A GUIDE to the development implementation

How to review the evidence: systematic identification and review of the scientific literature

and evaluation of clinical practice guidelines

NHMRC
National Health and Medical Research Council

How to review the evidence: systematic identification and review of the scientific literature

Handbook series on preparing clinical practice guidelines

Endorsed November 1999

NHMRC
National Health and Medical Research Council

ISBN 1864960329

The strategic intent of the National Health and Medical Research Council (NHMRC) is to work with others for the health of all Australians, by promoting informed debate on ethics and policy, providing knowledge-based advice, fostering a high quality and internationally recognised research base, and applying research rigour to health issues.

NHMRC documents are prepared by panels of experts drawn from appropriate Australian academic, professional, community and government organisations. NHMRC is grateful to these people for the excellent work they do on its behalf. The work is usually performed on an honorary basis and in addition to their usual work commitments.

This document is sold through Government Info Bookshops at a price that covers the cost of printing and distribution only. For publication purchases please contact AusInfo on their toll-free number 132 447, or through their Internet address: www.ausinfo.gov.au/general/gen_hottobuy.htm

Production by Biotext, Canberra

CONTENTS

Preface ix

INTRODUCTION 1

Development of evidence-based guidelines 1
Systematic literature reviews 1
Method 2
How much work is a systematic review? 4
About this handbook 5

PART 1 GENERAL METHODS 7

1 THE QUESTION 9

1.1 What types of questions can be asked? 9
1.1.1 Interventions 9
1.1.2 Frequency or rate 10
1.1.3 Diagnostic test performance 12
1.1.4 Aetiology and risk factors 12
1.1.5 Prediction and prognosis 13
1.1.6 Economics 13
1.2 What is the relevant question? 13
1.3 How focused should the question be? 13

2 FINDING RELEVANT STUDIES 15

2.1 Finding existing systematic reviews 15
2.2 Finding published primary studies 15
2.2.1 Breaking down the study question into components 15
2.2.2 Use of synonyms 16
2.2.3 Snowballing 18
2.2.4 Handsearching 18
2.2.5 Methodological terms 18
2.2.6 Methodological filters 19
2.2.7 Use of different databases 19
2.3 Finding unpublished primary studies 20
2.3.1 Searching relevant databases 20
2.3.2 Writing to experts 20
2.4 Publication bias — a crucial problem 20
2.4.1 What is publication bias? 20

2.4.2 Does this affect the results of the reviews? 21
2.4.3 What can we do about publication bias? 21
2.4.4 Duplicate publications 22

3 APPRAISING AND SELECTING STUDIES 23

3.1 Standardising the appraisal 23
3.1.1 What study features should be assessed? 23
3.1.2 Is it important to have a structured appraisal? 23
3.1.3 How many reviewers are required? 24
3.1.4 Is it necessary to do the appraisal 'blind' to the outcome of the study? 24
3.2 Using the quality appraisal 24

4 SUMMARISING AND SYNTHESISING THE STUDIES 27

4.1 Presenting the results of the studies (data extraction) 27
4.1.1 Tabular summary 27
4.1.2 Graphical presentation 27
4.2 Synthesis of study results 27
4.2.1 Fixed and random effects estimates 30
4.3 Assessing heterogeneity 31
4.3.1 Measures of heterogeneity 31
4.3.2 Individual patient data meta-analysis 33
4.4 Detecting publication bias 34
4.4.1 Funnel plots 34
4.4.2 Statistical tests 35
4.4.3 If publication bias is suspected, what can be done? 36

5 APPLICABILITY: RETURNING TO THE QUESTION 37

PART 2 QUESTION-SPECIFIC METHODS 39

6 INTERVENTIONS 41

6.1 The question 41
6.1.1 Study design 41
6.2 Finding relevant studies 41
6.2.1 Finding existing systematic reviews 41
6.2.2 Finding published primary studies 42
6.2.3 Finding unpublished primary studies 42
6.3 Appraising and selecting studies 43
6.3.1 Standardising the appraisal 43
6.4 Summarising and synthesising the studies 44
6.4.1 Presenting the results of the studies 44

6.4.2 Synthesis of study results 46
6.4.3 Assessing heterogeneity 48
6.5 Economic evaluation 49
6.6 Further information 50

7 FREQUENCY AND RATE 51

7.1 The question 51
7.1.1 Study design 52
7.2 Finding relevant studies 52
7.2.1 Finding existing systematic reviews 52
7.2.2 Finding published primary studies 52
7.2.3 Finding unpublished primary studies 53
7.3 Appraising and selecting studies 53
7.3.1 Standardising the appraisal 53
7.4 Summarising and synthesising the studies 54
7.4.1 Presenting the results of the studies 54
7.4.2 Synthesis of study results 55
7.4.3 Assessing heterogeneity 56

8 DIAGNOSTIC TESTS 57

8.1 The question 57
8.1.1 Study design 57
8.2 Finding relevant studies 58
8.2.1 Finding existing systematic reviews 58
8.2.2 Finding published primary studies 58
8.2.3 Finding unpublished primary studies 60
8.3 Appraising and selecting studies 60
8.3.1 Standardising the appraisal 60
8.4 Summarising and synthesising the studies 64
8.4.1 Presenting the results of the studies 64
8.4.2 Synthesis of study results 66
8.4.3 Assessing heterogeneity 68

9 AETIOLOGY AND RISK FACTORS 69

9.1 The question 69
9.1.1 Study design 69
9.2 Finding relevant studies 70
9.2.1 Finding existing systematic reviews 70
9.2.2 Finding published primary studies 70
9.2.3 Finding unpublished primary studies 71
9.3 Appraising and selecting studies 71
9.3.1 Standardising the appraisal 71
9.4 Summarising and synthesising the studies 74
9.4.1 Presenting the results of the studies 74
9.4.2 Synthesis of study results 75

9.4.3 Assessing heterogeneity 75
9.5 Judging causality 76

10 PREDICTION AND PROGNOSIS 79

10.1 The question 79
10.1.1 Why should we be interested in prediction? 79
10.1.2 Study design 80
10.2 Finding relevant studies 80
10.2.1 Finding existing systematic reviews 80
10.2.2 Finding published primary studies 80
10.2.3 Finding unpublished primary studies 80
10.3 Appraising and selecting studies 81
10.3.1 Standardising the appraisal 81
10.4 Summarising and synthesising the studies 81
10.4.1 Presenting the results of the studies 81
10.4.2 Synthesis of study results 82
10.4.3 Assessing heterogeneity 82

Appendix A Membership of production team for handbook 83

Appendix B Process report 85

Appendix C Literature searching methods 87

Appendix D Software for meta-analysis 91

Glossary 95

Acronyms and abbreviations 107

References 109

TABLES

Table 1.1 Types of clinical and public health questions, ideal study types and major appraisal issues 10

Table 1.2 Types of studies used for assessing clinical and public health interventions 11

Table 2.1 Using synonyms of components of the three-part question to devise a literature search 17

Table 2.2 Comparison of published and registered studies for multiagent versus single agent chemotherapy for ovarian cancer 22

Table 4.1 Some possible outcome measures of study effects 29

Table 6.1	Example summary table of quality features of a set of hypothetical intervention trials	46
Table 7.1	Example summary table of a set of hypothetical studies of frequency	54
Table 8.1	Example summary table of quality features of a set of hypothetical diagnostic accuracy trials	65

FIGURES

Preface	Flow chart showing the clinical practice guidelines development process	xi
Figure 2.1	Venn diagram for colorectal screening	16
Figure 2.2	Papers identified by different search methods in a systematic review of near-patient testing	21
Figure 4.1	Relative mortality from colorectal cancer in screened versus unscreened (control) groups from four randomised trials of faecal occult blood screening	28
Figure 4.2	Meta-analysis of 12 placebo-controlled trials of St John's wort for depression, showing significant heterogeneity	32
Figure 4.3	Hypothetical study showing combined and subgroup analysis	34
Figure 4.4	Funnel plot of 12 placebo-controlled trials of St John's wort showing some suggestion of 'missing' smaller negative trials	35
Figure 6.1	Placebo-controlled trials of treatment of epilepsy with the drug gabapentin	47
Figure 6.2	L'Abbe plot of the stroke risk in the treated group versus the stroke risk in the control group from a meta-analysis of 6 placebo-controlled trials of warfarin for nonvalvular atrial fibrillation.	48
Figure 7.1	Proportion of patients with antibiotic resistance in *Propionibacterium acnes* for four studies, listed by publication date	55
Figure 8.1	Plot of sensitivity versus specificity (with 95% confidence intervals) for 14 studies of carotid ultrasound for carotid stenosis	66

Figure 8.2 Receiver–operator curve (ROC) plotting true positive rate (sensitivity) against false positive rate (1 – specificity) for a meta-analysis of carotid ultrasound accuracy showing the individual study points and the fitted summary ROC (SROC) 67

Figure 8.3 Plot of *D* versus *S* for a meta-analysis of carotid ultrasound accuracy showing the individual study points and the fitted line 68

BOXES

Box 6.1 Checklist for appraising the quality of studies of interventions 45

Box 8.1 Checklist for appraising the quality of studies of diagnostic accuracy 62

Box 9.1 Checklist for appraising the quality of studies of aetiology and risk factors 73

PREFACE

Clinical practice guidelines are systematically developed statements that assist clinicians, consumers and policy makers to make appropriate health care decisions. Such guidelines present statements of 'best practice' based on a thorough evaluation of the evidence from published research studies on the outcomes of treatment or other health care procedures. The methods used for collecting and evaluating evidence and developing guidelines can be applied to a wide range of clinical interventions and disciplines, including the use of technology and pharmaceuticals, surgical procedures, screening procedures, lifestyle advice, and so on.

