2 constitute a success. Methods for dealing with this, and for combining data from scales which are related but have different definitions for their categories are available (Whitehead 1994).

8.6.7 Meta-analysis of counts and rates

Results may be expressed as **count data** when each participant may experience an event, and may experience it more than once. For example, ‘number of strokes’, or ‘number of hospital visits’ are counts. These events may not happen at all, but if they do happen there is no theoretical maximum number of occurrences for an individual.

As described in 8.5.4 Data extraction for counts and rates, count data may be analysed using methods for dichotomous (if so see Section 8.6.3), continuous (if so see Section 8.6.4) and time-to-event data (if so see Section 8.6.8) as well as being analysed as rate data.

Rate data occur if counts are measured for each participant along with the time over which they are observed. This is particularly appropriate when the events being counted are rare. For example, a woman may experience two strokes during a follow-up period of two years. Her **rate** of strokes is one per year of follow up (or, equivalently 0.083 per month of follow-up). Rates are conventionally summarised at the group level. For example, participants in the control group of a trial may experience 85 strokes during a total of 2836 person-years of follow-up. An underlying assumption associated with the use of rates is that the risk of an event is constant across participants and over time. This assumption should be carefully considered for each situation. For example, in contraception studies, rates have been used (known as Pearl indices) to describe the number of pregnancies per 100 women-years of follow-up. This is now considered inappropriate since couples have different risks of conception, and the risk for each woman changes over time. Pregnancies are now analysed more often using life tables or time to event methods that investigate the time elapsing before the first pregnancy.

Analysing count data as rates is not always the most appropriate approach and is uncommon in practice. This is because:

1. the assumption of a constant underlying risk may not be suitable; and
2. statistical methods are not as well developed as they are for other types of data.

The results of a trial may be expressed as a **rate ratio**, that is the ratio of the rate in the intervention group to the rate in the control group. Suppose A events occurred during X participant-years of follow-up in the intervention group, and C events during Y participant-years in the control group. The rate ratio is $(A/X)/(C/Y) = AY/CX$.

The (natural) logarithms of the rate ratios may be combined across trials using the generic inverse variance method (see 8.6.2.2 The generic inverse variance outcome type in RevMan 4.2). An approximate standard error of the log rate ratio is given by $\sqrt{(1/A + 1/C)}$. A correction of 0.5 may be added to each count in the case of zero events. Note that the choice of time unit (i.e. patient-months, women-years, etc) is irrelevant since it is cancelled out of the rate ratio and does not figure in the standard error. However the units should still be displayed when presenting the study results. An alternative means of estimating the rate ratio is through the approach of Whitehead and Whitehead (Whitehead 1991).

In a randomized trial rate ratios may often be very similar to relative risks obtained after dichotomising the participants, since the average period of follow-up should be similar in all

intervention groups. Rate ratios and relative risks will differ, however, if an intervention affects the likelihood of some participants experiencing multiple events.

It is possible also to focus attention on the rate difference, $(A/X) – (C/Y)$. An approximate standard error for the rate difference is $\sqrt{(A/X^2 + C/Y^2)}$. The analysis again requires use of the generic inverse variance method in RevMan. One of the only discussions of meta-analysis of rates, which is still rather short, is that by Hasselblad and McCrory (Hasselblad 1995).

8.6.8 Meta-analysis of time-to-event outcomes

Two approaches to meta-analysis of time-to-event outcomes are available in RevMan. Which is used will depend on what data have been extracted from the primary studies, or obtained from reanalysis of individual patient data.

If logrank 'O – E' and 'V' statistics have been obtained, either through re-analysis of individual patient data or from aggregate statistics presented in the study reports, trial results can be combined using a modified version of the Peto method for dichotomous data (available as the only analysis option for the Individual Patient Data outcome type in RevMan). In the output 'Odds Ratio' will actually mean 'Hazard Ratio'. This is a fixed effect analysis – no equivalent random effects analysis is available in RevMan.

Alternatively if estimates of log hazard ratios and standard errors have been obtained from results of Cox proportional hazards regression models trial results can be combined using the generic inverse variance method (available in RevMan 4.2 and later), see 8.6.2.2 The generic inverse variance outcome type in RevMan 4.2. Both fixed effect and random (DerSimonian and Laird) effects analyses are available.

If a mixture of logrank and Cox model estimates are obtained from the trials, all results can be combined using the generic inverse variance method as the logrank estimates can be converted into log hazard ratios and standard errors using the formulae given in 8.5.5 Data extraction for time-to-event data.

8.6.9 A summary of meta-analysis methods available in RevMan

RevMan includes the following options for statistical analysis:

TYPE OF DATA	SUMMARY STATISTIC	METHOD (F:fixed, R:random)
Dichotomous	odds ratio	Mantel-Haenszel (F)
		Peto (F)
		DerSimonian and Laird (R)
	risk ratio	Mantel-Haenszel (F)
		DerSimonian and Laird (R)
	risk difference	Mantel-Haenszel (F)
		DerSimonian and Laird (R)
Continuous	(weighted) mean difference	inverse variance (F)
		DerSimonian and Laird (R)
	standardised mean difference	inverse variance (F)
		DerSimonian and Laird (R)

Time to event (IPD)	odds/hazard ratio	Peto (F)
Generic inverse variance*	defined by author	inverse variance (F) DerSimonian and Laird (R)

*only available since RevMan 4.2

RevMan requires the author to select one preferred method for each outcome. If these are not specified then the software defaults to the fixed effect Mantel-Haenszel odds ratio for dichotomous outcomes, the fixed effect weighted mean difference for continuous outcomes and the fixed effect model for generic inverse variance outcomes. It is important that authors make it clear which method they are using when results are presented in the text of a review, since it cannot be guaranteed that a meta-analysis displayed to the user will coincide with the selected preferred method.

8.6.10 Use of vote counting for meta-analysis

Occasionally meta-analyses use “vote-counting” to compare the number of positive studies with the number of negative studies. Vote-counting is limited to answering the simple question “is there any evidence of an effect?” Two problems can occur with vote-counting, which suggest that it should be avoided whenever possible. Firstly, problems occur if subjective decisions or statistical significance are used to define “positive“ and “negative” studies (Cooper 1980, Antman 1992). To undertake vote counting properly the number of studies showing harm should be compared with the number showing benefit, regardless of the statistical significance or size of their results. A sign test can be used to assess the significance of evidence for the existence of an effect in either direction (if there is no effect the studies will be distributed evenly around the null hypothesis of no difference). Secondly, vote-counting takes no account of the differential weights given to each study. Vote-counting might be considered as a last resort in situations when standard meta-analytical methods cannot be applied (such as when there is no consistent outcome measure).

8.7 Heterogeneity

8.7.1 What is heterogeneity?

Inevitably, studies brought together in a systematic review will differ. Any kind of variability among studies in a systematic review may be termed heterogeneity. It can be helpful to distinguish between different types of heterogeneity. Variability in the participants, interventions and outcomes studied may be described as **clinical diversity** (sometimes called clinical heterogeneity), and variability in trial design and quality may be described as **methodological diversity** (sometimes called methodological heterogeneity). Variability in the treatment effects being evaluated in the different trials is known as **statistical heterogeneity**, and is a consequence of clinical and/or methodological diversity among the studies. Statistical heterogeneity manifests itself in the observed treatment effects being more different from each other than one would expect due to random error (chance) alone. We will follow convention and refer to **statistical heterogeneity** simply as **heterogeneity**.

Clinical variation will lead to heterogeneity if the treatment effect is affected by the factors that vary across studies – most obviously, the specific interventions or patient characteristics. In other words, the true treatment effect will be different in different studies.

Differences between trials in terms of methodological factors, such as use of blinding and concealment of allocation, or if there are differences between trials in the way the outcomes

are defined and measured, may be expected to lead to differences in the observed treatment effects. Significant statistical heterogeneity arising from methodological diversity or differences in outcome assessments suggests that the studies are not all estimating the same quantity, but does not necessarily suggest that the true treatment effect varies. In particular, heterogeneity associated solely with methodological diversity would indicate the studies suffer from different degrees of bias. Empirical evidence suggests that some aspects of design can affect the result of clinical trials, although this is not always the case. Further discussion appears in Section 6.

The scope of a review will largely determine the extent to which studies included in a review are diverse. Sometimes a review will include trials addressing a variety of questions, for example when several different interventions for the same condition are of interest. Trials of each intervention should be analysed and presented separately (see also 4.5 broad versus narrow questions). Meta-analysis should only be considered when a group of trials is sufficiently homogeneous in terms of participants, interventions and outcomes to provide a meaningful summary. It is often appropriate to take a broader perspective in a meta-analysis than in a single clinical trial. A common analogy is that systematic reviews bring together apples and oranges, and that combining these can yield a meaningless result. This is true if apples and oranges are of intrinsic interest on their own, but may not be if they are used to contribute to a wider question about fruit. For example, a meta-analysis may reasonably evaluate the average effect of a class of drugs by combining results from trials where each evaluates the effect of a different drug from the class.

There may be specific interest in a review in investigating how clinical and methodological aspects of trials relate to their results. Where possible these investigations should be specified a priori, i.e. in the systematic review protocol. It is legitimate for a systematic review to focus on examining the relationship between some clinical characteristic(s) of the studies and the size of treatment effect, rather than on obtaining a summary effect estimate across a series of trials (see 8.8 Investigating heterogeneity). Meta-regression may best be used for this purpose, although it is not implemented in RevMan (see 8.8.3 Meta-regression).

8.7.2 Identifying and measuring heterogeneity

It is important to consider to what extent the results of studies are consistent. If confidence intervals for the results of individual studies (generally depicted graphically using horizontal lines) have poor overlap, this generally indicates the presence of statistical heterogeneity. More formally, a statistical test for heterogeneity is available. This chi-squared test is included in the graphical output of Cochrane reviews. It assesses whether observed differences in results are compatible with chance alone. A low *p*-value (or a large chi-squared statistic relative to its degree of freedom) provides evidence of heterogeneity of treatment effects (variation in effect estimates beyond chance).

Care must be taken in the interpretation of the chi-squared test, since it has low power in the (common) situation of a meta-analysis when trials have small sample size or are few in number. This means that while a statistically significant result may indicate a problem with heterogeneity, a non-significant result must not be taken as evidence of no heterogeneity. This is also why a *P*-value of 0.10, rather than the conventional level of 0.05, is sometimes used to determine statistical significance. A further problem with the test, which seldom occurs in Cochrane reviews, is that when there are many studies in a meta-analysis, the test has high power to detect a small amount of heterogeneity that may be clinically unimportant.

Some argue that, since clinical and methodological diversity always occur in a meta-analysis, statistical heterogeneity is inevitable. Thus the test for heterogeneity is irrelevant to the choice of analysis; heterogeneity will always exist whether or not we happen to be able to detect it using a statistical test. Methods have been developed for quantifying inconsistency across studies that move the focus away from testing whether heterogeneity is present to assessing its impact on the meta-analysis. A useful statistic for quantifying inconsistency is $I^2 = [(Q –$

df)/Q] x 100%, where Q is the chi-squared statistic and df is its degrees of freedom (Higgins 2003, Higgins 2002). This describes the percentage of the variability in effect estimates that is due to heterogeneity rather than sampling error (chance). A value greater than 50% may be considered substantial heterogeneity.

8.7.3 Strategies for addressing heterogeneity

A number of options are available if (statistical) heterogeneity is identified among a group of trials that would otherwise be considered suitable for a meta-analysis.

1. Check again that the data are correct

Severe heterogeneity can indicate that data have been incorrectly extracted or entered into RevMan. For example, if standard errors have mistakenly been entered as standard deviations for continuous outcomes, this could manifest itself in overly narrow confidence intervals with poor overlap and hence substantial heterogeneity. Unit of analysis errors may also be causes of heterogeneity (see 8. 3 Study designs and identifying the unit of analysis).

2. Do not do a meta-analysis

A systematic review need not contain any meta-analyses (O'Rourke 1989). If there is considerable variation in results, and particularly if there is inconsistency in the direction of effect, it may be misleading to quote an average value for the treatment effect.

3. Explore heterogeneity

It is clearly of interest to determine the causes of heterogeneity among results of studies. This process is problematic since there are often many characteristics that vary across studies from which one may choose. Heterogeneity may be explored by conducting subgroup analyses (see 8.8.2 Undertaking subgroup analyses) or meta-regression (8.8.3 Meta-regression), though this latter method is not implemented in RevMan. Ideally, investigations of characteristics of trials that may be associated with heterogeneity should be pre-specified in the protocol of a review (see 8.1.5 Writing the analysis section of the protocol). Reliable conclusions can only be drawn from analyses that are truly pre-specified before inspecting the trials' results, and even these conclusions should be interpreted with caution. In practice, authors will often be familiar with some trial results when writing the protocol, so true pre-specification is not possible. Explorations of heterogeneity that are devised after heterogeneity is identified can at best lead to the generation of hypotheses. They should be interpreted with even more caution and should generally not be listed among the conclusions of a review. Also, investigations of heterogeneity when there are very few studies are of questionable value.

4. Ignore heterogeneity

Fixed effect meta-analyses ignore heterogeneity. The pooled effect estimate from a fixed effect meta-analysis is normally interpreted as being the best estimate of the treatment effect. However, the existence of heterogeneity suggests that there may not be a single treatment effect but a distribution of treatment effects. Thus the pooled fixed effect estimate may be a treatment effect that does not actually exist in any population, and therefore have a confidence interval that is meaningless as well as being too narrow, (see 8.7.4 Incorporating heterogeneity into random effects models). The *P*-value obtained from a fixed effect meta-analysis does however provide a meaningful test of the null hypothesis that there is no effect in every study.

5. Perform a random effects meta-analysis

A random effects meta-analysis may be used to incorporate heterogeneity among trials. This is not a substitute for a thorough investigation of heterogeneity. It is intended primarily for heterogeneity that cannot be explained. An extended discussion of this option appears below (8.7.4 Incorporating heterogeneity into random effects models).

6. Change the effect measure

Heterogeneity may be an artificial consequence of an inappropriate choice of effect measure. For example, when trials collect continuous outcome data using different scales or different units, extreme heterogeneity may be apparent when using the mean difference but not when the more appropriate standardised mean difference is used. Furthermore, choice of effect measure for dichotomous outcomes (odds ratio, relative risk, or risk difference) may affect the degree of heterogeneity among results. In particular, when control group event rates vary, homogeneous odds ratios or risk ratios will necessarily lead to heterogeneous risk differences, and vice versa. However, it remains unclear whether homogeneity of treatment effect in a particular meta-analysis is a suitable criterion for choosing between these measures (see also 8.6.3.4 Which measure for dichotomous outcomes?).

7. Exclude studies

Heterogeneity may be due to the presence of one or two outlying trials with results that conflict with the rest of the trials. In general it is unwise to exclude studies from a meta-analysis on the basis of their results as this may introduce bias. However, if an obvious reason for the outlying result is apparent, the study might be removed with more confidence. Since usually at least one characteristic can be found for any trial in any meta-analysis which makes it different from the others, this criterion is unreliable because it is all too easy to fulfil. It is advisable to perform analyses both with and without outlying trials as part of a sensitivity analysis (see 8.10 Sensitivity analysis). Whenever possible, potential sources of clinical diversity that might lead to such situations should be specified in the protocol.

8.7.4 Incorporating heterogeneity into random effects models

A fixed effect meta-analysis provides a result that may be viewed as a ‘typical treatment effect’ from the studies included in the analysis. In order to calculate a confidence interval for a fixed effect meta-analysis the assumption is made that the true effect of treatment (in both magnitude and direction) is the same value in every study (that is, fixed across studies). This assumption implies that the observed differences among study results are due solely to the play of chance: i.e. that there is no statistical heterogeneity.

When there is heterogeneity that cannot readily be explained, one analytical approach is to incorporate it into a random effects model. A random effects meta-analysis model involves an assumption that the effects being estimated in the different studies are not identical, but follow some distribution. The model represents our lack of knowledge about why real, or apparent, treatment effects differ by considering the differences as if they were random. The centre of this symmetric distribution describes the average of the effects, while its width describes the degree of heterogeneity. The conventional choice of distribution is a normal distribution. It is difficult to establish the validity of any distributional assumption, and this is a common criticism of random effects meta-analyses. The importance of the particular assumed shape for this distribution is not known.

Note that a random effects model does not ‘take account’ of the heterogeneity, in the sense that it is no longer an issue. It is always advisable to explore possible causes of heterogeneity, although there may be too few studies to do this adequately (see 8.8 Investigating heterogeneity).

For random effects analyses in RevMan, the pooled estimate and confidence interval refer to the centre of the distribution of treatment effects, and do not describe the width of the distribution. Often the pooled estimate and its confidence interval are quoted in isolation as an alternative estimate of the quantity evaluated in a fixed effect meta-analysis, which is inappropriate. Note that the confidence interval from a random effects meta-analysis describes uncertainty in the location of the mean of systematically different effects in the different studies. It does not describe the degree of heterogeneity among studies as may be commonly believed. For example, when there are many studies in a meta-analysis, one may obtain a tight confidence interval around the random effects estimate of the mean effect even when there is a large amount of heterogeneity. The range of the treatment effects observed in

the trials may be thought to give a rough idea of the spread of the distribution of true treatment effects, but in fact it will be slightly too wide as it also describes the random error in the observed effect estimates.

If variation in effects (statistical heterogeneity) is believed to be due to clinical diversity, the centre of the distribution should be interpreted differently from the fixed effect estimate since it relates to a different question. The random effects estimate and its confidence interval address the question 'what is the average treatment effect?' while the fixed effect estimate and its confidence interval addresses the question 'what is the best estimate of the treatment effect?' The answers to these questions coincide either when no heterogeneity is present, or when the distribution of the treatment effects is roughly symmetrical. When the answers do not coincide, the random effects estimate may not reflect the actual effect in any particular population being studied.

For any particular set of studies in which heterogeneity is present, a confidence interval around the random effects pooled estimate is wider than a confidence interval around a fixed effect pooled estimate. This will happen if the I^2 statistic is greater than zero, even if the heterogeneity is not detected by the chi-squared test for heterogeneity (Higgins 2003) (see 8.7.2 Identifying and measuring heterogeneity).

In a heterogeneous set of studies, a random effects meta-analysis will award relatively more weight to smaller studies than such studies would receive in a fixed effect meta-analysis. This is because small studies are more informative for learning about the distribution of effects across studies than for learning about an assumed common treatment effect. Care must be taken that random effects analyses are applied only when the idea of a 'random' distribution of treatment effects can be justified. In particular, if results of smaller studies are systematically different from results of larger ones, which can happen as a result of publication bias or low study quality bias, (Poole 1999) (Egger 1997b, Kjaergard 2001), then a random effects meta-analysis will exacerbate the effects of the bias. A fixed effect analysis will be affected less, although strictly it will also be inappropriate. In this situation it may be wise to present neither type of meta-analysis, or to perform a sensitivity analysis in which small studies are excluded.

Similarly, when there is little information, either because there are few trials or if the trials are small with few events, a random effects analysis will provide poor estimates of the width of the distribution of treatment effects. The Mantel-Haenszel method will provide more robust estimates of the average treatment effect, but at the cost of ignoring the observed heterogeneity.

RevMan implements a version of random effects meta-analysis that is described by DerSimonian and Laird (DerSimonian 1986). The attraction of this method is that the calculations are straightforward, but it has a theoretical disadvantage that the confidence intervals are slightly too narrow to encompass full uncertainty resulting from having estimated the degree of heterogeneity. Alternative methods exist that encompass full uncertainty, but they require advanced statistical software (see 8.X Bayesian meta-analysis, 8.X Hierarchical models). In practice, the difference in the results is likely to be small unless there are few studies.

8.8 Investigating heterogeneity

Does the treatment effect vary with different populations or treatment characteristics (such as dose or duration)? Such variation is known as interaction by statisticians and as effect modification by epidemiologists. Methods to search for such interactions include subgroup analyses and meta-regression. All methods have considerable pitfalls.

8.8.1 What are subgroup analyses?

Subgroup analyses involve splitting all the participant data into subgroups, often so as to make comparisons between them. Subgroup analyses may be done for subsets of participants (such as males and females), or for subsets of studies (such as different geographical locations). Subgroup analyses may be done as a means of investigating heterogeneous results, or to answer specific questions about particular patient groups, types of intervention or types of study.

Subgroup analyses of subsets of participants within trials are uncommon in systematic reviews of the literature because sufficient details to extract data about separate participant types are seldom published in reports. By contrast, such subsets of participants are easily analysed when individual patient data have been collected (see Appendix 11a).

Findings from multiple subgroup analyses may be misleading. Subgroup analyses are observational by nature and are not based on randomized comparisons. False negative and false positive significance tests increase in likelihood rapidly as more subgroup analyses are performed. If their findings are presented as definitive conclusions there is clearly a risk of patients being denied an effective intervention or treated with an ineffective (or even harmful) intervention. Subgroup analyses can also generate misleading recommendations about directions for future research that, if followed, would waste scarce resources.

It is useful to distinguish between the notions of 'qualitative interaction' and 'quantitative interaction' (Yusuf 1991). Qualitative interaction exists if the direction of effect is reversed, that is if an intervention is beneficial in one subgroup but is harmful in another. Qualitative interaction is rare. This may be used as an argument that the most appropriate result of a meta-analysis is the overall effect across all subgroups. Quantitative interaction exists when the size of the effect varies but not the direction, that is if an intervention is beneficial to different degrees in different subgroups.

Authors will find useful advice concerning subgroup analyses in Oxman and Guyatt (Oxman 1992) and Yusuf et al (Yusuf 1991). See also 8.8.5 Interpretation of subgroup analyses and meta-regressions.

8.8.2 Undertaking subgroup analyses

Subgroup analyses may be undertaken within RevMan. Meta-analyses within subgroups and meta-analyses that combine several subgroups are both permitted. It is tempting to compare effect estimates in different subgroups by considering the meta-analysis results from each subgroup separately. This should only be done informally by comparing the magnitudes of effect. Noting that either the effect or the test for heterogeneity in one subgroup is statistically significant whilst that in other subgroup is not statistically significant does not indicate that the subgroup factor explains heterogeneity. Since different subgroups are likely to contain different amounts of information and thus have different abilities to detect effects, it is extremely misleading simply to compare the statistical significance of the results.

8.8.2.1 Is the effect different in different subgroups?

Valid investigations of whether an intervention works differently in different subgroups involve comparing the subgroups with each other. No formal method is currently implemented in RevMan. When there are only two subgroups the overlap of the confidence intervals of the summary estimates in the two groups can be considered. Non-overlap of the confidence intervals indicates statistical significance, but note that the confidence intervals can overlap to a small degree and the difference still be statistically significant.

A simple approach for a significance test that can be used to investigate differences between two or more subgroups is described by Deeks et al, although some statistical help may be required (Deeks 2001a). This method uses information given by RevMan when subgroups

and totals are displayed. It is based on the test for heterogeneity chi-squared statistics that appear in the bottom left hand corner of the forest plots, and proceeds as follows. Suppose a chi-squared heterogeneity statistic, Q_{all}, is available for all of the trials, and that chi-squared heterogeneity statistics Q_1 up to Q_m are available for m subgroups (such that every trial is in one and only one subgroup). Then the new statistic $Q_{int} = Q_{all} - (Q_1 + \ldots + Q_m)$, compared with a chi-squared distribution with $m - 1$ degrees of freedom, tests for a difference among the subgroups. (Relevant details of the chi-squared distribution are available as appendices of many statistical textbooks, or using standard computer spreadsheet packages. For example typing **=chidist(5.2,2)** in any cell in a Microsoft Excel spreadsheet will give the P-value for a value of Q_{int} of 5.2 on 2 degrees of freedom). If the values of the heterogeneity chi-squared statistics are obtained from the continuous or generic inverse variance data types in RevMan then there are no problems in using this test. However, if the dichotomous data type is used, then the test will currently include a slight inaccuracy due to the way in which the heterogeneity chi-squared statistic is calculated in RevMan.

A more flexible alternative to testing for differences between subgroups is to use meta-regression techniques, in which residual heterogeneity (that is, heterogeneity not explained by the subgrouping) is allowed (see 8.8.3 Meta-regression).

8.8.3 Meta-regression

If studies are divided into subgroups (see 8.8.2 Subgroup analysis), this may be viewed as an investigation of how a categorical study characteristic is associated with the treatment effects in the meta-analysis. For example, studies in which allocation concealment was adequate may yield different results from those in which allocation concealment was inadequate. Here, allocation concealment, being either adequate or inadequate, is a categorical characteristic at the study level. Meta-regression is an extension to subgroup analyses that allows the effect of continuous, as well as categorical, characteristics to be investigated, and in principle allows the effects of multiple factors to be investigated simultaneously (although this is rarely possible due to inadequate numbers of trials) (Thompson 2002). Meta-regression should generally not be considered when there are fewer than 10 trials in a meta-analysis.

Meta-regressions are similar in essence to simple regressions, in which an **outcome variable** is predicted according to the values of one or more **explanatory variables**. In meta-regression, the outcome variable is the effect estimate (for example, a mean difference, a risk difference, a log odds ratio or a log risk ratio). The explanatory variables are characteristics of studies that might influence the size of treatment effect. These are often called 'potential effect modifiers' or covariates. Meta-regressions usually differ from simple regressions in two ways. First, larger studies have more influence on the relationship than smaller studies, since studies are weighted by the precision of their respective effect estimate. Second, it is wise to allow for the residual heterogeneity among treatment effects not modelled by the explanatory variables. This gives rise to the term 'random effects meta-regression', since the extra variability is incorporated in the same way as in a random effects meta-analysis (Thompson 1999).

The regression coefficient obtained from a meta-regression analysis will describe how the outcome variable (the treatment effect) changes with a unit increase in the explanatory variable (the potential effect modifier). The statistical significance of the regression coefficient is a test of whether there is a linear relationship between treatment effect and the explanatory variable. If the treatment effect is a ratio measure, the log-transformed value of the treatment effect should always be used in the regression model (see 8.2.6 Expressing treatment effects on log scales), and the exponential of the regression coefficient will give an estimate of the relative change in treatment effect with a unit increase in the explanatory variable.

Meta-regression can also be used to investigate differences for categorical explanatory variables as done in subgroup analyses. If there are m subgroups membership of particular

subgroups is indicated by using *m*-1 dummy variables (which can only take values of zero or one) in the meta-regression model (as in standard linear regression modelling). The regression coefficients will estimate how the treatment effect in each subgroup differs from a nominated reference subgroup. The *P*-value of each regression coefficient will indicate whether this difference is statistically significant.

Meta-regression is currently best performed using the 'metareg' macro in the Stata statistical package (Sterne 2001).

8.8.4 Selection of study characteristics for subgroup analyses and meta-regression

Authors need to be cautious about undertaking subgroup analyses, and interpreting any that they do. Some considerations are outlined here for selecting characteristics (also called explanatory variables, potential effect modifiers or covariates) which will be investigated for their possible influence on the size of the treatment effect. These considerations apply similarly to subgroup analyses and to meta-regressions. Further details may be obtained from Oxman and Guyatt (Oxman 1992) and Berlin and Antman (Berlin 1994).

8.8.4.1 Ensure that there are adequate studies to justify subgroup analyses and meta-regressions

It is very unlikely that an investigation of heterogeneity will produce useful findings unless there is a substantial number of studies. It is worth noting the typical advice for undertaking simple regression analyses: that at least ten observations (i.e. ten studies in a meta-analysis) should be available for each characteristic modelled.

8.8.4.2 Specify characteristics in advance

Authors should, whenever possible, pre-specify characteristics in the protocol that later will be subject to subgroup analyses or meta-regression. Pre-specifying characteristics reduces the likelihood of spurious findings, first by limiting the number of subgroups investigated and second by preventing knowledge of the trials' results influencing which subgroups are analysed. True pre-specification is difficult in systematic reviews, because the results of some of the relevant trials are often known when the protocol is drafted. If a characteristic was overlooked in the protocol, but is clearly of major importance and justified by external evidence, then authors should not be reluctant to explore it. However, such post hoc analyses should be identified as such.

8.8.4.3 Select a small number of characteristics

The likelihood of a false positive result among subgroup analyses and meta-regression increases with the number of characteristics investigated. It is difficult to suggest a maximum number of characteristics to look at, especially since the number of available studies is unknown in advance. If more than one or two characteristics are investigated it may be sensible to adjust the level of significance to account for making multiple comparisons. The help of a statistician is recommended (see 8.X Multiple comparisons and the play of chance).

8.8.4.4 Ensure there is scientific rationale for investigating each characteristic

Selection of characteristics should be motivated by biological and clinical hypotheses, ideally supported by evidence from sources other than the included studies. Subgroup analyses using characteristics that are implausible or clinically irrelevant are not likely to be useful and should be avoided. For example, a relationship between treatment effect and year of

publication is seldom in itself clinically informative, and if statistically significant runs the risk of initiating a post-hoc data dredge of factors that may have changed over time.

Prognostic factors are those that predict the outcome of a disease or condition, whereas effect modifiers are factors that influence how well a treatment works in affecting the outcome. Confusion between prognostic factors and effect modifiers is common in planning subgroup analyses, especially at the protocol stage. Prognostic factors are not good candidates for subgroup analyses unless they are also believed to modify the effect of treatment. For example, being a smoker may be a strong predictor of mortality within the next ten years, but there may not be reason for it to influence the effect of a drug therapy on mortality (Deeks 1998b). Potential effect modifiers may include the precise interventions (dose of active treatment, choice of comparison treatment), how the study was done (length of follow-up) or methodology (design and quality).

8.8.4.5 Be aware that the effect of a characteristic may not always be identified

Many characteristics that might have important effects on how well an intervention works cannot be investigated using subgroup analysis or meta-regression. These are characteristics of participants that might vary substantially within studies, but which can only be summarised at the level of the study. An example is age. Consider a collection of clinical trials involving adults ranging from 18 to 60 years old. There may be a strong relationship between age and treatment effect that is apparent within each study. However, if the mean ages for the trials are similar, then no relationship will be apparent by looking at trial mean ages and trial-level effect estimates. The problem is one of aggregating individuals' results and is variously known as aggregation bias, ecological bias or the ecological fallacy (Morgenstern 1982, Greenland 1987, Berlin 2002). It is even possible for the differences between trials to display the opposite pattern to that observed within each trial.

8.8.4.6 Think about whether the characteristic is closely related to another characteristic (confounded)

The problem of 'confounding' complicates interpretation of subgroup analyses and meta-regressions and can lead to incorrect conclusions. Two characteristics are confounded if their influences on the treatment effect cannot be disentangled. For example, if those studies implementing an intensive version of a therapy happened to be the studies that involved patients with more severe disease, then one cannot tell which aspect is the cause of any difference in effect estimates between these studies and others. In meta-regression, co-linearity between potential effect modifiers leads to similar difficulties as is discussed by Berlin and Antman (Berlin 1994). Computing correlations between trial characteristics will give some information about which trial characteristics may be confounded with each other.

8.8.5 Interpretation of subgroup analyses and meta-regressions

Appropriate interpretation of subgroup analyses and meta-regressions requires caution. For more detailed discussion see Oxman and Guyatt (Oxman 1992).

- Subgroup comparisons are observational

It must be remembered that subgroup analyses and meta-regressions are entirely observational in their nature. These analyses investigate differences between trials, and while individuals are randomised to one group or other within a trial, they are not randomised to go in one trial or another. Hence, subgroup analyses suffer the limitations of any observational investigation, including possible bias through confounding by other trial-level characteristics. Furthermore, even a genuine difference between subgroups is not necessarily due to the classification of the subgroups. As an example, a subgroup analysis of bone marrow transplantation for treating leukaemia might show a strong association between the age of a sibling donor and the success

of the transplant. However, this probably does not mean that the age of donor is important. In fact, the age of the recipient is probably a key factor and the subgroup finding would simply be due to the strong association between the age of the recipient and the age of their sibling.

- Was the analysis pre-specified or post hoc?

Authors should state whether subgroup analyses were pre-specified or undertaken after the results of the studies had been compiled (post hoc). More reliance may be placed on a subgroup analysis if it was one of a small number of pre-specified analyses. Performing numerous post hoc subgroup analyses to explain heterogeneity is data dredging. Data dredging is condemned because it is usually possible to find an apparent, but false, explanation for heterogeneity by considering lots of different characteristics.

- Is there indirect evidence in support of the findings?

Differences between subgroups should be clinically plausible and supported by other external or indirect evidence, if they are to be convincing.

- Is the magnitude of the difference practically important?

If the magnitude of a difference between subgroups will not result in different recommendations for different subgroups, then it may be better to present only the overall analysis results.

- Is there a statistically significant difference between subgroups?

To establish whether there is a different effect of an intervention in different situations, the magnitudes of effects in different subgroups should be compared directly with each other. In particular, statistical significance of the results within separate subgroup analyses (as presented in RevMan) should not be compared. See 8.8.2 Undertaking subgroup analyses.

- Are analyses looking at within-study or between-study relationships?

For patient and intervention characteristics, differences in subgroups that are observed within studies are more reliable than analyses of subsets of studies. If such within-study relationships are replicated across studies then this adds confidence to the findings.

8.8.6 Investigating the effect of baseline risk

One potentially important source of heterogeneity among a series of studies is when the underlying average risk of the outcome event varies between the studies. The baseline risk of a particular event may be viewed as an aggregate measure of case-mix factors such as age or disease severity. It is generally measured as the observed risk of the event in the control group of each trial (the control group risk (CGR) or control event rate (CER)). The notion is controversial in its relevance to clinical practice since baseline risk represents a summary of both known and unknown risk factors. Problems also arise because baseline risk will depend on the length of follow-up, which often varies across studies. However, baseline risk has received particular attention in meta-analysis because the information is readily available once dichotomous data have been prepared for use in meta-analyses. A full discussion of the subject appears in Sharp (Sharp 2000).

Intuition would suggest that participants are more or less likely to benefit from an effective treatment according to their risk status. However, the relationship between baseline risk and treatment effect is a complicated issue. For example, suppose a treatment is equally beneficial in the sense that for all patients it reduces the risk of an event, say a stroke, to 80% of the baseline risk. Then it is not equally beneficial in terms of absolute differences in risk in the sense that it reduces a 50% stroke rate by 10 percentage points to 40% (number needed to treat = 10), but a 20% stroke rate by 4 percentage points to 16% (number needed to treat = 25).

Use of different summary statistics (risk ratio, odds ratio and risk difference) will demonstrate different relationships with baseline risk. Summary statistics that show close to no

relationship with baseline risk are generally preferred for use in meta-analysis (see 8.6.3.4 Which measure for dichotomous outcomes?).

Investigating any relationship between effect estimates and the control group risk is also complicated by a technical phenomenon known as regression to the mean. This arises because the control group risk forms an integral part of the effect estimate. A high risk in a control group, observed entirely by chance, will on average give rise to a higher than expected effect estimate, and vice versa. This phenomenon results in a false correlation between effect estimates and control group risks. Methods are available, requiring sophisticated software, that correct for regression to the mean (McIntosh 1996, Thompson 1997). These should be used for such analyses and statistical expertise is recommended.

8.8.7 Dose-response analyses

The principles of meta-regression can be applied to the relationships between treatment effect and dose (commonly termed dose-response), treatment intensity or treatment duration (Greenland 1992, Berlin 1993). Conclusions about differences in effect due to differences in dose (or similar factors) are on strongest ground if participants are randomized to one dose or another within a study and a consistent relationship is found across similar studies. While authors should consider these effects, particularly as a possible explanation for heterogeneity, they should be cautious about drawing conclusions based on between-study differences. Authors should be particularly cautious about claiming that a dose-response relationship does not exist, given the low power of many meta-regression analyses to detect genuine relationships.

8.8.8 Indirect comparisons

Indirect comparisons are made between interventions in the absence of head-to-head randomized studies. Consider the situation in which some trials have compared the effectiveness of doctors versus dieticians in providing dietary advice, and others the effectiveness of nurses versus dieticians, but no trials have compared the effectiveness of doctors versus nurses. We might then wish to learn about the relative effectiveness of doctors versus nurses.

The problem can be considered as an investigation of a source of heterogeneity (different intervention) in a subgroup analysis. The trials should be considered in two separate subgroups, one of doctors versus dieticians and one of nurses versus dieticians. The difference between the summary effects in the two subgroups will be an estimate of the difference between doctors and nurses. The significance of this difference is best assessed by using meta-regression, although for this particular example the approach is equivalent to a simpler procedure described by Bucher (Bucher 1997). The validity of an indirect comparison relies on the two subgroups of trials being similar on average in other factors that may affect outcome.

One approach that should never be used is the direct comparison of the relevant single arms of the trials. For example, patients receiving advice from a nurse in the nurse versus dietician trials should not be compared directly with patients receiving advice from a doctor in the doctor versus dietician trials. This comparison ignores the potential benefits of randomization and suffers from the same (usually extreme) biases as a comparison of independent cohort studies.

Indirect comparisons are not randomized comparisons, and cannot be interpreted as such. They are essentially observational findings across trials, and may suffer the biases of observational studies, for example due to confounding (see 8.8.5 Interpretation of subgroup analyses and meta-regressions). In situations when both direct and indirect comparisons are available in a review, then unless there are design flaws in the head-to-head trials, the two

approaches should always be considered separately and the direct comparisons should take precedence as a basis for forming conclusions.

8.9 Presenting, illustrating and tabulating results

Several types of figures and tables are available for the presentation of results in a Cochrane review. This section reviews those available in RevMan, and describes how to incorporate results produced outside of RevMan. First we address some issues to consider when reporting results in the text of the review.

8.9.1 Presenting results in the text

The results of individual studies and meta-analyses in a Cochrane review are displayed in Figures and Tables. Each Figure and Table should be referred to in the results section of the review text. The results section should summarise the findings in a clear and logical order, and should explicitly address the objectives of the review. The section should be organised to follow the order of comparisons and outcomes specified in the protocol, and used as the data structure in RevMan.

Findings for the most important comparisons and/or outcomes should be prominent in the text of the review, even when little relevant data were available. Answers to post hoc analyses and less important questions for which there happen to be plentiful data should not be overemphasised. *Post hoc* analyses should always be identified as such.

The analytic methods that are used in a review should be described in the methods section. The author should also make clear in the results section the method of analysis used for each quoted result (in particular, the choice of effect measure, the direction of a beneficial effect and the meta-analysis model used). Results should always be accompanied by a measure of uncertainty, such as a 95% confidence interval.

Authors should consider presenting results in formats that are easy to interpret. For example, odds ratios and standardized mean differences do not lend themselves to direct application in clinical practice but can be re-expressed in more accessible forms. See 8.X Re-expressing standardised mean differences and 8.X Re-expressing meta-analysis results as NNTs.

The abstract should summarise findings for only the most important comparisons and outcomes, and not selectively report those with the most significant results. It is helpful also to indicate the amount of information (numbers of studies and participants) on which analyses were based.

Methods for meta-analysis allow quantification of direction of effect, size of effect and consistency of effect. If suitable numerical data are not available for meta-analysis, or if meta-analyses are considered inappropriate, then these domains may often still be examined to provide a systematic assessment of the evidence available (see 8.1 Planning the analysis).

8.9.2 Figures

Graphical displays provide a clear and systematic means of presenting results both from individual studies and from meta-analyses. However, reviews that contain large numbers of figures are often difficult to follow, especially when each figure contains very little information.

The standard graphic in Cochrane reviews is the forest plot, which doubles as both a Table and a Figure. The graphical section of a forest plot displays effect estimates and confidence intervals for both individual studies and meta-analyses. Each study is represented by a block at the point estimate of treatment effect with a horizontal line extending either side of the

block. The area of the block indicates the weight assigned to that study in the meta-analysis while the horizontal line depicts the confidence interval (usually with a 95% level of confidence). The area of the block and the confidence interval convey similar information, but both make different contributions to the graphic. The confidence interval depicts the range of treatment effects compatible with the study's result and indicates whether each was individually statistically significant. The size of the block draws the eye towards the studies with larger weight (narrower confidence intervals), which dominate the calculation of the pooled result.

8.9.2.1 Forest plots in RevMan

RevMan produces forest plots and similar plots are automatically incorporated into the published version of the Cochrane review. The different options for analyses, including the choice between fixed and random effects meta-analyses are available as options when forest plot figures are viewed in RevMan. Default analyses are displayed unless options are overridden. The defaults are Mantel-Haenszel odds ratios for dichotomous data, fixed effect meta-analyses of (weighted) mean differences for continuous data, Peto odds ratios for IPD outcomes and (in RevMan 4.2 and later) fixed effect meta-analyses for generic inverse variance outcomes. The author should override any default settings that do not correspond with results reported in the text when setting up or editing outcomes in RevMan. This ensures that the results displayed are consistent with what is described in the text. Note that some users of the Cochrane Database of Systematic Reviews will be able to select alternative summary statistics and meta-analysis models to those intended by the author when they view the results.

A past convention in CDSR has been that dichotomous outcomes have focussed on unfavourable outcomes, so that risk ratios and odds ratios less than one (and risk differences less than zero) indicate that an experimental intervention is superior to a control intervention. This would result in effect estimates to the left of the vertical line in a forest plot implying a benefit of the experimental intervention. The convention is no longer encouraged since it is not universally appropriate, and a much superior approach is to make it transparent which side of the line indicates benefit of which intervention by labelling the directions on the axis on the forest plots. RevMan allows authors to specify the labels used for 'treatment' and 'control' groups in each outcome. These labels are then used in the CDSR. Thus it is essential to know which way round figures are constructed and should be interpreted. This is particularly important for measurement scale data where it is not always apparent to a reader which direction on a scale indicates worsening health.

Presentation of data as a forest plot is discouraged where no study or only a single study is found for a particular outcome, except in circumstances where a blank forest plot makes a particular point about the lack of available data for an important outcome. To display outcomes noted only in single studies a forest plot using a subgroup for each outcome can be used (ensuring that the option to pool the data is disabled). Otherwise results of single studies may more conveniently be presented in an Additional Table (see 8.9.3 Tables).

Forest plots for dichotomous outcomes and IPD outcomes illustrate, by default:

1. The raw data (corresponding to the 2x2 tables) for each study;
2. Point estimates and confidence intervals for the chosen effect measure, both as blocks and lines and as text;
3. A meta-analysis for each subgroup using the chosen effect measure and chosen method (fixed or random effects), both as a diamond and as text;
4. The total numbers of participants in the experimental intervention and control intervention groups;
5. Heterogeneity statistics (the chi-squared test and the I^2 statistic);

6. A test for overall effect (overall average effect for random effects meta-analyses);
7. The total numbers of events in the experimental intervention and control intervention groups;
8. Percent weights given to each study.

Note that 3-8 are not displayed unless data are pooled. RevMan 4.2 separates 7 from 4, whereas earlier versions presented them together. This led to some confusion since it wrongly suggested to some users that the meta-analysis had been computed by comparing the totals of participants and events between experimental and control groups. For IPD outcomes it is also possible to enable display of the O – E and V statistics.

Forest plots for continuous outcomes illustrate, by default:

1. The raw data (means, standard deviations and sample sizes) for each arm in each study;
2. Point estimates and confidence intervals for the chosen effect measure, both as blocks and lines and as text;
3. A meta-analysis for each subgroup using the chosen effect measure and chosen method (fixed or random effects), both as a diamond and as text;
4. The total numbers of participants in the experimental and control groups;
5. Heterogeneity statistics (the chi-squared test and the I^2 statistic);
6. A test for overall effect (overall average effect for random effects meta-analyses);
7. Percent weights given to each study.

Note that 3-7 are not displayed unless the data are pooled.

Forest plots for the generic inverse variance method illustrate, by default:

1. The summary data for each study, as entered by the author (for ratio measures these will be on the log scale);
2. Point estimates and confidence intervals, both as blocks and lines and as text (for ratio measures these will be on the natural scale rather than the log scale);
3. A meta-analysis for each subgroup using the chosen method (fixed or random effects), both as a diamond and as text;
4. Heterogeneity statistics (the χ^2 test and the I^2 statistic);
5. A test for overall effect (overall average effect for random effects meta-analyses);
6. Percent weights given to each study.

Note that 3-6 are not shown unless data are pooled. It is possible additionally to enter sample sizes for experimental and control groups. These should be entered as appropriate for the design of the study. The sample sizes are not involved in the analysis, but if entered are displayed as:

7. Numbers of participants in the experimental and control group for each study;
8. The total numbers of participants in the experimental and control groups.

8.9.2.2 Additional figures

Additional figures may be attached to reviews in RevMan 4.2 and later. Examples of figures that authors may wish to include in a review include:

1. forest plots where each line represents a meta-analysis rather than a study (for example, to illustrate multiple subgroup analyses or sensitivity analyses);
2. funnel plots;

3. plots illustrating meta-regression analyses;
4. L'Abbé plots

Note that although funnel plots may be drawn using RevMan, they may only be included in the published review by attaching them as additional figures. Additional figures should not often be required, and should not be used to draw forest plots that can currently be drawn using RevMan. Where possible graphics should be produced using specialist statistical software packages such as Stata, SAS, SPSS, S-Plus or specialised meta-analysis software which produce appropriate publication quality graphics. General purpose spreadsheet programs may not provide suitable flexibility nor produce output of adequate quality.

A separate document (Appendix 8a) is available that provides extensive guidance on the content of additional figures that illustrate numerical data. The document includes descriptions and recommendations for the four plots listed above among others. Authors should refer to this document before submitting a review containing additional figures. All additional figures should be assessed by a statistical editor or advisor prior to submission of a Cochrane review for publication. Authors should be aware that additional figures can often be large and take up valuable storage space on the Cochrane Library. Guidance on technical aspects of additional figures is available at http://www.cc-ims.net.

Important results from all additional figures should be overviewed in the Results section of the review text. Wherever numerical results taken from a Figure are reported in the text of the review their meaning and derivation should be clear, and a reference to the relevant Figure should be provided.

8.9.3 Tables

RevMan supports three types of tables of results that can be linked to the Results text of the review.

1. Forest plots generated by RevMan present summary data and effect estimates alongside their graphical representation (see 8.9.2.1 Forest plots in RevMan).
2. The Table of Comparisons allows an outcome type of 'Other data'. Results of individual trials may be entered here as plain text. This option is well suited for entering summary data such as median values which cannot be pooled in a meta-analysis
3. A flexible way of creating tables is provided by the Additional Tables feature, allowing presentation of results of both trials and meta-analyses, and other meta-analytical investigations (such as meta-regression analyses).

For further information see the RevMan User Guide.

Note that descriptions of study characteristics (methods, participants, interventions and outcomes studied) should be presented in the Table of Characteristics of Included Studies. Study results should not be included in this table.

The ability to incorporate additional figures in RevMan 4.2 and later technically allows authors to attach further additional tables as graphics files. Authors are discouraged from doing this due to the high volume of storage space taken up by graphics files. Authors are instead asked to use the Additional Tables function, which is provided for this purpose.

Important results from all tables should be discussed and summarised in the Results section of the review text. When numerical results are reported in the text of the review a reference to the relevant Table should be provided.

8.10 Sensitivity analyses

Because there are different approaches to conducting a systematic review, authors should ask: How sensitive are the results of the analysis to changes in the way it was done? This provides authors 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 author must identify how the key decisions and assumptions 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 for the types of study (e.g. using different methodological cut-points), participants, interventions or outcome measures
- including or excluding studies where there is some ambiguity as to whether they meet the inclusion criteria
- reanalysing the data using a reasonable range of results for studies where there may be some uncertainty about the results (e.g. 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 (e.g. using a random effects model instead of a fixed effect model, or *vice versa*)

The same cautions discussed for subgroup analyses apply to sensitivity analyses. In particular, since many sensitivity analyses involve between study subgroup comparisons, these findings need to be interpreted very carefully.

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

8.11 Special topics

8.11.1 Publication bias and funnel plots

As discussed in section 5, a particularly important component of a review is the identification of relevant studies. Publication bias has long been recognised as a problem in this regard since it means that the likelihood of finding studies is related to the results of those studies (Begg 1988, Begg 1989, Easterbrook 1991, Dickersin 1992b). One way to investigate whether a review is subject to publication bias is to prepare a ‘funnel plot’ and examine this for signs of asymmetry. RevMan 4.0 includes a facility to produce such a graph. However, if there is asymmetry, reasons other than publication bias should also be considered.

Funnel plots were first used in educational research and psychology (Light 1984a). They are simple scatter plots of the treatment effects estimated from individual studies (on the x axis) against some measure of each study’s sample size (y axis). The name ‘funnel plot’ arises from the fact that precision in the estimation of the true treatment effect increases as the sample size of the component studies increases. Effect estimates from small studies will therefore scatter more widely at the bottom of the graph, with the spread narrowing among larger studies. In the absence of bias the plot should resemble a symmetrical inverted funnel (see panel A of the figure).

Relative measures of treatment effect (such as relative risks and odds ratios) are plotted on a logarithmic scale. This ensures that effects of the same magnitude but opposite directions (for example relative risks of 0.5 and 2) are equidistant from 1.0 (Galbraith 1988). Treatment effects have generally been plotted against sample sizes. However, the statistical power of a trial is determined both by its total sample size and the number of participants experiencing the event of interest. For example, a study with 100,000 patients and 10 events is less likely to show a statistically significant treatment effect than a study with 1000 patients and 100 events. The standard error (SE) or the variance of the effect estimate, rather than total sample size, have therefore been increasingly used for the y axis in funnel plots. RevMan 4.0 uses 1/SE, plotted against the effect size calculated by the statistical method chosen by the reviewer for the particular outcome.

If there is bias, for example because smaller studies without statistically significant effects (shown as open circles in the figure) remain unpublished, this will lead to an asymmetrical appearance of the funnel plot with a gap in a bottom corner of the graph (panel B). In this situation the effect calculated in a meta-analysis will overestimate the treatment effect (Villar 1997, Egger 1997b). The more pronounced the asymmetry, the more likely it is that the amount of bias will be substantial.

Publication bias has long been associated with funnel plot asymmetry (Light 1984a). However the funnel plot should be seen as a generic means of examining whether the smaller studies in a meta-analysis tend to show larger treatment effects and this may be due to reasons other than publication bias (Egger 1998b, Egger 1998a). Some of these are shown in the table:

Possible sources of asymmetry in funnel plots

1. Selection biases
    - Publication bias
    - Location biases
        - Language bias
        - Citation bias
        - Multiple publication bias
2. Poor methodological quality of smaller studies
    - Poor methodological design
    - Inadequate analysis
    - Fraud
3. True heterogeneity
    - Size of effect differs according to study size (for example, due to differences in the intensity of interventions or differences in underlying risk between studies of different sizes)
4. Artefactual
5. Chance

Even if a study has been published, the probability of finding it is also influenced by its results. For example, language bias (the preferential publication of studies without significant findings in languages other than English), makes it less likely that such ‘negative’ studies will be found (Grégoire 1995, Egger 1997c). Citation bias leads to ‘negative’ studies being referred to less often and they are therefore more likely to be missed when searching for relevant trials (Gotzsche 1987, Gotzcshe 1989, Ravnskov 1992). Conversely, results of

'positive' trials are sometimes reported more than once, increasing the probability that they will be located (multiple publication bias) (Gotzcshe 1989, Huston 1996, Tramèr 1997).

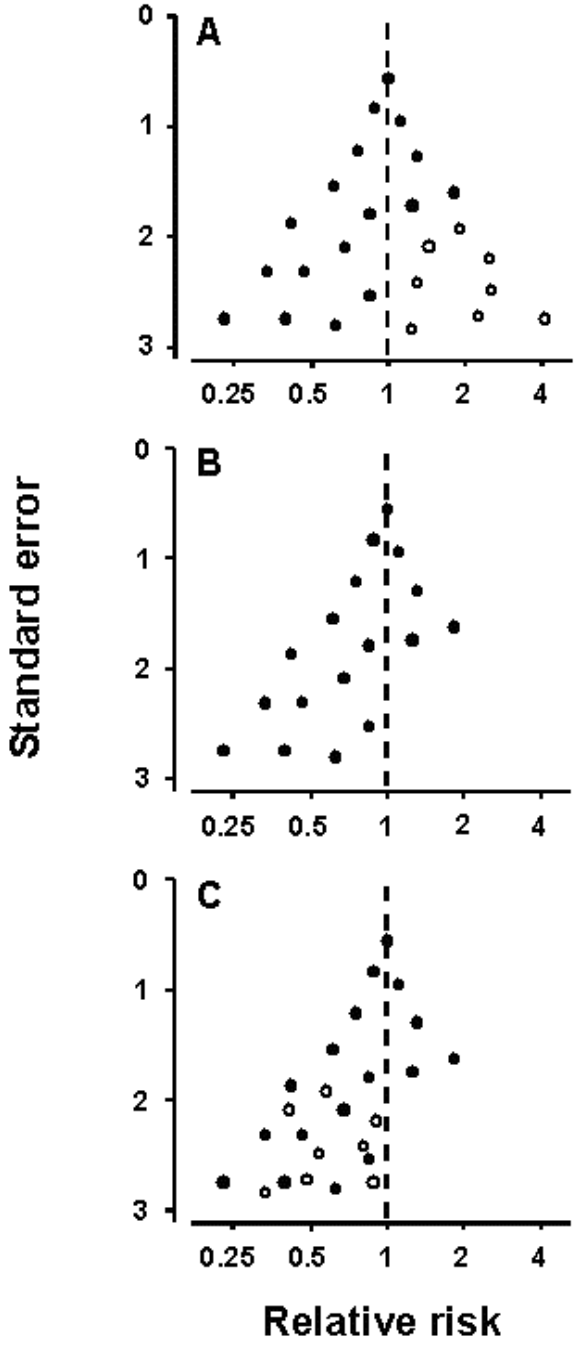
Another source of asymmetry arises from differences in methodological quality. Smaller studies are, on average, conducted and analysed with less methodological rigour than larger studies. Trials of lower quality also tend to show larger treatment effects (Schulz 1995, Moher 1998). Trials which, if conducted and analysed properly, would have been 'negative' may thus become 'positive' (panel C).

True heterogeneity in treatment effects may also lead to funnel plot asymmetry. For example, substantial benefit may be seen only in patients at high risk for the outcome which is affected by the intervention and these high risk patients are usually more likely to be included in early, small studies (Davey Smith 1994, Glasziou 1995). In addition, small trials are generally conducted before larger trials are established and in the intervening years standard treatment may have improved. Furthermore, some interventions may have been implemented less thoroughly in larger trials and, therefore, have resulted in smaller estimates of the treatment effect (Stuck 1998). It has also been argued that funnel plot asymmetry may be artefactual (Irwig 1998), but simulation studies have shown that this will occur infrequently, if the overall treatment effect is very large and the outcome of interest is rare (Sterne 2000). Finally, it is, of course, possible that an asymmetrical funnel plot arises merely by the play of chance.

Symmetry or asymmetry is generally defined informally, through visual examination, but the visual interpretation of funnel plots may vary between observers (Villar 1997). More formal statistical methods to examine associations between the study effects and size have been proposed (Begg 1994, Egger 1997b). At present there is debate regarding the statistical properties, potentials and limitations of these tests (Naylor 1997, Irwig 1998, Seagrott 1998, Egger 1998b). No such tests are available in RevMan 4.0. Methodological work examining these issues is currently underway, but it is clear that both visual examination and statistical analysis of funnel plots have limited power to detect bias if the number of studies is small.

Authors should look at the relevant funnel plot whenever they do a meta-analysis. If asymmetry is present, likely reasons should be explored. The power of this method is, however, at its most limited in those situations when bias is most likely to distort the results of the meta-analyses: when it comprises only a few small studies. Finally, it should be remembered that although funnel plots may alert authors to a problem which needs considering, they do not provide a solution to this problem. The only satisfactory way to address reporting bias and the inadequate quality of individual trials is through prospective registration of trials (Simes 1986, Dickersin 1988, Anonymous 1991, Dickersin 1992a) and improvements in the quality of the conduct, analysis and reporting of studies, meta-analyses and systematic reviews (Begg 1996, Moher 1995b).

Legend to figure: Hypothetical funnel plots. Panel A: symmetrical plot in the absence of bias; Panel B: asymmetrical plot in the presence of reporting bias, Panel C: asymmetrical plot in the presence of bias due to low methodological quality of smaller studies.



8.11.2 Cluster-randomized trials

In cluster-randomized trials, groups of individuals rather than individuals are randomized to different interventions. Cluster-randomized trials are also known as group-randomized trials. We say the ‘unit of allocation’ is the cluster, or the group. The groups may be, for example, schools, villages, medical practices or families. Such trials may be done for one of several reasons. It may be to evaluate the group effect of an intervention, for example herd-immunity of a vaccine. It may be to avoid ‘contamination’ across interventions when trial participants are managed within the same setting, for example in a trial evaluating a dietary intervention,

families rather than individuals may be randomized. A cluster-randomized design may be used simply for convenience.

One of the main consequences of a cluster design is that participants within any one cluster often tend to respond in a similar manner, and thus their data can no longer be assumed to be independent of one another. Many of these studies, however, are incorrectly analysed as though the unit of allocation had been the individual participants. This is often referred to as a 'unit of analysis error' (Whiting-O'Keefe 1984) because the unit of analysis is different from the unit of allocation. If the clustering is ignored and cluster trials are analysed as if individuals had been randomized, resulting P values will be artificially small. This can result in false positive conclusions that the intervention had an effect. In the context of a meta-analysis, studies in which clustering has been ignored will have overly narrow confidence intervals and will receive more weight than is appropriate in a meta-analysis. This situation can also arise if participants are allocated to interventions that are then applied to parts of them (for example, to both eyes or to several teeth), or if repeated observations are made on a patient. If the analysis is by the individual units (for example, each tooth or each observation) without taking into account that the data are clustered within participants, then a unit of analysis error can occur.

There are several useful sources of information on cluster-randomized trials (Donner 2000)(Donner 2001a)(Murray 1995). A detailed discussion of incorporating cluster-randomized trials in a meta-analysis is available (Donner 2002), as is a more technical treatment of the problem (Donner 2001b). Special considerations for analysis of standardised mean differences from cluster-randomised trials are discussed by White and Thomas (White 2005).

8.11.2.1 Methods of analysis for cluster-randomized trials

One way to avoid unit of analysis errors is to conduct the analysis at the same level as the allocation, using a summary measurement from each cluster. Then the sample size is the number of clusters and analysis proceeds as if the trial was individually randomized (though the clusters become the individuals). However, this might considerably, and unnecessarily, reduce the power of the study, depending on the number and size of the clusters.

Alternatively, statistical methods now exist that allow analysis at the level of the individual while accounting for the clustering in the data. The ideal information to extract from a cluster-randomized trial is a direct estimate of the required effect measure (for example, an odds ratio with its confidence interval) from an analysis that properly accounts for the cluster design. Such an analysis might be based on a 'multilevel model', a 'variance components analysis' or may use 'generalised estimating equations (GEEs)', among other techniques. Statistical advice is recommended to determine whether the method used is appropriate. Effect estimates and their standard errors from correct analyses of cluster-randomized trials may be meta-analysed using the generic inverse variance method in RevMan version 4.2 or later.

8.11.2.2 Approximate analyses of cluster-randomized trials for a meta-analysis: Effective sample sizes

Unfortunately, many cluster-randomized trials have in the past failed to report appropriate analyses. They are commonly analysed as if the randomization was performed on the individuals rather than the clusters. If this is the situation, approximately correct analyses may be performed if the following information can be extracted:

- the number of clusters (or groups) randomized to each intervention group; or the average (mean) size of each cluster
- the outcome data ignoring the cluster design for the total number of individuals (for example, number or proportion of individuals with events, or means and standard deviations).

- an estimate of the intracluster (or intraclass) correlation coefficient (ICC)

The ICC is an estimate of the relative variability within and between clusters (Donner 1980). It describes the 'similarity' of individuals within the same cluster. In fact this is seldom available in published reports. A common approach is to use external estimates obtained from similar studies (Ukoumunne 1999). ICCs may appear small compared with other types of correlations: values lower than 0.05 are typical. However, even small values can have a substantial impact on confidence interval widths (and hence weights in a meta-analysis), particularly if cluster sizes are large. An example below provides an illustration.

An approximately correct analysis proceeds as follows. The idea is to reduce the size of each trial to its 'effective sample size' (Rao 1992). The effective sample size of a single intervention group in a cluster-randomized trial is its original sample size divided by a quantity called the 'design effect'. The design effect is $1+(m-1)r$, where m is the average cluster size and r is the intracluster correlation coefficient. A common design effect is usually assumed across intervention groups. For dichotomous data both the number of participants and the number experiencing the event can be divided by the same design effect. Since the resulting data must be rounded to whole numbers for entry into RevMan this approach may be unsuitable for small trials. For continuous data only the sample size need be reduced; means and standard deviations should remain unchanged.

8.11.2.3 Example of incorporating a cluster-randomized trial

As an example, consider a cluster-randomized trial that randomized 10 school classrooms with 295 children into an intervention group and 11 classrooms with 330 children into a control group. The numbers of successes among the children, ignoring the clustering, are

Intervention: 63/295

Control: 84/330

Imagine an intracluster correlation coefficient of 0.02 has been obtained from a reliable external source. The average cluster size in the trial is (295+330)/(10+11) = 29.8. The design effect for the trial as a whole is then $1+(m-1)r = 1 + (29.8 - 1)\times 0.02 = 1.576$. The effective sample size in the intervention group is 295 / 1.576 = 187.2 and for the control group is 330 / 1.576 = 209.4.

Applying the design effects also to the numbers of events produces the following results:

Intervention: 40.0/187.2

Control: 53.3/209.4

Once trials have been reduced to their effective sample size, the data may be entered into any version of RevMan as, for example, dichotomous outcomes or continuous outcomes. Results from the example trial may be entered as

Intervention: 40/187

Control: 53/209

8.11.2.4 Approximate analyses of cluster-randomized trials for a meta-analysis: Inflating standard errors

A clear disadvantage of the above approach is the need to round the effective sample sizes to whole numbers. A slightly more flexible approach, which is equivalent to calculating effective sample sizes, is to multiply the standard error of the effect estimate (from an analysis ignoring clustering) by the square root of the design effect. The standard error may be calculated from a confidence interval (see 8.5.6 Obtaining standard errors from confidence intervals and P-values). Standard analyses of dichotomous or continuous outcomes may be used to obtain these confidence intervals using RevMan. The meta-analysis using the inflated

variances may be performed using RevMan (version 4.2 or later) using the generic inverse variance method.

8.11.2.5 Issues in the incorporation of cluster-randomized trials

Cluster-randomized trials may, in principle, be combined with individually randomized trials in the same meta-analysis. Consideration should be given to the possibility of important differences in the effects being evaluated between the different types of trial. There are often good reasons for performing cluster-randomized trials and these should be examined. For example, in the treatment of infectious diseases an intervention applied to all individuals in a community may be more effective than treatment applied to select (randomized) individuals within the community since it may reduce the possibility of re-infection.

Authors should always identify any cluster-randomized trials in a review and explicitly state how they have dealt with the data. They should conduct sensitivity analyses to investigate the robustness of their conclusions, especially when ICCs have been borrowed from external sources (see Section 8.10 Sensitivity analyses). Statistical support is recommended.

8.11.3 Cross-over trials

Parallel group trials allocate each participant to a single intervention for comparison with one or more alternative interventions. In contrast, cross-over trials allocate each participant to a sequence of interventions. A simple randomized cross-over design is an AB/BA design in which participants are randomized initially to intervention A or intervention B, and then 'cross over' to intervention B or intervention A, respectively. It can be seen that data from the first period of a cross-over trial represent a parallel group trial, a feature referred to below.

Cross-over designs offer a number of possible advantages over parallel group trials. Among these are (i) that each participant acts as his or her own control, eliminating among-patient variation; (ii) that, consequently, fewer participants are required to obtain the same power; and (iii) that every participant receives every intervention, which allows the determination of the best intervention or preference for an individual patient. A readable introduction to cross-over trials is given by Senn (Senn 2002). More detailed discussion of meta-analyses involving cross-over trials is provided by Elbourne et al (Elbourne 2002).

8.11.3.1 Assessing suitability of cross-over trials

Cross-over trials are suitable for evaluating interventions with a temporary effect in the treatment of stable, chronic conditions. They are employed, for example, in the study of interventions to relieve asthma and epilepsy. They are not appropriate when an intervention can have a lasting effect that compromises entry to subsequent periods of the trial, or when a disease has a rapid evolution. The advantages of cross-over trials must be weighed against their disadvantages. The principal problem associated with cross-over trials is that of carry-over (a type of period-by-intervention interaction). Carry-over is the situation in which the effects of an intervention given in one period persist into a subsequent period, thus interfering with the effects of a different subsequent intervention. Many cross-over trials include a period between interventions known as a washout period as a means of reducing carry-over. If a primary outcome is irreversible (for example mortality, or pregnancy in a subfertility study) then a cross-over study is generally considered to be inappropriate. Another problem with cross-over trials is the risk of drop-out due to their longer duration compared with comparable parallel group trials. The analysis techniques for crossover trials with missing observations are limited.

In considering the inclusion of cross-over trials in meta-analysis, authors should first address the question of whether a cross-over trial is a suitable method for the condition and intervention in question. For example, although they are frequently employed in the field, one

group of authors decided cross-over trials were inappropriate for studies in Alzheimer's disease due to the degenerative nature of the condition, and included only data from the first period (Qizilbash 1998). The second question to be addressed is whether there is a likelihood of serious carry-over, which relies largely on judgement since the statistical techniques to demonstrate carry-over are far from satisfactory. The nature of the interventions and the length of any washout period are important considerations.

It is only justifiable to exclude cross-over trials from a systematic review if the design is inappropriate to the clinical context. Very often, however, it is difficult or impossible to extract suitable data from a cross-over trial. Below we outline some considerations and suggestions for including cross-over trials in a meta-analysis.

8.11.3.2 Methods of analysis for cross-over trials

If carry-over is not thought to be a problem then the appropriate analysis of continuous data from an AB/BA cross-over trial is a paired t-test. This evaluates the value of 'measurement on intervention A' minus 'measurement on intervention B' separately for each patient. The mean and standard error of these difference measures are the building blocks of a statistical test and an effect estimate. The effect estimate may be included in a meta-analysis using the generic inverse variance method, which may be performed using RevMan version 4.2 or later.

A paired analysis is possible if the data in any one of the following bullet points is available:

• individual patient data from the paper or by correspondence with the trialist;

• the mean and standard deviation (or standard error) of the patient-specific differences between intervention A and intervention B measurements;

• the mean difference (or difference between means) and one of the following: (i) a t-statistic from a paired t-test; (ii) a P-value from a paired t-test; (iii) a confidence interval from a paired analysis;

• a graph of measurements on intervention A and intervention B from which individual data values can be extracted, as long as matched measurements for each individual can be identified as such.

For details see Elbourne et al (Elbourne 2002).

8.11.3.3 Methods for incorporating cross-over trials into a meta-analysis

Unfortunately, the reporting of cross-over trials has been very variable, and the data required to include a paired analysis in a meta-analysis are often not published. A common situation is that means and standard deviations (or standard errors) are available only for measurements on A and B separately. A simple approach to incorporating cross-over trials in a meta-analysis is thus to take all measurements from intervention A periods and all measurements from intervention B periods and analyse these as if the trial were a parallel group trial of A versus B. This approach gives rise to a unit of analysis error (8.3 Study designs and identifying the unit of analysis) and should be avoided unless it can be demonstrated that the results approximate those from a paired analysis, as described above. The reason for this is that confidence intervals are likely to be too wide, and the trial will receive too little weight, with the possible consequence of disguising clinically important heterogeneity. Nevertheless, this incorrect analysis is conservative, in that studies are under-weighted rather than over-weighted. The unit of analysis error is not as serious as some other types of unit of analysis error.

A second approach to incorporating cross-over trials is to include only data from the first period. This is particularly recommended if carry-over is thought to be a problem, or if a cross-over design is considered inappropriate for other reasons. However, it is possible that available data from first periods constitute a biased subset of all first period data. This is

because reporting of first period data may be dependent on the trialists having found statistically significant carry-over.

A third approach to incorporating inappropriately reported cross-over trials is to attempt to approximate a paired analysis, by imputing a measure describing the similarity of outcomes within each participant. We address this approach in detail in the next section.

Cross-over trials with dichotomous outcomes require more complicated methods and consultation with a statistician is recommended (Elbourne 2002).

8.11.3.4 Approximate analyses of cross-over trials for a meta-analysis

A 'hidden' number known as the correlation coefficient describes how similar the measurements on interventions A and B were within a participant. Here we describe (1) how to estimate the correlation coefficient from a study that is reported in considerable detail and (2) how to approximate a paired analysis in a different study, making use of an imputed correlation coefficient. Note that the methods in (2) are applicable both to correlation coefficients obtained using (1) and to correlation coefficients obtained in other ways (for example, by reasoned argument).

(1) Suppose a study involving n participants is available that presents the following information:

Intervention A (sample size *n*)	*mean(A), SD(A)*
Intervention B (sample size *n*)	*mean(B), SD(B)*
Difference between A and B (sample size *n*)	*mean(*diff*), SD(*diff*)*

We can use the data from this study to estimate the correlation coefficient, *r*, as follows. This assumes that the mean and standard deviation of measurements for intervention A is the same when it is given in the first period as when it is given in the second period (and similarly for intervention B).

$$r = \frac{SD(\mathrm{A})^2 + SD(\mathrm{B})^2 - SD(\mathrm{diff})^2}{2 \times SD(\mathrm{A}) \times SD(\mathrm{B})},$$

Correlation coefficients lie between –1 and 1. If zero or a negative number is obtained, then there is no benefit of using a cross-over design over using a parallel group design. Before imputation is undertaken it is recommended that correlation coefficients are computed for as many studies as possible and it is noted whether or not they are consistent. Imputation should be done only as a very tentative analysis if correlations are inconsistent.