In 1995, recognising the need for a clear and widely accessible guide for groups wishing to develop clinical practice guidelines, the National Health and Medical Research Council (NHMRC) published a booklet to assist groups to develop and implement clinical practice guidelines. In 1999 a revised version of this booklet was published called *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999), which includes an outline of the latest methods for evaluating evidence and developing and disseminating guidelines.

The emerging guideline processes are complex, however, and depend on the integration of a number of activities, from collection and processing of scientific literature to evaluation of the evidence, development of evidence-based recommendations or guidelines, and implementation and dissemination of the guidelines to relevant professionals and consumers. The NHMRC has therefore decided to supplement the information in the guideline development booklet (NHMRC 1999) with a series of handbooks with further information on each of the main stages involved. Experts in each area were contracted to draft the handbooks. An Assessment Panel was convened in June 1999 to oversee production of the series. Membership of the Assessment Panel and the writing group for this handbook are shown at Appendix A.

Each of the handbooks in the series focuses on a different aspect of the guideline development process (review of the literature, evaluation of evidence, dissemination and implementation, consumer publications, economic assessment and so on). This handbook describes how to systematically identify scientific literature relevant to a particular question, select and review the most important (highest quality) studies and summarise and present the results for further consideration by the committee that will develop the clinical practice guidelines.

The way in which the guidance provided in this handbook fits into the overall guideline development process recommended by the NHMRC is

shown in the flowchart on page xi. Other handbooks that have been produced in this series so far are:

How to Use the Evidence: Assessment and Application of Scientific Evidence

How to Put the Evidence into Practice: Implementation and Dissemination Strategies

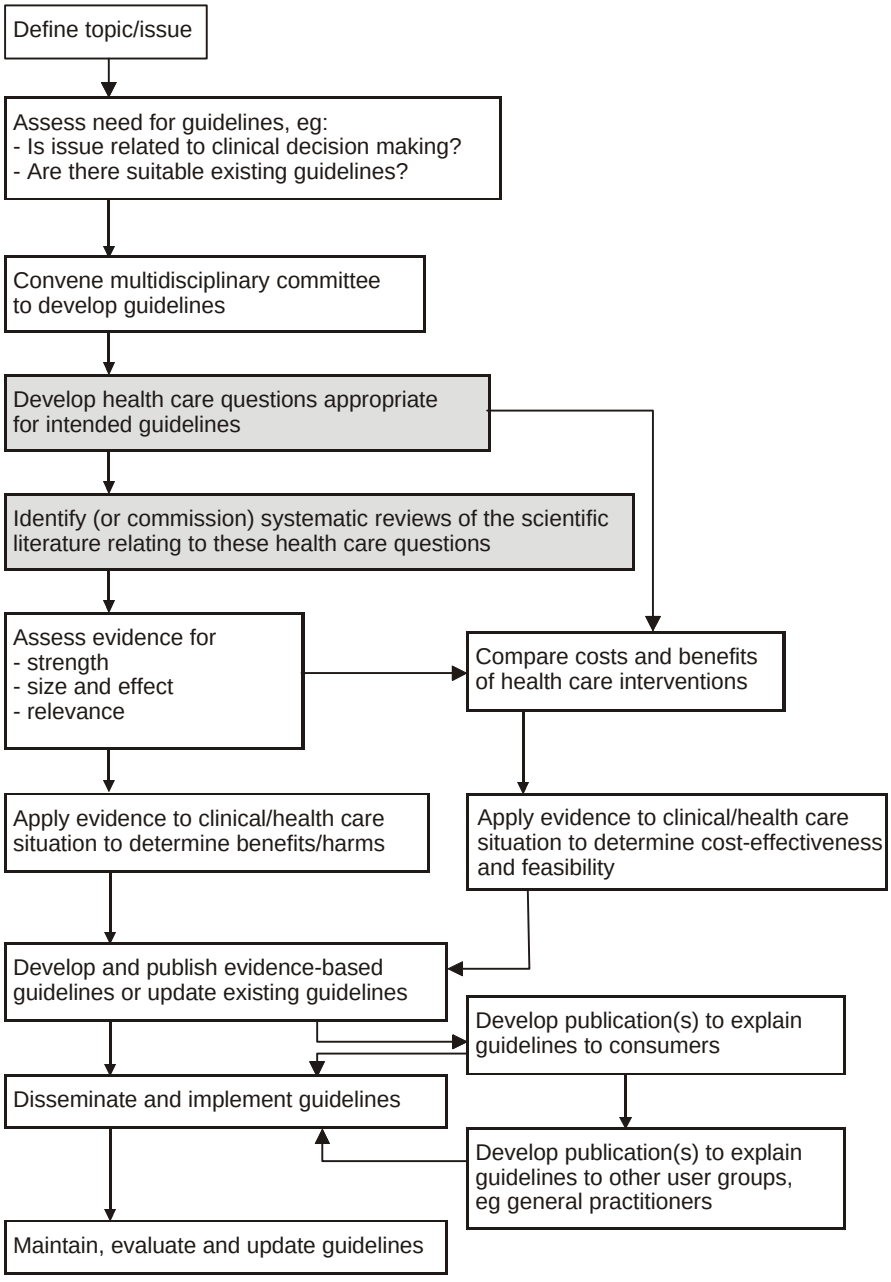
How to Present the Evidence for Consumers: Preparation of Consumer Publications

How to Compare the Costs and Benefits: Evaluation of the Economic Evidence

The series may be expanded in the future to include handbooks about other aspects of the guideline development process, as well as related issues such as reviewing and evaluating evidence for public health issues.

Flow chart showing the clinical practice guidelines development process

(Shaded boxes show the stages described in this handbook)



INTRODUCTION

Development of evidence-based guidelines

The process for clinical practice guideline development is described in the National Health and Medical Research Council (NHMRC) publication *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999).

This recommends that guidelines should be developed by a multidisciplinary guideline development committee, the initial task of which is to determine the need for and scope of the guidelines, define the purpose and target audience and identify the health outcomes that will improve as a result of their implementation.

The membership of the guideline development committee will depend on the nature of the particular guidelines being developed but will include clinicians, health professionals, consumers, health policy analysts, economists and regulatory agency representatives, industry representatives and bioethicists (see NHMRC 1999 for a full list and further discussion of the multidisciplinary committee).

One of the main principles of guideline development is that they should be based on the best available evidence. Therefore, the next task of the guideline development committee is to commission or undertake a systematic review of the scientific literature to find out what is known about the benefits and harms of the intervention under consideration or about other health outcomes related to the particular guidelines that the committee is developing. This handbook outlines the principles and methods involved in such a review. It has been prepared to assist reviewers with the task and to help the guideline development committee interpret the review when it is received. However, it should be noted that the systematic review of scientific literature is a specialised task and the guideline development committee should ensure that a reviewer is engaged with the necessary skills and experience to undertake the task.

Systematic literature reviews

Methods for reviewing and evaluating the scientific literature range from highly formal, quantitative information syntheses to subjective summaries of observational data. The guideline development committee must select the most rigorous and systematic review methods practicable.

The purpose of a systematic literature review is to evaluate and interpret all available research evidence relevant to a particular question. In this approach a concerted attempt is made to identify all relevant primary research, a standardised appraisal of study quality is made and the studies of acceptable quality are systematically (and sometimes quantitatively) synthesised. This differs from a traditional review in which previous work is described but not systematically identified, assessed for quality and synthesised.

Advantages

There are two major advantages of systematic reviews (or 'meta-analysis'). Firstly, by combining data they improve the ability to study the consistency of results (that is, they give 'increased power'). This is because many individual studies are too small to detect modest but important effects (that is, they have insufficient power). Combining all the studies that have attempted to answer the same question considerably improves the statistical power.

Secondly, similar effects across a wide variety of settings and designs provide evidence of robustness and transferability of the results to other settings. If the studies are inconsistent between settings, then the sources of variation can be examined.