(2) The point estimate for the paired analysis is the same as for a parallel group analysis, since the mean of the differences is equal to the difference in means:

$$MD = mean(\mathrm{diff}) = mean(\mathrm{A}) - mean(\mathrm{B}).$$

To impute the standard deviation of these differences, we use an imputed value R for the correlation coefficient. The value R might be imputed from another study in the meta-analysis (using the method in (1) above), it might be imputed from elsewhere, or it might be

hypothesised based on reasoned argument. In all of these situations, a sensitivity analysis should be undertaken, trying different values of R, to determine whether the overall result of the analysis is robust to the use of imputed correlation coefficients.

To obtain a standard deviation of the differences, use

$$SD(\text{diff}) = \sqrt{SD(\text{A})^2 + SD(\text{B})^2 - (2 \times R \times SD(\text{A}) \times SD(\text{B}))}\,.$$

Finally, the standard error of the mean difference is obtained as

$$SE(\text{MD}) = SD(\text{diff}) \big/ \sqrt{n}$$

The quantities MD and SE(MD) may be entered into RevMan (version 4.2 or later) under the generic inverse variance outcome type.

8.11.3.5 Example of incorporating a cross-over trial

As an example, suppose a cross-over trial reports the following data:

Intervention A (sample size 10)	*mean*(A) = 7.0, *SD*(A) = 2.38
Intervention B (sample size 10)	*mean*(B) = 6.5, *SD*(B) = 2.21

The estimate of the mean difference is MD = 7.0 – 6.5 = 0.5. Using an imputed correlation coefficient of 0.68, we can impute the standard deviation of the differences as:

$$SD(\text{diff}) = \sqrt{2.38^2 + 2.21^2 - (2 \times 0.68 \times 2.38 \times 2.21)} = 1.84$$

and the standard error of the mean difference is

$$SE(\text{MD}) = 1.84 \big/ \sqrt{10} = 0.58\,.$$

The numbers 0.5 and 0.58 may be entered into RevMan as the estimate and standard error of a mean difference. Correlation coefficients other than 0.68 might be used as part of a sensitivity analysis.

8.11.3.6 Issues in the incorporation of cross-over trials

Cross-over trials may, in principle, be combined with parallel group trials in the same meta-analysis. Consideration should be given to the possibility of important differences in other characteristics between the different types of trial. For example, cross-over trials may have shorter intervention periods or may include participants with less severe illness. It is generally advisable to meta-analyse parallel-group and cross-over trials separately irrespective of whether they are also combined together.

Authors should explicitly state how they have dealt with data from cross-over trials and should conduct sensitivity analyses to investigate the robustness of their conclusions, especially when correlation coefficients have been borrowed from external sources (see Section 8.10 Sensitivity analyses). Statistical support is recommended.

8.12 Contributions

Contributing authors: Doug Altman, Deborah Ashby, Jacqueline Birks, Michael Borenstein, Marion Campbell, Jon Deeks, Matthias Egger, Julian Higgins, Joseph Lau, Keith O'Rourke, Rob Scholten, Jonathan Sterne, Simon Thompson, Anne Whitehead

Comments on drafts (statistical): Doug Altman, Deborah Ashby, Jesse Berlin, Joseph Beyene, Jacqueline Birks, Michael Bracken, Marion Campbell, Chris Cates, Mike Clarke, Albert Cobos, Francois Curtin, Roberto D'Amico, Keith Dear, Jon Deeks, Heather Dickinson, Diana Elbourne, Simon Gates, Paul Glasziou, Peter Herbison, Julian Higgins, Sally Hollis, David Jones, Steff Lewis, Nathan Pace, Craig Ramsey, Keith O'Rourke, Rob Scholten, Guido Schwarzer, Jonathan Sterne, Simon Thompson, Andy Vail, Clarine van Oel, Paula Williamson, Fred Wolf

Comments on drafts (non-statistical): Bodil Als-Nielsen, Wendong Chen, Esther Coren, Christian Gluud, Philippa Middleton, Jack Sinclair

8.13 References

Agresti 1996. Agresti A. An Introduction to Categorical Data Analysis. New York: Wiley, 1996.

Altman 1996. Altman DG, Bland JM. Detecting skewness from summary information. BMJ 1996; 313: 1200-1200.

Anonymous 1991. Anonymous. Making clinical trialists register. Lancet 1991; 338:244-5.

Antman 1992. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. JAMA 1992; 268: 240-248.

Begg 1988. Begg CB, Berlin JA. Publication bias: a problem in interpreting medical data. J Roy Stat Soc A 1988; 151:419-63.

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

Begg 1994. Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088-99.

Begg 1996. Begg CB, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637-639.

Berg 1988. Berg L. Clinical Dementia Rating (CDR). Psychopharm Bull 1988;24:637-639.

Berlin 1993. Berlin JA, Longnecker MP, Greenland S. Meta-analysis of epidemiologic dose-response data. Epidemiology 1993; 4: 218-228.

Berlin 1994. Berlin JA, Antman EM. Advantages and limitations of metaanalytic regressions of clinical trials data. Online J Curr Clin Trials 1994; Doc No 134.

Berlin 2002. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman KA. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. Stat Med 2002; 21: 371-387.

Bucher 1997. Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta- analysis of randomized controlled trials. J Clin Epidemiol 1997; 50: 683-691.

Chinn 2000. Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. Stat Med 2000; 19: 3127-3131.

Collett 1994. Collett D. Modelling Survival Data in Medical Research. London: Chapman and Hall, 1994.

Cooper 1980. Cooper HM, Rosenthal R. Statistical versus traditional procedures for summarizing research findings. Psychol Bull 1980; 87: 442-449.

Davey Smith 1994. Davey Smith G, Egger M. Who benefits from medical interventions? Treating low risk patients can be a high risk strategy. Br Med J 1994;308:72-4.

Deeks 1997a. Deeks J. Are you sure that's a standard deviation? (part 1). Cochrane News 1997; Number 10; 11-12.

Deeks 1997b. Deeks J. Are you sure that's a standard deviation? (part 2). Cochrane News 1997; Number 11: 11-12.

Deeks 1998a. Deeks JJ, Bradburn MJ, Localio R, Berlin J. Much ado about nothing: Meta-analysis for rare events. 6th Cochrane Colloquium, Baltimore, 1998.

Deeks 1998b. Deeks JJ. Systematic reviews of published evidence: Miracles or minefields? Annals of Oncology 1998; 9: 703-709.

Deeks 2001a. Deeks JJ, Altman DG, Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publication Group, 2001.

Deeks 2001b. Deeks JJ, Altman DG. Effect measures for meta-analysis of trials with binary outcomes. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publication Group, 2001.

Deeks 2002. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002; 21: 1575-1600.

DerSimonian 1986. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clin Trials 1986; 7: 177-188.

Dickersin 1988. Dickersin K. Report from the panel on the Case for Registers of Clinical Trials at the Eighth Annual Meeting of the Society for Clinical Trials. Controlled Clin Trials 1988; 9:76-81.

Dickersin 1992a. Dickersin K. Keeping posted. Why register clinical trials? - revisited. Controlled Clin Trials 1992; 13:170-7.

Dickersin 1992b. 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.

Donner 1980. Donner A, Koval JJ. The estimation of intraclass correlation in the analysis of family data. Biometrics 1980; 36: 19-25

Donner 2000. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. London: Arnold, 2000

Donner 2001a. Donner A, Klar N. Special Issue: Design and analysis of cluster randomized trials. Statistics in Medicine 2001; 20: 329-496

Donner 2001b. Donner A, Piaggio G, Villar J. Statistical methods for the meta-analysis of cluster randomized trials. Statistical Methods in Medical Research 2001; 10: 325-338

Donner 2002. Donner A, Klar N. Issues in the meta-analysis of cluster randomized trials. Statistics in Medicine 2002; 21: 2971-80.

Early Breast Cancer Trialists' Collaborative Group 1990. Early Breast Cancer Trialists' Collaborative Group. Treatment of Early Breast Cancer. Volume 1: Worldwide Evidence 1985-1990 . Oxford: Oxford University Press, 1990.

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

Egger 1997a. Egger M, Davey Smith G, Phillips AN. Meta-analysis: principles and procedures. BMJ 1997;315:1533-7.

Egger 1997b. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997; 315: 629-634.

Egger 1997c. Egger M, Zellweger-Zähner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997;350:326-329.

Egger 1998. Egger M, Davey Smith G, Minder C. Authors' reply. Br Med J 1998;316:471.

Elbourne 2002. Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: Methodological issues. International Journal of Epidemiology 2002; 31, 140-149.

Follman 1992. Follmann D, Elliott P, Suh I, Cutler J. Variance imputation for overviews of clinical trials with continuous response. J Clin Epidemiol 1992; 45: 769-773.

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

Glasziou 1995. Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. BMJ 1995;311:1356-9.

Gotzsche 1987. Gotzsche PC. Reference bias in reports of drug trials. Br Med J 1987;295:654-6.

Gotzsche 1989. Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis [published erratum appears in Controlled Clin Trials 1989;50:356]. Controlled Clin Trials 1989; 10:31-56.

Greenland 1987. Greenland S. Quantitative methods in the review of epidemiologic literature. Epidemiol Rev 1987; 9 : 1-30.

Greenland 1992. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. Am J Epidemiol 1992; 135: 1301-1309.

Greenland 1985. Greenland S, Robins J. Estimation of a common effect parameter from sparse follow-up data. Biometrics 1985; 41: 55-68.

Grégoire 1995. Grégoire G, Derderian F, LeLorier J. Selecting the language of the publications included in a meta-analysis: is there a Tower of Babel bias? J Clin Epidemiol 1995;48:159-63.

Hasselblad 1995. Hasselblad VIC, Mccrory DC. Meta-analytic tools for medical decision making: A practical guide. Med Decis Making 1995; 15: 81-96.

Higgins 2002. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med 2002; 21: 1539-1558.

Higgins 2003. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327: 557-560.

Hollis 1999. Hollis S, Campbell F. What is meant by intention to treat analysis? BMJ 1999; 319: 670-674.

Huston 1996. Huston P, Moher D. Redundancy, disaggregation, and the integrity of medical research. Lancet 1996;347:1024-6.

Irwig 1998. Irwig L, Macaskill P, Berry G. Graphical test is itself biased [letter]. Br Med J 1998;316:470.

Kjaergard 2001. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001; 135: 982-989.

Laupacis 1988. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. New Engl J Med 1988; 318: 1728-1733.

Lewis 1993. Lewis JA, Machin D. Intention to treat--who should use ITT? Br J Cancer 1993; 68: 647-650.

Light 1984. Light RJ, Pillemer DB. Summing up. The science of reviewing research. Cambridge, Massachusetts, and London, England: Harvard University Press, 1984.

Mantel 1959. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 1959; 22: 719-748.

McIntosh 1996. McIntosh MW. The population risk as an explanatory variable in research synthesis of clinical trials. Stat Med 1996; 15: 1713-1728.

Moher 1998. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609-13.

Morgenstern 1982. Morgenstern H. Uses of ecologic analysis in epidemiologic research. Am J Public Health 1982; 72: 1336-1344.

Murray 1995. Murray DM, Short B. Intraclass correlation among measures related to alcohol-use by young-adults - estimates, correlates and applications in intervention studies. Journal of Studies on Alcohol 1995; 56: 681-694.

Naylor 1997. Naylor CD. Meta-analysis and the meta-epidemiology of clinical research. Br Med J 1997;315 :617-9.

Newell 1992. Newell DJ. Intention-to-treat analysis: implications for quantitative and qualitative research. Int J Epidemiol 1992; 21: 837-841.

O'Rourke 1989. O'Rourke K, Detsky AS. Meta-analysis in medical research: strong encouragement for higher quality in individual research efforts. J Clin Epidemiol 1989; 42: 1021-1026.

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

Parmar 1998. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med 1998; 17: 2815-2834.

Poole 1999. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. Am J Epidemiol 1999; 150: 469-475.

Qizilbash 1998. Qizilbash N, Whitehead A, Higgins J, Wilcock G, Schneider L, Farlow M. Cholinesterase inhibition for Alzheimer disease - A meta-analysis of the tacrine trials. JAMA 1998; 280: 1777-1782.

Rao 1992. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. Biometrics 1992; 48: 577-585.

Ravnskov 1992. Ravnskov U. Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. Br Med J 1992;305:15-9.

Sackett 1996. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! Evidence Based Medicine 1996; 1: 164-166.

Sackett 1997. Sackett DL, Richardson WS, Rosenberg W, Haynes BR. Evidence-Based Medicine: How to Practice and Teach EBM. Edinburgh: Churchill Livingstone, 1997.

Schulz 1995. Schulz KF, Chalmers I, Hayes RJ, Altman D. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408-12.

Seagrott 1998. Seagrott V, Stratton I. Test had 10% false positive rate [letter]. Br Med J 1998;316:470.

Senn 2002. Senn S. Cross-Over Trials in Clinical Research. 2nd edition. Chichester: John Wiley and sons, 2002

Sharp 2000. Sharp SJ. Analysing the relationship between treatment benefit and underlying risk: precautions and practical recommendations. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publications Group, 2000.

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

Sinclair 1994. Sinclair JC, Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. J Clin Epidemiol 1994; 47: 881-889.

Sterne 2000. Sterne JA, Gavaghan D, Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. J Clin Epidemiol 2000;53(11):1119-29.

Sterne 2001. Sterne JAC, Bradburn MJ, Egger M. Meta-analysis in StataTM. In: Egger M, Davey Smith G, Altman DG (Eds). Systematic Reviews in Health Care: Meta-Analysis in Context (2nd edition). London: BMJ Publication Group, 2001.

Stuck 1988. Stuck AE, Rubenstein LZ, Wieland D. Bias in meta-analysis detected by a simple, graphical test. Asymmetry detected in funnel plot was probably due to true heterogeneity [letter]. Br Med J 1998;316:469-71.

Thompson 1997. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 1997; 16: 2741-2758.

Thompson 1999. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Stat Med 1999; 18: 2693-2708.

Thompson 2002. Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? Stat Med 2002; 21: 1559-1574.

Tramèr 1997. Tramèr MR, Reynolds DJM, Moore RA, McQuay HJ. Impact of covert duplicate publication on meta-analysis: a case study. Br Med J 1997;315:635-40.

Ukoumunne 1999. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JA, Burney PG. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. Health Technology Assessment 1999; 3: iii-92.

Unnebrink 2001. Unnebrink K, Windeler J. Intention-to-treat methods for dealing with missing values in clinical trials of progressively deteriorating diseases. Stat Med 2001; 20: 3931-3946.

Villar 1997. Villar J, Piaggio G, Carroli G, Donner A. Factors affecting the comparability of meta-analyses and largest trials results in perinatology. J Clin Epidemiol 1997;50:997-1002.

White 2005. White IR,Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. Clinical Trials 2005; 2: 141–151.

Whitehead 1991. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. Stat Med 1991; 10: 1665-1677.

Whitehead 1994. Whitehead A, Jones NMB. A meta-analysis of clinical trials involving different classifications of response into ordered categories. Stat Med 1994; 13: 2503-2515.

Whiting-O'Keefe 1984. Whiting-O'Keefe QE, Henke C, Simborg DW (1984). Choosing the correct unit of analysis in medical care experiments. Med Care 1984; 22:1101-14.

Yusuf 1985. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Progress in Cardiovascular Diseases 1985; 27: 335-371.

Yusuf 1991. Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991; 266: 93-98.

8.14 Sections under construction

You may have been directed to the following sections, which are currently under construction.

8.X Issues in interpretation

8.X Other types of study

8.X Missing data

8.X Investigating and dealing with bias

8.X Where to go for help

8.X Re-expressing meta-analysis results as NNTs

8.X Rare events (including zero frequencies)

8.X Re-expressing standardised mean differences

8.X Trials with more than two treatment groups

8.X Sensitivity analyses

8.X Bayesian meta-analysis

8.X Hierarchical models

8.X Multiple comparisons and the play of chance

These parts of Section 8 are from an earlier version and will be replaced soon.

8.10 Sensitivity analyses

8.11.1 Publication bias and funnel plots

9 Interpreting results

Although it can be argued that the results of a systematic review should stand on their own, many people faced with a decision look to the Discussion and Authors' Conclusions for help interpreting the results. Indeed, many people prefer to go directly to the conclusions before looking at the rest of the review.

Discussion and conclusions about the following issues can help people to make decisions:

- The strength of the evidence
- The applicability of the results
- Other information, such as considerations of costs and current practice, that might be relevant to someone making a decision
- Clarification of any important trade-offs between the expected benefits, harms and costs of the intervention

Because Cochrane reviews have an international audience, the discussion and authors' (reviewers') conclusions should, so far as possible, assume a broad international perspective, rather than addressing specific national or local circumstances. Authors should be particularly careful to bear in mind that different people might make different decisions based on the same evidence. The primary purpose of the review should be to present information, rather than to offer advice. The discussion and conclusions should be to help people to understand the implications of the evidence in relationship to practical decisions. Recommendations that depend on assumptions about resources and values should be avoided.

9.1 Strength of evidence

A good starting point for the discussion section of a review is to address any *important* methodological limitations of the included trials and the methods used in the review that might affect practical decisions about healthcare or future research. This should not be a detailed discussion of study or review methods. Information provided in the section of the review on methodological quality need not be repeated here.

It is often helpful to discuss how the included studies fit into the context of other evidence that is not included in the review. For example, for reviews of drug therapy it may be relevant to refer to dosage studies or non-randomised studies of the risk of rare adverse events. It should be stated clearly whether the other evidence that is referenced was systematically reviewed when other types of evidence are cited.

One type of evidence that can be helpful in considering the likelihood of a cause-effect relationship between an intervention and an important outcome is indirect evidence of a relationship. This includes evidence relating to intermediate outcomes (such as physiological or biochemical measures that are markers for risk of the outcome of interest), evidence from studies of different populations (including animal studies) and evidence from analogous relations (i.e. similar interventions).

Because conclusions regarding the strength of inferences about the effectiveness of an intervention are essentially causal inferences, authors might want to consider guidelines for assessing the strength of a causal inference, such as those put forward by Hill (Hill 1971). In the context of a systematic review of clinical trials, 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?
- Have other plausible competing explanations of the observed effects (eg. bias or co-intervention) been ruled out?

More or less explicit approaches to grading the strength of evidence underlying a conclusion are available (CTFPHE 1979, Cook 1992, Gyorkos 1994 Guyatt 1995, US PSTF 1996), although there is no single approach that is universally accepted as being appropriate for the wide range of reviews included in the *Cochrane Database of Systematic Reviews*. A Collaborative Review Group (CRG) may elect to use a standard approach to grading the strength of evidence across its reviews. Over time, it may be possible for the Cochrane Collaboration as a whole to develop a more consistent and explicit approach to drawing conclusions about the overall strength of evidence for the main conclusions of a review. However, it is currently up to individual authors, in consultation with others in their CRG, to select an approach to summarising the strength of evidence that is appropriate for the question being reviewed.

9.2 Applicability

'A leap of faith is always required when applying any study findings to the population at large' or to a specific person. *'In making that jump, one must always strike a balance between making justifiable broad generalizations and being too conservative in one's conclusions.'* (Friedman 1985)

Users of Cochrane reviews must decide, either implicitly or explicitly, how applicable the evidence is to their particular circumstances. To do this, they must first decide whether the review provides valid information about potential benefits and harms that are important to them. To the extent that this is the case, they then need to decide whether the participants and settings in the included studies are reasonably similar to their own situation. In addition, it will often be important for them to consider the characteristics of the interventions or additional care provided in the included studies in reaching conclusions about the applicability of the evidence.

Decisions about applicability depend on knowledge of the particular circumstances in which decisions about healthcare are being made. In addressing the applicability of the results of a review, authors should be cautious not to assume that their own circumstances, or the circumstances reflected in the included studies are necessarily the same as those of others. Authors can, however, help people to make decisions about applicability by drawing attention to the spectrum of circumstances to which the evidence is likely to be applicable, circumstances where the evidence is not likely to be applicable, and predictable variation in effects across different circumstances.

Generally, rather than rigidly applying the inclusion and exclusion criteria of studies to specific circumstances, it is better to ask whether there are compelling reasons why the evidence should not be applied under certain circumstances (Guyatt 1994, Dans 1996). Authors can sometimes help people making specific decisions by identifying important variation where divergence might limit the applicability of results, including:

- biologic and cultural variation
- variation in compliance
- variation in baseline risk

In addressing these issues, authors cannot be expected to be aware of, or address the myriad differences in circumstances around the world. They can, however, address differences of known importance to many people and, importantly, they should avoid assuming that other people's circumstances are the same as their own in discussing the results and drawing conclusions.

9.2.1 Biologic and cultural variation

Issues of biologic variation that might be considered include divergence in pathophysiology (e.g. biologic differences between women and men that are likely to affect responsiveness to a treatment) and divergence in a causative agent (e.g. for infectious diseases such as malaria). For some healthcare problems, such as psychiatric problems, cultural differences can sometimes limit the applicability of results.

9.2.2 Variation in compliance

Variation in the compliance of the recipients and providers of care can limit the applicability of results. Predictable differences in compliance can be due to divergence in economic conditions or attitudes that make some forms of care not accessible or not feasible in some settings, such as in developing countries (Dans 1996).

9.2.3 Variation in baseline risk

The net benefit of any intervention depends on the risk of adverse outcomes without intervention, as well as on the effectiveness of the intervention. Therefore, variation in baseline risk is almost always an important consideration in determining the applicability of results. However, it is important to distinguish between two issues. First, whether the relative benefits and harms are applicable. For example, there might be reasons to doubt whether results obtained in high-risk patients are applicable to low-risk patients, or whether they are applicable to patients with co-morbid conditions. If there is not a compelling reason to assume that the relative benefits and harms are applicable, it is possible to estimate the expected effect of an intervention (e.g. the number needed to treat) by applying the estimated relative effect of an intervention to a specific baseline risk. The second issue related to baseline risk that warrants consideration is the extent of variation that can be expected in the impact of the intervention. For example, it can be useful to consider the number needed to treat for the range of baseline risk observed in the control groups of the studies included in the review.

9.2.4 Variation in the results of the included studies

In addition to identifying limitations of the applicability of the results of their review, authors should discuss and draw conclusions about important variation in results within the circumstances to which the results are applicable. Is there predictable variation in the relative effects of the intervention, and are there identifiable factors that may cause the response or effects to vary? These might include:

- patient features, such as age, sex, biochemical markers
- intervention features, such as the timing or intensity of the intervention
- disease features, such as hormone receptor status

These features should be examined even if there is not statistically significant heterogeneity. This should be done by testing whether there is an interaction with treatment, and not by subgroup analysis. As discussed in section 8.7, differences between subgroups, particularly those that correspond to differences between studies, need to be interpreted cautiously. Some chance variation between subgroups is inevitable, so unless there is strong evidence of an interaction then it should be assumed there is none.

9.3 Other relevant information

It can be helpful for authors to discuss the results of a review in the context of other relevant information, such as epidemiological data about the magnitude and distribution of a problem, information about current clinical practice from administrative databases or practice surveys, and information about costs. However, this is often beyond the scope of Cochrane reviews and can be done better on a national or regional basis; for example, by people developing clinical practice guidelines or undertaking a technology assessment. It must be kept in mind that evidence about the effects of healthcare is essential for well informed decisions, but it is not sufficient. Cochrane reviews cannot and should not be expected to provide all of the information that is needed for people making decisions. On the other hand, authors can help people by clarifying other information, that might vary widely, which is likely to be important in making a decision.

9.4 Adverse effects

The discussion and conclusions of a review should note the strength of the evidence on adverse effects including the estimates of their seriousness and frequency in different circumstances. In particular, the causal relationship of an adverse effect to a particular intervention should be critically assessed, bearing in mind that under-ascertainment and under-reporting of adverse and unexpected effects are common. Authors may wish to comment on how adverse effects should be further investigated in their Implications for Research section.

9.5 Trade-offs

In addition to considering the strength of evidence underlying any conclusions that are drawn, authors should be as explicit as possible about any judgements about preferences (the values attached to different outcomes) that they make. Healthcare 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 trade-offs, either implicitly or explicitly, between the estimated benefits and the estimated costs and harms (Eddy 1990b). It is beyond the scope of most Cochrane reviews to incorporate formal economic analyses (although they might well be used for such analyses) (Mugford 1989, Mugford 1991) and this is discussed in Appendix 9. However, authors 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.

9.6 Implications

The above cautions about drawing conclusions not withstanding, CRGs (and users of Cochrane reviews) may find it useful to categorise interventions into one of six mutually exclusive categories. This has been done by the Pregnancy and Childbirth Group (Enkin 1994), 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

B) Forms of care that should be abandoned in light of the available evidence

C) Forms of care that involve important trade-offs between known benefits and known adverse effects

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

B) Forms of care that have not been shown to have the effects expected from them, but which may require further attention

C) Forms of care with reasonable evidence that they are not effective for the purpose for which they have been used

9.7 Common errors in reaching conclusions

A common mistake when there is inconclusive evidence is to confuse 'no evidence of an effect' with 'evidence of no effect'. When there is inconclusive evidence, it is wrong to claim that it shows that an intervention has ‘no effect’ or is ‘no different’ from the control intervention. It is safer to report the data, with a confidence interval, as being compatible with either a reduction or an increase in the outcome. When there is a ‘positive’ but statistically non-significant trend authors commonly describe this as ‘promising’, whereas a ‘negative’ effect of the same magnitude is not commonly described as a ‘warning sign’. Authors should be careful not to do this. Another mistake is to frame the conclusion in wishful terms. For example, authors might write ‘the included studies were too small to detect a reduction in mortality’ when the included studies showed a statistically non-significant increase in mortality. One way of avoiding errors such as these is to consider the results blinded; i.e. consider how the results would be presented and framed in the conclusions if you reversed the direction of the results. If the confidence interval for the estimate of the difference in the effects of the interventions overlaps the null value, the analysis is compatible with both a true beneficial effect and a true harmful effect. If one of the possibilities is mentioned in the conclusion, the other possibility should be mentioned as well.

Another common mistake is to reach conclusions that go beyond the evidence that is reviewed. Often this is done implicitly, without referring to the additional information or judgements that are used in reaching conclusions about the implications of a review for practice. Even when conclusions about the implications of a review for practice are supported by additional information and explicit judgements, the additional information that is considered is rarely systematically reviewed and implications for practice are often dependent on specific circumstances and values that must be taken into consideration (see section 9.5). Authors should always be cautious about reaching conclusions about implications for practice and they should avoid making recommendations.

In reaching conclusions about implications for research, platitudes like "more research is needed" should also be avoided. Authors 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.

9.8 References

CTFPHE 1979. Canadian Task Force on the Periodic Health Examination. The periodic health examination. Can Med Assoc J 1979; 121:1193-254.

Cook 1992. 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.

Dans 1996. Dans AL, Dans LF. Users' guides for the applicability of the results of clinical trials, perspective of clinicians from developing countries. Manila: UP College of Medicine, 1996 (unpublished manuscript).

Eddy 1990b. Eddy DM: Anatomy of a decision. JAMA 1990; 263:4413.

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

Friedman 1985. Friedman LM, Furberg CD, DeMets DL. Fundamentals of Clinical Trials. 2nd edition. Littleton, MA: John Wright PSG Inc, 1985.

Guyatt 1994. Guyatt GH, Sackett DL, Cook DJ, for the Evidence-Based Medicine Working Group. Users' guides to the medical literature, II: how to use an article about therapy or prevention, B: what were the results and will they help me in caring for my patients. JAMA 1994; 271:59-63.

Guyatt 1995. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ, for the Evidence-Based Medicine Working Group. Users’ guide to the medical literature IX: A method for grading health care recommendations. JAMA 1995; 274:1800-4.

Gyorkos 1994. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, Oxman AD, Scott EA, Millson ME, et al. An approach to the development of practice guidelines for community health interventions. Can J Public Health 1994; 85(Suppl 1):S8-S13.

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

Mugford 1989. 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.

Mugford 1991. 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.

US PSTF 1996. US Preventive Services Task Force. Guide to Clinical Preventive Services. 2nd edition. Baltimore: Williams & Wilkins, 1996; xxxix-lv.

10 Improving and updating reviews

If Cochrane reviews are to be useful to those who want to take more informed decisions in healthcare and research, then they must be up-to-date and trustworthy, and transparently so. As made clear throughout the Handbook, the Collaboration uses 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* (*CDSR*) will vary in the level of methodological quality that it has been possible for the review authors (reviewers) to achieve. The 'gold standard' will continue to be represented by systematic reviews, conducted by the responsible investigators, that are based on individual patient data for all patients entered into all of the trials meeting the entry criteria for the review (see section 11). Such reviews require not only substantial resources (including time), they also depend on the success of negotiations among the investigators. 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 include:

- Attracting dedicated participants and avoiding conflicts of interest
- Consumer involvement
- Ensuring access to studies
- Improving access to unpublished data
- Establishing and developing standards and guidelines
- Using rigorous review methods
- Software and informatics support
- Training
- Ongoing and open peer review
- Keeping reviews up-to-date

10.1 Ensuring access to studies

Because of the disarray of the medical literature, considerable efforts are required to locate the research that addresses the questions posed by a review (see section 5). The Collaboration is helping to ensure that relevant, valid studies are located by review authors and included in their reviews by:

- Hand-searching the world's healthcare literature to identify trials
- Facilitating and supporting the development and maintenance of specialised registers by CRGs
- Providing training and support to those undertaking searching activities
- Developing the Cochrane Central Register of Controlled Trials (CENTRAL) to facilitate the transfer of trials between CRGs and other Cochrane entities, and to facilitate access to studies from other sources contributed to this register

- Working with the US National Library of Medicine to improve the coding of trials in MEDLINE and to develop an ancillary database of reports of trials not included in MEDLINE
- Developing and evaluating strategies to improve the coding and classification of trials

This work involves a large number of people engaged in a variety of activities through CRGs, Cochrane Centres, Fields and Methods Groups. The CENTRAL/CCTR Advisory Group, the New England Cochrane Center, Providence Office and the Information Retrieval Methods Group have key responsibilities for co-ordinating and guiding these activities.

10.2 Improving access to unpublished data

Improved access to unpublished data is needed to overcome problems with missing information in published reports and to protect against publication bias. In addition to the efforts undertaken by each CRG to help ensure access to relevant unpublished data within their scope, the Collaboration as a whole is working to develop strategic alliances with the pharmaceutical industry and others, and is actively promoting ethical standards that clarify the unacceptability of withholding unpublished data.

10.3 Using rigorous review methods

It is neither feasible nor desirable to dictate the decisions that a review author 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 review author. However, in general, the validity of Cochrane reviews is ensured by:

- Searching as thoroughly as possible for studies meeting the inclusion criteria of a review, relying as much as possible on the Collaboration's efforts to ensure the thoroughness and efficiency with which randomised trials are identified
- Use of explicit criteria for selecting studies for inclusion in a review and for assessing the quality of these studies
- Application of these critieria by more than one review author 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 availablity of resources and data
- Collection of individual patient data from investigators 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
- Carefully drawn conclusions, including implications for practice and research, based on cautious interpretation of results - taking into account the limitations of the review and variability in the values and conditions of people whose decisions might be influenced by the review
- 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 how these might affect the validity of the results of a review.

10.4 Peer review and the Criticism Management System

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. 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 in a paper journal or book, 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 has created a Criticism Management System through which successive versions of each review can be updated to reflect not only the emergence of new data, but also valid criticisms. Successive versions of a particular review, together with any intervening criticisms, will be archived electronically.

10.4.1 Refereeing

Each CRG is required to publish a statement describing its pre-publication peer-reviewing policy in the ‘Editorial process’ section of their module in *The Cochrane Library.*

The main issues to consider when the title for a review is being considered for registration are whether there is any overlap or potential duplication of effort with another review author either within or outside the CRG; objectives are clearly phrased and include all of the components of a well-formulated question; and the review is likely to be feasible. This refereeing stage can often be accomplished quickly by a CRG's editorial team.

Refereeing protocols can be more time-consuming than the refereeing of the full review. This is done to ensure that background information is rational and clearly presented, and that appropriate methods are planned for identifying, collecting and synthesising data. Peer review at this stage is particularly important to prevent methodological errors that may not be easily remediable at later stages of the review. The refereeing of the full review will include a second critique of the review's methods as well as a critique of the actual results, presentation of results, discussion and conclusion.

Prior to publication, all reviews must be refereed by at least two people external to the editors of the CRG. The CRG editors should appoint a referee (or contact) editor(s) for each review. If they inform the San Francisco Cochrane Center (sfcc@sirius.com) of their choice, the Center can train and support this person. It is recommended that these referees have 1) methodological expertise, 2) content area expertise, and/or 3) are a potential consumer of the review. The two referees should be selected on the basis of having differing viewpoints. Referees should include people without direct financial or personal conflicts of interest concerning the topic addressed. The referees should be asked to submit courteous and constructive comments on the Review that identify its weaknesses or fatal flaws, as well as ways of improving it. They should also be requested to return these comments to the Referee Editor within, at most, a month.

Explicit standardised methods and checklists aimed at ensuring comprehensiveness and limiting bias should be encouraged among peer authors. Specific areas to address at each stage of peer review vary. Differences among referees' critiques should be elucidated and reconciled whenever possible. This could be done by arbitration by the CRG editors or the use of an additional independent referee. The referee editor should monitor the timeliness of

returned comments, grade the quality of the comments, and, if necessary, appoint backup referees. They forward the comments from the referees, together with their own comments (if any), to the authors of the review or to the CRG Co-ordinator for distribution to the authors of the review and, if appropriate, the other editors. The referee editor, in concert with the editorial team, approves the final version of the review before it is published in The Cochrane Library.

The referee editor should keep records of all materials received and sent out during the refereeing process. An electronic refereeing system for keeping electronic records of these exchanges is being developed. Copies of the records will be requested and studied periodically by the San Francisco Cochrane Center in order to improve the refereeing process.

10.4.2 Checklist for peer authors

Preparing a review involves judgements at each step in the review process. Both systematic and random errors can occur. Several checklists are available for peer authors to use as guides for detecting important errors in the review process. Some points to keep in mind are shown below. These have been extracted from multiple citations (Jackson 1980, Cooper 1982, Light 1984a, L'Abbe 1987, Mulrow 1987, Sacks 1987, Oxman 1988, Oxman 1994a, Oxman 1994b, Cook 1995)

Problem Formulation

- Are review questions well formulated with specified key components?
- Are any changes to the protocol well documented and justified?

Study Identification

- Is there a thorough search for relevant data using appropriate sources?
- Are the search strategies appropriate to the question posed?

Study Selection

- Are appropriate inclusion and exclusion criteria used to select studies?
- Are selection criteria applied in a manner that limits bias?