Thus while some people see the mixing of 'apples and oranges' as a problem of systematic reviews, it can be a distinct advantage because of its ability to enhance the generalisability and transferability of data.

Disadvantages

Without due care, however, the improved power can also be a disadvantage. It allows the detection of small biases as well as small effects. All studies have flaws, ranging from small to fatal, and it is essential to assess individual studies for such flaws. The added power of a systematic review can allow even small biases to result in an 'apparent' effect. For example, Schulz et al (1995) showed that unblinded studies gave, on average, a 17% greater risk reduction than blinded studies.

Method

A systematic review generally requires considerably more effort than a traditional review. The process is similar to primary scientific research and involves the careful and systematic collection, measurement, and synthesis of data (the 'data' in this instance being research papers). The term 'systematic review' is used to indicate this careful review process and is preferred to 'meta-analysis' which is usually used synonymously but which has a more specific

meaning relating to the combining and quantitative summarising of results from a number of studies.

It may be appropriate to provide a quantitative synthesis of the data but this is neither necessary nor sufficient to make a review 'systematic'.

A systematic review involves a number of discrete steps:

- question formulation;
- finding studies;
- appraisal and selection of studies;
- summary and synthesis of relevant studies;
- determining the applicability of results; and
- reviewing and appraising the economics literature.

Before starting the review, it is advisable to develop a protocol outlining the question to be answered and the proposed methods. This is required for all systematic reviews carried out by Cochrane reviewers (Mulrow and Oxman 1997).

Question formulation

Getting the question right is not easy. It is important to recognise that devising the most relevant and answerable question may take considerable time. Repeatedly asking 'why is this important to answer?' is helpful in framing the question correctly.

For example, are you really interested in the accuracy of the new test per se? Or would it be better to know whether or not the new test is more accurate than the current standard? If so, are you clear about what the current standard is?

Question formulation also involves deciding what type of question you are asking. Is it a question about an intervention, diagnostic accuracy, aetiology, prediction or prognosis, or an economic question? The multiple perspectives of health service providers, consumers and methodologists may be helpful in getting the question right.

Finding studies

The aim of a systematic review is to answer a question based on all the best available evidence — published and unpublished. Being comprehensive and systematic is important in this critical, and perhaps most difficult, phase of a systematic review. Finding some studies is usually easy — finding all relevant studies is almost impossible. However, there are a number of methods and resources that can make the process easier and more productive.

Appraisal and selection of studies

The relevant studies identified usually vary greatly in quality. A critical appraisal of each of the identified potentially relevant studies is therefore needed, so that those of appropriate quality can be selected. To avoid a selection that is biased by preconceived ideas, it is important to use a systematic and standardised approach to the appraisal of studies.

Summary and synthesis of relevant studies

Although a quantitative synthesis is often desirable, a comprehensive and clear summary of the high quality studies relevant to a particular question may be sufficient for synthesis and decision making. The initial focus should be on describing the study's design, conduct and results in a clear and simple manner — usually in a summary table. Following this, some summary plots are helpful, particularly if there are a large number of studies. Finally, it may be appropriate to provide a quantitative synthesis. However, as indicated above, this is neither a sufficient nor a necessary part of a systematic review.

Determining the applicability of results

Following the summary and synthesis of the studies, the next step is to ask about the overall validity, strength and applicability of any results and conclusions. How and to whom are the results of the synthesis applicable? How will the effects vary in different populations and individuals? This final step will be only briefly mentioned in this guide because it is the principal focus of another handbook in this series (*How to Use the Evidence: Assessment and Application of Scientific Evidence,* NHMRC 2000a).

Reviewing and appraising the economics literature

Review and appraisal of the economics literature is an essential step in conducting an economic evaluation for clinical practice guidelines. This specialist area of evaluation is essential in developing a clinical practice guideline and is dealt with in detail in a separate handbook (*How to Compare the Costs and Benefits: Evaluation of the Economic Evidence,* NHMRC 2000b).

How much work is a systematic review?

An analysis of 37 meta-analyses done by Allen and Olkin (1999) of MetaWorks, a company based in Massachusetts (USA) that specialises in doing systematic reviews, showed that the average hours for a review were 1139 (median 1110) — or about 30 person-weeks of full-time work — but this ranged from 216 to 2518 hours. The breakdown was:

- 588 hours for protocol development, searching and retrieval;
- 144 hours for statistical analysis;
- 206 hours for report writing; and
- 201 hours for administration.

However the total time depended on the number of citations. A systematic review has a fixed component, even if there were no citations, and a variable component, which increases with the number of citations. A regression analysis of the MetaWorks analyses gives a prediction of the number of hours of work as:

$$721 + 0.243x - 0.0000123x^2$$

where: x = number of potential citations before exclusion criteria were applied

About this handbook

The remainder of this handbook is divided into two parts:

- **Part 1** — includes general information on methods relevant to all systematic reviews irrespective of the type of question.
- **Part 2** — includes issues specific to six different question types:
  - effects of an intervention;
  - frequency or rate of a condition or disease;
  - diagnostic accuracy;
  - aetiology and risk factors;
  - prediction and prognosis; and
  - economics.

Appendixes C and D also include details of search procedures and a listing of available software.

Part 1

General methods

1 THE QUESTION

1.1 What types of questions can be asked?

In this handbook we will examine six types of health care questions associated with the following issues:

- *interventions:* 'What are the effects of an intervention?'
- *frequency or rate of a condition or disease:* 'How common is a particular condition or disease in a specified group in the population?'
- *diagnostic test performance:* 'How accurate is a sign, symptom, or diagnostic test in predicting the true diagnostic category of a patient?'
- *aetiology and risk factors:* 'Are there known factors that increase the risk of the
- *prediction and prognosis:* 'Can the risk for a patient be predicted?'
- *economics:* 'What are the overall costs of using the procedure?'

Answering each type of question requires different study designs, and consequently different methods of systematic review. A thorough understanding of the appropriate study types for each question is therefore vital and will greatly assist the processes of finding, appraising, and synthesising studies from the literature. A summary of the appropriate study types for each question and also the issues that are important in the appraisal of the studies is given in Table 1.1. A summary of the possible study types for questions of intervention effectiveness are shown in Table 1.2 and are described in more detail in another handbook in this series (*How to Use the Evidence: Assessment and Application of Scientific Evidence*, NHMRC 2000a). General information on how to find and review studies is given in the remainder of Part 1 with further details for each question type in Part 2.

1.1.1 Interventions

An intervention will generally be a therapeutic procedure such as treatment with a pharmaceutical agent, surgery, a dietary supplement, a dietary change or psychotherapy. Some other interventions are less obvious, such as early detection (screening), patient educational materials, or legislation. The key characteristic is that a person or their environment is manipulated in order to benefit that person.

To study the effects of interventions, it is necessary to compare a group of patients who have received the intervention (study group) with a comparable group who have not received the intervention (control group). A randomised controlled trial (RCT), which is a trial in which subjects are randomly allocated to the study or control group, is usually the ideal design.

Table 1.1 Types of clinical and public health questions, ideal study types and major appraisal issues

Question	Study types	Major appraisal issues
1. Intervention	Systematic review RCTs Cohort study Case-control study	Randomisation Follow-up complete Blinding of patients and clinicians
2. Frequency / rate (burden of illness)	Systematic review Cohort study Cross-sectional study	Sample frame Case ascertainment Adequate response / follow-up achieved
3. Diagnostic test performance	Systematic review Cross-sectional study (random or consecutive sample)	Independent, blind comparison with 'gold standard' Appropriate selection of patients
4. Aetiology and risk factors	Systematic review Cohort study Case-control study	Groups only differ in exposure Outcomes measurement Reasonable evidence for causation
5. Prediction and prognosis	Systematic review Cohort / survival study	Inception cohort Sufficient follow-up

RCT = randomised controlled trial

1.1.2 Frequency or rate

This question asks how common a particular feature or disease is in a specified group in the population. This is measured as the frequency (proportion, or prevalence) or rate (incidence) of the feature or disease; for example, the prevalence of osteoarthritis with ageing, or the rate of new cases of human immunodeficiency virus (HIV). The appropriate study design in this case is a cross-sectional survey with a standardised measurement in a representative (eg random) sample of people; for a rate, the sample would need to be followed over time. If, instead of a single frequency, we become interested in the causes of variation of that frequency, then this becomes a question of risk factors or prediction (see below).