Assessment of Studies

- Is the validity of individual studies addressed in a reliable manner?
- Are important parameters (e.g., setting, study population, study design) that could affect study results systematically addressed?

Data Collection

- Is there a minimal amount of missing information regarding outcomes and other variables considered key to interpretation of results?

Data Synthesis

• Are reasonable decisions made concerning whether and how to combine data?

• Are important factors, such as study designs, considered in the synthesis?

• Are results sensitive to changes in the way the analysis was done?

• Is the precision of results reported?

Discussion

- Are limitations of studies and the review process stated?
- Are review findings integrated within the context of relevant indirect evidence?

Author's Conclusions

- Are conclusions supported by the content of the review?

- Are plausible competing explanations of observed effects addressed?
- Is any interpretation of inconclusive evidence (i.e. no evidence of effect) and/or of evidence that a particular strategy did not work (i.e. evidence of no effect) appropriate?
- Are important considerations for decision-makers identified, including values and contextual factors that might influence decisions?

10.5 Updating reviews

When registering a review with the Cochrane Collaboration, authors agree to keep it up-to-date. This entails repeating, at periodic intervals, the steps involved in the original review. Some of the steps will require minimal effort (e.g. reviewing the research question to make sure it is still relevant) while others may require a substantial investment of time and effort.

The most logistically demanding aspect of keeping a review up-to-date is the identification of new studies. For CRGs that are sufficiently organised and funded, the periodic identification of relevant new studies is an ongoing function of the editorial team (usually the CRG's Coordinator or Trial Search Coordinator). In other instances, authors and editors must work out collaborative mechanisms to periodically identify new studies. At a minimum, strategies to identify new studies should include periodically checking the CRG's specialised register, CENTRAL and MEDLINE. The Cochrane Collaboration has a Criticism Management System which continues to develop and allows users of Cochrane reviews to provide comments and criticisms of reviews, and this is discussed further in the next section. It is likely to provide an additional source of studies to be considered for the review.

Original data collection forms should be used to abstract new research evidence. If new research evidence addresses important variables that were not included in the original collection form, these may be modified. For example, if authors had originally only abstracted morbidity and mortality outcomes in trials addressing treatment of advanced cancer, and recent studies routinely report quality of life outcomes, the collection form could be amended. In such instances, authors may need to recheck whether any of their earlier identified studies had such information that was overlooked.

Occasionally, authors may decide to include a new analysis strategy in their updated review; for example, using statistical methods not previously available in RevMan. In general, new analysis strategies will represent substantive changes that merit editorial critique through the CRG's established editorial process.

How often reviews need updating will vary depending on the production of valid new research evidence. Authors should work with their editorial team to establish guides addressing when new research evidence is substantive enough to warrant a major update or amendment. The dates of such amendments must be recorded in the What's New section of the review. It is Collaboration policy that reviews should either be updated within two years or should have a commentary added to explain why this is done less frequently. It is also Collaboration policy that protocols that have not been converted into full reviews within two years should generally be withdrawn from the CDSR. Even if no substantive new evidence is found on annual review and no major amendment is indicated, this information should still be used to update the review by adding the date of the latest search for evidence to the review.

If a review needs to be suspended or withdrawn, this should be noted in the Published Notes section of the review. The review containing this suspension/withdrawal notice should be submitted for publication in each issue of the CDSR, until the content of review is judged to be satisfactory by the authors and their CRG. If a review is merged with another review, a notice should be included in its Published Notes section to explain that it has been withdrawn for this reason.

10.6 Responding to criticisms

The electronic format of the *CDSR* offers a unique opportunity to respond to, and incorporate criticism from the users of Cochrane reviews. This will greatly increase the quality of the reviews, and also allow users to be brought into the reviewing process.

The reader should use the 'Comments/Criticisms' button to make constructive and courteous comments. These are automatically sent to the Criticism Editor of the relevant CRG. In an effort to prevent redundancy, the user should read criticisms that have already been received before sending in their own criticism. They can do this by visiting the 'Current Comments and Criticisms' Internet page: http://www.update-software.com/comcrit.htm.

When they receive the feedback, the Criticism Editor should summarize it and send a copy to the authors and to the San Francisco Cochrane Center so that it can be posted on the 'Current Comments and Criticisms' web page. The authors are responsible for responding to all criticisms in a timely fashion. They should provide a written response by using the criticism section of RevMan and update their review if appropriate.

10.7 References

Bastian 1998. Bastian H. Speaking up for ourselves: the evolution of consumer advocacy in health care. International Journal of Technology Assessment in Health Care 1998;14:3-23.

Cook 1995. Cook DJ, Sackett DL, Spitzer WO. Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. J Clin Epidemiol 1995; 48:167-71.

Cooper 1982. Cooper HM. Scientific guidelines for conducting integrative research reviews. Rev Educ Res 1982; 52:291-302.

Jackson 1980. Jackson GB. Methods for integrative reviews. Rev Educ Res 1980; 50:438-60.

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

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

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

Oxman 1988. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. Can Med Assoc J 1988; 138:697-703.

Oxman 1994a. Oxman AD, Cook DJ, Guyatt GH. Users' guide to the medical literature, VI. How to use an overview. Evidence-Based Medicine Working Group. JAMA 1994; 272:1367-71.

Oxman 1994b. Oxman AD. Checklists for review articles. BMJ 1994; 309:648-51.

Sacks 1987. Sacks HS, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta-analyses of randomized controlled trials. N Engl J Med 1987; 316:450-5.

Smith 1994. Smith R. Conflict of interest and the BMJ. BMJ 1994; 308:4-5.

11 Reviews using IPD

11.1 Rationale

If a systematic review is to contain a meta-analysis in which the results of separate studies will be brought together in a statistical synthesis, then the data for this could be collected in a variety of ways. These include extraction from published reports, collection of aggregate data from the responsible investigators or collection of individual patient data (IPD) from the investigators. The latter has been used in large-scale collaborative overviews in which data from all randomised trials in a particular disease area are brought together (EBCTCG 1992) and also in more restricted reviews in which data from a relatively small number of trials assessing a specific healthcare intervention are collected and combined (Jeng 1995). Systematic reviews based on IPD have been described as the yardstick against which all reviews should be measured (Chalmers 1993). Although they can require more time, resources and expertise than other forms of review, the process brings with it a number of advantages. Authors should consider the importance of these advantages to their particular systematic review when deciding whether to embark on such a project. Examples of IPD reviews are available in the *Cochrane Database of Systematic Reviews*, including some that were originally published in paper journals.

11.2 Methods Group on Individual Patient Data Reviews

To try to help with this decision and with the logistics of such projects, a Cochrane Collaboration Methods Group has been established to provide guidance to those wishing to conduct an IPD meta-analysis. This group (co-convened by Lesley Stewart and Mike Clarke) was formed following a UK Cochrane Centre sponsored workshop in April 1994 at which representatives of research groups involved in such projects were brought together for the first time. This allowed for discussion of areas such as protocol use and development, methods of data-checking, and resource requirements. And a detailed report from the workshop was published in Statistics in Medicine in October 1995 (Stewart 1995). This report is included in full in this Handbook (Appendix 11a) with permission of the publisher.

11.3 What an IPD meta-analysis is and is not

As with any systematic review the fundamental principle for one which uses IPD is that as much as possible of the relevant, valid evidence is included. This means that the process of trial identification must be as thorough as possible and that the attempts to collect data must be equally thorough. The ultimate aim should be that all randomised participants, and no non-randomised participants, from all relevant studies are included and that they are analysed using the intention-to-treat principle. In this way, systematic biases and chance effects will be minimised. To this end, the data collection should be kept simple and straightforward, with the minimum amount of data being collected for the required analyses. It should be as easy as possible for the investigators to supply their data since this should increase the likelihood that data will be received for all relevant studies. In addition, investigators should know that any data supplied for the review will be held in confidence and will not be used for any other purpose without their permission, and that the reports of the review will be published in the names of the collaborating investigators rather than the central co-ordinators.

The predominant difference between an IPD meta-analysis and meta-analysis based on aggregate data (whether extracted from published reports or supplied direct by investigators) is that the combined study results come from a central re-analysis of the raw data from each study. The necessary data items are sought and, after central processing, any inconsistencies

or problems are discussed and hopefully resolved by communication with the responsible investigators. The finalised data for each study are then analysed separately to obtain summary statistics, which are combined to give an overall estimate of the effect of treatment. In this way, participants are only directly compared with others in the same study and the entire dataset is not pooled as though it came from a single, homogeneous study.

11.4 How can an IPD meta-analysis help?

If a systematic review relies solely on data from published studies, it is open to a number of problems. The most obvious of these is that unpublished studies will not be included, but the published data may be inadequate for other reasons also. For example, there may be insufficient information on the types of patient or outcome of interest in the review, the data are 'frozen-in-time' when important findings may come from longer follow-up or more detailed study, and the intention-to-treat principle may not have been followed (and, occasionally, this might not be clear from the published report). Collection of either aggregate or individual patient data from investigators will resolve some of these problems: unpublished trials can be included, updated data on specific types of participant and outcome can be requested, and whether the data are based on the randomised allocations can be clarified (if the studies are randomised trials).

Collecting IPD rather than aggregate data brings additional advantages. These include the ability to undertake survival and other time-to-event analyses; to undertake analyses using commonly defined subgroups to test and generate hypotheses; to ensure the quality of the randomisation and follow-up data used in the meta-analysis through detailed data checking and iterative correction of errors by communication with the investigators; and to update follow-up information through patient record systems (such as mortality registers) where available. In addition, it might be easier for an investigator to send individual patient, rather than aggregate, data particularly if they do not have sufficient data-management or statistical support to prepare the necessary tables. It will also be easier for a small amount of extra information to be supplied. For example, if further follow-up becomes available on some participants, the investigator can simply send these details instead of preparing new tables.

Furthermore, as IPD meta-analyses involve the collaboration of the investigators, they can have other benefits, some of which may also be found if the investigators are contacted for aggregate data. These include more complete identification and understanding of the studies; better compliance with providing missing data; more balanced interpretation of the results of the review; wider endorsement and dissemination of these results; a broader consensus on the implications for future practice and research; and possible collaboration in such research.

11.5 Where is the evidence?

One of the aims of the Methods Group on IPD based meta-analysis was to establish, and encourage the tackling of, a research agenda to investigate this approach to systematic reviews. Limited empirical evidence already exists for some of the advantages of IPD reviews over other types of review. Typically, these have involved comparison of the results from an IPD meta-analysis with those from a meta-analysis based on published material. They have shown the importance of the former in helping control publication bias, in ensuring the use of the intention-to-treat principle in the analysis, and in obtaining a fuller picture of the effects of different treatments over time (Stewart 1993, Pignon 1993, Jeng 1995, Clarke 1997).

11.6 Converting reviews that used individual patient data into Cochrane reviews

The conversion into Cochrane reviews of relevant, pre-existing reviews that have used individual patient data should be encouraged, unless a Cochrane review of higher quality can be prepared in some other way. However, these conversions can present particular challenges to authors (reviewers) and Collaborative Review Groups (CRGs).

IPD meta-analyses have generally been carried out by large**,** collaborative groups of trialists. Sometimes more than 100 people will be involved, including the trialists who provided their source data for re-analysis, an organisational secretariat and, in some cases, an advisory committee. However, the size of these groups, the social politics involved and prior paper publication can make it difficult to comply with certain Cochrane procedural, style and format recommendations. In particular:

- For pre-existing IPD reviews, a protocol can usually not be provided retrospectively and CRGs should not require one before accepting the review. However, IPD authors should try to submit protocols for ongoing projects at an early stage.
- The text of the IPD review will usually have been through many drafts and circulated to all members of the collaborative group for comment. Agreement on wording within such large groups is not always easy to achieve and so it may be difficult to change the text of a review for inclusion in the Cochrane Database of Systematic Reviews. The editors and peer authors of CRGs should be sympathetic to this constraint.
- The Study Identifiers or labels will usually have been chosen in collaboration with the trialists and it is unlikely to be possible to change these to reflect particular conventions.
- For a pre-published IPD review, the secretariat will already have obtained sufficient declarations of contribution and consent to authorship from each member of the collaborative group to satisfy publication of the review in a journal. It would be resource-intensive to further obtain Cochrane authorship contribution forms for each "author" and the authorship declarations submitted to the journal should be accepted by the CRG as an alternative.

IPD meta-analyses should be peer reviewed by the CRG's normal peer review process. However, the difficulties of making changes (discussed above) should be made clear to the peer authors. They should also bear in mind that pre-existing IPD reviews will probably have been through an extensive peer review process prior to submission to the CRG. Manuscripts will have been scrutinised by the trialists' group, secretariat and advisory committee for the review, as well as by the peer review process of the journal in which the IPD review was published. As with all reviews, the final decision on whether an IPD review is acceptable for publication as a Cochrane review in the Cochrane Database of Systematic Reviews must rest with the editorial group of the CRG.

CRGs who would like advice relating to IPD reviews, for example in regard to their peer review, should contact the IPD meta-analysis Methods Group for help.

11.7 Prospective meta-analysis

Prospective meta-analysis are a special form of IPD meta-analysis. In these projects, a group of investigators agree, in advance of knowing the results of their studies, to pool their data in the future. A Cochrane Collaboration Methods Group has been established to address this issue and will provide training and support in the conduct of these projects (Appendix 11b).

11.8 Further information

Many of the topics discussed here are expanded on in Stewart 1995 (Appendix 11a). That report also contains examples of how IPD meta-analyses have been conducted previously, which may be useful to authors planning one now. If Cochrane authors would like further information they should contact the Methods Group. In addition, a slide show that is used by the Methods Group in training workshops is available from the Collaboration's Internet site (http://www.cochrane.org/cochrane/training.htm).

11.9 References

Chalmers 1993. Chalmers I. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. Ann N Y Acad Sci 1993; 703: 156-65.

Clarke 1997. Clarke M, Stewart L. Individual patient data or published data meta analysis: a systematic review [abstract]. *Proceedings of the Fifth Cochrane Collaboration Colloquium* 1997; 94, abstract 019.04.).

EBCTCG 1992. Early Breast Cancer Trialists' Collaborative Group. Systematic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. Lancet 1992; 339:1-15, 71-85.

Jeng 1995. Jeng GT, Scott JR, Burmeister LF. A comparison of meta-analytic results using literature vs individual patient data: paternal cell immunization for recurrent miscarriage. JAMA 1995; 274: 830-6.

Pignon 1993. Pignon JP, Arriagada R. Meta-analysis. Lancet 1993; 341:964-5.

Stewart 1993. Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is ther a difference? Lancet 1993; 341:418-22.

Stewart 1995. Stewart L, Clarke M, for the Cochrane Collaboration Working Group on meta-analyses using indivudal patient data. Practical methodology of meta-analyses (overviews) using updated individual patient data. Stat Med 1995; 14:2057-79.

Appendices

APPENDIX 5a. Cochrane and National Library of Medicine randomized controlled trial and controlled clinical trial criteria

5a.1 Cochrane criteria for randomized controlled trials (RCTs) and controlled clinical trials (CCTs)

Records identified for inclusion should meet the eligibility criteria devised and agreed in November 1992, which were first published, in 1994, in Section 5 of the Cochrane Reviewer's Handbook. According to these eligibility criteria:

A trial is eligible if, on the basis of the best available information (usually from one or more published reports), it is judged that:

- the individuals (or other units) followed in the trial were definitely or possibly assigned prospectively to one of two (or more) alternative forms of health care using
    - random allocation or
    - some quasi-random method of allocation (such as alternation, date of birth, or case record number)

A trial is eligible if, on the basis of the best available information (usually from one or more published reports), it is judged that:

- the individuals (or other units) followed in the trial were definitely or possibly assigned prospectively to one of two (or more) alternative forms of health care using
- random allocation or
- some quasi-random method of allocation (such as alternation, date of birth, or case record number)

Trials eligible for inclusion are classified according to the reader's degree of certainty that random allocation was used to form the comparison groups in the trial. If the author(s) state explicitly (usually by some variant of the term 'random' to describe the allocation procedure used) that the groups compared in the trial were established by random allocation, then the trial is classified as an 'RCT' (randomized controlled trial). If the author(s) do not state explicitly that the trial was randomized, but randomization cannot be ruled out, the report is classified as a 'CCT' (controlled clinical trial). The classification 'CCT' is also applied to quasi-randomized studies, where the method of allocation is known but is not considered strictly random, and possibly quasi-randomized trials. Examples of quasi-random methods of assignment include alternation, date of birth, and medical record number.

The classification as RCT or CCT is based solely on what the author has written, not on the reader's interpretation; thus, it is not meant to reflect an assessment of the true nature or quality of the allocation procedure. For example, although double-blind trials are nearly always randomized, many trial reports fail to mention random allocation explicitly and should therefore be classified as 'CCT'.

Relevant reports are reports published in any year, of studies comparing at least two forms of health care (healthcare treatment, healthcare education, diagnostic tests or techniques, a preventive intervention, etc.) where the study is on either living humans or parts of their body or human parts that will be replaced in living humans (e.g., donor kidneys). Studies on cadavers, extracted teeth, cell lines, etc. are not relevant. *Searchers should identify all controlled trials meeting these criteria regardless of relevance to the entity with which they are affiliated.*

The highest possible proportion of all reports of controlled trials of health care should be included in CENTRAL. Thus, those searching the literature to identify trials should give reports the benefit of any doubts. Reviewers will decide whether to include a particular report in a review.

5a.2 National Library of Medicine definitions for Publication Type terms: RANDOMIZED CONTROLLED TRIAL, CONTROLLED CLINICAL TRIAL

RANDOMIZED CONTROLLED TRIAL:

A clinical trial that involves at least one test treatment and one control treatment, concurrent enrollment and follow-up of the test- and control-treated groups, and in which the treatments to be administered are selected by a random process, such as the use of a random numbers table. Treatment allocations using coin flips, odd-even numbers, patient social security numbers, days of the week, medical record numbers, or other such pseudo- or quasi-random processes, are not truly randomized and a trial employing any of these techniques for patient assignment is designated simply a CONTROLLED CLINICAL TRIAL.

CONTROLLED CLINICAL TRIAL:

A clinical trial involving one or more test treatments, at least one control treatment, specified outcome measures for evaluating the studied intervention, and [an intended to be bias-free] method of assigning patients to the test treatment. The treatment may be drugs, devices, or procedures studied for diagnostic, therapeutic, or prophylactic effectiveness. Control measures include placebos, active medicine, no-treatment, dosage forms and regimens, historical comparisons, etc. When randomization using mathematical techniques, such as the use of a random numbers table, is employed to assign patients to test or control treatments, the trial is characterized as a RANDOMIZED CONTROLLED TRIAL. However, trials employing treatment allocation methods such as coin flips, odd-even numbers, patient social security numbers, days of the week, medical record numbers, or other such pseudo- or quasi-random processes are simply designated as controlled clinical trials.

APPENDIX 5b: Highly sensitive search strategies for identifying reports of randomized controlled trials in MEDLINE:

In the SilverPlatter and Ovid search strategies below, upper case denotes controlled vocabulary and lower case denotes free-text terms. Cochrane Reviewers wishing to run these search strategies are recommended to seek the advice of their Review Group's Trials Search Co-ordinator. Others should seek the advice of a trained medical librarian.

5b.1 Format for MEDLINE on SilverPlatter WinSPIRS 4.0 (checked and updated February 2004):

phase 1:

#1 RANDOMIZED-CONTROLLED-TRIAL in PT
#2 CONTROLLED-CLINICAL-TRIAL in PT
#3 RANDOMIZED-CONTROLLED-TRIALS
#4 RANDOM-ALLOCATION
#5 DOUBLE-BLIND-METHOD
#6 SINGLE-BLIND-METHOD
#7 #1 or #2 or #3 or #4 or #5 or #6
#8 TG=ANIMALS not (TG=HUMANS and TG=ANIMALS)
#9 #7 not #8

phase 2:

#10 CLINICAL-TRIAL in PT
#11 explode CLINICAL-TRIALS
#12 (clin* near trial*) in TI
#13 (clin* near trial*) in AB
#14 (singl* or doubl* or trebl* or tripl*) near (blind* or mask*)
#15 (#14 in TI) or (#14 in AB)
#16 PLACEBOS
#17 placebo* in TI
#18 placebo* in AB
#19 random* in TI
#20 random* in AB
#21 RESEARCH-DESIGN

#22 #10 or #11 or #12 or #13 or #15 or #16 or #17 or #18 or #19 or #20 or #21
#23 TG=ANIMALS not (TG=HUMANS and TG=ANIMALS)
#24 #22 not #23
#25 #24 not #9

phase 3:
#26 TG=COMPARATIVE-STUDY
#27 explode EVALUATION-STUDIES
#28 FOLLOW-UP-STUDIES
#29 PROSPECTIVE-STUDIES
#30 control* or prospectiv* or volunteer*
#31 (#30 in TI) or (#30 in AB)
#32 #26 or #27 or #28 or #29 or #31
#33 TG=ANIMALS not (TG=HUMANS and TG=ANIMALS)
#34 #32 not #33
#35 #34 not (#9 or #25)

#36 #9 or #25 or #35 (to combine all 3 phases)

Note: if you require both phases 1 and 2 but not phase 3, type as line #26: #9 or #25

5b.2 Format for MEDLINE on Ovid web version (checked and updated February 2004):

phase 1:
1 RANDOMIZED CONTROLLED TRIAL.pt.
2 CONTROLLED CLINICAL TRIAL.pt.
3 RANDOMIZED CONTROLLED TRIALS.sh.
4 RANDOM ALLOCATION.sh.
5 DOUBLE BLIND METHOD.sh.
6 SINGLE BLIND METHOD.sh.
7 or/1 6
8 ANIMALS.sh. not HUMANS.sh.
9 7 not 8

phase 2:
10 CLINICAL TRIAL.pt.
11 exp CLINICAL TRIALS/
12 (clin$ adj25 trial$).ti,ab.

13 ((singl$ or doubl$ or trebl$ or tripl$) adj25 (blind$ or mask$)).ti,ab.

14 PLACEBOS.sh.

15 placebo$.ti,ab.

16 random$.ti,ab.

17 RESEARCH DESIGN.sh.

18 or/10 17

19 18 not 8

20 19 not 9

phase 3:

21 COMPARATIVE STUDY.sh.

22 exp EVALUATION STUDIES/

23 FOLLOW UP STUDIES.sh.

24 PROSPECTIVE STUDIES.sh.

25 (control$ or prospectiv$ or volunteer$).ti,ab.

26 or/21 25

27 26 not 8

28 27 not (9 or 20)

29 9 or 20 or 28 (to combine all 3 phases)

Note: if you require both phases 1 and 2 but not phase 3, type as line 21: 9 or 20

5b.3 Format for PubMed (checked and updated February 2004):

Phase 1

(randomized controlled trial [pt] OR controlled clinical trial [pt] OR randomized controlled trials [mh] OR random allocation [mh] OR double-blind method [mh] OR single-blind method [mh]) NOT (animals [mh] NOT humans [mh])

Phases 1 and 2

(randomized controlled trial [pt] OR controlled clinical trial [pt] OR randomized controlled trials [mh] OR random allocation [mh] OR double-blind method [mh] OR single-blind method [mh] OR clinical trial [pt] OR clinical trials [mh] OR ("clinical trial" [tw]) OR ((singl* [tw] OR doubl* [tw] OR trebl* [tw] OR tripl* [tw]) AND (mask* [tw] OR blind* [tw])) OR (placebos [mh] OR placebo* [tw] OR random* [tw] OR research design [mh:noexp]) NOT (animals [mh] NOT humans [mh])

All Phases

(randomized controlled trial [pt] OR controlled clinical trial [pt] OR randomized controlled trials [mh] OR random allocation [mh] OR double-blind method [mh] OR single-blind method [mh] OR clinical trial [pt] OR clinical trials [mh] OR ("clinical trial" [tw]) OR ((singl* [tw] OR doubl* [tw] OR trebl* [tw] OR tripl* [tw]) AND (mask* [tw] OR blind* [tw])) OR (placebos [mh] OR placebo* [tw] OR random* [tw] OR research design [mh:noexp] OR comparative study [mh] OR evaluation studies [mh] OR follow-up studies [mh] OR prospective studies [mh] OR control* [tw] OR prospectiv* [tw] OR volunteer* [tw]) NOT (animals [mh] NOT humans [mh])

Note: Subject specific terms (MeSH and textwords) should be ORed together, enclosed within parentheses, then ANDed with the appropriate version of the Cochrane highly sensitive search strategy.

APPENDIX 5c. Example of a search strategy for electronic databases

(from the following Cochrane Review: Wilkinson C. Interventions for asymptomatic retinal breaks and lattice degeneration for preventing retinal detachment (Cochrane Review). In: *The Cochrane Library*, Issue 1, 2003. Oxford: Update Software.)

Search strategy for identification of studies

See: Collaborative Review Group search strategy

Trials were identified by electronic searches of the Cochrane Controlled Trials Register - CENTRAL (which includes the Cochrane Eyes and Vision Group specialized register), MEDLINE and EMBASE.

The following strategy was used to search CENTRAL Issue 1 2001 [search conducted January 5, 2001]:

#1 RETINAL-DETACHMENT:ME

#2 (RETINA* near ((((DETACH* or BREAK*) or PERFORATION*) or TEAR*) or HOLE*))

#3 (LATTICE near DEGENERAT*)

#4 RETINAL-PERFORATIONS:ME

#5 ((VITREO* near DETACH*) and POSTERIOR)

#6 ((VITREORETINAL or VITREO-RETINAL) near DEGENERAT*)

#7 (((((#1 or #2) or #3) or #4) or #5) or #6)

#8 LASER-COAGULATION*:ME

#9 LIGHT-COAGULATION:ME

#10 CRYOTHERAPY*1:ME

#11 ((LASER or LIGHT) near COAGULAT*)

#12 (LASER near PHOTOCOAGULAT*)

#13 CRYOPTHERAP*

#14 (((((#8 or #9) or #10) or #11) or #12) or #13)

#15 PROPHYLA*

#16 (#7 and (#14 or #15))

The following strategy was used to search MEDLINE to December 2000 [search conducted January 5, 2001]:

SilverPlatterASCII 3.0DOSN

#1 "RETINAL-DETACHMENT"/ all subheadings

#2 "RETINAL-PERFORATIONS"/ all subheadings

#3 "VITREOUS-DETACHMENT"/ all subheadings

#4 RETINA* near (DETACH* or BREAK* or PERFORATION* or TEAR* or HOLE*)
#5 (LATTICE near DEGENERAT*)
#6 VITREO?RETINAL next DEGENERAT*
#7(VITREO* near DETACH*) and POSTERIOR
#8 (#4 or #5 or #6 or #7) in TI,AB
#9 #1 or #2 or #3 or #8
#10 explode "LIGHT-COAGULATION"/ all subheadings
#11 explode "CRYOTHERAPY"/ all subheadings
#12(LASER or LIGHT) near COAGULAT*
#13 LASER near PHOTOCOAGULAT*
#14 CRYOTHERAP*
#15 (#12 or #13 or #14) in TI,AB
#16 #10 or #11 or #15
#17 PROPHYLA* in TI,AB
#18 #9 and (#16 or #17)

To identify randomized controlled trials, this search was combined with the Cochrane Highly Sensitive Search Strategy phases one and two as contained in the Cochrane Reviewer's Handbook (Clarke 2000).

The following strategy was used to search EMBASE to February 2001 [search conducted February 2, 2001]:

SilverPlatterASCII 3.0DOSN
#1 explode "RETINA-DETACHMENT"/ all subheadings
#2 "VITREOUS-BODY-DETACHMENT"/ all subheadings
#3 "VITREORETINAL-DEGENERATION"/ all subheadings
#4 RETINA* near (DETACH* or BREAK* or PERFORATION* or TEAR* or HOLE*)
#5 (LATTICE near DEGENERAT*)
#6 VITREO?RETINAL near DEGENERAT*
#7 (VITREO* near DETACH*) and POSTERIOR
#8 #4 or #5 or #6 or #7
#9 #1 or #2 or #3 or #8
#10 explode "LASER-COAGULATION"/ all subheadings
#11 "CRYOTHERAPY"/ all subheadings
#12 (LASER or LIGHT) near COAGULAT*
#13 LASER near PHOTOCOAGULAT*
#14 CRYOTHERAP*
#15 (#12 or #13 or #14) in TI,AB
#16 #10 or #11 or #15
#17 "PROPHYLAXIS"/ all subheadings

#18 PROPHYLA* in TI,AB

#19 #9 and (#16 or #17 or #18)

To identify randomized controlled trials, this search was combined with the following search:

SilverPlatterASCII 3.0DOSNEMBASE (R) 1998/07-1998/12

#1 "RANDOMIZED-CONTROLLED-TRIAL"/ all subheadings

#2 "RANDOMIZATION"/ all subheadings

#3 "CONTROLLED-STUDY"/ all subheadings

#4 "MULTICENTER-STUDY"/ all subheadings

#5 "PHASE-3-CLINICAL-TRIAL"/ all subheadings

#6 "PHASE-4-CLINICAL-TRIAL"/ all subheadings

#7 "DOUBLE-BLIND-PROCEDURE"/ all subheadings

#8 "SINGLE-BLIND-PROCEDURE"/ all subheadings

#9 #1 or #2 or #3 or #4 or #5 or #6 or #7 or #8

#10 (RANDOM* or CROSS?OVER* or FACTORIAL* or PLACEBO* or VOLUNTEER*) in TI,AB

#11 (SINGL* or DOUBL* or TREBL* or TRIPL*) near (BLIND* or MASK*) in TI,AB

#12 #9 or #10 or #11

#13 HUMAN in DER

#14 (ANIMAL or NONHUMAN) in DER

#15 #13 and #14

#16 #14 not #15

#17 #12 not #16

APPENDIX 6a. Reviews including non-randomised studies

6a.1. Rationale

The Cochrane Collaboration builds on ten principles, two of which are to minimise bias and to ensure relevance. In order to minimise bias, reviewers may choose to include only randomised controlled trials (RCTs) in their reviews. While this approach minimises bias it may not always ensure relevance. The challenge facing reviewers is this: How far is it possible to achieve a higher level of relevance by including evidence other than that derived from RCTs without violating the central principle: minimising bias?

6a.2. What might be the advantages and dangers of including non-randomised studies in systematic reviews?

If a systematic review relies solely on data from randomised trials, it is open to a number of problems. The most obvious of these is that certain important health care problems have not been studied, or are impossible or very difficult to study in randomised trials. But randomised trials may be inadequate for other reasons also. For example, there may be insufficient information on the types of participant or outcome which are of relevance to the review (e.g. rare side effects), or the data may only contain short term follow-up when important findings depends on longer follow-up. Inclusion of evidence from non-randomised studies may resolve some of these problems, but it also poses problems and threats to validity as unexpected biases may creep in and invalidate the conclusions.

Some examples already exist where inclusion of non-randomised evidence in systematic reviews have been helpful. For example the possible causal relationship between prone sleeping position and cot death which was strongly supported by meta-analyses of observational studies (Beal 1991) was subsequently corroborated by national intervention programmes leading to a reduced rate for cot deaths (Wennergren 1997). A recent example of the opposite might be the many systematic reviews of observational studies of hormone replacement therapy in postmenopausal women showing a dramatic and highly significant decrease in mortality but contradicted by an ensuing large randomised trial showing no significant difference with a fairly narrow confidence interval (Petitti 1998).

Several empirical studies of the possible biases in non-randomised studies have been published recently (Britton 1998; Reeves 1998; Kunz 1998; Benson 2000; Concato 2000). The foci, the quality assessments and the conclusions of these studies vary and have led to some confusion and discussion. High quality research projects with prespecificied protocols are needed.

6a.3. Guidelines for inclusion of non-randomised studies in Cochrane reviews

The Cochrane Non-Randomised Studies Methods Group (NRSMG) was registered in November 1999 and is currently developing guidelines for the inclusion of non-randomised studies in Cochrane reviews. The following guideline chapters are planned and under development:

1. Introduction

2. Types of study design

2.1. Scope and terminology of the NRSMG guidelines

2.2. What types of study designs should be included in a Cochrane review?

2.3. What types of research questions are expected to benefit from the inclusion of non-randomised evidence?

3. Searching for non-randomised studies

4. Quality assessment

5. Data extraction

6. Analysis

7. Interpretation

The draft chapters will be made available at www.cochrane.dk/nrsmg/ as they reach a useable form (during 2000 and 2001). The chapters will be approved by the NRSMG as they reach their final form. The full set of guidelines is expected to be ready by the end of 2001.

6a.4. Further information

This appendix was prepared by Ole Olsen on behalf of the Cochrane Non-Randomised Studies Methods Group. Further information can be found in the NRSMG module in The Cochrane Library or at www.cochrane.dk/nrsmg/.

6a.5. References

Beal 1991. Beal SM, Finch CF. An overview of retrospective case-control studies investigating the relationship between prone sleeping position and SIDS. J Paediatr Child Health 1991;27:334-9.

Benson 2000. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000;342:1878-86.

Britton 1998. Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. Health Technol Assessment 1998;2(13).

Concato 2000. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887-92.

Kunz 1998. Kunz R, Oxman A. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998;317:1185-90.

Petitti 1998. Petitti DB. Hormone replacement therapy and heart disease prevention: experimentation trumps observation. JAMA 1998;280:650-2.

Reeves 1998. Reeves BC, MacLehose RR, IM Harvey, TA Sheldon, IT Russell, AMS Black. Comparisons of effect size estimates derived from randomised and non-randomised studies. In: Black N, et al, editors. Health Services Research Methods: a guide to best practice. London: BMJ Publishing Group, 1998; 73-85.

Wennergren 1997. Wennergren G, Alm B, Oyen N, Helweg-Larsen K, Milerad J, Skjaerven R, Norvenius SG, Lagercrantz H, Wennborg M, Daltveit AK, Markestad T, Irgens LM. The decline in the incidence of SIDS in Scandinavia and its relation to risk-intervention campaigns. Nordic Epidemiological SIDS Study. Acta Paediatr 1997;86:963-8.

APPENDIX 6b. Including adverse effects

Edited by Yoon K Loke, Deirdre Price and Andrew Herxheimer on behalf of the Cochrane Adverse Effects Subgroup.

6b.1. Introduction

The policy of The Cochrane Collaboration has always been that reviews look at all relevant outcomes of a healthcare intervention. In practice, however, review authors have frequently avoided studying unintended effects, and have concentrated instead on the intended, beneficial outcomes. Methodological guidance on how to review adverse effects has also been lacking. This section provides guidance from the Adverse Effects Subgroup of the Non-randomised Studies Methods Group.

Every healthcare intervention comes with the risk, great or small, of harmful or adverse effects. A Cochrane review that considers only the favourable outcomes of the interventions that it examines, without also assessing the adverse effects, will lack balance and may make the intervention look more favourable. This source of bias, like others, should be minimised. All reviews should therefore include some evaluation of adverse effects.