Table 1.2 Types of studies used for assessing clinical and public health interventions (question 1 in Table 1.1)

Study design	Protocol
Systematic review	Systematic location, appraisal and synthesis of evidence from scientific studies.
Experimental studies	
Randomised controlled trial	Subjects are randomly allocated to groups either for the intervention/treatment being studied or control/placebo (using a random mechanism, such as coin toss, random number table, or computer-generated random numbers) and the outcomes are compared.
Pseudorandomised controlled trial	Subjects are allocated to groups for intervention/treatment or control/placebo using a nonrandom method (such as alternate allocation, allocation by days of the week, or odd–even study numbers) and the outcomes are compared.
Clustered randomised trial	Subjects are randomised to intervention or control in groups (eg families, communities, hospitals).
Comparative (nonrandomised and observational) studies	
Concurrent control or cohort	Outcomes are compared for a group receiving the treatment/intervention being studied, concurrently with control subjects receiving the comparison treatment/intervention (eg usual or no care).
Case-control	Subjects with the outcome or disease and an appropriate group of controls without the outcome or disease are selected and information is obtained about the previous exposure to the treatment/intervention or other factor being studied.
Historical control	Outcomes for a prospectively collected group of subjects exposed to the new treatment/intervention are compared with either a previously published series or previously treated subjects at the same institutions.
Interrupted time series	Trends in the outcome or disease are compared over multiple time points before and after the introduction of the treatment/intervention or other factor being studied.

Table 1.2 (contd)

Study design	Protocol
Other observational studies	
Case series	A single group of subjects are exposed to the treatment/intervention.
– post-test	Only outcomes after the intervention are recorded in the case series, so no comparisons can be made.
– pretest/post-test	Outcomes are measured in subjects before and after exposure to the treatment/intervention for comparison (also called a 'before-and-after' study).

Note: A definition of cross-sectional study, which is not included here as it is not useful for assessing intervention studies, is given in the Glossary.

1.1.3 Diagnostic test performance

If there is good randomised trial evidence that an intervention for a particular condition works then it may be necessary to assess how accurately the condition can be diagnosed from a sign, symptom, or diagnostic test. To do this, the ideal study design is a representative sample of people in whom the new test is compared with an appropriate 'gold standard' or reference standard (cross-sectional study). The most commonly used measures of test performance are the sensitivity and specificity of the test.

If we move from an interest in test performance to an interest in the effects on patient outcomes, then the question becomes one of intervention (that is, the effects on patients of using or not using the test, as is the case for population screening). However, diagnostic test performance can often be used as a surrogate to predict the benefits to patients.

1.1.4 Aetiology and risk factors

This type of question is concerned with whether a particular factor, such as patient characteristic, laboratory measurement, family history, etc, is associated with the occurrence of disease or adverse outcomes. To answer this question a clear association between the factor and the disease must first be established. The most appropriate study types are a long-term follow-up of a representative inception cohort or an approximation to this through sampling for a case-control study (cohort or case-control study).

If a clear association is shown, the next stage is to determine whether that association is causal, that is, whether the factor under consideration causes the disease or outcome of interest or is merely associated with it for other reasons.

This involves issues beyond the degree of association, such as the dose–response relationship and biological plausibility.

1.1.5 Prediction and prognosis

This question seeks to determine the risk to the person by putting together several risk factors and using the combined information to decide the level of risk to the person. Unlike the question of aetiology, causation is not so crucial. Strongly predictive risk markers are also useful. The most appropriate study type is a long-term follow-up of a representative inception cohort (cohort or survival study).

1.1.6 Economics

In all of the previous questions, one of the outcomes of interest is often the cost. For example, the costs of an intervention and potential downstream cost may be offset by improved patient outcomes with reduced need for further medical treatment.

The issues of economic evaluation and cost-effectiveness are discussed briefly in Section 6.5 of this handbook and in greater detail in another handbook in this series (*How to Compare the Costs and Benefits: Evaluation of the Economic Evidence*, NHMRC 2000b).

1.2 What is the relevant question?

A well-formulated question generally has three parts:

- the study factor (eg the intervention, diagnostic test, or exposure);
- the population (the disease group or a spectrum of the well population); and
- the outcomes.

Since we will often be interested in all outcomes, the first two parts of the question may be sufficient (see Section 2.2).

1.3 How focused should the question be?

The question should be sufficiently broad to allow examination of variation in the study factor (eg intensity or duration) and across populations. For example:

> '*What is the mortality reduction in colorectal cancer from yearly faecal occult blood screening in 40–50-year-old females?*' is too narrow.

However:

> '*What is the effect of cancer screening on the general population?*' is clearly too broad and should be broken down into cancer-specific screening questions.

A better question may be:

> '*What is the mortality reduction in colorectal cancer from faecal occult blood screening in adults?*' which allows the effects of screening interval, age group and gender to be studied.

2 FINDING RELEVANT STUDIES

Finding all relevant studies that have addressed a single question is not easy. There are currently over 22,000 journals in the biomedical literature. MEDLINE indexes only 3700 of these, and even the MEDLINE journals represent over 200 metres of journals per year.

Beyond sifting through this mass of literature, there are problems of duplicate publications and accessing the 'grey literature', such as conference proceedings, reports, theses and unpublished studies. A systematic approach to this literature is essential in order to identify all the best evidence available that addresses the question.

As a first step, it is helpful to find out if a systematic review has already been done or is under way. If not, published original articles need to be found.

2.1 Finding existing systematic reviews

Published reviews may answer the question, or at least provide a starting point for identifying the studies. Finding such reviews takes a little effort. A general MEDLINE search strategy by Hunt and McKibbon (1997) relevant to all question-types is given in Appendix C. However, for interventions, a check should also be made of the Cochrane Library for a completed Cochrane review, a Cochrane protocol (for reviews under way), or a nonCochrane review (in the Database of Abstracts and Reviews [DARE] in the Cochrane Library, compiled by the Centre for Reviews and Dissemination at York (United Kingdom).

2.2 Finding published primary studies

It is usually easy to find a few relevant articles by a straightforward literature search, but the process becomes progressively more difficult as we try to identify additional articles. Eventually, you may sift through hundreds of articles in order to identify one further relevant study.

There are no magic formulae to make this process easy, but there are a few standard tactics, which, together with the assistance of a librarian experienced in the biomedical literature, can make your efforts more rewarding.

2.2.1 Breaking down the study question into components

A central tactic is to take a systematic approach to breaking down the study question into components using a Venn diagram. The Venn diagram for the

question *What is the mortality reduction in colorectal cancer from faecal occult blood screening in adults?* is shown in Figure 2.1.

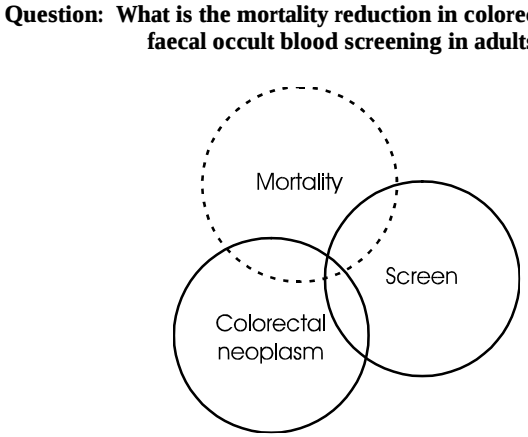


Figure 2.1 Venn diagram for colorectal screening

Once the study question has been broken into its components, they can be combined using 'AND' and 'OR'. For example, in Figure 2.1:

> **(mortality AND screen)** — represents the overlap between these two terms — retrieves only articles that use both terms.
>
> **(screen AND colorectal neoplasm AND mortality)** — represents the small area where all three circles overlap — retrieves only articles with all three terms.

Complex combinations are possible. For example, the following combination captures all the overlap areas between the circles:

> **(mortality AND screen) OR (mortality AND colorectal neoplasms) OR (screen AND colorectal neoplasms)**

Although the overlap of all three parts will generally have the best concentration of relevant articles, the other areas may still contain many relevant articles. Hence, if the disease AND study factor combination (solid circles in figure) is manageable, it is best to work with this and not further restrict by, for example, using outcomes (dotted circle in figure).

2.2.2 Use of synonyms

When the general structure of the question is developed it is worth looking for synonyms for each component. This process is illustrated in Table 2.1.

Table 2.1 Using synonyms of components of the three-part question to devise a literature search

Question: What is the mortality reduction in colorectal cancer from faecal occult blood screening in adults?		
Question part	**Question term**	**Synonyms**
Population/setting	Adult, human	–
Study factor	Screening, colorectal cancer	Screen, early detection, bowel cancer
Outcome [a]	*Mortality*	*Death*, survival*
Ideal design [a]	*Methodological terms*	

* = wildcard symbol (finds words with the same stem)

[a] Both outcome and design are options needed only when the search results are unmanageable

Thus a search string might be:

(screen* OR early detection) AND (colorectal cancer OR bowel cancer) AND (mortality OR death* OR survival)

The term 'screen*' is shorthand for words beginning with screen, for example, screen, screened, screening. (Note: the 'wildcard' symbol varies between systems, eg it may be an asterisk [*], or colon [:].)