An intervention may have many potential adverse effects. The systematic assessment of adverse effects can make substantial demands on time and resources. This needs to be considered in the early stages of the protocol design. Many adverse effects are too uncommon to be observed in randomized controlled trials, which are most appropriate for the assessment of common, known effects. Full evaluation of adverse effects, therefore, often requires other types of evidence.

The extent and nature of the adverse effects analysis should be formulated based on the principles laid out in Section 2.3.1 of the main Handbook. It is worth highlighting two aspects that are of special relevance here:

1. The selected adverse outcomes should be those that are important in guiding the decisions of healthcare providers, researchers, policymakers and consumers;
2. There is often a major trade-off between comprehensiveness and the quality of the adverse effects data included in a review. It may not help to include evidence that is likely to be biased, even if no better evidence exists. Nevertheless, it is recognized that important information on rare, serious harms may only be available from sources that are susceptible to bias. In these instances, the limitations of the data should be rigorously appraised and critically discussed.

If the review will not evaluate adverse effects, this should be stated explicitly and a reason given.

6b.1.1 Definitions

Many terms are used to describe harmful effects of healthcare interventions; several of these are defined in Table 1. Published papers often use the terms ‘adverse effect’, ‘adverse drug reaction’, ‘side effect’, ‘toxic effect’, ‘adverse event’ and ‘complications’ loosely and interchangeably.

Table 1. Definitions of terms related to adverse outcomes

Adverse event	An unfavourable outcome that occurs during or after the use of a drug or other intervention but is not necessarily caused by it. It can be defined as “any abnormal sign, symptom, or laboratory test, or any syndromic combination of such abnormalities, any untoward or unplanned occurrence (for example, an accident or unplanned pregnancy), or any unexpected worsening or improvement in a concurrent illness” (Aronson 2005).

Adverse effect	An adverse event for which the causal relation between the drug/intervention and the event is at least a reasonable possibility. This term applies to all interventions.
Adverse drug reaction (ADR)	This term is used only with drugs. The terms ADR and adverse effect are used interchangeably with respect to drugs (Edwards 2000).
Complications	This term is widely used to describe adverse events following surgical and other invasive interventions. It can be considered to be synonymous with 'adverse event' or 'adverse effect'.
Seriousness and intensity (or severity) of the adverse effect	Often confused with seriousness, severity is better termed-'intensity'. WHO terminology differentiates between the terms 'serious' and 'severe' in this way: 'serious' refers to adverse effects that have significant medical consequences, e.g. lead to death, permanent disability or prolonged hospitalisation. A review should state whether 'serious' is defined in this way, or whether it also includes other effects that the patient considers serious. In contrast, 'severe' refers to the intensity of a particular adverse effect. For example, a non-serious adverse effect, such as headache, may be severe in intensity (as opposed to mild or moderate).
Side effect	This is any unintended effect of a pharmaceutical product that occurs at doses normally used for therapeutic purposes in humans and is related to the pharmacological properties of the drug. While some side effects may be harmful (and can thus be considered adverse effects), there are also side effects that are beneficial.
Safety	This word usually refers to (the relative lack of) serious adverse reactions, such as those that threaten life, require or prolong hospitalization, result in permanent disability, or cause birth defects. But, serious, indirect adverse effects, such as traffic accidents, violence, and damaging consequences of mood change, can also be categorized by this term. They may or may not be detected in trials (depending on participant numbers, intensity of monitoring, and length of follow up), and data on such adverse effects may be available only from non-randomised studies.
Tolerability	The term is usually used in referring to medically less important, that is, without serious or permanent sequelae, but unpleasant adverse effects of drugs. These include symptoms such as dry mouth, tiredness, etc, that can affect a person's quality of life and willingness to continue the treatment. As these adverse effects usually develop early and are relatively frequent, RCTs may yield reliable data on their incidence.

6b.2. Formulating the problem

6b.2.1 Scope of an assessment of adverse effects

It would be impractical for review authors to carry out exhaustive safety analyses for every intervention. Table 2 describes some specific therapeutic situations in which a detailed evaluation of adverse effects is warranted.

The scope of the adverse effects evaluation needs to be defined during protocol development as the subsequent direction of the review depend critically on the chosen approach, which may be:

1. *Assess intended and unintended (adverse) effects together, applying common inclusion criteria (in terms of types of studies, types of participants and types of interventions).*

   Here a single search strategy would be used. The critical issue is how the review authors intend to deal with the three datasets that may potentially arise:

   (a) studies that report both the intended effects and adverse effects of interest
   (b) studies that report intended effects but not adverse effects
   (c) studies that report adverse effects, but not the beneficial outcomes of interest

   A review based on the first dataset (a) is relatively easy to perform, and has the important advantage that benefits and harms can be compared directly since the data are derived from the same population and setting. Furthermore, evidence on benefits and harms arise from studies with similar designs and quality. However, data on adverse effects may be very limited and biased towards short-term harms.

   Evaluation of benefit and harm using some combination of the three datasets (rather than (a) alone) will increase the amount of information available. For instance, datasets (a) and (b) could be used to evaluate beneficial effects, while (a) and (c) could be used to assess adverse effects. However, as the studies addressing adverse effects are different from studies addressing beneficial effects, authors should note benefits and harms cannot be easily compared directly.

2. *Assess intended and unintended (adverse) effects together but use different inclusion criteria for selecting studies that address unintended (adverse) effects*
   The application of different inclusion criteria is a method of specifically addressing the problem that most experimental studies (such as RCTs) are insufficient to evaluate rare, long-term or previously unrecognized adverse effects. The approach allows a more rigorous evaluation of adverse effects, but is more costly in time and resources, tends to increase the quantity of data with higher risk of bias, and means that benefits and harms can often not be compared directly.
3. *Undertake a separate review only of adverse effects*
   A separate review might be considered for an intervention that is given for a variety of diseases or conditions, yet whose adverse effect profile might be expected to be similar in different populations and settings. For example, aspirin is used in a wide variety of patients, such as those with stroke, or peripheral vascular disease, and also in those with coronary artery disease. The main effects of aspirin would typically be addressed in separate Cochrane reviews, but adverse effects (such as intracerebral or gastrointestinal bleeding) are probably similar within the different disease groups and might be addressed together in an independent review. Indeed, unless trials exist on combined populations, such a question would be difficult to address in any other way. This approach might reduce the workload.

Table 2. Contexts and examples warranting detailed examination of adverse effects

When there is a narrow margin between benefit and harm	
Treatment is of modest or uncertain benefit, with some possibility of harm.	• Aspirin for prevention of cardiovascular events in a healthy patient; increase in haemorrhage. • Antibiotics for sore throat and respiratory tract infections; risk of rash and diarrhoea. • Finasteride for the treatment of male pattern baldness; causes erectile dysfunction. • Urgent direct current cardioversion in patients with new atrial fibrillation who are cardiovascularly stable; risk of stroke from cardioversion
Treatment potentially highly beneficial, but there are major safety concerns	• Aspirin for a patient with a stroke, but who has a past history of gastrointestinal haemorrhage. • Carotid endartarectomy in elderly patients with ischaemic heart disease who present with stroke
Treatment potentially beneficial in long-term, or to community, but no immediate direct benefit to individual.	• Improving uptake of a vaccine to promote herd immunity, while trying to assuage fears about early serious neurological adverse effects.
When there are a number of efficacious treatments with differing safety profiles	
Treatments are of equivalent efficacy, but they have different safety profiles	• Antiepileptic drugs for women with epilepsy who plan on becoming pregnant • A new insulin injection device is thought to cause less pain than the existing device
The balance of benefits and harms differ substantially e.g. the most efficacious intervention may have serious adverse effects, while the less effective intervention is potentially safer.	• Warfarin or aspirin in a healthy middle aged man with lone atrial fibrillation. • Disease-modifying drug in erosive rheumatoid arthritis e.g. using hydroxychloroquine (relatively safe) or methotrexate (potentially more effective, but less safe). • Radical mastectomy for breast cancer as opposed to limited, breast-conserving surgery
When adverse effects deter a patient from continuing on an efficacious treatment	
Treatment is of considerable benefit but adverse effects threaten patient's adherence.	• Patient with severe heart failure has responded well to an ACE inhibitor, but now complains of cough. Which is the best option - stopping the medication altogether, trying a lower dose, or changing to an angiotensin receptor blocker?

6b.2.2 What types of outcomes?

Selection of adverse outcomes can be difficult. Specific adverse effects associated with an intervention may be known in advance of the review, others will not. It may not be possible to identify beforehand exactly which effects will be most relevant to the review. The following

general strategies may be used depending on the study question and the therapeutic or preventive context.

Narrow focused:

A detailed analysis of one or two known or a few of the most serious adverse effects that are of special concern to patients and health professionals;

Pros: Easiest approach, especially with regard to data extraction. Can focus on important adverse effects and reach a meaningful conclusion on issues that have a major impact on the treatment decision (McIntosh 2004).

Cons: Scope may be too narrow. Method is only really suitable for adverse events that are known in advance.

Broad sweep:

To detect a variety of adverse effects, whether known or previously unrecognized, in the included studies.

Pros: Wider coverage, and can evaluate new adverse effects that we may not have previously been aware of.

Cons: Potentially large volume of work with particular difficulties in the data extraction process. Some researchers have found broad, non-specific evaluations to be very resource-intensive, with little useful information to show for the effort expended (McIntosh 2004). These researchers also point out that detection of previously unrecognized adverse effects may be best addressed through primary surveillance (see Section 3.3), rather than in a systematic review.

In order to address adverse effects in a more organized manner, review authors may choose to narrow down the broad sweep into some of the following areas:

- the five to ten most frequent adverse effects
- all adverse effects that either the patient or the clinician considers to be serious
- the most common adverse effects that lead the patient to stop using the intervention (caution – see also section 5.4 in this chapter);
- By category, for example:
    - diagnosed by clinician (e.g. gastrointestinal haemorrhage)
    - diagnosed by lab results (e.g. hypokalaemia)
    - patient-reported symptoms (e.g. pain).
    - biomarkers that may be early indicators of possible adverse effects (for example, abnormal liver enzymes); offering a means of collecting relevant information even from short-term studies.

This is not a comprehensive list, but the use of any of the above strategies should help authors approach the adverse effects analysis in a systematic, manageable and clinically useful fashion.

6b.2.3 What types of studies?

The decisions on what types of studies to include will be based primarily on the research question, balancing the elements of comprehensiveness, type of adverse effect(s) of interest, as well as the time and resources available.

Although most Cochrane systematic reviews focus on RCTs, which provide the most reliable estimates of effect, rare adverse events are unlikely to be observed in clinical trials, and a thorough investigation may require the inclusion of cohort studies, case-control studies and even case series.

6b.3. Locating and selecting studies

6b.3.1 Choice of search method

The scope of the review (see Section 2.1 Scope of an assessment of adverse effects) determines the nature of the search strategy. The general approaches for searching and selection are:

(i) Apply a standard search strategy as recommended by the author's Collaborative Review Group (CRG). Check all retrieved studies to identify those that report the intended effects and/or unintended effects of interest. This strategy is relatively simple and less resource intensive. However, it is likely to lead to different lists of potentially relevant studies for intended and unintended effects, with some overlap between them. Further evaluation of these lists depends on the proposed scope of the review, as described in Section 2.1.

(ii) Conduct a separate, additional adverse effects search to supplement the standard strategy. This is far more comprehensive but is likely to be time-consuming and resource intensive.

6b.3.2 Additional adverse effects search

The optimal search strategy for specifically identifying reports of adverse effects has yet to be established, although work on this area is ongoing (Golder 2004a, Golder 2004b). Two main approaches can be used, both of which have their own limitations and so a combination of these approaches is advisable to maximise sensitivity (the likelihood of not missing studies that might be relevant):

Searching electronic databases using index terms (also called controlled vocabulary or thesaurus terms)

Index terms such as MeSH or Medical Subject Headings in MEDLINE and EMTREE in EMBASE are assigned to records in electronic databases in order to describe the studies. Subheadings can also be added to index terms to describe specific aspects for example, side effects of drugs, or complications of surgery. There are differences in index terms used to denote data on adverse effects in the major databases (MEDLINE and EMBASE), for example:

Aspirin/adverse effects (MEDLINE)

Acetylsalicylic-acid/ adverse-drug-reaction (EMBASE)

In the above example, Aspirin is the MeSH term and adverse effects is the subheading; Acetylsalicylic-acid is the EMTREE term and adverse-drug-reaction is the subheading.

Within a database, studies may be (i) indexed under the name of the intervention together with a subheading to denote that adverse effects occurred, for example, Aspirin/adverse effects or Mastectomy/complications; or (ii) the adverse event itself may be indexed, together with the nature of the intervention, for example, Gastrointestinal Hemorrhage/ and Aspirin/ or Lymphedema/ and surgery/; or (iii) occasionally, an article may be indexed only under the adverse event, for example, Hemorrhage/chemically-induced.

Thus, no single index or subheading search term can be relied on to identify all data on adverse effects, although a combination of index terms and subheadings is useful in detecting reports of major adverse effects which are likely to be considered of significance by the indexers (Derry 2001).

Subheadings which may prove useful in MEDLINE are:

/adverse effects

/poisoning

/toxicity

/chemically induced

/contraindications

/complications

Subheadings which may prove useful in EMBASE are:

/side effect

/adverse drug reaction

/drug toxicity

/complication

Searching electronic databases using free text terms (also called text words)

Free text terms are used by authors in the title and abstract of their studies when published as journal articles and these terms are then searchable in the title and abstract of electronic records in databases. There are two important problems that severely limit the usefulness of free text searching:

- there is a wide range of terms used by authors to describe adverse effects, both in a general sense (toxicity, side-effect, adverse effects) and more specifically (for example, lethargy, tiredness, malaise may be used synonymously). Therefore, as many relevant synonyms as possible should be included in the search.
- adverse effects that are not mentioned in the title or abstract of the study and are, therefore, not included in the electronic record (even though they are described in the full report), will not be detected using the free text search (Derry 2001).

A highly sensitive free text search should incorporate this potentially wide variety of synonymous terms used to denote data on adverse effects in studies while also taking into account different conventions in spelling and variations in the endings of terms to include, for example, singular and plural terms, for example, adverse or side or hemorrhage or haemorrhage or bleed or bleeding or blood loss. These terms used to describe adverse effects should then be combined with free text terms used to describe the intervention of interest, for example (aspirin or acetylsalicylic acid) and (adverse or side or hemorrhage or haemorrhage or bleed or bleeding or blood loss).

It is clear that no single approach can be relied on to yield all the studies that have data on adverse effects of an intervention. The search, therefore, needs to combine index terms and free text terms and is likely to take several iterations. For instance, it may be necessary to repeat the electronic search incorporating additional index terms, subheadings and free text terms derived from the terms used to index and describe the studies initially identified as relevant. In deciding which combination of terms to use, authors will need to balance comprehensiveness (sensitivity) against precision. For example, an electronic search that retrieves 20,000 studies is likely to contain the majority of all relevant studies but if only 300 are relevant (1.5%), then it is very imprecise and will have a cost implication in terms of time and resources. (See Section 4).

6b.3.3 Additional sources of information

Review authors who are planning an exhaustive search may wish to consider checking the following sources:

- Standard reference books on adverse effects such as Meyler's Side Effects of Drugs and its annual update, the Side Effects of Drugs Annuals, and screening the papers they summarise.
- Regulatory agencies, for example:
    - in Australia the Australian Adverse Drug Reactions Bulletin (http://www.tga.gov.au/adr/aadrb.htm)

- and the European Public Assessment Reports from the European Medicines Evaluation Agency (http://www.emea.eu.int/#).
- in the UK Current Problems in Pharmacovigilance (http://medicines.mhra.gov.uk/ourwork/monitorsafequalmed/currentproblems/cpprevious.htm)
- in the US, MedWatch, the Food and Drug Administration Safety information and Adverse Events Reporting Program (http://www.fda.gov/medwatch/elist.htm)

Authors can also apply to the WHO Uppsala Monitoring Centre (UMC; http://www.who-umc.org) for special searches of their spontaneous reporting database; this was for example done for melatonin (Herxheimer 2002). However, frequencies of adverse effects calculated from UMC data may differ from the figures derived from a meta-analysis of double-blind, randomized controlled trials (Loke 2004).

Information on the safety of medical devices and surgical interventions is also available from a variety of regulatory authorities. Some examples include:

- UK National Joint Registry, which records details of hip and knee replacement operations in England and Wales (http://www.njrcentre.org.uk)
- The Medical Devices section of the UK Medicines and Healthcare Products Regulatory Agency (http://devices.mhra.gov.uk/)
- the US Food and Drug Administration, MedWatch for devices (http://www.fda.gov/medwatch/index.html)

6b.4. Assessment of study quality

The usual tools for assessment of methodological quality (see Section 6. Assessment of study quality) should identify more rigorous studies with results closer to the 'truth' – presumably for both therapeutic and adverse effects. However, we lack empirical evidence for the relevance of quality tools to adverse effect analysis. The author should use the standard quality assessment tools cautiously as the study quality assessed may apply only to the primary focus of the study, which would usually be the intended effects of the intervention. For example, the primary outcome measure of an intervention may have been studied in a placebo controlled, triple-blind, adequately concealed randomized trial, with standard laboratory measurements. In contrast, the adverse effects of the same treatment may be collected retrospectively, when treatment allocation is known to one or more of the parties (patients, clinician, analyst) via a self-assessment questionnaire. Although a high quality grade may be given to the primary portion of the study, the design to monitor the harmful affects of the treatments falls far short of this standard.

However, there is evidence that the methods used in monitoring or detecting adverse effects have a major influence on adverse effect frequencies. For example, in a group of hypertensive patients, passive monitoring based on spontaneous reports yielded rates of 16%, while active surveillance using specific questioning found a rate of 62% (Olsen 1999). Studies in which adverse effects are carefully sought will report a higher frequency than studies in which they are sought less carefully. Different methods of monitoring adverse effects will yield different results, which may make comparisons between studies, or a formal meta-analysis, impossible (Edwards 1999).

Selective reporting of results is also a particular problem with adverse effects (Ioannidis 2001, Loke 2001). For example:

- Certain categories only may be reported (for example, the study states that they looked for events defined by: several body systems, methods of collection, time periods (3, 6, 12 months), dose (20mg, 40mg, 80mg), but report only laboratory results for neurological disorders after 6 months with the 40mg dose).

- Adverse event categories may not be clearly defined (for example, ‘system = cardiovascular’ but, without indicating seriousness, intensity, duration, diagnostic method, or final outcome).
- Treatment groups may be combined (for example, “x participants withdrew from the study because of adverse effects”).
- Generic statements (for example, “no unexpected adverse effects were seen”/“there was no difference between the groups in adverse effects reported’/’the drugs were well tolerated”).

In many instances (particularly with the generic statements above), authors may have to take greater account of what was left unsaid rather than what was actually reported. Authors will have to choose either to exclude the study from the adverse effect analysis, or to include the study on the assumption that there were indeed no adverse effects (this should be the exception rather than the rule).

Thus authors should take into account two important aspects in assessing the quality of adverse effects:

- How rigorous were the methods used in detecting adverse effects?
- How good is the quality of reporting?

Examples of potentially useful questions in each area are:

On conduct:

Are definitions of reported adverse effects given?

How were adverse effects data collected: prospective/routine monitoring, spontaneous reporting, patient checklist/ questionnaire/diary; systematic survey of patients?

On reporting:

Were any patients excluded from the adverse effects analysis?

Were the methods used for monitoring adverse effects reported?

Did the report provide numerical data by intervention group?

Which categories of adverse effects were reported by the investigators?

Did the investigators report on all important or serious adverse effects?

Finally, non-randomized studies are prone to biases, which can be hard to identify and deal with and authors planning to include such data should seek guidance from the Cochrane Non-randomised Studies Methods Group.

6b.5. Collecting data

6b.5.1 Terms

We suggest that information falling under any of these terms ‘adverse effect’, ’adverse drug reaction‘, ’side effect‘, ’toxic effect‘, and ’adverse event’ be considered as being potentially suitable for data extraction when evaluating the harmful effects of a treatment. For further details see Glossary and Table 1, Section 1 above.

6b.5.2 Exclusions

Remember that no mention of adverse effects does not necessarily mean that no adverse effects occurred. It is usually safest to assume that they were not ascertained or not recorded: authors have to choose between excluding the study from the adverse effect analysis, and including it on the assumption that the incidence was zero (that should be the exception).

6b.5.3 Data collection forms

Authors may find it useful to design and use a separate data collection form for safety outcomes. Some reviews may include additional studies beyond those included in the therapeutic portion of a review.

6b.5.4 Outcome characteristics

The definition of a particular adverse effect may vary between studies, as can definitions of intensity. For example, in a review of aspirin and gastrointestinal haemorrhage, some trials simply reported "gastrointestinal bleeds"; others reported specific categories of bleeding, such as haematemesis, melaena, and proctorrhagia, (Derry 2000). The definition and reporting of severity of the haemorrhages (for example, major, severe, requiring hospital admission) also varied considerably among the trials. (Zanchetti 1999).

Moreover, a particular adverse effect may be described and/or measured in different ways among the trials – take for example, tiredness, fatigue or lethargy, all of which might be terms used in adverse effects reports. Authors may also use different thresholds for 'abnormal' results (for example, hypokalaemia diagnosed at a serum potassium concentration of 3.0 mmol/l or 3.5 mmol/l).

Are the adverse effects terms comparable across studies? Authors will need to decide which categories are similar enough to collect data on and justify lumping together in the analysis. For example, gastrointestinal bleed, haematemesis, and melaena were included in the aspirin analysis, but proctorrhagia was excluded.

There are a number of initiatives aimed at harmonizing adverse effects terms (Bankowski 1999), and the National Cancer Institute set of toxicity criteria is an example of a standardized scheme for judging severity of adverse effects across trials of cancer therapy. (http://ctep.cancer.gov/reporting/CTC-3.html). The WHO uses the system-organ class categories (http://www.who-umc.org/pdfs/ardguide.pdf) which allows authors to collate adverse effects data into one of several system-organ classes such as 'gastrointestinal system disorders' or 'vision disorders' (MacLehose 2003). However, some researchers have found that the standard 'preferred terms' used by regulators and industry can distort descriptions in the original reports of adverse events and blur distinctions between them (Medawar 2003).

Withdrawal or drop-outs as outcome measure

These outcome measures are often seen in trial reports. We urge authors to be cautious in interpreting such data as surrogate markers for safety or tolerability because of the potential for bias:

- The attribution of reason(s) for discontinuation is complex and may be due to mild but irritating side effects, toxicity, lack of efficacy, non-medical reasons, or a combination of causes (Ioannidis 2004).
- The pressures on patients and investigators under trial conditions to keep the number of withdrawals and drop-outs low can result in rates that do not reflect the experience of adverse events within the study population.
- Unblinding of treatment assignment often takes place prior to the decision to withdraw. This can lead to an over-estimate of the intervention's effect on patient withdrawal. For example, symptoms of patients in the placebo arm are less likely to lead to discontinuation. Conversely, patients in the active intervention group who complained of symptoms suggestive of adverse effects would have been more readily withdrawn.

Quality of Life Indicators

These are usually general measures that do not look specifically at particular adverse effects of the intervention. While quality of life scales can be used to gauge the overall well-being, they should not be regarded as substitutes for a detailed evaluation of safety and tolerability.

6b.6. Analysing and presenting results

In addition to the advice given in Section 8, there are number of issues especially relevant to the analysis of adverse effects.

If different types of studies are being used to evaluate beneficial and harmful effects, then an author must consider how to analyse potentially disparate datasets where studies reporting intended effects are different from those that report adverse effects. Special techniques might be used to synthesise data from a diverse range of sources (Wald 2003, Jefferson 2003).

The analysis of zero events in either arm (for example, “the drug was safe”, and “no serious adverse effects were seen”) needs careful consideration. Data of this type need to be evaluated in the following contexts:

- How thorough were the methods used to detect adverse effects?
- How many patients were studied and for how long?

It is not possible (based on zero events detected) to conclude that a drug does not cause a suspected adverse effect. However, we can use the rule of 3, which states that the 95% confidence intervals of zero are 0-3 events in the observed sample, to estimate an upper limit for the frequency of the adverse effect (Eyspach 1995). For example, if no adverse effects occur in 300 participants, then any adverse effects associated with the intervention might be as frequent as 1 in 100, but are unlikely to be more frequent. Note that studies with no events in either arm can be included in a meta-analysis of risk differences, although they cannot be included in a meta-analysis of odds ratios or risk ratios.

It is important to remember that a systematic review is not synonymous with a meta-analysis. There may be occasions when adverse effect information is best summarised in a qualitative or descriptive manner. For instance, data derived from divergent sources (for example, different study design, different populations, different data collection methods) cannot be combined. It may not be possible to compare benefits and harms directly. In practice this means that adverse effects from RCTs, case reports, case series, cohorts, and case controls cannot all be pooled together using standard meta-analysis principles. Moreover, the data from non-randomised studies are more prone to bias, and are often heterogeneous; combining them to produce a summary statistic may not be appropriate.

6b.7. Interpreting results

6b.7.1 Applicability

Many RCTs are restricted to carefully selected subgroups of the population, and it is generally inappropriate to extrapolate adverse effects data from such studies to the wider population, which includes more vulnerable people, for example, with co-morbidities, co-medications. In interpreting adverse effects data, authors must take into account the inclusion and exclusion criteria used during recruitment of participants.

6b.7.2 Trade-offs

Including studies beyond those included in the analysis of intended effects means that the analysis of harm is carried out in studies whose participants may differ from those included in the studies used in the analysis of benefit. This creates potential difficulties in assessing the trade-off

between benefits and harms. Review authors will need to consider how much, if at all, the participants in the additional studies can differ from those in the benefit studies, and remain comparable.

For example, in a study of the benefits and harms of aspirin used as an antiplatelet drug to reduce cardiovascular events, a review author might want to include in the adverse effect analysis a study in which aspirin was used as an antiplatelet drug to reduce scarring after mastectomy. Predefined inclusion criteria, other than indication for treatment (for example, dose, duration of treatment, reporting of adverse effects), would need to be met. The decision to include the study or not should depend on whether there is evidence that these women differ systematically in their risk of gastrointestinal haemorrhage from people who take the drug to prevent cardiovascular problems.

Extending the review to observational studies and anecdotal case reports can create additional difficulties in evaluating the benefit: harm trade-off. Authors will need to consider how efficacy data from high-quality trials can be weighed up against adverse effects from low quality studies.

6b.8. Contributions

Contributing authors: Jeff Aronson, Anne-Marie Bagnall, Andrea Clarke, Sheena Derry, Andrew Herxheimer, Yoon Loke, Heather McIntosh, Harriet MacLehose, Deirdre Price, Nerys Woolacott

Comments on drafts: Phil Alderson, Jon Deeks, Anne Eisinga, Su Golder, Sally Green, Julian Higgins, Tom Jefferson, Carol Lefebvre, Philippa Middleton

Editors: Yoon Loke, Deirdre Price and Andrew Herxheimer on behalf of the Cochrane Adverse Effects Subgroup.

6b.9. References

Aronson 2005. Aronson JK, Ferner RE. Clarification of terminology in drug safety. Drug Safety. 2005; in press

Derry 2001. Derry S, Kong Loke Y, Aronson JK. Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. BMC Medical Research Methodology 2001; 1: 7

Derry 2000. Derry S, Loke YK. Risk of gastrointestinal haemorrhage with long term use of aspirin: meta-analysis. BMJ 2000; 321: 1183-7

Edwards 2000. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. Lancet 2000; 356: 1255-1259

Edwards 1999. Edwards JE, McQuay HJ, Moore RA, Collins SL. Reporting of adverse effects in clinical trials should be improved: lessons from acute postoperative pain. Journal of Pain and Symptom Management 1999; 18: 427-37

Eypasch 1995. Eypasch E, Lefering R, Kum CK, Troidl H. Probability of adverse events that have not yet occurred: a statistical reminder. BMJ 1995; 311: 619-20

Golder 2004a. Golder S, Duffy S, Glanville J, McIntosh H, Miles J. Developing efficient search strategies to identify papers on adverse events. A: testing and precision sensitivity [abstract]. In: 12th Cochrane Colloquium 2004 Oct 2-6; Ottawa, Ontario, Canada:75-6

Golder 2004b. Golder S, Duffy S, Glanville J, McIntosh H. Developing efficient search strategies to identify papers on adverse events. B: using statistical analysis [abstract]. In: 12th Cochrane Colloquium 2004 Oct 2-6; Ottawa, Ontario, Canada:75.

Herxheimer 2002. Herxheimer A, Petrie KJ. Melatonin for the prevention and treatment of jet lag (Cochrane review). The Cochrane Database of Systematic Reviews 2002, Issue 2 Art. No.: CD001520. DOI: 10.1002/14651858.CD001520

I**oannidis 2004.** Ioannidis JP, Evans SJ, Gotzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Annals of Internal Medicine 2004; 141: 781-8

Ioannidis 2001. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. Journal of the American Medical Association 2001; 285: 437-43

Jefferson 2003. Jefferson T, Demicheli V. Balancing benefits and harms in health care: observational data on harm are already included in systematic reviews. BMJ 2003; 327: 750

Loke 2004. Loke YK, Derry S, Aronson JK. A comparison of three different sources of data in assessing the frequencies of adverse reactions to amiodarone. British Journal of Clinical Pharmacology 2004; 57: 616-21

Loke 2001. Loke YK, Derry S. Reporting of adverse drug reactions in randomised controlled trials - a systematic survey. BMC Clinical Pharmacology 2001; 1: 3

McIntosh 2004. McIntosh HM, Woolacott NF, Bagnall AM. Assessing harmful effects in systematic reviews. BMC Medical Research Methodology 2004; 4: 19

MacLehose 2003. MacLehose H, Klaes D, Garner P. Amodiaquine: a systematic review of adverse events. Version 1, March 2003. WHO. Available at: http://www.who.int/medicines/oranization/par/edl/expcom13/expcom03add.shtml.

Medawar 2003. Medawar C, Herxheimer A. A comparison of adverse drug reaction reports from professionals and users, relating to risk of dependence and suicidal behaviour with paroxetine. International Journal of Risk and Safety in Medicine 2003; 16: 5-19

Olsen 1999. Olsen H, Klemetsrud T, Stokke HP, Tretli S, Westheim A. Adverse drug reactions in current antihypertensive therapy: A general practice survey of 2586 patients in Norway. Blood Pressure 1999; 8: 94-101

Wald 2003. Wald NJ, Morris JK. Teleoanalysis: Combining data from different types of study. BMJ 2003; 327: 616-8

Zanchetti 1999. Zanchetti A, Hansson L. Risk of major gastrointestinal bleeding with aspirin. Lancet 1999; 353: 148-50

APPENDIX 8a. Considerations and recommendations for figures in Cochrane reviews: Graphs of statistical data

Date this version prepared: 4 December 2003

8a.1 Introduction

Historically, graphical illustrations of data in Cochrane Reviews have been generated using MetaView, an analysis program from Update Software that is used in conjunction with Review Manager (versions up to and including 4.1), and with the Cochrane Library. From version 4.2, RevMan uses a program called RevMan Analyses instead of MetaView, although MetaView is still currently used to present some output on The Cochrane Library. MetaView and RevMan Analyses perform and display meta-analyses of dichotomous data, continuous data and 'O – E' statistics from time-to-event data (Alderson 2004, Deeks 2001). In addition, RevMan Analyses will perform meta-analyses from a variety of data types using the generic inverse variance option. The Information Management System Group (an advisory group to the Steering Group) agreed in December 2000 the need for additional figures to be available in Cochrane Reviews. The purpose of this document is to provide recommendations from the Statistical Methods Group (SMG) of the Cochrane Collaboration regarding the content of graphical displays. It is intended to cover forest plots as displayed by MetaView and RevMan Analyses and additional figures that reviewers may wish to include in a Cochrane Review.

8a.1.1 Graphs and Cochrane Reviews

The purpose of a graph is to present numerical data in visual form. Graphs enable the identification of overall patterns, correlations and outlying observations that might be overlooked in tables of data. Graphs are especially valuable when a table is not an option (for example, presenting numerous data in a scatter diagram) and/or where there is some possible trend to look for. They can save the reader considerable time and effort in absorbing the findings of a systematic review, and can facilitate the comparison of data across different scenarios. However, if poorly designed they can frustrate and even mislead the reader.

There are many ways of analysing and displaying data arising from a systematic review, a meta-analysis or indeed a single study included in a systematic review. Graphical displays for meta-analysis have been discussed by Galbraith (Galbraith 1988), Light et al (Light 1994), Pettiti (Petitti 1994) and Sutton et al (Sutton 1998). It is expected that the majority of figures deemed appropriate for inclusion in Cochrane Reviews will be forest plots. Facilities for drawing forest plots are available within Cochrane review-writing software, and these should be used in preference to other facilities whenever possible.

This document has been developed by members of the Statistical Methods Group to address the following:

- General considerations and recommendations for graphs in systematic reviews
- Recommendations and examples for forest plots
- Recommendations and examples for the following types of plots that might, on occasion, be appropriately included in Cochrane Reviews as additional figures
    - Summary forest plots
    - Funnel plots
    - Relationship between treatment effect and a single covariate (meta-regression)

        - Graphical displays particular to dichotomous outcome data (L'Abbé plots and plots relating treatment effect to "underlying risk")
    - Considerations for the following plots that are not specifically encouraged in Cochrane Reviews
        - Galbraith (radial) plots
        - Relationship between treatment effect and two or more covariates (meta-regression)
        - Survival curves
        - Cumulative meta-analysis
        - Other graphical displays

The SMG has developed recommendations as guidelines and not as rules. On occasion there may be good reason to approach a graph differently. Further, the types of graph addressed in this document are not a comprehensive list of those that may usefully be included in a systematic review. Given the almost limitless possibilities available to a reviewer, we place high emphasis on the following general recommendation.

General recommendation

1.1. Every graphical display of data should be assessed by a statistician as part of the editorial process within the relevant Collaborative Review Group, before being submitted as part of a Cochrane Review. The assessment should cover appropriateness, clarity and obvious errors. Ideally it should also cover correctness of the data and/or analyses being presented. Establishing correctness of data may require examination of original reports from the included studies.

A key characteristic of meta-analyses included in Cochrane Reviews has been the ready availability of the data being analysed. This allows the interested reader to investigate alternative ways of analysing the data. In fact, RevMan Analyses and MetaView allow the reader to re-analyse the data using different measures of treatment effect and different models for the meta-analysis. As a general rule, it should be possible for the interested reader to duplicate analyses included in all graphs.

General recommendation

1.2. Data represented in a graph should be tabulated whenever it is reasonable to do so (this may not be suitable for scatter plots, for example). Such data may appear within the graph, or elsewhere such as in 'Other data' tables or 'Additional tables' within the Cochrane Review.