In looking for synonyms you should consider both textwords and keywords in the database. The MEDLINE keyword system, known as MeSH (Medical Subject Heading), has a tree structure that covers a broad set of synonyms very quickly. The 'explode' (exp) feature of the tree structure allows you to capture an entire subtree of MeSH terms within a single word. Thus for the colorectal cancer term in the above search, the appropriate MeSH term might be:

colonic neoplasm (exp)

with the 'explode' incorporating all the MeSH tree below colonic neoplasm, viz:

colorectal neoplasms
- **colonic polyps**
  - **adenomatous polyposis coli**
- **colorectal neoplasms**
  - **colorectal neoplasms, hereditary nonpolyposis**
- **sigmoid neoplasms**

While the MeSH system is useful, it should supplement rather than usurp the use of textwords so that incompletely coded articles are not missed.

2.2.3 Snowballing

The process of identifying papers is an iterative one. It is best to devise a strategy on paper initially, as illustrated in Table 2.1. However, this will inevitably miss useful terms, and the process will need to be repeated and refined. The results of the initial search are used to retrieve relevant papers, which can then be used in two ways to identify missed papers:

- the bibliographies of the relevant papers can be checked for articles missed by the initial search; and

- a citation search, using the Science Citation Index,[1] can be conducted to identify papers that have cited the identified relevant studies, some of which may be subsequent primary research.

These 'missed' papers are invaluable — they provide clues on how the search may be broadened to capture further papers (eg by studying the MeSH keywords that have been used). The whole procedure may then be repeated using the new keywords identified. This iterative process is sometimes referred to as 'snowballing'.

2.2.4 Handsearching

If the relevant articles appear in a limited range of journals or conference proceedings, it may be feasible and desirable to search these by hand. This is obviously more important for unindexed or very recent journals, but may also pick up relevant studies not easily identified from title or abstracts. Fortunately, the Cochrane Collaboration is systematically handsearching a number of journals to identify controlled trials and a master list is maintained on the Internet.[2] This should be checked before undertaking your own handsearch. However, for other question and study types there has been no such systematic search.

2.2.5 Methodological terms

MEDLINE terms cover not only specific content but also a number of useful terms on study methodology. For example, if we are considering questions of therapy, many randomised trials are tagged in MEDLINE by the specific methodological term:

[1] www.isinet.com/products/basic

[2] www.cochrane.org/

randomized-controlled-trials[3] **in [publication type]**

or as:

controlled-clinical trials in [publication type]

However, many studies do not have the appropriate methodological tag. The Cochrane Collaboration and the United States National Library of Medicine (NLM) are working on correctly retagging the controlled trials, but this is not so for other study types.

2.2.6 Methodological filters

An appropriate 'methodological filter' may help confine the retrieved studies to primary research. For example, if you are interested in whether screening reduces mortality from colorectal cancer (an intervention), then you may wish to confine the retrieved studies to controlled trials. The idea of methodological terms may be extended to multiple terms that attempt to identify particular study types. One very useful tool for a noncomprehensive but good initial search is available using the NLM's free Internet version of MEDLINE PubMed — the Clinical Queries section,[4] which has inbuilt search filters based on methodological search techniques developed by Haynes et al (1994). The filters are described in Appendix C. They offer four study categories (aetiology, prognosis, treatment, diagnosis) and the choice of emphasising sensitivity or specificity in the search. Other methodological filters are discussed in Part 2 for each type of question.

2.2.7 Use of different databases

There are a number of other databases apart from MEDLINE; selection depends on the content area and the type of question being asked. For example, there are databases for nursing and allied health studies such as CINHAL and for psychological studies such as Psyclit. If it is a question of intervention, then the Controlled Trials Registry within the Cochrane Library is a particularly useful resource. This issue is further discussed in the specific question types in Part 2 of this handbook.

[3] Use of 'ize' spelling of randomised is necessary when using MEDLINE

[4] www.ncbi.nlm.nih.gov/PubMed/clinical.html

2.3 Finding unpublished primary studies

To reduce publication bias (see Section 2.4), it is important to search for unpublished studies. There are two approaches to finding unpublished studies: searching relevant databases and contacting experts.

2.3.1 Searching relevant databases

An appendix in the Cochrane Handbook (available on the Cochrane Library CD) contains a list of about 30 clinical trials registries with completed and ongoing studies registered in specialised areas such as acquired immune deficiency syndrome (AIDS) and cancer.

For other question types, information will be more difficult to find, but any available databases should be checked — in particular, research funding bodies may be able to provide a list of research. However, this has rarely been systematically compiled outside controlled trials. An exception is the International Agency for Research on Cancer (IARC) bibliography of ongoing cancer epidemiology research (Sankaranarayanan et al 1996).

2.3.2 Writing to experts

Another option is to contact the principal investigators of relevant studies directly, asking whether they know of additional studies.

However, the usefulness of writing to experts varies. An analysis of a recent review of the value of near-patient testing (that is, diagnostic tests that can be done entirely at the clinic, such as dipstick urine tests) (McManus et al 1998) showed that of 75 papers eventually identified, nearly one-third were uniquely identified by contacting experts. The data are shown in Figure 2.2, which also illustrates the general point that it is worth using multiple sources. However, near-patient testing is an area of emerging technology, and a larger proportion than usual of papers were possibly unpublished, published in less common sources, or presented at conferences.

2.4 Publication bias — a crucial problem

2.4.1 What is publication bias?

If ‘positive’ studies are more likely to be published than ‘negative’ studies then any review (traditional or systematic) of the published literature must be biased towards a ‘positive’ result. This is the essence of publication bias — the positive correlation between the results of the study and our ability to find that study. For example, a follow-up of 737 studies approved by the Institutional Review Board at Johns Hopkins University found that the odds ratio for the likelihood

of publication of positive compared with negative studies was 2.5 (Dickersin et al 1992). Interestingly, most nonpublication was because authors failed to submit, rather than because journals rejected 'negative' studies (Stern and Simes 1997).

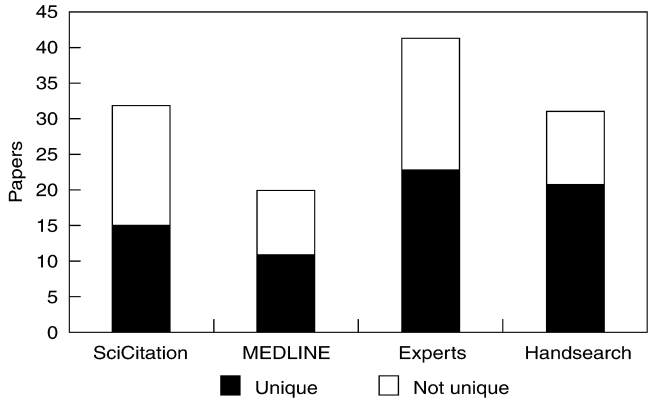


Figure 2.2 Papers identified by different search methods in a systematic review of near-patient testing

2.4.2 Does this affect the results of the reviews?

Systematic exclusion of unpublished trials from a systematic review introduces bias if the unpublished studies differ from the published, for example, because of the statistical significance or the direction of results. In a review of multiagent versus single agent chemotherapy for ovarian cancer, Simes (1987) found statistically and clinically different results for 16 published studies and 13 studies that had been registered in a clinical trials register, some of which were published and some not (see Table 2.2). Since the registered trials were registered at inception rather than completion, their selection for inclusion in the review is not influenced by the outcome of the study — therefore they constitute an incomplete *but unbiased* set of studies.

2.4.3 What can we do about publication bias?

It is vital that eventually all clinical trials are registered at their inception so that systematic reviews and recommendations about therapy can be made on the basis of all relevant research, and not a biased subset. In the meantime we must settle for making our best efforts at retrieving the grey literature.

Table 2.2 Comparison of published and registered studies for multiagent versus single agent chemotherapy for ovarian cancer

	Published studies	Registered studies[a]
Number of studies	16	13
Survival ratio	1.16	1.05
95% confidence interval	1.06–1.27	0.98–1.12
Probability (*P*-value)	0.02	0.25

[a] Studies registered in a clinical trial registry at initiation (ie before the results were known)
Source: Simes (1987)

These methods include using existing clinical trials registries (a list of registries and their contact details are available in the Cochrane Handbook in the Cochrane Library), scanning major conference proceedings, and contacting experts and researchers working in the area of a particular question to ask if they know of other relevant published or unpublished research. In Section 4, on synthesis of the studies, we will also describe some methods of identifying the potential significance of publication bias based on the identified studies, such as 'funnel' plots. However, these only help to diagnose the problem of bias, not to cure or prevent it.

2.4.4 Duplicate publications

The converse of an unpublished study is a study that is published several times. This is often, but not always, obvious. For example, in a review of the effect of the drug ondansetron on postoperative vomiting, Tramer et al (1997) found 17% of trials had duplicate reports. Nine trials of oral ondansetron were published as 16 reports, and 19 trials of intravenous ondansetron were published as 25 reports. One multicentre trial had published four separate reports with different first authors. Most surprisingly, four pairs of identical trials had been published that had nonoverlapping authorships!

Unfortunately, there is no simple routine means of detecting such duplicates except by some careful detective work. Occasionally, it will be necessary to write to the authors. Clearly, if duplicate publications represent several updates of the data, then the most recent should be used.

3 APPRAISING AND SELECTING STUDIES

Readers will naturally wish to know how good the reviewed research is and why you have excluded some studies that address the question at issue. In both situations you need to explain your judgments, which will usually be based on your assessment of study quality and applicability.

3.1 Standardising the appraisal

Providing an explicit and standardised appraisal of the studies that have been identified is important for two reasons. First, a systematic review should try to base its conclusions on the highest quality evidence available. To do this requires a valid and standardised procedure to select from the large pool of studies identified so that only the relevant and acceptable quality studies are included in the review. Second, it is important to convey to the reader the quality of the studies included as this indicates the strength of evidence for any recommendation made.

3.1.1 What study features should be assessed?

Overall the study features that are most important to assess are those that involve selection and measurement bias, confounding, and follow-up of participants. In Part 2 these features are examined for each question type under the following headings:

A. Has selection bias (including allocation bias in RCTs) been minimised?

B. Have adequate adjustments been made for residual confounding?

C. Was follow-up for final outcomes adequate?

D. Has measurement or misclassification bias been minimised?

3.1.2 Is it important to have a structured appraisal?

If unstructured appraisals are made, there is a tendency to look more critically at the studies whose conclusions we dislike. For example, 28 reviewers were asked to assess a single (fabricated) 'study' but were randomly allocated to receive either the 'positive' or 'negative' version (Mahoney 1977). The identical methods section of these fabricated versions was rated significantly worse by the reviewers of the 'negative' study compared with the 'positive' study. Hence, it is essential to appraise all papers equally. This can be done in part by using a standardised checklist. Part 2 of this handbook outlines the important appraisal

issues for the different question types outlined in Section 1.1 and shows specific checklists for some of the question types.

These standardised checklists allow assessment of how important measurement and selection biases were avoided.

3.1.3 How many reviewers are required?

Using more than one reviewer is rather like getting a second opinion on a medical diagnosis. Because of the importance of appropriately selecting studies, at least two reviewers should be used. Each reviewer should independently read and score each of the studies that can potentially be included in the review. They should then meet to resolve any discrepancies between the scoring of the paper by open discussion about their justification for each of the scores. This discussion is a useful educational procedure in itself, which probably increases the consistency and accuracy of the appraisals of the paper.

3.1.4 Is it necessary to do the appraisal 'blind' to the outcome of the study?

Some meta-analysts have suggested that all appraisals should be done blind to the results of the individual study. This requires removing identification of the authors and journal, and all reference to any results from the paper. Generally, the methods and the results section of the paper are sufficient to provide the information necessary for the appraisal (with the explicit outcomes 'blackened

However, this approach is very time consuming. The effect has been examined in two empirical studies, which suggest that the benefit, if any, in bias reduction by using the blinding process is small (Berlin 1997). At present there is not a consensus about whether the gain is worth the effort. However, for particularly controversial and important issues, such a blinded appraisal should be considered.

3.2 Using the quality appraisal

The first and most important use of the quality appraisal will be to decide whether the study is included at all in the main analysis. For example, with a question of treatment, only RCTs may be selected. Deciding whether a study is randomised or not can be difficult, and hence it is very valuable to have reviewers to look carefully at the paper and come to a conclusion about this.

After the decision to include or exclude the study has been made, there are three further uses for the appraisal scores or 'quality weights', as follows.

Grouping or sorting by design and/or quality [RECOMMENDED]

It is useful to consider an exploratory analysis on the design or quality features of studies. Studies can be categorised by design (eg randomised, cohort, case-control) or by important quality features (eg blinded versus unblinded) and then plotted in subgroups, with or without summary estimates provided for each of these design or quality groups. Does this make a difference to the results seen? For example:

- Do the blinded studies give different results to the unblinded studies?
- Do the studies with good randomisation procedures give different results from those with doubtful randomisation procedures?

A sensitivity analysis on quality has been suggested by Detsky et al (1992): a cumulative meta-analysis is done looking at the best single study, the best two single studies combined, the best three studies combined, etc. However, recent empirical work (Juni et al 1999) showed that different summary quality scores give highly inconsistent results. Since flaws in one feature, such as follow-up, may not give a similar size or direction of bias to another design feature, such as blinding, analysing summary scores is problematic. Hence, we suggest the main focus should be on individual quality features.

Meta-regression on quality items [OPTIONAL]

It is possible to extend this further by looking at all the features of quality simultaneously in a so-called meta-regression. However, because there will usually be a limited number of studies, such techniques are probably not justified in most meta-analyses.

Weighting by quality [NOT RECOMMENDED]

Some analysts have suggested using the quality score to weight the contribution of particular studies to the overall estimate. This is inappropriate — it neither adjusts for nor removes the bias of poor studies, but merely reduces it slightly.

Further information on the appraisal for each question type is given in Part 2. The Journal of the American Medical Association 'Users' Guides' series (Guyatt and Rennie 1993; Guyatt et al 1993,1994) is also a good source of further information.

4 SUMMARISING AND SYNTHESISING THE STUDIES

4.1 Presenting the results of the studies (data extraction)

4.1.1 Tabular summary

It is helpful to produce tabular and graphical summaries of the results of each of the individual studies. An example of a summary table for an intervention question is shown in Part 2 of this handbook (Section 6).

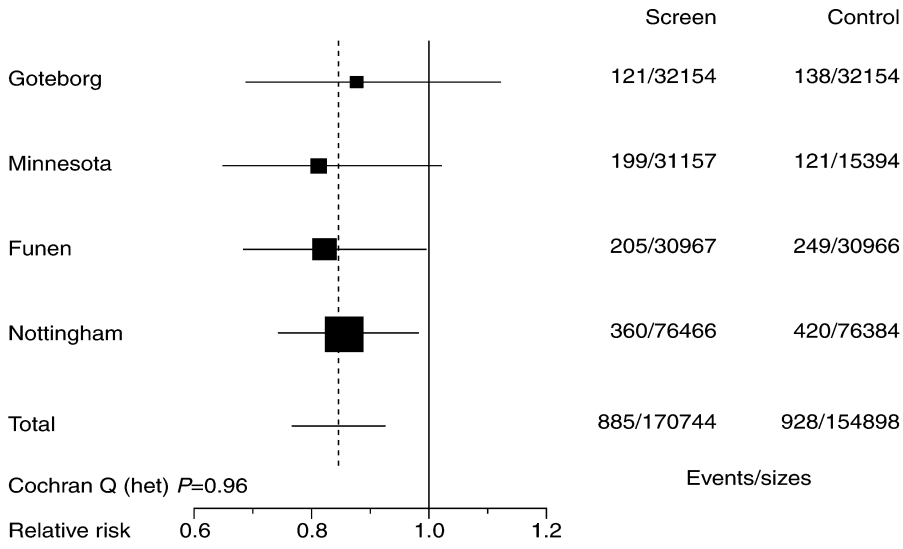
4.1.2 Graphical presentation

The most common and useful graphical presentation of the results of individual studies is a point estimate plot with the 95% confidence interval (CI) for each study (known as a 'forest plot'). A value of less than 1.0 indicates that the intervention studied is beneficial. A forest plot can be done for the relative risk reduction or a specific measure such as reduction in blood pressure. Studies should be sorted from those with the broadest to those with the narrowest confidence interval. If there is a summary estimate, this should be nearest the studies with the narrowest confidence intervals.

In addition, because studies with broad confidence intervals draw greater visual attention, it is useful to indicate the contribution of the study visually by the size of the symbol at the summary estimate. The area of the symbol should be made proportional to the precision of the study (more specifically to the inverse of the variance of the study's estimate). This means that the diameter of each symbol is proportional to the inverse of the standard error of the study's estimate. These principles are illustrated in Figure 4.1, which shows the results of the systematic review of colorectal cancer screening (Towler et al 1998).

4.2 Synthesis of study results

Except in rare circumstances, it is not advisable to pool the results of the individual studies as if they were one common large study. This can lead to significant biases because of confounding by the distribution of the study factor and the outcome factor.



Notes:
Dotted vertical line = the combined estimate
Total = 95% confidence interval of the combined estimate

Figure 4.1 Relative mortality from colorectal cancer in screened versus unscreened (control) groups from four randomised trials of faecal occult blood screening (from Towler et al 1998)

However, if the studies are considered sufficiently homogeneous in terms of the question and methods, and this is supported by a lack of evidence of statistical heterogeneity (see Section 4.3), then it may be appropriate to combine the results to provide a summary estimate. The method for combining studies will vary depending upon the type of questions asked and the outcome measures used. Outcome measures are described in detail in another handbook in this series (*How to Use the Evidence: Assessment and Application of Scientific Evidence*, NHMRC 2000a) and are summarised in Table 4.1. A further measure, the summary receiver–operator curve (ROC) is a measure of diagnostic test accuracy and is described in Section 8.