8a.2 Principles of graphing data

Five principles, discussed in detail by Cleveland (Cleveland 1994), provide a useful framework for creating, selecting or refining a graph. They are (i) accuracy, (ii) simplicity, (iii) clarity, (iv) appearance, and (v) a well-defined structure. A reviewer or statistician creating graphs for inclusion in a Cochrane Review should also remember that a high proportion of the readership have had no training in research methods or statistics.

There are certain criteria that all graphical displays of data should fulfil. The list below represents an ideal, and incorporates advice drawn from various external sources (Arkin 1940, Simmonds 1980, Schmid 1983, Cleveland 1994). It may not be possible for a reviewer to control all of these aspects within their chosen software.

Recommendations for all graphical displays

Titles, captions and scales

2.1. The graph should be supplied with a brief, comprehensive title. It may be helpful to supplement this with a caption, that is a sentence or two to aid understanding and interpretation of the picture. The graph, along with its associated title and caption should generally be understandable outside the context of the rest of the document.

2.2. Explanatory variables (variables used to 'predict' changes in other variables) should be on the horizontal axis. This general rule is not followed in some common representations of meta-analysis, and we discuss it further in the context of specific graph types below.

2.3. Every axis should be labelled, identifying both the quantity and its units (using SI units where applicable).

2.4. Ranges of scales should be chosen so that all (or nearly all) the range of the data is included, and so as to maximise use of available space. However, they should not be chosen so that unimportant variation is exaggerated.

2.5. Excluded data (through curtailing axes or other reasons) should be mentioned in a caption to the graph.

2.6. It is generally desirable but not always necessary that key reference values are included on an axis (for example, 0 for a difference measure of treatment effect; 1 for a ratio measure of treatment effect, 0% and 100% for percentages)

2.7. If two or more graphs are to be compared directly (e.g. for subgroups), identical scales should be used.

2.8. There should not be an excessive number of tick marks or gridlines, and these should not interfere with data.

2.9. Sufficient tick marks should be labelled to allow the reader to interpolate values between them. There should be at least 3 tick marks on any axis. A "0." should be placed in front of decimal points.

2.10. When a log scale is used, the tick marks should be labelled on the original (un-logged scale)

2.11. A reference line should be considered for an important value (for example, a meta-analysis result), though such a line should not interfere with other components of the graph.

Representing data

2.12. The data should stand out so that main trends can be seen at a glance. Superfluous contents should be removed.

2.13. The weight (or thickness) of lines for data should be equal to, or exceed, that for the axes.

2.14. Clear and prominent symbols should be used to show data. Different plotting symbols should be distinguishable, especially if they may overlap.

2.15. Notes or keys should be used to define the meaning of different styles of lines or symbols. Direct labelling of lines or symbols is preferable. Notes and keys may be placed inside or outside the graphing area or within the caption. They should be placed inside the graphing area only when they do not interfere with data or clutter the graph.

2.16. It is important that variability and uncertainty are fully expressed when presenting results, but care must be taken when providing this information on a graph. Error bars may cause confusion or obscure the main data. Some possibilities are to present variability or uncertainty in separate tables; to use different sized plotting symbols; to extend error bars to one side only; or to plot points off-centre so that error bars do not overlap. All representations of variability or uncertainty must be explained, stating exactly which quantity (for example, standard error, weight, X% confidence interval) is being illustrated.

Perseverance of information

2.17. Graphs (including text within them) should be robust to reproduction and reduction. In particular, information must not be lost if the graph is reproduced in black and white. Whereas colour may be used to enhance the appearance of a graph, it must not be relied upon to distinguish different components.

2.18. Use of different line types can enhance visual impact.

8a.3 Principles of meta-analysis

Two of the principles underlying meta-analysis of healthcare intervention studies are as follows.

i. **Compare like with like**. Since studies are undertaken in different populations often using different variations of interventions, with different definitions of outcomes and using different designs, it is appropriate for experimental and control groups to be compared within studies and not across studies. The within-study comparisons ('treatment effects', or 'effect sizes') are combined across studies in the meta-analysis.

ii. **Not all studies are of equal importance**. The amount of weight awarded to each study in a meta-analysis reflects the amount of information in the study.

In using graphical methods for presenting meta-analyses, one would therefore generally expect that

i. *studies* (rather than, say, patients, treatments or single arms of studies) will be the unit of interest (the points being plotted); and

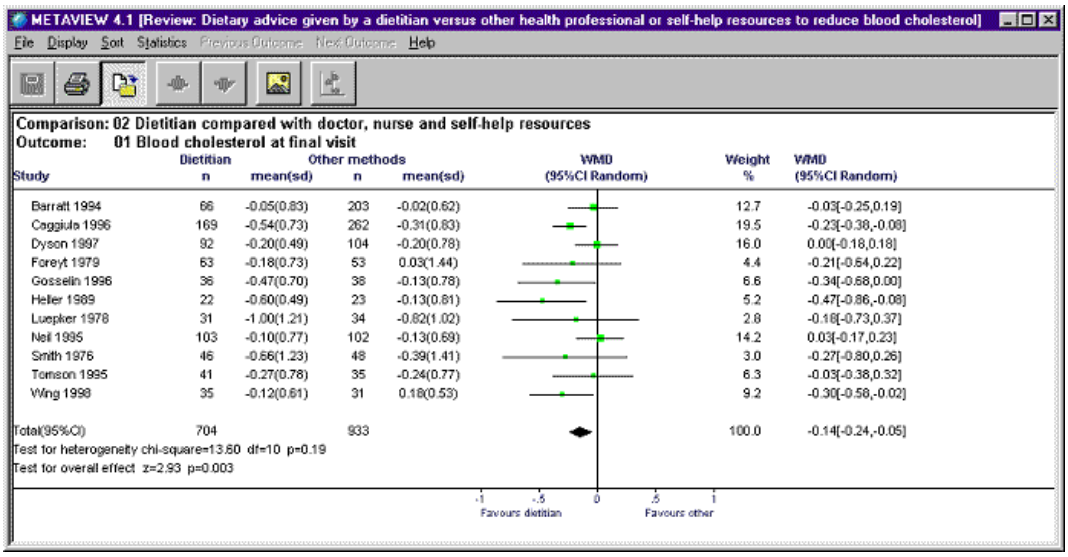
ii. the amount of information contained in each study will be reflected in the graph.

When creating graphical displays that are not addressed in this document, it may be helpful to bear these considerations in mind.

8a.4 Forest plots

Forest plots are also known as confidence interval plots. More informal terms include 'blocks and lines plots' and 'blobbograms'. They are the standard means of presenting results of individual studies and meta-analyses (Egger 1997a, Lewis 2001). A forest plot displays results (that is, estimates of treatment effect) and confidence intervals for individual studies and/or meta-analyses. Graphs produced by RevMan Analyses or MetaView are forest plots. An example is given in Figure 1. Each study is represented by a square at the point estimate of treatment effect and a horizontal line extending either side of the block. The area of the block is proportional to the weight assigned to that study in the meta-analysis, and the horizontal line gives a confidence interval (with specified level of confidence). The area of the block and the confidence interval convey similar information, but both have important contributions to the graph. The confidence interval provides a range of treatment effects compatible with the study's result. If it does not pass through the line of no effect this indicates that the result was individually statistically significant. The size of the block draws the eye towards the studies with larger weight (smaller confidence intervals). Failure to use this second device may result in unnecessary attention to those smaller studies with wider confidence intervals that put more ink on the page (or more pixels on the screen).

Figure 1: Forest plot from a Cochrane Review of dietary advice for cholesterol reduction (from Thompson 2001)



Forest plots may include meta-analyses, normally at the bottom of the graph. A variety of methods is available for conducting the meta-analysis, including both classical and Bayesian methods. Forest plots for Bayesian (or empirical Bayes) meta-analyses may include both the original and ‘shrunk’ estimates of treatment effect for each study. These would normally appear together.

It is conventional to represent all information relevant to each study (or meta-analysis) within a row. This means the horizontal axis of the graph denotes the size of treatment effect (the outcome, or dependent variable). This convention breaks the general rule that independent variables be plotted along the horizontal axis, and several authors (mainly statisticians) have thus drawn such graphs the other way round (Bailey 1987). However, we believe that the break with the general rule is justified, and offers advantages, for the following three reasons. We therefore incorporate the convention into our recommendations.

i. The ‘study’ axis is not a numerical scale, so the recommendation is of lesser importance. There is also a ‘natural break’ between a list of studies and a meta-analytic summary, which may be visually clearer when they are plotted one above the other.

ii. The convention enables written details of each study to be presented alongside the results. As a minimum, an identifier for the study (such as its Study ID) can be included without resorting to vertical or inclined text. Other information such as raw data, study characteristics and the numerical results being plotted may also be presented.

iii. The convention complements the typical presentation of tables of studies, in which studies appear in rows, and characteristics (or results) in columns.

Recommendations for forest plots

3.1. If a forest plot may appropriately be drawn using RevMan, it should be. All remaining recommendations are consistent with forest plots drawn using RevMan.

3.2. Forest plots should be referred to as ‘forest plots’ in preference to other names.

3.3. The treatment effect measure should be along the horizontal axis.

3.4. Ratio measures of treatment effect (such as odds ratios, relative risks, hazard ratios and rate ratios) should be plotted on the log scale. The labels on the axis, however, should be on the original (anti-logged) scale (Galbraith 1988).

3.5. A reference line should be drawn at the position of no treatment effect.

3.6. Another, usually dashed, line can be added to indicate the estimated pooled effect

3.7. Treatment effect estimates and confidence intervals should be plotted for each study and each meta-analysis.

3.8. The level of confidence for confidence intervals should be stated (for example, 95%, 99%). The levels of confidence need not be the same for individual studies and overall effect, though any differences must be clearly labelled.

3.9. The directions of effect should be clearly shown, preferably directly below the plot (for example, ‘Favours aspirin’ and ‘Favours placebo’ or ‘Aspirin better’ and ‘Aspirin worse’).

3.10. Treatment effect estimates and confidence intervals, or results sufficient to calculate these, must be presented numerically somewhere in the review.

Individual studies

3.11. The size of the block representing a point estimate from a study should usually relate to the amount of information in the study. If a meta-analysis is included, that information should be the weight apportioned to the study in the meta-analysis. If no meta-analysis is included, that information may be the weight that would be apportioned to that study in a meta-analysis, or the total sample size in the study. Note that weights depend not only on sample size, but also on the choice of treatment effect measure. (Thus, for example, relative weights are different on the odds ratios scale compared with the risk difference scale).

3.12. It should be possible to identify from which trial each result belongs. This will normally be achieved by including the ‘Study ID’ alongside the result.

3.13. Additional information such as the summary data and/or the numerical results being plotted can be helpful (Light 1994). This information is presented by default on meta-analyses generated using RevMan (see Figure 1).

3.14. The minimum number of studies appropriate for display in a forest plot is 2. In rare cases the number of studies will be very large, so that the plot cannot be read properly. It may be helpful to present a summary forest plot (see below).

3.15. Studies should have a meaningful order. Often this is alphabetical by study identifier, or according to date of publication. However, it may be helpful to order by some other characteristic, such as duration or dose of treatment.

Meta-analyses

3.16. The method used to perform a meta-analysis should be stated in the plot, in the title or in the caption. For example, it should be clear whether a fixed effect or random effects model has been used.

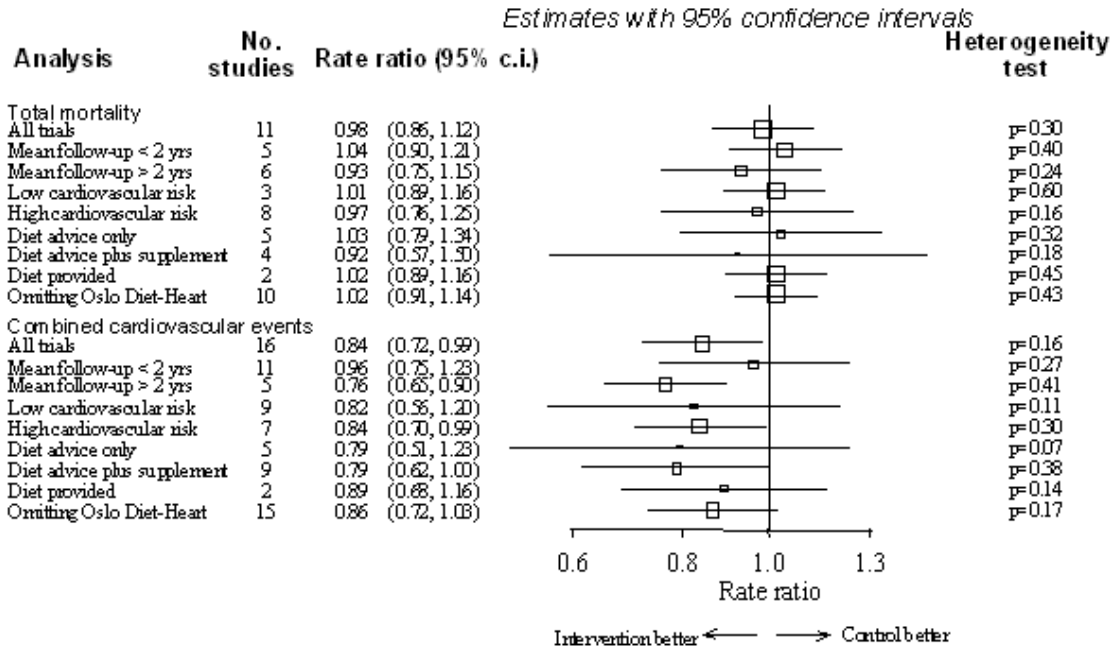
3.17. If both meta-analyses and individual studies are plotted, a meta-analysis should be plotted in a different style. For example, using a diamond (stretching the width of the confidence interval), or using an unfilled block (with accompanying confidence interval line).

3.18. If a meta-analysis is considered to be inappropriate, unhelpful, misleading or erroneous it should not be included in a forest plot.

8a.5 Summary forest plots

Forest plots may also be used to illustrate results of meta-analyses in the absence of individual study results, for example to enable the comparison of different outcomes, subgroup analyses or sensitivity analyses (see Figure 2). This is a particularly useful form of graph, and we propose the name ‘summary forest plot’ to indicate that the individual points represent meta-analyses rather than studies.

Figure 2: Forest tops plot of subgroup analyses and sensitivity analyses from a review of trials of reduction/modification of dietary fat or cholesterol (data from Hooper 2001)



Recommendations for summary forest plots

4.1. Recommendations 3.1 to 3.10 for forest plots, and 3.16 to 3.18 for meta-analyses within forest plots, should be followed.

4.2. The reviewer should consider carefully whether points should be drawn with equally sized blocks, or blocks according to total weight in each meta-analysis. For subgroup analyses and sensitivity analyses, block sizes according to total weight are recommended. When meta-analyses of different outcomes are presented in the same plot it may be more appropriate to use equally sized blocks.

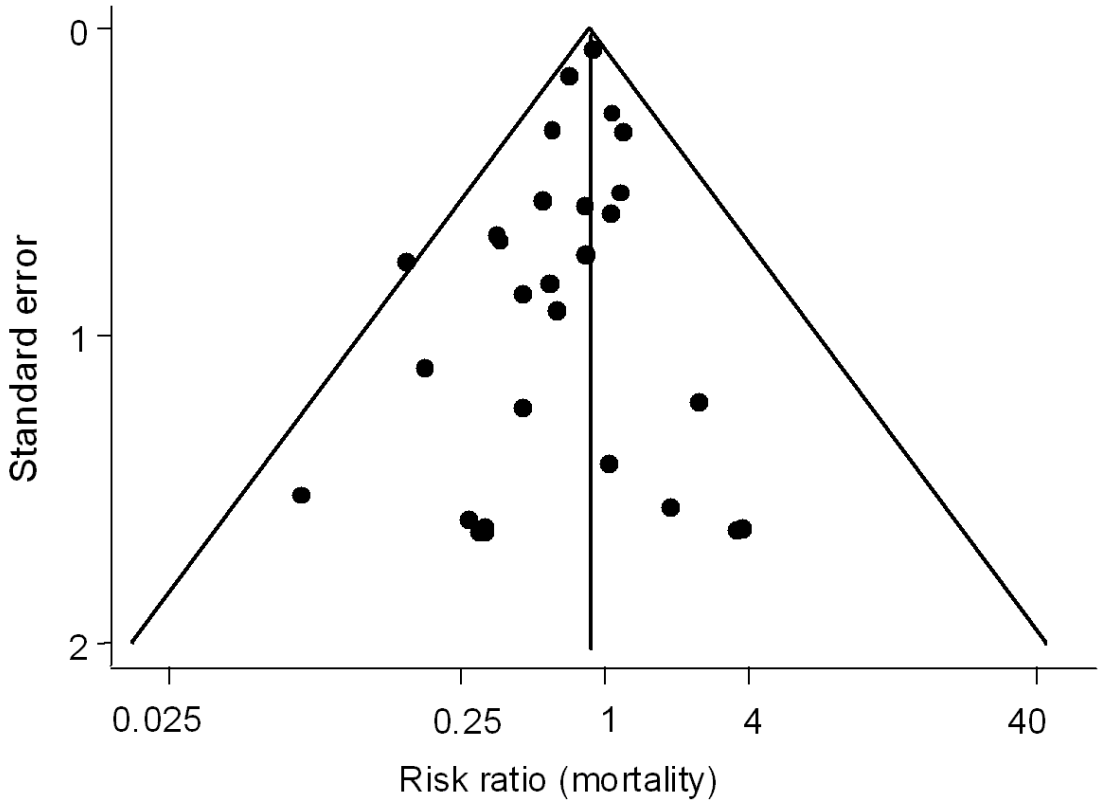
8a.6 Funnel plots

Funnel plots, introduced by Light and Pillemer (Light 1994) and discussed in detail by Egger and colleagues (Egger 1997b, Sterne 2001a), are useful adjuncts to meta-analyses. A funnel plot is a scatter plot of treatment effect against a measure of study size. It is used primarily as a visual aid to detecting bias or systematic heterogeneity. A symmetric inverted funnel shape arises from a 'well-behaved' data set, in which publication bias is unlikely. An asymmetric funnel indicates a relationship between treatment effect and study size. This suggests the possibility of either publication bias or a systematic difference between smaller and larger studies ('small study effects'). Asymmetry can also arise from use of an inappropriate effect measure. Whatever the cause, an asymmetric funnel plot leads to doubts over the appropriateness of a simple meta-analysis and suggests that there needs to be investigation of possible causes.

A variety of choices of measures of 'study size' is available, including total sample size, standard error of the treatment effect, and inverse variance of the treatment effect (weight). Sterne and Egger have compared these with others, and conclude that the standard error is to be recommended (Sterne 2001b). When the standard error is used, straight lines may be drawn to define a region within which 95% of points might lie in the absence of both heterogeneity and publication bias (Sterne 2001b).

In common with confidence interval plots, funnel plots are conventionally drawn with the treatment effect measure on the horizontal axis, so that study size appears on the vertical axis, breaking with the general rule. Since funnel plots are principally visual aids for detecting asymmetry along the treatment effect axis, this makes them considerably easier to interpret. We therefore feel this is justifiable and to be recommended. An example of a funnel plot appears in Figure 3. Funnel plots can be drawn within Review Manager version 4.

Figure 3: Funnel plot of trials of ACE inhibitors (data from Sterne 2001b)



Recommendations for funnel plots

5.1. The treatment effect measure should be along the horizontal axis.

5.2. Ratio measures of treatment effect (such as odds ratios, relative risks, hazard ratios and rate ratios) should be plotted on the log scale. The ticks and labelled values on the axis, however, should be on the original (anti-logged) scale.

5.3. The measure of study size (on the vertical axis) should generally be the standard error of the treatment effect estimate. A trick to invert the graph so that bigger trials appear at the top is to plot the negative standard error and override (or edit) the axis labels to remove the minus signs (Sterne 2001b).

5.4. Points should all be the same size, since the size of a study is already described using the vertical axis.

5.5. 95% limit lines may be included. If so they should usually be centred around a fixed effect meta-analysis.

5.6. Funnel plots may not be useful for small numbers of studies (for example, a small study effect may difficult to spot among fewer than ten studies)

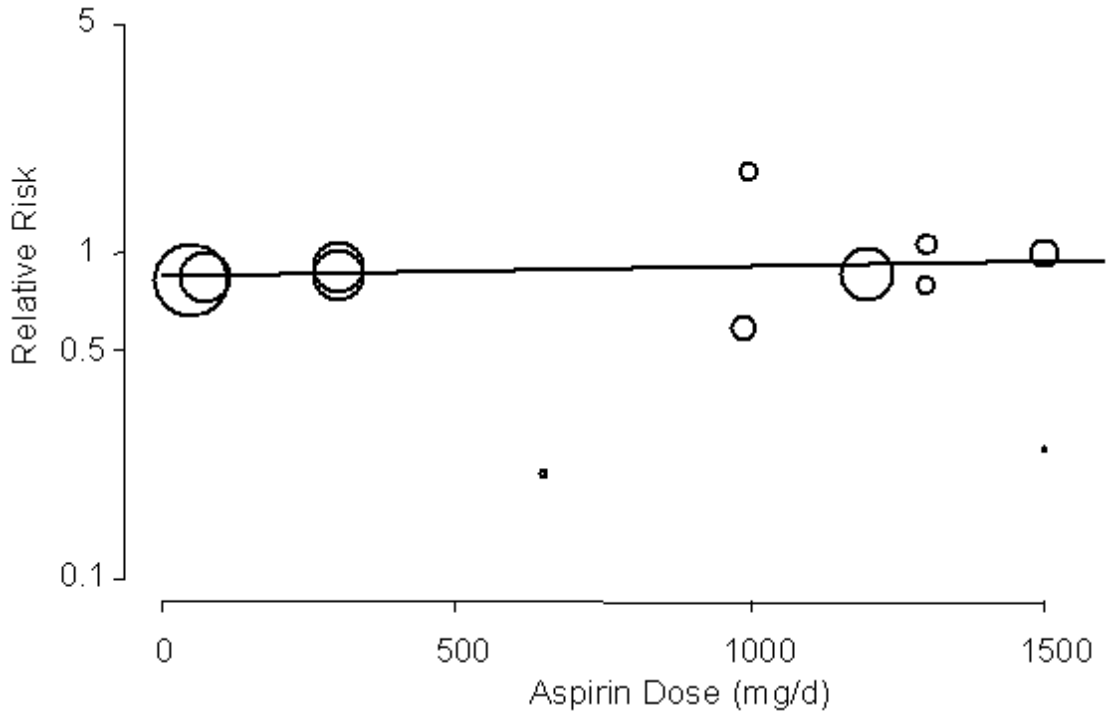
5.7. Treatment effect estimates and their standard errors, or results sufficient to calculate these, must be presented numerically somewhere in the review.

8a.7 Relationship between treatment effect and a single covariate (meta-regression)

It has been argued that sources of heterogeneity in a meta-analysis should be investigated (Thompson 1994). Often a source of heterogeneity can be summarized as a trial-level covariate, that is some varying characteristic of the trials. A scatter plot with the covariate along the horizontal axis and the treatment effect along the vertical axis provides a convenient visual impression of the relationship (Thompson and Higgins 2002). Such scatter plots have commonly followed the convention of plotting the covariate (explanatory variable) along the horizontal axis and the treatment effect (outcome variable) on the vertical axis.

Meta-regression is the statistical analysis of the association between treatment effect and the value of one, or more, trial-level covariate(s). The analysis yields a regression line that may be superimposed on the scatter plot. A particular application is when the treatment affects a continuous surrogate endpoint, such as blood pressure or serum cholesterol, in which case it may be hypothesized that the benefit of treatment, say on mortality, would be related to the success in modifying the surrogate. An example of a meta-regression analysis appears in Figure 4.

Figure 4: Relationship between relative risk and aspirin dose in 12 trials of aspirin for secondary prevention of stroke (data from Johnson 1999)



Recommendations for single variable ‘meta-regression’ plots

6.1. The covariate (trial-level characteristic) should be along the horizontal axis.

6.2. The treatment effect should be up the vertical axis.

6.3. A reference line at the position of no treatment effect may be useful.

6.4. Ratio measures of treatment effect (such as odds ratios, relative risks, hazard ratios and rate ratios) should be plotted on the log scale. The labels on the axis, however, should be on the original (anti-logged) scale.

6.5. Points should be of a size proportional to weight or trial size (preferably weight).

6.6. Trial weights or sample sizes should not be illustrated using confidence intervals alone (these draw attention to trials with small weights rather than those with large weights).

6.7. A meta-regression line may be plotted.

6.8. Confidence or prediction lines either side of the meta-regression line may be useful. Note that these are unlikely to be parallel to the meta-regression line.

6.9. For dichotomous outcome data, plots of treatment effect against underlying risk (as measured by observed control group event rate) is usually misleading and should be avoided (see below).

6.10. Treatment effect estimates, their standard errors and the covariate values, or results sufficient to calculate these, must be presented numerically somewhere in the review.

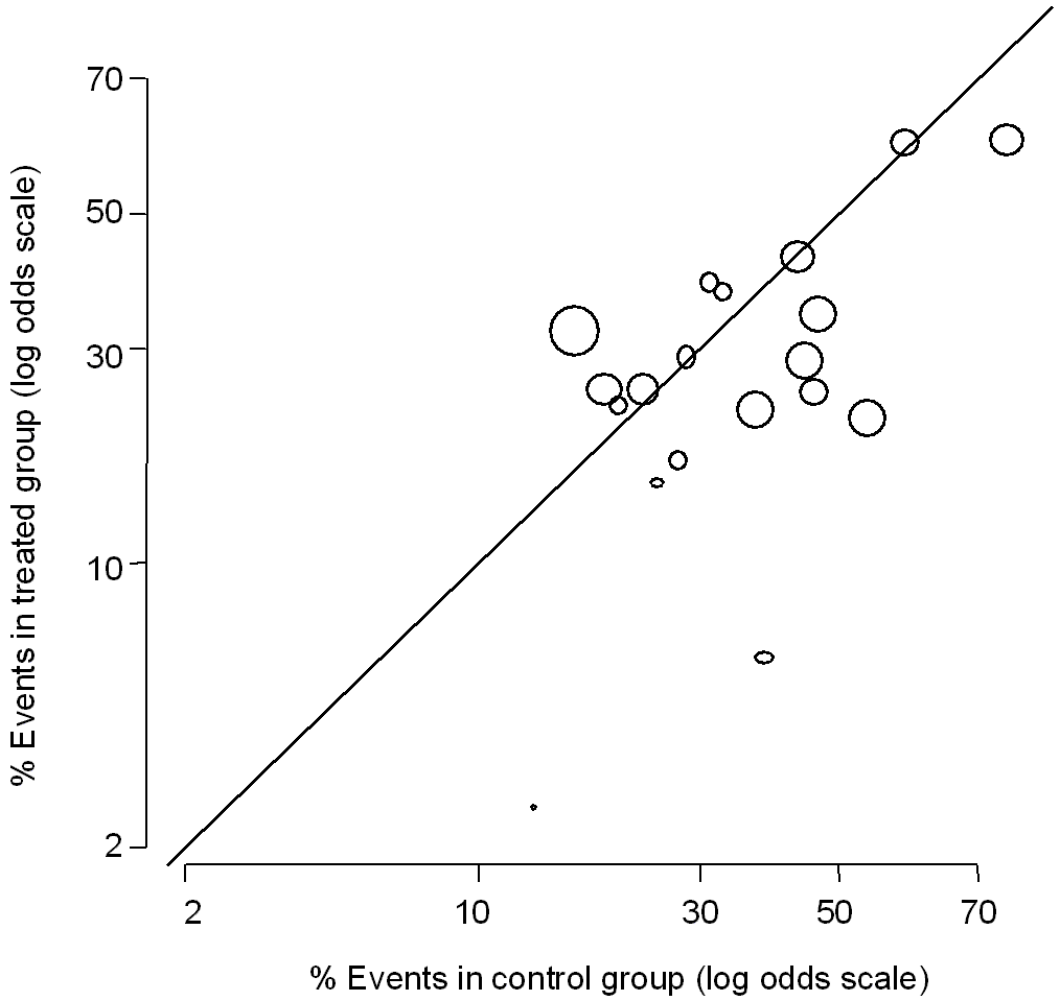
8a.8 Graphical displays particular to dichotomous outcome data

8a.8.1 L'Abbé plots

Results of multiple clinical trials with dichotomous outcomes may be represented in a L'Abbé plot, after a paper by L'Abbé and colleagues (L'Abbé 1987). This is a plot showing for each study the observed event rate in the experimental group plotted against observed event rate in the control group. L'Abbé plots may be used to view the range of event rates among the trials, to highlight excessive heterogeneity, and, on occasion, to indicate which treatment effect measure may be most consistent across trials. Naïve regression analyses based on L'Abbé plots are misleading, however, since they do not account for sampling error in both observed event rates (Sharp 1996).

L'Abbe plots may be drawn on the scale of the risk (the event rate), the log(risk) or the log(odds) (see Van Houwelingen 1993 for examples of the first and last). At present no advice is available on whether any is preferable in general. The first, however, is most likely to be interpretable by clinicians. An example appears in Figure 5.

Figure 5: L'Abbé plot of 19 trials of sclerotherapy (data from Sharp 1996)



Recommendations for L'Abbé plots

7.1. Where treatments are experimental and standard/control, the experimental event rate should be plotted on the vertical axis. When there is no such asymmetry it does not matter which way the plot is done.

7.2. A line indicating no treatment effect should be added.

7.3. Regression lines should not be added (unless they are derived using techniques that account for sampling error in both variables)

7.4. It may be useful to plot points at a size proportional to weight or trial size (preferably weight).

7.5. If the software permits, the graph should be square.

7.6. The raw data (information sufficient to create a 2´2 table from each trial) should be available somewhere in the review.

8a.8.2 Relating treatment effect to 'underlying risk'

A special case of meta-regression is to assess the dependence of treatment effect on control group event rate, on the assumption that the control group event rates reflect the underlying risks of participants in the studies. As Sharp et al. explain (Sharp 1996), such regressions may be highly misleading since they can be affected by regression to the mean. Techniques are available that overcome this problem (Sharp 2000). Simple scatter plots of treatment effect against control group event rate may be misleading, also due to regression to the mean. We recommend that such plots

are not presented unless the results of a suitable analysis of the relationship is obtained and superimposed on the plot.

Recommendations for relationship between treatment effect and underlying risk

8.1. Plots should follow recommendations for single variable meta-regression

8.2. The regression line from an analysis specifically designed for underlying risk meta-regression should be superimposed on the plot.

8.3. The raw data (information sufficient to create a 2´2 table from each trial) should be available somewhere in the review.

8a.9 Other graphical displays

In this section we outline two types of graph that have statistical merit but are less familiar to users of Cochrane Reviews, and two types of graph in common use but with unproven or poor statistical grounding. These types of graph are not encouraged as part of a Cochrane Review, and if used should be accompanied with a sound justification. We close with a brief mention of some other graphs that have been proposed for use within systematic reviews.

8a.9.1 Galbraith (radial) plots

Galbraith has described an alternative to the confidence interval plot for visualising results of studies and meta-analyses (Galbraith 1988, Galbraith1994). His graph has been enthusiastically received by statisticians (Whitehead 1991, Thompson 1993) but may be less readily interpreted by non-statisticians. The plot provides the basis of a simple graphical test for funnel plot asymmetry (Egger 1997). Galbraith plots facilitate examinations of heterogeneity, including detection of outliers.

A Galbraith plot is a plot of a standardized treatment effect (treatment effect divided by its standard error) against the reciprocal of the standard error. Imprecise estimates of effect lie near the origin, and precise estimates further away, giving the correct impression of being more informative. Vertical variation in points describes the extent of heterogeneity. The plot may be interpreted in terms of lines through the origin. Linear regression through the origin of the standardized treatment effects on their inverse standard errors yields a slope equal to the fixed effect meta-analysis estimate. A ‘radial’ scale (an arc of a circle) allows the determination of any slope, and hence provides details of the unstandardized effect estimates.

Egger et al’s test for funnel plot asymmetry is based on the linear regression (not confined to passing through the origin) of standardized treatment effects on their inverse standard errors. Statistical significance of the intercept provides a test for funnel plot asymmetry, since under ideal conditions the regression line should pass through the origin.

8a.9.2 Relationship between treatment effect and two or more covariates (meta-regression)

On occasion it may be of interest to investigate the relationship between treatment effect and two or more covariates. Illustration of such a relationship requires three or more dimensions. Lau et al. have described the use of response surfaces for the illustration of relationships with two covariates (Lau 1998). Response surface plots and 3-dimensional histograms/bar charts are not encouraged in Cochrane Reviews. Two dimensional scatter plots illustrating the relationships between treatment effect and each covariate, and between covariates, may be helpful.

8a.9.3 Survival curves

A standard representation of time-to-event outcomes from clinical trials is a Kaplan Meier curve. These illustrate the survival times of participants in the trial while acknowledging that some were not observed, so that appropriate comparison of the different treatment groups can be made. Kaplan Meier plots from individual trials are suitable for inclusion in Cochrane Reviews, though they may easily become too numerous.

Kaplan Meier plots for all pooled participants across trials in a meta-analysis have previously been presented in medical journals. This practice breaks with the principle of comparing like with like. For this reason, until further discussions have taken place the Statistical Methods Group is unable to recommend inclusion of such plots in Cochrane Reviews.

8a.9.4 Cumulative meta-analysis

Cumulative meta-analysis (Lau 1995) plots accumulations of studies: this suffers from a lack of independence of points, which could mislead a naïve reader (Antman 1992).

8a.9.5 Further graphical displays

Numerous other graphical displays can sometimes add useful insights to reports of systematic reviews. For example, sequential/prospective meta-analysis (Whitehead 1997, Pogue 1998) may be used to illustrate the accumulation of data with respect to some a priori desirable amount of information. Other suggestions for graphics relevant to meta-analyses include box plots (Light 1994, Petitti 1994), plots related to model diagnostics (Olkin 1995, Hardy 1998), illustrations of distributions (including prior and posterior distributions for Bayesian meta-analyses (Carlin 1992)) and plots to illustrate two-dimensional uncertainty (Thompson 1993, Hardy 1996). Finally, ‘odd-man-out meta-analysis’ (Walker 1988) is a proposal for illustrating summary confidence regions.

8a.10 Contributions

This appendix was prepared by Julian Higgins on behalf of the Cochrane Statistical Methods Group. The help of the following is particularly appreciated: Doug Altman, Deborah Ashby, Jon Deeks, Gordon Dooley, Diana Elbourne, Sally Hollis, Steff Lewis, Keith O’Rourke, Jonathan Sterne, Simon Thompson and members of the Cochrane Information Management System Group

8a.11 References

Alderson 2004. Alderson P, Green S, Higgins JPT, editors. Cochrane Reviewers’ Handbook 4.2.1 [updated December 2003]. In: The Cochrane Library, Issue 1, 2004. Chichester, UK: John Wiley & Sons, Ltd.