An estimate of the effect for each individual study should be obtained, along with a measure of random error (variance or standard error). The individual studies can then be combined by taking a weighted average of the estimates for each study, with the weighting being based on the inverse of the variance of each study's estimate. For example, Figure 4.1 shows, for colorectal cancer screening, the combined estimate (the dotted vertical line) and its 95% CI (the horizontal line marked 'total').

Table 4.1 Some possible outcome measures of study effects

Outcome measures	Description
Continuous outcomes	
Difference between group means	Difference between treatment and control groups in mean values of outcome variable.
Standardised difference	Differences between the treatment and control group means for each study, standardised by an estimate of the standard deviation of the measurements in that study. This removes the effect of the scale of measurement, but can be difficult to interpret.
Weighted difference in means	Average (pooled) difference between treatment and control groups in mean values across a group of studies using the same scale of measurement for the outcome (eg blood pressure measured in mm Hg).
Standardised weighted mean difference	Average (pooled) standardised difference between treatment and control groups across a group of studies, where the outcome was measured using different scales with no natural conversion to a common measure (eg different depression scales or different quality-of-life instruments).
Binary outcomes	
Risk difference (RD)	Difference (absolute) between treatment and control groups in the proportions with the outcome. If the outcome represents an adverse event (such as death) and the risk difference is negative (below zero), this suggests that the treatment reduces the risk. In this situation the risk difference, without the negative sign, is called the *absolute risk reduction*.
Relative risk or risk ratio (RR)	Ratio of the proportions in the treatment and control groups with the outcome. This expresses the risk of the outcome in the treatment group relative to that in the control group. If the relative risk is below 1, an adverse outcome, this suggests that the treatment reduces the risk, and its complement (1 – relative risk) or *relative risk reduction* is often used.
Odds ratio (OR)	Ratio of the odds of the outcome in the treatment group to the corresponding odds in the control group. Again, for an adverse outcome, an odds ratio below 1 indicates that the treatment reduces the risk. In some studies (eg population-based case-control studies) the odds ratio is a reasonable estimate of the relative risk. It is not a good estimate when the outcome is common or is measured as a prevalence.
Hazard ratio (HR)	Ratio of the hazards in the treatment and control groups (when time to the outcome of interest is known); where the hazard is the probability of having the outcome at time t, given that the outcome has not occurred up to time t. Sometimes, the hazard ratio is referred to as the *relative risk*. For an adverse outcome, a hazard ratio less than unity indicates that the treatment reduces the risk.
Number needed to treat (NNT)	Number of patients who have to be treated to prevent one event. It is calculated as the inverse of the risk difference without the negative sign (NNT = 1/RD). When the treatment increases the risk of the harmful outcome, then the inverse of the risk difference is called the number needed to harm (NNH = 1/RD).

Note: Further discussion of outcome measures is given in the handbook *How to Use the Evidence: Assessment and Application of Scientific Evidence* in this series (NHMRC 2000a).

Although this principle is straightforward, a number of statistical issues make it more complicated. For example, the measures of effect have to be on a scale that provides an approximate normal distribution to the random error (eg by using the log odds ratio rather than just the odds ratio). Allowance must also be made for zeros in the cells of tables cross-classifying the study factor and the outcome factor, or outliers in continuous measurements. Most of the available meta-analysis software provides such methods (see Appendix D for examples of available software). Details of the properties of the various alternative statistical methods are given in Rothman and Greenland (1998). This handbook addresses only the general principles.

Further details of methods of synthesis for the different question types are given in Part 2. There is no single source of information for statistical methods of synthesis. The most comprehensive book currently available is the *Handbook of Research Synthesis* (Cooper and Hedges 1994), which is particularly strong on synthesis of studies but also covers finding and appraising studies.

4.2.1 Fixed and random effects estimates

Two major categories of summary estimates are the fixed and random effects estimates. That is, is the true value a single value or does it vary across populations and circumstances?

- A *fixed effect model* assumes that there is a single 'true' value, which all studies are attempts to measure but with some imprecision; the fixed effect summary is a weighted average with weights proportional only to each study's precision.

- A *random effects model* assumes that the 'true' value varies and attempts to incorporate this variation into the weightings and the uncertainty around the summary estimate. To do this, the model first estimates the underlying study-to-study variation (which is often designated as 'tau'), which is then included in the weighting for each study.

Mathematically, the fixed effects weights are $1/d^2$ (where d^2 is the variance of the studies estimate); the random effects weights are $1/(d^2 + tau^2)$. From this we can see that:

- if between-study variance is small (tau is near 0) then fixed and random effects models are similar; and

- if the between-study variance is large (tau is much greater than d) then the weights for each study become almost equal.

So which model should be used? This is best answered indirectly: if there is minimal between-study variation, the choice does not matter; but if there is considerable between-study variation then an explanation should be sought.

If no cause for the variation is found, then, although both models offer information, neither model is clearly 'correct'. The fixed effects model assumes no variation when it demonstrably exists. The random effects model assumes the studies are a representative (or random) sample for the population of situations to which the results will be applied — a fairly unlikely assumption. So the emphasis should be not on incorporating variation but on explaining it, which is discussed further in Section 4.3. However, if the variation cannot be explained, then, if pooling is still relevant, both fixed and random effects models should be presented.

4.3 Assessing heterogeneity

The variation between studies is often considered a weakness of a systematic review but, if approached correctly, it can be a considerable strength. If the results are consistent across many studies, despite variation in populations and methods, then we may be reassured that the results are robust and transferable. If the results are inconsistent across studies then we must be wary of generalising the overall results — a conclusion that a single study cannot usually reach. However, any inconsistency between studies also provides an important opportunity to explore the sources of variation and reach a deeper understanding of its causes and control (Thompson 1995).

The causes of a variation in results may be due to personal factors such as gender or genes, disease factors such as severity or stage, variation in the precise methods of the intervention or diagnostic test, differences in study design or conduct, such as duration and completeness of follow-up, or the quality of measurements.

4.3.1 Measures of heterogeneity

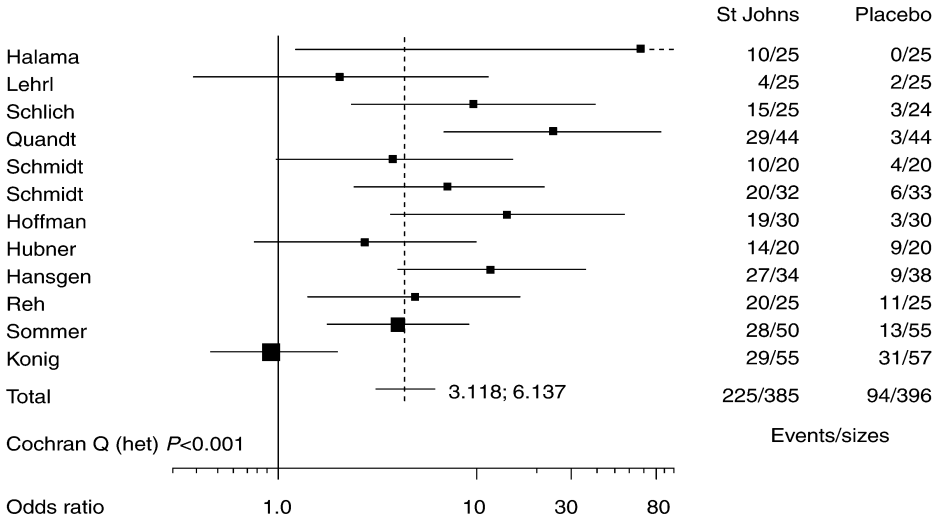
Generally, statistical tests of heterogeneity have low power. Some variation is inevitable, and we are really more interested in the degree and causes of variation. The best current measure of the degree of variation is the between-study variance (or tau), which is estimated when fitting a random effects model. This has the advantage of being in the same 'units' as the results measure; for example, if the meta-analysis looked at weight change in kilograms, then the tau is the between-study variance in kilograms. An alternative is to test for heterogeneity using the Cochran chi-square (Cochran Q) and Q divided by the degrees of freedom (df), where values greater than 1 indicate heterogeneity, as follows.

- *Definite heterogeneity.* If the Cochran Q is statistically significant, heterogeneity must be explored. If it cannot be explained, the significant heterogeneity must be clearly stated.

- *Possible heterogeneity.* If the Cochran Q is not statistically significant but Q/df is greater than 1, it is still important to explore heterogeneity.