Antman 1992. Antman EM, Lau J, Kupelnick B, Mosteller F, and Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: Treatments for myocardial infarction. Journal of the American Medical Association 1992; 268: 240-248.

Arkin 1940. Arkin H and Colton RR. Graphs: How to Make and Use Them. New York: Harper and Brothers, 1940.

Bailey 1987. Bailey KR. Inter-study differences - how should they influence the interpretation and analysis of results. Statistics in Medicine 1987; 6: 351-360.

Carlin 1992. Carlin JB. Meta-analysis for 2 x 2 tables: a Bayesian approach. Statistics in Medicine 1992; 11: 141-158.

Cleveland 1994. Cleveland WS. The Elements of Graphing Data. Summit, New Jersey: Hobart Press, 1994.

Deeks 2001. Deeks JJ, Altman DG and Bradburn MJ. Statistical methods for examining heterogeneity and combining results from several studies in meta-analysis. In: Egger M, Davey Smith G and Altman DG (eds). Systematic Reviews in Health Care: Meta-analysis in Context. London: BMJ Books, 2001.

Egger 1997a. Egger M, Davey Smith G, and Phillips AN. Meta-analysis: principles and procedures. British Medical Journal 1997; 315: 1533-1537.

Egger 1997b. Egger M, Smith GD, Schneider M, and Minder C. Bias in meta-analysis detected by a simple, graphical test. British Medical Journal 1997; 315: 629-634.

Galbraith 1988. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. Statistics in Medicine 1988; 7: 889-894.

Galbraith 1994. Galbraith RF. Some applications of radial plots. Journal of the American Statistical Association 1994; 89: 1232-1242.

Hardy 1996. Hardy RJ and Thompson SG. A likelihood approach to meta-analysis with random effects. Statistics in Medicine 1996; 15: 619-629.

Hardy 1998. Hardy RJ and Thompson SG. Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine 1998; 17: 841-856.

Hooper 2001. Hooper L, Summerbell CD, Higgins JPT, Thompson RL, Capps NE, Davey Smith G, Riemersma RA, and Ebrahim S. Dietary fat intake and prevention of cardiovascular disease: systematic review. British Medical Journal 2001; 322: 757-763.

Johnson 1999. Johnson ES, Lanes SF, Wentworth III CE, Satterfield MH, Abebe BL, and Dicker LW. A metaregression analysis of the dose-response effect of aspirin on stroke. Archives of Internal Medicine 1999; 159: 1248-1253.

L'Abbe 1987. L'Abbe KA, Detsky AS, and O'Rourke K. Meta-analysis in clinical research. Annals of Internal Medicine 1987; 107: 224-233.

Lau 1995. Lau J, Schmid CH, and Chalmers TC. Cumulative meta-analysis of clinical trials: Builds evidence for exemplary medical care. Journal of Clinical Epidemiology 1995; 48: 45-57.

Lau 1998. Lau J, Ioannidis JP, and Schmid CH. Summing up evidence: one answer is not always enough. The Lancet 1998; 351: 123-127.

Lewis 2001. Lewis S and Clarke M. Forest plots: trying to see the wood and the trees. British Medical Journal 2001; 322: 1479-1480.

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

Light 1994. Light RJ, Singer JD and Willett JB. The visual presentation and interpretation of meta-analyses. In: Cooper H and Hedges LV (eds). The Handbook of Research Synthesis. New York: Russell Sage, 1994.

Olkin 1995. Olkin I. Statistical and theoretical considerations in meta-analysis. Journal of Clinical Epidemiology 1995; 48: 133-146.

Petitti 1994 Petitti DB. Meta-Analysis, Decision Analysis and Cost-Effectiveness Analysis. New York: Oxford University Press, 1994.

Pogue 1998. Pogue J and Yusuf S. Overcoming the limitations of current meta-anlysis of randomised controlled trials. Lancet 1998; 351: 47-52.

Schmid 1983. Schmid CF. Statistical Graphics: Design Principles and Practices. New York: John Wiley and Sons, 1983.

Sharp 1996. Sharp SJ, Thompson SG, and Altman DG. The relation between treatment benefit and underlying risk in metaanalysis. British Medical Journal 1996; 313: 735-738

Sharp 2000. Sharp SJ and Thompson SG. Analysing the relationship between treatment benefit and underlying risk in meta-analysis: comparison and development of approaches. Statistics in Medicine 2000; 19: 3251-3274.

Simmonds 1980. Simmonds D (ed). Charts and Graphs: Guidelines for the Visual Presentation of Statistical Data in the Life Sciences. Lancaster, UK: MTP Press, 1980.

Sterne 2001a. Sterne JAC, Egger M and Davey Smith G. Investigating and dealing with publication bias and other biases. In: Egger M, Davey Smith G and Altman DG (eds). Systematic Reviews in Health Care: Meta-analysis in Context. London: BMJ Books, 2001.

Sterne 2001b. Sterne JAC and Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. Journal of Clinical Epidemiology 2001; 54: 1046-1055.

Sutton 1998. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, and Song F. Systematic reviews of trials and other studies. Health Technology Assessment 1998; 2.

Thompson 1993. Thompson SG. Controversies in meta-analysis: the case of the trials of serum cholesterol reduction. Statistical Methods in Medical Research 1993; 2: 173-192.

Thompson 1994. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. British Medical Journal 1994; 309: 1351-1355.

Thompson 2001. Thompson RL, Summerbell CD, Hooper L, Higgins JPT, Little PS, Talbot D, and Ebrahim S. Dietary advice given by a dietitian versus other health professional or self-help resources to reduce blood cholesterol. Cochrane Database of Systematic Reviews. Oxford: Update Software 2001 (Issue 4).

Thompson 2002. Thompson SG and Higgins JPT. How should meta-regression analyses be undertaken and interpreted Statistics in Medicine 2002;21:1559-73.

van Houwelingen 1993. Van Houwelingen HC, Zwinderman KH, and Stijnen T. A bivariate approach to meta-analysis. Statistics in Medicine 1993; 12: 2273-2284.

Walker 1988. Walker AM, Martin-Moreno JM, and Artalejo FR. Odd man out: a graphical approach to meta-analysis. American Journal of Public Health 1988; 78: 961-966.

Whitehead 1991. Whitehead A and Whitehead J. A general parametric approach to the meta-analysis of randomised clinical trials. Statistics in Medicine 1991; 10: 1665-1677.

Whitehead 1997. Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. Statistics in Medicine 1997; 16: 2901-2913.

APPENDIX 8b. Calculating the number needed to treat (NNT)

NNTs are a useful way to re-express the results of a study but some caution is needed when they are used in reviews. NNTs are specific to a particular length of follow-up since they are based on the number of people who will benefit within a certain period of time who otherwise would not benefit. Systematic reviews tend to combine trials of varying follow-up periods, which could make an NNT difficult to interpret (Smeeth 1999). NNTs should only be calculated when the follow-up periods are similar.

When summarising results, the 'control event rate' (the rate of events in the control group) can be substituted for the 'patient expected event rate' (the baseline risk). In practice, individual patients' expected event rate might differ importantly from the control event rate in the studies in a review.

The following abbreviations are used in this appendix:

CER = control event rate

EER = experimental event rate

PEER = patient expected event rate

NNT = Number needed to treat

RD = risk difference (or absolute risk reduction, ARR)

RR = relative risk

RRR = relative risk reduction

OR = odds ratio

Then:

RD = CER – EER

RR = EER/CER

RRR = RD/CER = 1 – RR

The RRR can be calculated from the OR using

$$RRR = CER - \frac{OR \times CER/(1 + CER)}{[OR \times CER/(1 + CER)]}$$

The NNT can then be calculated with either

NNT = 1/RD

NNT = 1/(CER - RR x CER)

NNT = 1/(RRR x CER)

If the CER is very small, say less than 5%, the OR is approximately equal to the RR and the RRR is approximately equal to (1 – OR). However, as the CER (or PEER) increases, the difference between the OR and the RR increases.

If the average CER across studies is used in the above formulae, the NNT will be for the average baseline risk observed across the included studies. Since the PEER (baseline risk) often varies across studies and is likely to vary across patient groups, it is general important to specify the baseline risk for which an NNT is reported and to report NNTs for a range of PEERs. For example, the range of CERs in the included studies can be used, giving NNTs based on the lowest, the average and the highest of these. However, this assumes that the RRR is the same for different baseline risks. Although this assumption is often correct, it is not always (Sharp 1996, Ioannidis 1997, Smith 1997, Thompson 1997, Smeeth 1999).

Confidence limits for NNTs should be calculated by using the upper and lower confidence limits for the summary statistic that is used to calculate the NNT (RR, OR or RD). For further discussion about NNTs and their calculation see (Sackett 1996, Senn 1998, Altman 1998).

8b.1 References

Altman 1998. Altman DG. Confidence intervals for the number needed to treat. BMJ. 1998; 317:1309-12.

Ioannidis 1997. Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol 1997; 50: 1089-98.

Sackett 1996. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! Evidence Based Medicine, 1996 Sept-Oct; 1:164.

Senn 1998. Senn S, Walter S, Olkin I, Altman D, Deeks J, Sackett DL. Odds ratios revisited. Evidence-Based Medicine. 1998 May-June;71.

Smeeth 1999. Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses – sometimes informative, usually misleading. Brit Med J 1999;318:1548-51.

Sharp 1996. Sharp SJ, Thompson SG, Altman DG. The relation between treatment benefit and underlying risk in meta-analysis. BMJ 1996;313:735-8.

Smith 1997. Smith GD, Egger M, Phillips AN. Meta-analysis. Beyond the grand mean? BMJ 1997;315: 1610-4.

Thompson 1997. Thompson SG, Smith TC, Sharp SJ. Investigating underlying risk as a source of heterogeneity in meta-analysis. Stat Med 1997; 16: 2741-58.

APPENDIX 9. Incorporating economic evaluation into the Cochrane review process

The Cochrane Collaboration's main role of 'preparing, maintaining and making accessible reviews of the effects of healthcare' is motivated by an underlying aim to help people make decisions about healthcare. However, in the face of limited resources, decision makers need to consider further evidence when deciding how to act on the evidence from Cochrane reviews. Nearly every healthcare decision has an impact, not only on health and social welfare, but also on the use of resources. Therefore, to make the best decisions about alternative interventions, information is needed on resource use and costs as well as health effects.

The process of incorporating economic evaluation into Cochrane Reviews is not straight forward. As with many areas of scientific inquiry, the methodology is still developing. A particular challenge in the context of Cochrane Reviews is ensuring that economic information and analyses contained in reviews is relevant to people working in widely varying circumstances. For those who are considering addressing economic questions as part of their review, or along side of a Review, advice can be found in the module for the Cochrane Health Economics Methods Group in *The Cochrane Library*.

APPENDIX 11a. Practical Methodology of meta-analyses using updated individual patient data

11a.1 Front page

PRACTICAL METHODOLOGY OF META-ANALYSES (OVERVIEWS) USING UPDATED INDIVIDUAL PATIENT DATA

LESLEY A. STEWART

MRC Cancer Trials Office, 5 Shaftesbury Road, Cambridge CB2 2BW, U.K.

AND

MICHAEL J. CLARKE

University of Oxford, Clinical Trial Service Unit and ICRF Cancer Studies, Radcliffe Infirmary, Oxford OX2 6HE, U.K.

on behalf of the

COCHRANE WORKING GROUP ON META-ANALYSIS USING INDIVIDUAL PATIENT DATA

(Originally published in Statistics in Medicine, Vol. 14, 2057-2079, 1995)

11a.2 Further information

For further information on the Cochrane Working Group on meta-analysis using individual patient data, please contact one of the authors:

Lesley A Stewart
MRC Cancer Trials Office
5 Shaftesbury Road
Cambridge CB2 2BW
UK
Phone: +44-1223-311110
Fax: +44-1865-58817
e-mail: LS @ cto.mrc.ac.uk

Michael J Clarke
UK Cochrane Centre
Summertown Pavilion
Middle Way
Oxford OX2 7LG
United Kingdom
Phone: +44-1865-516300
Fax: +44-1865-516311
e-mail: mclarke@cochrane.co.uk

11a.3 Workshop participation

Doug Altman, Colin Baigent, Marc Buyse, Iain Chalmers, Mike Clarke, Rory Collins, Carl Counsell, Jack Cuzick, Rob Edwards, Tricia Elphinstone, Vaughan Evans, Richard Gray, Liz Greaves, Francois Gueyffier, Heather Halls, Rob Henderson, Jini Hetherington, Sally Hunt, Peter Langhorne, Carol Lefebvre, David Machin, Silvia Marsoni, Veronique Mosseri, Lennarth Nyström, Mandy Ogier, Andy Oxman, Max Parmar, Richard Peto, Jean-Pierre Pignon, Sue Richards, Carmen Ruiz, Paul Seed, Michael Sextro, Lena Specht, Sally Stenning, Lesley Stewart, Annet te Velde, Jayne Tierney, Harm van Tinteren, Valter Torri, Paul Weston, Keith Wheatley, Chris Williams.

11a.4 Summary

Meta-analyses using updated individual patient data may provide the most reliable means of combining data from similar randomised controlled trials and the benefits of this approach to systematic review are described. Guidance, based on the experience of several groups who have undertaken such projects is given. This includes practical advice on initiating and maintaining collaboration, the time and resource required to undertake these usually international projects and methods of data checking and validation. Example proforma are included.

11a.5 Introduction

Systematic reviews using meta-analysis to combine the results of related randomised controlled trials are increasingly common, and the number of associated publications has mushroomed. Although there is a burgeoning literature on the statistical methods of meta-analysis, less has been published on the practical methods of carrying out such projects. These can include calculations based solely on information presented in a few published papers, more detailed analysis of aggregate data supplied by individual trialists, and time-to-event analysis of thoroughly checked and updated individual patient data. The last of these has been described as the 'yardstick' against which all systematic reviews should be measured (1), and current limited empirical evidence shows that meta-analyses which rely solely on data extracted from published reports can give estimates of treatment effects, and of their significance, which are not confirmed when all of the relevant evidence is analysed (2, 3, 4). Given that the central collection, checking and analysis of individual patient data from all relevant trials can require a considerable amount of time, personnel and financial resource, further research is needed to determine when it is most appropriate to adopt this approach and what the most appropriate alternatives are if sufficient resources are not available. Irrespective of this, the additional benefits of meta-analyses based on individual patient data (IPD) when compared with meta-analyses based on published aggregate date include the ability to:

- Undertake survival and other time-to-event analyses
- Undertake subgroup analyses for important hypotheses about differences in effect
- Carry out detailed data checking and ensure the quality of randomisation and follow-up
- Ensure the appropriateness of analyses
- Update follow-up information

Further, as IPD meta-analyses require the collaboration of the investigators who conducted the trials, other benefits (which may also be found if trialists are approached for aggregate data) may include:

- More complete identification of relevant trials

- Better compliance with providing missing data
- More balanced interpretation of the results
- Wider endorsement and dissemination of the results
- Better clarification of further research
- Collaboration on further research

This paper provides guidance on the conduct of IPD meta-analyses, which aim to collect data on each randomised patient entered in all randomised trials addressing a particular question. The patient data are checked, collated and analysed centrally by a secretariat. Subsequent publication is generally made by the collaborative group of trialists, often following a meeting of this group at which the results and their implications are discussed. Until now, almost no information on either the techniques or the resources needed for such a project has been readily available. Thus each of the groups who have undertaken them has generally had to develop their own means of data collection, checking and analysis.

In the hope that this situation could be improved, a workshop (under the auspices of the Cochrane Collaboration) was convened in April 1994 to discuss the practicalities of meta-analyses based on individual patient data. This was attended by nearly 40 participants (Appendix A), all of whom had been involved in the planning or conduct of this form of meta-analysis. The aim was to discuss all practical aspects of such projects; to identify areas of agreement and disagreement on the methods used; and to prepare published guidance available to anyone contemplating using this technique in a systematic review. Participants did not discuss whether or not meta-analyses using individual patient data are indeed a 'gold standard' or statistical methodology.

11a.5 Running a meta-analysis based on individual patient data

The steps involved in a meta-analysis of individual patient data are shown in figure 1 along with some very approximate guidance on the time required for these. The majority of effort is required to plan, initiate, set up and manage the study and, although much has been written about the statistical methodology of meta-analysis, this can often represent the least time consuming and difficult aspect of the project. Nurturing collaboration and careful checking of incoming data generally consume much more time and resource, since the ultimate aim is to obtain accurate, up to date and complete data from all patients in all relevant randomised trials.

Figure 1. Stages of an Individual Patient Based Systematic Review

NB: All estimates of time are necessarily very approximate and will depend on the size of the meta-analysis and the complexity of the data requested

(1) Development

Identify need for IPD meta-analysis

Devise questions

Identify trials (continues throughout project)

Refine questions

Meta-analysis of published data (if appropriate)

Write Protocol

Initial contact with trialists

Typically requires approximately 3-6 months minimum (3-4 person months minimum effort)

(2) Data Collection and Checking

Assess feasibility

Set up database

Request data

Check data

Analyse trials individually

Finalise data

Requires approximately one year (15 person months for 50 trials, 4-5 person months for 5 trials)

(3) Analysis and Dissemination of Results

Analyse data

Present results to trialists

Discuss results and implications with trialists

Draft manuscript

Requires approximately 6-9 months (10 -12 person months for 50 trials, 5-6 person months for 5 trials)

(4) Future Projects

Future updates

New projects

- extend scope of meta-analysis
- initiate new trials

The total time required for the meta-analysis is approximately 2 - 3 years (approximately 30 person months for a meta-analysis of 50 trials and 15 person months for 5 trials).

11a.6 Resource requirements

It is perhaps not generally appreciated just how much time and effort is involved in performing an IPD meta-analysis. It is not something to be undertaken lightly, and since a variety of clinical, scientific, statistical, computing and data management skills are required, it is generally not something to be undertaken by a single individual. Of necessity, projects usually take a few years from initiation to first publication. Although some of this time can be saved by involving more personnel, the project duration will be constrained by the time taken to secure the full involvement of the collaborating trialists. This collaboration is the main way of ensuring that the data to be analysed are as complete, accurate and reliable as possible.

Financial

Based on estimates provided by those attending the workshop, the average cost of running an IPD meta-analysis was approximately £1,000 per trial or £5-£10 per patient (£ Sterling, 1994), whichever was the less. However, these estimates, which did not include the costs associated with a Collaborative Group Meeting, were very approximate and retrospective and varied greatly depending on the size and complexity of the project. In addition, most estimates did not include the hidden costs associated with administration. Interestingly, those meta-analyses funded by direct grants, where presumably a more detailed record of costs was required, were considerably more expensive. Previous projects have been financed by both core and grant-based funding. The first cycle of project initiation, data collection and analysis is well suited to one-off grant applications because of its structure and timescale, although many IPD meta-analyses will require subsequent updating which may at first seem less attractive to some funders. Funding could be sought from a variety of sources: Government bodies, research organisations, charities and industry.

Staff

Most of the estimated costs were associated with staff, typically representing around 80% of the total budget. As discussed above, a range of skills are required and the involvement of the various personnel will vary over time. It is therefore usual for some groups co-ordinating IPD meta-analyses to be simultaneously involved in several projects, scheduling them so that the workload of the clinical, scientific, statistical, computing, data management, administrative and secretarial staff is evenly distributed.

Time

Figure 1 includes very approximate estimates of the minimum time required to complete the various stages of a meta-analysis, both on an absolute time scale and in terms of person months. It should be noted, though, that the actual time taken may vary considerably depending on the circumstances of each project. In most circumstances it is unlikely that an IPD meta-analysis could reach first publication in much less than three years.

11a.7 Planning the meta-analysis

As with a clinical trial, a good deal of planning and organisation is required before a meta-analysis can be launched and trialists are asked to provide data. After the identification of a suitable question, the first step is to identify all relevant randomised trials and to plan the conduct of the meta-analysis. In most cases this will involve developing a protocol or written plan of the proposed investigation. A good deal of resource is involved in this pre-data collection planning stage, which may take several months. There is therefore a potential problem in that several groups may independently embark upon the same or similar projects, representing both a duplication of effort and an annoyance to the trialist who receives multiple requests for the same data. One way to help avoid this is through the prospective registration of these meta-analyses with the Cochrane Collaboration, in the same way that systematic reviews using other techniques can be registered.

11a.7.1 Establishing a Secretariat

At the earliest stages of the meta-analysis a secretariat to co-ordinate the project should be established. It is likely that this will consist of the scientific, statistical and data management staff who will do most of the work on the project, and also appropriate clinical experts. A larger Steering Group may also be formed to advise the secretariat on strategic issues and analyses. This is likely to be made up of members of the secretariat, trialists and independent experts.

11a.7.2 Methods of Identifying Trials

It is of the utmost importance that as high a proportion as possible of all relevant trials are identified, regardless of their results or publication status. Any trials that are missing should not be too numerous or unrepresentative to affect the results of the meta-analysis in any important way. This is true of any systematic review, irrespective of the analytical methods to be adopted, and searching for trials should continue throughout the duration of the project.

The first step towards identifying trials is usually to perform a computerised bibliographic search. However, such searches may miss a significant proportion of published trials. For example, it has been shown that electronic searching for randomised clinical trials using the US National Library of Medicine's database MEDLINE, might yield only around half of the relevant studies that are actually contained in the database (5). Further, MEDLINE indexes only 3,700 out of around 16,000 medical journals published worldwide (5). The coding of articles within MEDLINE is currently being revised to improve the retrieval of future RCTs and the Cochrane Collaboration is working with the National Library of Medicine (NLM) in the retrospective tagging of all previously published randomised trials. Other databases, for example, CancerLit, Current Contents, Excerpta Medica, The Index of Scientific and Technical Proceedings, Dissertation Abstracts and the Index to UK Theses, may be useful additions or alternatives to MEDLINE, but further research is required to determine which are most efficient in the various areas of medicine.

In order to make full use of the current computerised databases, it is important that efficient search strategies are used. An inexperienced searcher should seek as much help as possible. Optimal strategies for searching MEDLINE are currently under development (5) and these should be adopted as part of any systematic review. The latest version is shown in Appendix B. This strategy does not include subject specific searching so that individual searchers will need to add further steps, for example, adding terms such as the disease and therapy in question.

At present, problems will remain even with the best computer search strategy. Some relevant articles in the databases will be missed because of lack of clarity in the published reports or indexing errors, and the majority of medical journals are not covered by any literature database. Until all published randomised trials are accessible through MEDLINE, it is essential that electronic searches are supplemented by some hand-searching. This will need to include those journals that are most likely to contain relevant reports which cannot be identified in the existing databases, and also those meeting abstracts which are not available in any electronic form.

This aspect of any meta-analysis can be both time-consuming and labour intensive. Even a refined literature search strategy is likely to yield many more articles than will eventually prove relevant to the meta-analysis. A fair number of the unnecessary articles will have to be obtained as full papers in order to determine whether or not they are relevant. In addition, the thorough handsearching of journals and meeting abstract books requires a substantial amount of care, time and effort. The Cochrane Collaboration is attempting to coordinate such searching and it would be worthwhile for anyone planning to do such a search to communicate first with the Collaboration to avoid duplication of effort.

An additional problem is that trials with positive results are more likely to be published than those with negative or inconclusive results (6, 7, 8, 9), thus skewing the published literature in favour of the positive. It is therefore extremely important that, whenever possible, unpublished trials are sought and included in meta-analyses (especially where the results of a trial might have influenced the decision on whether it would be published). Although data from unpublished trials have not

been subject to peer review, obtaining the trial protocol and individual patient data enables thorough checking both of the data supplied and the trial design, allowing, in fact, a much more detailed review than is generally possible prior to the publication of a trial. Moreover, even if a trial has been published in a prestigious journal, this cannot be taken as guarantee of the quality of the actual data. All trials, both published and unpublished, should be subject to the same degree of careful checking prior to inclusion in an IPD meta-analysis.

The main reason for non-publication of a trial is failure by the authors to prepare a report (6, 7, 8), and these trials are usually small single institution studies. Finding such trials can therefore be difficult. Trial registers, which prospectively register trials at inception, are the best solution to this problem (10) and the conduct of all systematic reviews should be much simplified when the use of such registers becomes widespread (11). However, while it is to be hoped that increasing numbers of new trials will be registered, many existing trials will still not be included and the identification of these trials will continue to be a major part of most meta-analyses. As the collaborative group is likely to consist of international experts with a good knowledge of potentially relevant or otherwise unidentified trials, the direct contact with trialists that is an integral part of a meta-analysis based on individual patient data can be a rich source of information. In addition, the circulation of a list of all identified trials at appropriate clinical meetings may bring to light trials, as well as trialists, previously unknown to the secretariat and collaborative group. Other potential sources of information include pharmaceutical companies and regulatory authorities.

11a.7.3 Developing a written plan or protocol

As with any formal research, some form of written plan or protocol should be produced for the meta-analysis. Examples of formats that have been used successfully in previous projects include a two page summary sheet and a longer document similar to the protocol for a clinical trial.

Table I. Possible items to include in a written plan or protocol for an individual patient data based meta-analysis

RATIONALE

Underlying biology

Review of trials

Preliminary meta-analysis

OBJECTIVES

Inclusion or eligibility criteria

Search strategies

Data to be collected

Brief description of data checking procedures

Main analyses to be performed

Publication policy

Suggested timetable for the meta-analysis

Provisional list of trials to be included

Table 1 shows some of the items that might be considered for inclusion in such a document. As a minimum, trialists being asked to participate in the project should be provided with some guidance on the proposed analyses along with a statement on publication policy and the confidentiality of data. The most difficult item is perhaps the inclusion of a meta-analysis based on data other than individual patient data which may have been performed as part of the planning stage of the IPD meta-analysis, as this may give the impression to some potential collaborators that the review has already been done and that they need not go to the trouble of supplying individual patient data. It is important, therefore, that if a preliminary meta-analysis is included it is accompanied by a suitable explanation of why it is not felt to be adequate and why individual patient data are being sought. For example, if the meta-analysis simply relied on data that could be easily abstracted from publications which had been identified by an inadequate MEDLINE search, it should be noted that such an analysis might be biased by a failure to include trials whose data could not be abstracted from the identified publications, published trials which were not found in the MEDLINE search, and trials which had not been published. In such a case it should be noted that, as well as helping to rectify these potential problems, collecting IPD allows the published data to be updated. The reasons for requesting individual, rather than aggregate, data should also be given. If the IPD meta-analysis has been preceded by a thorough meta-analyses using aggregate data then just this information, to indicate why individual patient data was now felt to be necessary, would be required (12).

Developing a written plan or protocol makes setting up a meta-analysis more rigorous by helping to identify problems and clarify issues early in the project. Specifying inclusion criteria means that trials can be evaluated for suitability at an early stage. Although there may be a temptation to request data from all trials at the outset of the meta-analysis, a more measured approach makes it less likely that trials will have to be withdrawn or excluded after the trialists have started to prepare and provide data. Time spent at this stage more than makes up for itself later, although it does mean that initiating collaboration may be delayed.

11a.8 Initiating collaboration

Having decided on the therapeutic questions to be addressed, identified the relevant trials and done the appropriate planning, the trialists need to be contacted and persuaded to participate. Generally this will involve inviting them to join the collaborative group and to provide the data required for the analysis. Occasionally it will also involve seeking the advice of the trialists on the data to be collected. Establishing collaboration can take some time, especially if a trial was done many years ago and the appropriate personnel have moved since their trial was published or registered. In this case it pays to be persistent and to write to all authors.

In the initial correspondence the secretariat should emphasise the collaborative nature of the project and state that publication of the meta-analysis results will be made in the name of the collaborative group and stress that any data supplied will be held securely and treated as confidential. It is also useful to reassure trialists that data collection will be as simple and flexible as possible. Including a written plan or protocol in this initial mailing may help in explaining the project to trialists, and also demonstrate the seriousness with which it is being tackled. Enclosing a reply form may help in getting a prompt reply containing the basic trial information and ascertaining what data items the trialists would be able to provide. A trial protocol and other documentation including information on the method of treatment assignment (including details on stratification factors and block size) should also be requested at this stage. An example of an initial form inviting collaboration is given in Appendix C. However, it may take several letters or telephone calls and even, in a few extreme cases, meetings with the trialists to secure their participation in the meta-analysis.

11a.9 Data collection

Once a decision has been taken that the meta-analysis is indeed feasible, what is often the most labour intensive aspect of the project, both for the secretariat and the trialists supplying data, can begin. On average a minimum of one or two person weeks of secretariat time is required to collect the data, convert it to a standard format, check, query and rectify the inevitable problems for any one trial. However, this may vary considerably depending on the complexity of the data collected. Thus, depending on the size of the meta-analysis, completing this stage can take several months. Fewer trials will, of course, mean less work at this stage and increasing the number of staff working on the project can speed the checking process. However, the absolute amount of time taken will ultimately be determined by how long it takes trialists to provide the data and respond to queries. In most instances, therefore, it is unlikely that this stage can be completed in less than a year.

11a.9.1 Deciding which data items to collect

The minimum data that can be collected for an IPD meta-analysis are the patient identifier, treatment allocated and outcome(s), together with the date of randomisation and date of outcome if time to event is to be calculated. It is, however, often important to collect additional baseline variables, even when subgroup analyses are not planned, because these data are extremely useful in checking the integrity of the randomisation process. The collection of additional outcome data might also be advisable.

The decision on which data items to collect can be made by the secretariat, steering group or by the collaborative group. Obviously this last option will be time consuming and may lead to potential disagreements if suggestions are conflicting or if some are rejected. Whichever approach is adopted, it is essential that clinical as well as statistical input is sought. The final list of suggested variables should be sent to trialists early in the project to check that each variable will be available from a large enough proportion of trials to justify its request and collection.

11a.9.2 Data Collection

Specifying the desired format for data, suggesting codes where appropriate and providing data collection forms may help trialists. However, it is important that trialists should be allowed to supply data in whatever way is most convenient to them, whereupon the secretariat take responsibility for converting the data to the required format. In such instances, it is very important that there is a clear understanding between the secretariat and trialists as to the content of their non-standard data. At this stage it may be useful to identify a single individual (generally the person responsible for preparing the data) to whom all queries can be addressed, as this can simplify and speed the process considerably. Examples of forms and formats for data that have been used in the past are given in Appendix D1, D2, D3 and D4.

11a.9.3 Unavailable data

It must be appreciated that provision of data may entail considerable work for the trialist and so good communication is essential both to persuade them of the worth of the project and to explain what is required of them. Every effort should be made to reduce the burden on the trialist or data centre providing the information. On initial contact, some trialists may report that the data from their study are not available. Although in instances where data have been destroyed or lost, the trial may not be recoverable, it is often worth pursuing negative replies in case an alternative source of data can be found. For example, other people within a trial group may be more willing or able to supply the data. More usually the problem is one of insufficient resource, so that offers of assistance (usually in the form of sending someone to retrieve the data) are often effective. An invitation to the collaborators' meeting has often acted as an incentive to collaborate.

The aim of the meta-analysis should be to obtain individual data from all randomised patients in all relevant trials. If, despite all efforts to secure collaboration, data from one or more trials are not available, the question of how to deal with this arises. When a large proportion of the total randomised evidence (perhaps 90-95%) has been collected, the missing data may be considered unlikely to alter importantly the meta-analysis results. Nonetheless the unavailability of trials should be made clear in the published report of any meta-analysis.

If individual patient data are not available, aggregate data provided by trialists or data extracted from publications could be used. However, it is not clear whether or not the use of data extracted from published reports is desirable, given the potential problems with such data compared to data (aggregate or individual) supplied directly by the trialist. In addition, an explanation of why this was deemed acceptable for some trials would have to be given to those trialists who had put a great deal of effort into supplying individual patient data. The use of published data might therefore discourage some trialists from providing any data. Where a trialist is unable to supply individual patient data but can provide aggregate data, this would be more acceptable than published data alone, but, again, such data will preclude the specific advantages of individual patient data and this should be noted. However, completely excluding trials from the meta-analysis because individual patient data were not obtainable might cause problems through the omission of randomised evidence. Whenever the IPD meta-analysis is supplemented with trial results that are not based on the provision of individual patient data, this should be made clear. One option might be to conduct sensitivity analyses comparing a purely individual patient data based meta-analysis with one that incorporates whatever data are available on all relevant trials.

11a.9.4 Data Checking

The main aims of data checking procedures should be to ensure the accuracy of data, integrity of randomisation and completeness of follow up. For any one trial, it is important that the results of all the data checks should be considered together to build up an overall picture of that trial and any associated problems. Where there are concerns, these should be brought to the attention of the trialist and sympathetic attempts made to resolve them. This can often be done by letter or phone but may, occasionally, involve a visit to the trialist to help clarify and if necessary to rectify matters. The vast majority of cases will be resolved satisfactorily - often by the insertion of data that were not supplied initially. Although errors in data are common, having seen the patient data from hundreds of trials, the experience of the groups represented at the workshop is that fraud is very rare.

11a.9.5 Checking data accuracy

All data supplied should be subject to the sort of range and consistency checks that would be used in a prospective trial. This should be irrespective of whether data were supplied electronically or had to be entered manually into the meta-analysis database (when it is vitally important to check the accuracy of data input). Any missing data, obvious errors, inconsistencies between variables or extreme values should be queried and rectified as necessary by the trialist. If details of the trial have been published these also should be checked against the raw data and any inconsistencies queried. All of the changes made to the data originally supplied by the trialists, and the reasons for these changes, should be recorded.

11a.9.6 Checking the integrity of randomisation and follow up procedures

It is very important that the analysis should be based on the 'intention-to-treat' principle and therefore that data should be collected, and analyses based, on **all** randomised patients. Any

randomised patients that have been excluded from the trial should, wherever possible, be reintroduced to the analyses.

As part of the checking process prognostic variables should be checked for balance across treatment arms. It is, however, important to remember that imbalances may occur by chance alone especially for non-stratified variables and when trials are small. Other checks that can be done include looking at the weekday of randomisation. For example, in the UK we would expect very few non-acute randomisations at the weekend (although, in studies from other countries it is important to appreciate cultural differences in working patterns). Similarly, randomisations in trials of acute disease would be expected to spread throughout the week. A visual display of the chronological sequence of randomisations can be illuminating. For example, figure 2, which is included with the trialist's permission, shows such a curve from an unpublished trial of radiotherapy versus chemotherapy in multiple myeloma. In this trial the radiotherapy equipment was unavailable for six months during the trial but patients continued to enter the chemotherapy arm. It was only when the individual patient data were provided for a meta-analysis that this problem was brought to the attention of the trialist who agreed that the appropriate solution was to exclude this small number of non-randomised chemotherapy patients from the analysis. Similarly, looking at chronological accrual may reveal a period at the beginning or end of a trial when full randomisation was not taking place.

11a.9.7 Follow up

Where survival (or other time dependent variable) is the primary outcome it may be important that trial follow up is as up to date as possible since an increased follow-up may see a reduction in the treatment effect if the survival curves are converging (2, 13) or an increased treatment effect if the curves are diverging (14). Thus, where appropriate, data should be checked to ensure that follow up is up to date and to ensure that it is balanced across treatment arms. Balance in follow up can be checked by selecting all patients outcome-free and using the date of censoring as the event to carry out a 'reverse Kaplan-Meier' analysis producing censoring curves which should be the same for all arms of the trial. Any imbalance should be brought to the attention of the trialist and updated information should be sought. However, the trialist might not be able to provide updated follow up on all their patients. In such cases it may be possible for the secretariat to take responsibility for obtaining the additional follow up. For example, if death is a primary outcome, mortality information might be available from national death registers, provided that sufficient information is available to identify the patient. Some sources of this information are shown in Appendix E.

However, not all countries run such schemes and tracing the fate of patients especially those from older trials is not necessarily straightforward (16). In addition the cause of death information available from these sources might not be sufficiently accurate to use for analysis of cause-specific mortality (in those relatively few cases where such analyses are done as a supplement to the more usual analyses of death by all causes).

Figure 2. Entry of patients to randomized trial showing accrual of patients to chemotherapy (and radiotherapy) treatment group.

Not currently available

11a.9.8 Analysis of individual trials

Trials should be analysed individually and the trialists should be sent a copy of any such analyses as well as a printout of their data as included in the meta-analysis database. This allows

verification and also provides the trialist with an updated analysis of their own study which they may find useful for other purposes including further reports of their trial.

11a.9.9 What to do if a trial cannot be used

If a trial fails the checking procedures and the responsible trialist is unable to rectify the data or to explain the observed anomalies, the question arises of what to do next. Ultimately the decision on whether or not a particular aspect of a trial indicates a serious bias is a subjective one and the best solution may be to bring the problems to the attention of the trialist, and then to make a joint decision on whether to include or exclude it from the meta-analysis. If it is decided that a trial has to be excluded, this should be reported when the results of the meta-analysis are published. This is best done sympathetically, for example by noting simply that the trial had not been randomised properly. It is not the role of a meta-analysis group to oversee or to police the conduct of clinical trials and to be too explicit in the rejection of a trial could endanger the goodwill and collaborative spirit necessary for future meta-analyses.

11a.10 The collaborators' meeting

A collaborators' meeting is an important and integral part of the meta-analysis. It ensures that collaborators are the first to see the results of the meta-analysis and that they have a chance to question and discuss these results and their implications before they become available to a wider audience. These discussions and any conclusions that arise may lead to further analyses and they can then be incorporated into the published report of the meta-analysis. In addition, having the meta-analysis debated and endorsed by an internationally recognised group of experts may help with dissemination of results, which is a vital part of any systematic review. Finally the assembly of this international group also provides an excellent opportunity for discussing and possibly deciding the areas of treatment which require clarification or further research. In particular it can provide a good opportunity to discuss and propose future trials. The goodwill engendered is invaluable in completing, updating and publishing the analysis and the existence of the meeting may serve as an incentive to collaborate. Such meetings are also valuable in setting a deadline to which the secretariat and trialists supplying data have to work.

The planning and organisation of such a meeting requires considerable resource and its date must be planned well in advance to fit with the overall timetable for the meta-analysis. The meeting can be scheduled for various stages of the project. If held at a reasonably early stage, when a good deal of data may be outstanding, it acts as a good incentive for trialists who have not supplied data to do so as soon as possible. Alternatively, if it is held at a later point in time, after the majority of data has been assembled and analysed, the results presented are very similar to those that will be used finally, and the time between the meeting and publication will be minimised.

The main purpose of the meeting should be to present the results of the meta-analysis and to discuss the methods, results and implications with the trialists so that they can take a full and active role in this process. The meeting should probably have a structured format and there should be ample time for discussion. Equal proportions of presentation and discussion time might be a good balance. The meeting is also the appropriate place to discuss the future of the Collaborative Group, for example whether to update the IPD meta-analysis in the future.

All those present at the Oxford workshop who had organised such Collaborative Group meetings had provided accommodation free of charge to participating trialists. Some had provided either travel funds for all participants or for those who would otherwise be unable to attend. The provision of such funds obviously depends largely on circumstance: the number of people involved and whether it would be possible to generate sufficient sponsorship to pay for expenses.

One possible approach is to secure full funding for the first collaborators' meeting but for trialists to pay for their own travel to subsequent meetings. The cost of holding a one-day meeting without the provision of travel funds was approximately £100 per delegate increasing to around £600 per delegate when travel was provided, although of course this is very dependent on how far participants had to travel to the meeting.

11a.11 Publication

IPD meta-analyses should aim to publish the results as soon as possible after the Collaborative Group Meeting. Primary publications should be in the name of the collaborative group responsible for the meta-analysis rather than individual authors, the secretariat or steering group. This emphasises the collaborative nature of the project and engenders continued collaboration. As IPD meta-analyses are usually international projects and since trialists may wish to place varying emphasis on the interpretation of the results, it is wise for the publication to concentrate on the presentation of the results leaving detailed interpretation to separate commentaries by independent experts.

11a.12 Research agenda

The methodology described in this report stems from the collective experience of many groups who have already conducted meta-analyses using individual patient data. Such projects provide the most reliable and informative type of systematic review by collecting and analysing all of the relevant randomised evidence. Although some aspects of IPD meta-analyses cannot be done in any other way, for example time to event analyses, they are also particularly time and resource consuming. It is therefore important that additional empirical evidence of the relative values of the different techniques involved in such reviews should be sought and published. With this in mind, the Cochrane Working Group on meta-analysis using individual patient data has initiated a research agenda (Appendix F). In addition to questions directly related to the conduct of meta-analyses, trial information collected as an integral part of these projects is a useful resource which would allow research into randomised controlled trials generally. Some of the topics listed have already been investigated (2, 17, 18, 19, 20, 21) and we would be interested to learn of any other relevant past, current or planned research.

11a.13 Conclusions

Meta-analyses based on updated individual patient data provide the most comprehensive and reliable means of assessing the results of existing randomised clinical trials. It is the only reasonable way of performing time to event analyses, the best way of performing subgroup analyses and allows the review to use common prognostic and outcome variables. The detailed checking of data possible with this approach also improves the accuracy of the data included in the meta-analysis, allowing the integrity of the randomisation and follow up procedures to be assessed centrally. However, considerable expertise, time, effort and resource are required to carry out meta-analyses using individual patient data. They should not be undertaken lightly and might best be carried out by a secretariat on behalf of an international collaborative group. We hope that the guidance contained in this report will prove useful to such people.

11a.14 Appendix A: Participants at the Cochrane Collaboration workshop on Meta-Analysis Using Individual Patient Data, Oxford, 1994

BELGIUM

International Institute for Drug Development, Brussels

Marc Buyse

DENMARK

Danish National Study Group, Herlev

Lena Specht

FRANCE

Clinical Trials and Meta-Analysis - Clinical Pharmacology Unit, Lyon

Francois Gueyffier

Institut Curie, Paris

Veronique Mosseri

Institut Gustave-Roussy, Villejuif

Jean-Pierre Pignon

GERMANY

German Hodgkin's Disease Study Group, Köln

Michael Sextro

ITALY

Mario Negri Institute, Milano

Silvia Marsoni, Valter Torri

NETHERLANDS

Antoni van Leeuwenhoek huis Institute, Amsterdam

Harm van Tinteren, Annet te Velde

SWEDEN

Umeå University Hospital , Umeå

Lennart Nyström

UK

Academic Section of Geriatric Medicine, University of Glasgow

Peter Langhorne

Clinical Trial Service Unit, Oxford

Colin Baigent, Mike Clarke, Rory Collins, Tricia Elphinstone, Vaughan Evans, Richard Gray, Liz Greaves, Heather Halls, Mandy Ogier, Richard Peto, Sue Richards, Keith Wheatley

CRC Wessex Medical Oncology Unit, Southampton
Chris Williams

Department of Clinical Neurosciences, University of Edinburgh
Carl Counsell

Imperial Cancer Research Fund, London
Doug Altman, Jack Cuzick, Rob Edwards

London School of Hygiene and Tropical Medicine, London
Paul Seed

MRC Cancer Trials Office, Cambridge
David Machin, Max Parmar, Sally Stenning, Lesley Stewart, Jayne Tierney, Paul Weston

University College Hospital, London
Carmen Ruiz

UK Cochrane Centre, Oxford (observers)
Iain Chalmers, Jini Hetherington, Sally Hunt, Carol Lefebvre, Andy Oxman

Witthenshawe Hospital, Manchester
Rob Henderson

11a.15 Appendix B: Medline search strategies for optimal sensitivity in identifying randomised clinical trials

Format shown is for SilverPlatter version 3.10. Upper case denotes controlled vocabulary. Lower case denotes free-text terms. Those wishing to run this search strategy are recommended to seek the advice of a trained medical librarian.

#1 RANDOMIZED-CONTROLLED-TRIAL in PT
#2 RANDOMIZED-CONTROLLED-TRIALS
#3 RANDOM-ALLOCATION
#4 DOUBLE-BLIND-METHOD
#5 SINGLE-BLIND-METHOD
#6 #1 or #2 or #3 or #4 or #5
#7 TG=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#8 #6 not #7

#9 CLINICAL-TRIAL in PT
#10 explode CLINICAL-TRIALS
#11 (clin* near trial*) in TI
#12 (clin* near trial*) in AB

#13 (singl* or doubl* or trebl* or tripl*) near (blind* or mask*)
#14 (#13 in TI) or (#13 in AB)
#15 PLACEBOS
#16 placebo* in TI
#17 placebo* in AB
#18 random* in TI
#19 random* in AB
#20 RESEARCH-DESIGN
#21 #9 or #10 or #11 or #12 or #14 or #15 or #16 or #17 or #18 or #19 or #20
#22 TG=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#23 #21 not #22
#24 #23 not #8

#25 TG=COMPARATIVE-STUDY
#26 explode EVALUATION-STUDIES
#27 FOLLOW-UP-STUDIES
#28 PROSPECTIVE-STUDIES
#29 control* or prospectiv* or volunteer*
#30 (#29 in TI) or (#29 in AB)
#31 #25 or #26 or #27 or #28 or #30
#32 TG=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#33 #31 not #32
#34 #33 not (#8 or #24)

Reproduced with kind permission from Carol Lefebvre, UK Cochrane Centre

11a.16 Appendix C: Form supplied with invitation to collaborate in an individual patient-based meta-analysis

LOCALISED SOFT TISSUE SARCOMA META-ANALYSIS

Name:

Did we get your title, affiliation and address correct? If not please give correct details:

Telephone:	Fax:
(*area code and number*)	E-mail:

Please give your own reference or protocol number for this study.

Are the details concerning your study correct?	Yes	No
Is the most recent publication cited in the protocol reference list?		

If no please give details:

Are you willing to take part in this overview?	Yes	No

If yes please confirm that you would be able to supply survival information for each patient randomised

Yes	No		Yes	No	
		Patient identifier			Date of randomisation
		Date of birth or age at randomisation			Survival status
		Sex			Cause of death
		Disease status			Date of death/last follow up
		Disease site			Local recurrence status
		Histology			Date of local recurrence
		Histologic Grade			Distant recurrence status
		Tumour size			Date of distant recurrence
		Primary treatment			Whether excluded from own analysis
		Treatment allocated			Reason for exclusion
		Extent of resection			

How will you supply data?

Floppy disk:	E-mail:	Computer print-out:	Sealed envelope:

Please give the method of randomisation used in this study

Central telephone call Other (please specify): Sealed envelope:

Please state stratification factors used (if any):

What proportions was this study designed to have in each arm? (eg 1:1

Please give the name and address of the appropriate contact for collection of data:

Please five details of any relevant publications or trials you may know of not listed in the tables or Appendix A of the protocol:

Signed Date

Please note that any information supplied will be treated in strict confidence and used only for the purpose of the overview

11a.17 Appendix D1: Example coding and formatting instructions for data supplied electronically

LOCALISED SOFT TISSUE SARCOMA META-ANALYSIS

Suggested Coding: Individual Patient Data

- Disks should be formatted for the DOS operating system.
- Files should be in DBASE, FoxPro (.dbf files) or ASCII format with fields separated by spaces. However, it would be preferable if you did not use spaces to denote unknown values (see below).
- You may code the data in whichever way is most convenient to you, although it would be helpful if you adopted the coding suggested on this sheet. If you are unable to do this, please supply full details of the coding system used.

Please list fields in the following order using the suggested coding:

Patient identifier	Type	Character	**Treatment Allocated**	Type	numeric
	Width	15		Width	1
	Any alphanumeric string up to15 characters			Code	1=treatment

					2=control
Date of birth (DOB)	Type	date	**Extent of Resection**	Type	numeric
	Width	-			1
	Code	date in dd/mm/yy format			1=well clear
		unknown day=15/mm/y			2=close/marginal
		unknown month=157067yy			3=macroscopically involved
		unknown date=0/01/01			9=unknown
Age	Type	numeric	**Date of Randomisation (DOR)**	Type	date
	Width	3		Width	-
	Code	age in years		Code	date in dd/mm/yy format
		unknown=999			
Sex	Type	numeric	**Survival Status**	Type	numeric
	Width	1		Width	1
	Code	1=female		Code	0=alive
		2=male			1=dead
		9=unknown		If survival status is unknown code as 0, the patient being censored at the date of the last follow up	
Disease status (at randomisation)	Type	numeric	**Cause of death**	Type	numeric
	Width	1		Width	1
	Code	1=primary		Code	1=soft tissue sarcoma
		2=recurrent			2=chemotherapy related
		3=metastatic			3=other
		9=unknown			8=not applicabe
					9=unknown
Disease site	Type	numeric	**Date of death/Last follow up**	Type	date

	Width	1		Width	-
	Code	1=extremity		Code	date in dd/mm/yy format
		2=trunk			unknown day=15/mm/yy
		3=head and neck			unknown month=15/06/yy
		4=breast			unknown date=01/01/01
		5=uterus			
		6=retroperitoneum	**Local Recurrence Status**	Type	numeric
		7=viscera/abdomen		Width	1
		9=unknown		Code	0=no recurrence
					2=leiomyosarcoma
					9=unknown
Histology	Type	numeric			
	Width	1			
	Code	1=MFH	**Date of Local Recurrence**	Type	date
		1=recurrence		Width	-
		3=liposarcoma		Code	date in dd/mm/yy format
		4=synovial			unknown day=15/mm/yy
		5=malignant schwannoma			unknown month=15/06/yy
		6=alveolar or embryonal rhabdomyosarcoma/Ewing's/PNET			unknown date=01/01/01
		7=AIDS-related sarcoma	**Distant Recurrence Status**	Type	numeric
		8=other		Width	1
		9=unknown		Code	0=no recurrence
					1=recurrence
Grade	Code as convenient, but please supply full details of the coding system used				9=unknown

			Date of Distant Recurrence	Type	date
Tumour size	Type	numeric		Width	-
	Width	2		Code	date in dd/mm/yy format
	Code	Give the size of the largest single dimension in centimetres			unknown day=15/mm/yy
		unknown=99			unknown month=15/06/yy
					unknown date=01/01/01
					-
Primary Treatment	Type	numeric	**Excluded**	Type	numeric
	Width	3		Width	1
	Code	*1st digit (pre-op treatment)*		Code	0=included in analysis
		0=non			1=excluded from analysis
		1=radiotherapy			9=unknown
		2=induction chemotherapy			
		3=radiotherapy + induction chemotherapy	**Reason for Exclusion**	Type	character
		9=unknown		Width	15
				Code	short string giving reason for exclusion ornumeric codes with code meanings provided
		2nd digit (surgery)			
		1=amputation			
		2=excision			
		3=biopsy only			
		9=unknown			
		3rd digit (post-op treatment)			
		0=no radiotherapy			
		1=radiotherapy			

2=unknown

11a.18 Appendix D2: Example of a form that could be used to supply data manually

Colectoral Cancer Collaboration (CCC) overview of mortality by randomly-allocated treatment in resectable colorectal cancer trials: provision of one line line of CONFIDENTIAL data for each patient ever randomised (INCLUDING any ineligible, withdrawn, inevaluable, lost or 'protocol deviant' patients).

Name of trialist of trial group:

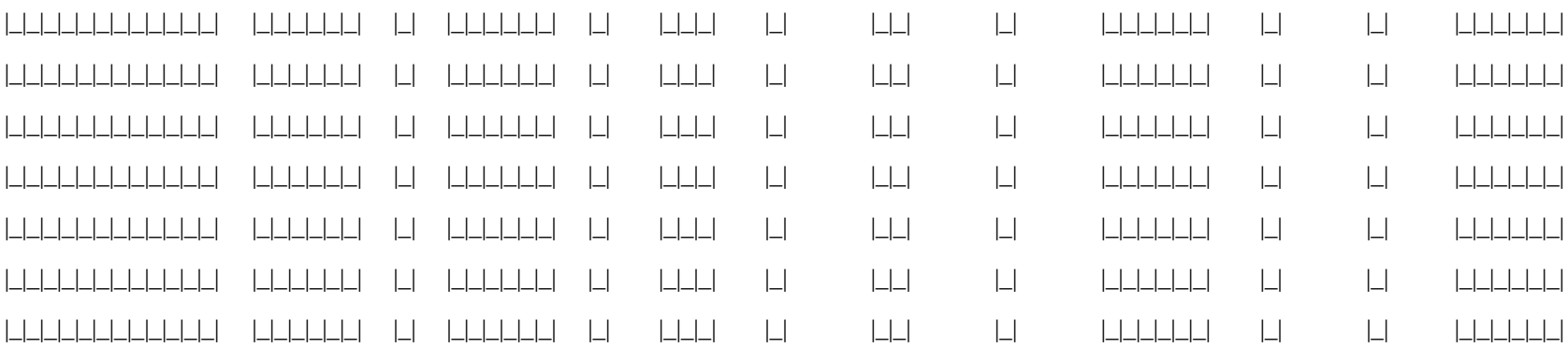
Data Sheet No.:

Name of trial:

Staging system used: Dukes: |_| Astler-Coller: |_| TNM: |_| Other: |_| (please specify)

Treatment group 1 = ; Trt. gp. 2 = ; Trt. gp. 3= ; Trt. gp. 4=

Patient Identifier	Date randomi-sed	Trf. gp.	Date of surgery	Tum.site	Tumour stage	Gender	Entry age	Rec-ur?	Approx. date of 1st rec.	Site rec.	Dead/ other	Date died/last traced	Cause of death if died without recurrence
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	
|_|_|_|_|_|_|_|_|_|_|_|_|_|	|_|_|_|_|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|_|_|	|_|	|_|_|	|_|	|_|_|_|_|_|_|	|_|	|_|	|_|_|_|_|_|_|	



Leave blanks if not applicable or if data not (or not yet) conveniently available

11a.19 Appendix D3: Coding scheme that was used with the form for supplying data manually

GUARANTEE OF CONFIDENTIALITY OF DATA: ANY INFORMATION PROVIDED OVERLEAF TO THE

CC SECRETARIAT WILL BE HELD SECURELY AND IN STRICT CONFIDENCE

NOTES OF FORMAT OF DATA REQUESTED OVERLEAF:

- Special coding conventions:

  Please accompany these forms by an explanatory letter about any special coding conventions (e.g. on tumour site, tumour staging or cause of death) you have used, plus notes on any special features of the study(s) to which you wish to draw attention.

- Dates that are not (or not yet) known exactly:

  either leave DAY blank and give (approximate or provisional) month and year;

  or leave DAY and MONTH blank, and just give approximate year.

BASELINE DATA:

Patient identifier:

Any convenient convention you wish, in case any correspondence becomes necessary. (If reporting several trials, please try to use a system that implicitly specifies both the trial and the patient.)

Date randomised:

Please describe ALL patients EVER randomised, including even lost, ineligible or withdrawn patients, and ignore all non-randomised patients.

Trt. gp. allocated:

Treatment group number: 1 or 2 only, for 2-group trials, or a wider range for trials with more arms, as defined by you at the top of the form. N.B: even if, in reality, some quite different (or even opposite!) treatment was inadvertently given, what is wanted is the originally-allocated treatment. (For patients erroneously entered more than once, give only the first allocation.)

Date of surgery:

See note above on approximate dates.

Tumour site:

0 = unspecified; 1 = colon; 2 = rectum; 3 = colon and rectum. If you prefer to use your own classification of tumour site (e.g. in order to code sigmoid tumours separately) please do so, and send us details of it.

Tumour stage:

Please use your own classification and send us details of it, or use the Dukes classification (A = lesion confined to muscularis propria; B = lesion extends through muscularis propria with negative nodes; C = positive nodes), or any other standard system (e.g. Astler-Coller modification, TNM etc). Extra codes: D = metastatic disease; X = benign tumour (eg adenoma); and Y = inoperable disease.

Gender:

1 = male; 2 = female.

Entry age:

Age at randomisation.

FOLLOW-UP DATA:

Recur?:

Any recurrence? 1 = none recorded; 2 = some recurrence (local or distant or both).

Approx. date of 1st recur.:

Give the best estimate you can: see note above on approximate dates.

Site of 1st recurrence:

0 = unknown; 1 = local only; 2 = local and distant; 3 = distant only.

Dead/other:

1 = alive when last traced; 2 = known to be dead; 3 = lost despite extensive inquires, but still alive when last traced.

Date died/last traced:

Date of death, or date last known to be alive, as accurately as possible: see note above on approximate dates.

Death cause:

If the patient died without reported recurrence, give underlying cause of death. Either state the cause in words, use an ICD code or use your own classification and send us details of it.

11a.20 Appendix D4: Example of instructions that could be used to create a formated electronic file

MACH-NC

Meta-analysis of Chemotherapy in Head and Neck Cancer

	Column	
Patient Identifier	2-11	10 characters
Date of birth	13-18	dd/mm/yy, 999999=Unknown
or age	17-18	2 digits (13-16 blanks) 99=Unknown
Sex	20	1=Male, 2=Female, 9=Unknown

Site of primary	22	1=Oral cavity, 2=Oropharynx, 3=Nasopharynx, 4=Larynx, 5=Hypopharynx, 6=Cervical node(s) without primary, 7=Others, 9=Unknown
T	24	O=TO, X=TX, S=Tis, 1=T1, 2=T2, 3=N3, 9=Unknown
N	25	O=NO, X=NX, 1=N1, 2=N2, 3=N3, 9=Unknown
M	26	O=MO, 1=M1, 9=Unknown
or stage	26	1 digit (24-25 blanks), 9=Unknown

(The aim of the next four questions is to identify presenting characteristics at the time of randomisation)

Recurrence at randomisation	28	0=No, 1=Yes
Second primary at randomisation	30	0=No, 1=Yes
Squamous cell	32	0=No, 1=Yes
Type of histology if not squamous cell	34-45	12 characters (*blanks for squamous cell)*
Treatment allocated	47	1=No chemotherapy, 2=Chemotherapy
Date of randomisation	49-54	dd/mm/yy, 999999=Unknown
Received at least one cycle of chemotherapy	56	0=No, 1=Yes, 9=Unknown
Date of last follow-up	58-63	dd/mm/yy, 999999=Unknown
Survival status	65	0=Alive, 1=Dead
Death related to treatment	67	0=No, 1=Yes
Complete response at the end of treatment (including salvage treatment)	69	0=No, 1=Yes *(collected for computation of disease-free survival)*
Recurrence of second primary	71	0=No, 1=Yes *(only for complete responders)*
Date of first event	73-78	dd/mm/yy, 999999=Unknown
Type of first event	80	1=locoregional, 2=metastasis, 3=locoregional + metastasis, 4=second primary without recurrence, 9=Unknown
Excluded from your analysis	82	0=No, 1=Yes
Reasons for exclusion	84-95	12 characters

11a.21 Appendix E: Sources of mortality information for individual patients

England and Wales

The Chief Medical Statistician (Dept MR) Health Statistics
OPCS
St Catherine House
10 Kingsway
London WC2B 6JP

France

INSEE
Département de Démographie
Division Répertoire et Mouvement de la Population
18, Bd Adolphe Pinard
75675 PARIS
Cedex 14

Service d'information sur les causes médicale de décès
INSERM SC8
55, Chemin de Rorde
BP 34
78100 LE VESINET

Isle of Man

Isle of Man Health Services Board
Registration Department
Markwell House
Market Street
Douglas
Isle of Man

Northern Ireland

The Central Services Agency
27 Adelaide Street
Belfast BT2 8SH

Norway

Statistisk Sentralbyraa
Skippergt. 15
PB 8131 Dep
N-0033 Oslo
Norway

Scotland

Departmental Record Officer

General Register Office for Scotland

New Register House

Edinburgh EH1 3YT

USA

National Death Index

Division of Vital Statistics

National Centre for Health Statistics

6525 Belcrest Road

Hyattsville, MD 20782

USA

11a.22 Appendix F: Research agenda proposed by Cochrane Working Group on Individual Patient Based Meta-Analyses

Although some aspects of IPD meta-analyses cannot be done in any other way, for example time to event analyses, these projects also particularly time and resource consuming. It is therefore important that additional empirical evidence of the relative values of the different techniques involved in such reviews should be sought and published.

A Research relating to individual patient-based meta-analysis

1. ***Comparison of individual patient data with summary data supplied by trialists*:** At least two individual patient based meta-analyses have been conducted following the collection of summary data from the same set of trials. These are in Hodgkin's disease and in antiplatelet therapy

2. ***Comparison of individual patient data with published data*:** This has been done for cisplatin-based therapy in ovarian cancer but most of the individual patient data meta-analyses could repeat those analyses. This would allow the evidence to be extended to other disease and therapy areas.

3. ***Comparison of individual patient data after extensive data-checking with individual patient data supplied initially*:** There are different levels of data-checking - from finding and querying missing or inconsistent data variables, to detailed investigation of the integrity of the randomisation and follow-up procedures. Detailed data-checking is resource-intensive and time-consuming and may delay the publication of the meta-analysis results, so empirical evidence of its value would be useful.

4. ***Comparison of trial quality as assessed using the individual patient data with quality as assessed from the published report:*** Does the individual patient data reveal problems in the randomisation or follow-up procedures that were not mentioned in the published report?

B Research relating to all types of meta-analysis and to RCTs

5. *Method of randomisation:* Sensitivity analyses could be performed using the method of randomisation (eg envelope, central computer, 'blinded' date of birth) to distinguish between RCTs. Stratification, minimisation and block size could also be investigated.

6. ***Size of RCTs*:** Sensitivity analyses could be performed to take into account the size of RCTs. This could also investigate whether there are important differences in the results from multi-centre or single institute trials.

7. ***Chronology of RCTs:*** Sensitivity analyses could be performed distinguishing between RCTs by their place in time - perhaps the early RCTs have the more striking results. A RCT's place in time could be defined in various ways (start date, finish date, publication date) and cumulative meta-analyses could be done ordered in these ways. Sensitivity analyses could also be performed distinguishing between RCTs published before the systematic review was conducted and those published afterwards.

8. ***Place of publication:*** Sensitivity analyses could be performed distinguishing between RCTs which have been published as full papers, as abstracts or are unpublished. This will also investigate whether there are any important differences between RCTs published in journals indexed by medical literature databases; between RCTs in those databases which would or would not have been found by a simple search strategy; between RCTs in or not in the 'major' journals identified by these databases;and between RCTs published in different languages.

9. ***Speed of publication:*** The variation in the speed of publication among trials with differing results could be investigated, especially with regard to changes in their results with further follow-up.

10. ***Repeated publications*:** RCTs may be reported several times and it is often difficult to know that reports are of the same trial, and so may be included more than once in a meta-analysis. It has been suggested that positive trials are more likely to be published repeatedly. This could be investigated.

11. ***Fate of RCTs published as abstracts:*** Sensitivity analyses could be performed distinguishing between RCTs which were published as abstracts and then did or did not publish as full papers.

12. ***Citation bias:*** To investigate whether the RCTs in the meta-analyses selectively cite other RCTs with similar results. This could also investigate (using the Science Citation Index) which RCT publications are cited most often to see if their results are representative of the overall conclusion as shown by the meta-analysis or are they at an extreme?

13. ***Source of trial funding:*** Sensitivity analyses could be performed using the source of funding (eg drug company, government, charity, local) to distinguish between RCTs.

11a.23 Acknowledgements

We are grateful to the UK Cochrane Centre, in particular to Iain Chalmers for suggesting the Oxford Workshop and Caroline Caldicott for helping to organise it. We also thank Linda Baulk for typing this manuscript.

11a.24 References

1. Chalmers I. The Cochrane Collaboration: preparing, maintaining and disseminating systematic reviews of the effects of healthcare. Ann N Y Acad Sci 1993; **703**: 156-65.

2. Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data:is there a difference? Lancet 1993; **341**: 418-22.

3. Pignon JP, Arriagada R. Meta-analysis. Lancet 1993; **341**: 964-5.

4. Pignon JP, Ducreux M, Rougier P. Meta-analysis of adjuvant chemotherapy in gastric cancer: a critical reappraisal. J Clin Oncol 1994; **12:** 877-9.

5. Dickersin K, Scherer R, Lefebvre C. Identification of relevant studies for systematic review. Br Med J 1994; **309**: 1286-91.

6. Cook DJ, Guyatt GH, Ryan et. Should unpublished data be included in meta-analyses. JAMA 1993; **269:** 2749-53.

7. Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. JAMA 1992; **267**: 374-8.

8. Dickersin K, Min YI. NIH Clinical trials and publication bias. Online J Curr Clin Trials (serial online) 1993. Doc No 50.

9. Easterbrook PJ, Berlin JE, Gopalan R, Matthews DR. Publication bias in clinical trials. Lancet 1991; **337**: 865-72.

10. Dickersin K. Keeping posted. Why register clinical trials? - Revisited. Cont Clin Trials 1992; **13**: 170-7.

11. Easterbrook PJ. Directory of registries of clinical trials. Stats in Med 1992; **11:** 345-423.

12. Clarke MJ, Stewart LA. Obtaining data from randomised controlled trials: how much do we need in order to perform reliable and informative meta-analyses? Br Med J 1994; **309:**1007-10.

13. Advanced Ovarian Cancer Trialists Group. Chemotherapy in advanced ovarian cancer: an overview of randomised clinical trials. Br Med J 1991; **303:** 884-93.

14. Early Breast Cancer Trialists' Collaborative Group. Systemic treatment of early breast cancer by hormal, cytotoxic or immune therapy. Lancet 1992; **339:** 1-15, 71`-85.

15. Early Breast Cancer Trialists' Collaborative Group. Treatment of early breast cancer: Vol 1, Worldwide evidence 1990 Oxford: Oxford University Press, 1990.

16. Haugh MC, Cornu C, Boissel JP. Long-term survival follow-up in international clinical trials 1993; **14:** 416.

17. Scherer RW, Dickersin K, Langenberg P. Full publication of results initially presented in abstracts. A meta-analysis. JAMA 1994; **272**: 158-62.

18. Rochon PA, Gurwitz JH, Cheung CM, Hayes HA, Chalmers TC. Evaluating the quality of articles published in journal supplements compared with the quality of those published in the parent journal. JAMA 1994; **272**: 108-13.

19. Gotzsche PC. Multiple publication of reports of drug trials. Eur J Clin Pharmacol 1989; **36**: 429-32.

20. Gotzsche PC. Reference bias in reports of drug trials. Br Med J 1987; **295:** 654-9.

21. Antman EM, Lau J, Kupelnick B, Mostellar F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. JAMA 1992; **268:** 240-8. (Impact of meta-analyses).

22. Bobbio M, Demichelis B, Giustetto G. Completeness of reporting trials results: effect on physician' willingness to prescribe. Lancet 1994; **343:** 1209-11. (Absolute versus relative difference for the report of treatment effect).

APPENDIX 11b. Prospective meta-analysis

A systematic review should, ideally, define the question to be addressed in advance of the identification of potentially eligible studies. However, these projects are by their nature, retrospective, since the studies included are usually identified after they have been completed and reported (Pogue 1998, Zanchetti 1998). The reviewer's knowledge of the results of the study may influence:

- the criteria for study selection
- the definition of a systematic review question
- the interventions and participant groups evaluated
- the outcomes to be assessed in the review

In contrast, a systematic review which is conducted as a prospective meta-analysis includes studies that were identified, evaluated and determined to be eligible for inclusion before their results became known. It is a method that has been used in recent years in cardiovascular disease (Simes 1995, CTTC 1995, WHO-ISHBPL 1998) and childhood leukaemia. (Shuster 1996, Valsecchi 1996) and can help to overcome some of the problems of traditional systematic reviews by enabling:

- hypotheses to be specified *a priori*, ignorant to the results of individual studies
- prospective application of selection criteria
- *a priori* statements of intended analyses, including subgroup analyses, to be made before the results of individual studies are known. This avoids potentially unreliable data-dependent emphasis on particular subgroups.

A Methods Group has been established to investigate methodological issues around such projects and to offer guidance on their conduct. For example, because studies should not be included in a prospective meta-analyses if their results are known before the decision is taken to include them, PMA will not always include all studies of a particular question. Research is needed to investigate the impact of this on systematic reviews.

To register a PMA as a Cochrane review, investigators need to submit a protocol to the relevant Collaborative Review Group (CRG). The protocol will then undergo the same peer review process as any Cochrane review. The decision as to whether or not a PMA should be a Cochrane Review rests with the CRG. If a CRG decides it does not have the expertise necessary to determine whether or not the submitted protocol meets the requirements of a PMA, members of the PMA Methods Group will be available to review the protocol.

11b.1 References

CTTC 1995. Protocol for a prospective collaborative overview of all current and planned randomized trials of cholesterol treatment regimens. Cholesterol Treatment Trialists' (CTT) Collaboration. American Journal of Cardiology 1995; 75: 1130-4.

Pogue 1998. Pogue J, Yusuf S. Overcoming the limitations of current meta-analysis of randomised controlled trials. Lancet. 1998; 351: 47-52.

Shuster 1996. Shuster JJ. Gieser PW. Meta-analysis and prospective meta-analysis in childhood leukemia clinical research. Annals of Oncology 1996; 7: 1009-14.

Simes 1995. Simes RJ. Prospective meta-analysis of cholesterol-lowering studies: the Prospective Pravastatin Pooling (PPP) Project and the Cholesterol Treatment Trialists (CTT) Collaboration. American Journal of Cardiology 1995; 76: 122C-126C.

Valsecchi 1996. Valsecchi MG. Masera G. A new challenge in clinical research in childhood ALL: the prospective meta-analysis strategy for intergroup collaboration. Annals of Oncology 1996 ; 7: 1005-8.

WHO-ISHBPL 1998. Protocol for prospective collaborative overviews of major randomized trials of blood-pressure-lowering treatments. World Health Organization-International Society of Hypertension Blood Pressure Lowering Treatment Trialists' Collaboration . Journal of Hypertension 1998; 6: 127-37.

Zanchetti 1998. Zanchetti A, Mancia G. Searching for information from unreported trials--amnesty for the past and prospective meta-analyses for the future. Journal of Hypertension 1998; 16: 125.