- *No heterogeneity.* If the Cochran Q is not statistically significant and Q/df is less than 1, important heterogeneity is very unlikely.

Figure 4.2 shows the results of 12 placebo-controlled trials of the effect of taking St John's wort (*Hypericum perforatum*) on depression (Linde et al 1996). The Cochran Q is 34.7; an alternative chi-square (Breslow and Day 1987) is 37.9. Since there are 11 degrees of freedom (df), the Q/df ratio is 3.2 (34.7/11), indicating important heterogeneity. The *P*-value for a chi-square of 34.7 on 11 df is <0.001.



Notes:
Dotted vertical line = the combined estimate
Total = 95% confidence interval of the combined estimate
Cochran Q (odds ratio) of 34.7 on 11 degrees of freedom (df) gives $P < 0.001$

Figure 4.2 Meta-analysis of 12 placebo-controlled trials of St John's wort for depression, showing significant heterogeneity

Even without the heterogeneity test, the graph is suspicious because the confidence interval of the largest trial (Konig) does not overlap with the confidence interval of the summary estimate.

Before exploring other sources of variation, it is important to consider whether variation may be an artefact of the outcome measure. For example, is the

'effect' a proportional or absolute effect? If it is proportional, then measures such as the relative risk (or odds ratio or hazard ratio) or the percentage reduction (eg in cholesterol or blood pressure) will be appropriate. If it is absolute, then absolute risk or absolute risk reduction (risk difference) may be appropriate measures.

This question is partly biological and partly empirical. In a recent analysis of 115 meta-analyses, it was found that the absolute risk was clearly inappropriate in 30% of studies; the relative risk fared better but was still clearly inappropriate in 13% of studies (Schmid et al 1998). Hence, an initial check of the appropriateness of the common scale used is essential. In the St John's wort example, the Cochran Qs were: 34.7 for the odds ratio, 39.0 for the relative risk, and 41.6 for the risk difference. Hence, the odds ratio minimises the Q, and appears the best choice, but clearly important heterogeneity remains.

The ideal way to study causes of true biological variation (or 'effect modification') is within rather than between studies, because the variation in incidental design features confounds our ability to look at true causes of effect modification (Gelber and Goldhirsch 1987). For example, if there was one study in older men and one in younger women, then the effect of gender is confounded by the effect of age. If there was one short-term study in Caucasians and one long-term study in Chinese, then the effect of ethnicity is confounded by study duration. Looking across studies can provide a useful initial exploratory analysis, but confirmation by combining the within-studies analysis across all studies is then desirable (see information on individual patient data meta-analysis, below).

In general, the approach to subgroup analysis and effect modification should be to assume similarity unless a difference can be demonstrated. Thus individual subgroups should NOT be tested for significance of their main effects, but should be tested for interaction to see whether the subgroups differ significantly.

The problem is illustrated in Figure 4.3, which shows a hypothetical study that is clearly statistically significant overall (the confidence interval does not cross the relative risk of 1.0). If this is now split into two subgroups (1 and 2, which each have the identical estimate), group 1 is no longer statistically significant. The correct approach here is to first test whether groups 1 and 2 are significantly different. In this case, where their point estimates are the same, it is clear that they will not differ significantly.

4.3.2 Individual patient data meta-analysis

Obtaining the original data from each study makes a number of analyses possible that are difficult or impossible if based only on summary measures from each study.

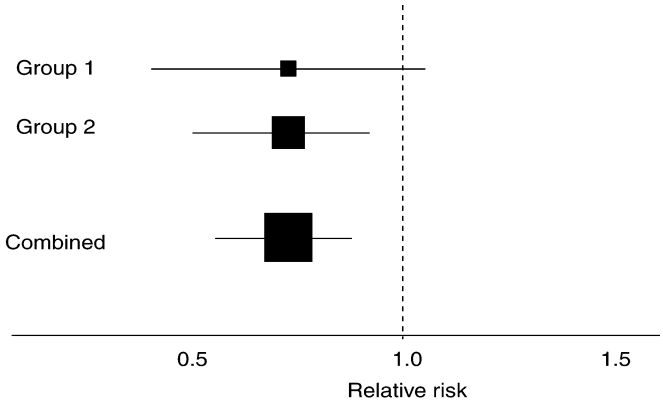


Figure 4.3 Hypothetical study showing combined and subgroup analysis: subgroups 1, 2 and the combined effect are all equivalent, but only group 2 and the combined groups are statistically significant

For example, combined survival analysis is best done using individual patient data (Whitehead and Whitehead 1991). As mentioned above, the ideal approach to subgroup analysis is using individual patient data. However, this usually entails much more work and collaboration, and may not be feasible. Such pooling of trial data has worked best when there is an ongoing collaboration between the trialists involved (EBCTCG 1992).

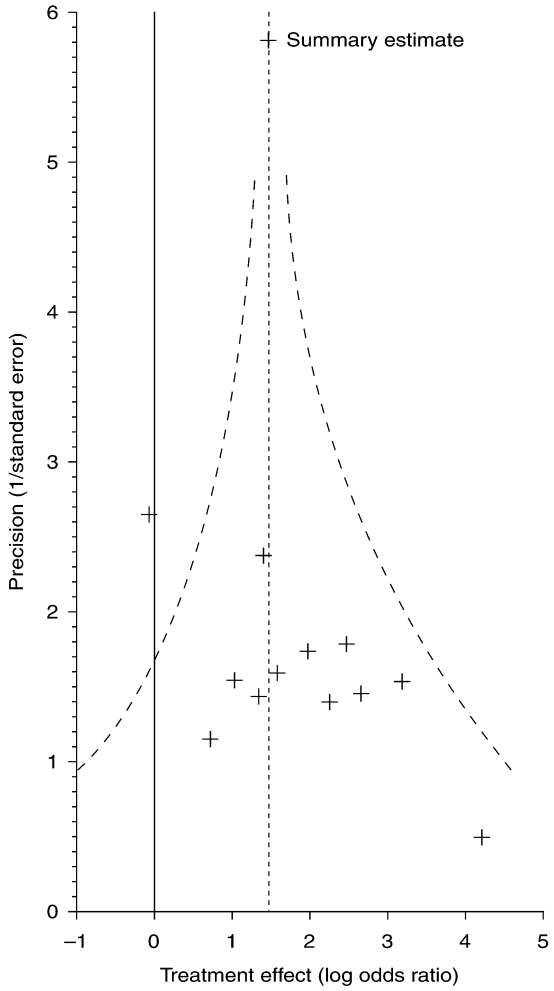
4.4 Detecting publication bias

Publication bias is best avoided by improved literature searching and use of study registries (see Section 2.4). However, there are some useful diagnostic plots and statistics available that can help detect, and to some extent adjust for, publication bias.

4.4.1 Funnel plots

Smaller single-centre trials are less likely to be published, as these are more likely to be 'negative' (not statistically significant). This may be made apparent from a funnel plot that plots the size of the treatment effect against the precision of the trial (1/standard error), which is a statistical measure of the size of the study that takes into account study numbers, duration, etc. Without publication bias, this plot should be funnel shaped — the 'neck' of the funnel showing little spread among the larger trials, and the base of the funnel showing a wider spread among the smaller trials. With publication bias, one tail or other of the funnel may be weak or missing because the small negative trials are not present.

This may be the case in Figure 4.4 of the trials of St John's wort for depression, where there is some suggestion of a publication bias. Unfortunately this technique requires a large number of trials with a spread of sizes to provide an adequate 'funnel', and hence will not be helpful in many meta-analyses.



Note: Dashed outer lines show boundaries of an 'ideal' funnel; if there is no heterogeneity the points are distributed evenly on either side of the summary estimate.

Figure 4.4 Funnel plot of 12 placebo-controlled trials of St John's wort showing some suggestion of 'missing' smaller negative trials

4.4.2 Statistical tests

A statistical test that is a direct analogue of the funnel plot has been developed (Begg and Mazumdar 1994). This provides a *P*-value for the degree of apparent bias. However, as with the graphical approach, it requires large numbers of studies — at least 25 being required for modest power. For the St John's wort

example (Figure 4.4), there is a trend to bias: *P*-value of 0.14, but this is unreliable as it is based on only 12 studies.

4.4.3 If publication bias is suspected, what can be done?

If publication bias is suspected, the ideal method would be to estimate the degree to which bias has occurred and correct the summary estimate accordingly. Egger et al (1997) has suggested a regression on an analogue of the funnel plot in which the regression parameters estimate the degree of publication bias and allow a correction to be made. This is a promising line of analysis, but is unfortunately subject to a number of problems and cannot currently be recommended.

The 'File Drawer Number'

An alternative to correcting for publication bias is a sensitivity analysis to estimate its potential impact on the conclusions. One way of doing this is to estimate the number of unpublished neutral trials of equivalent average size that would be required to make the result no longer statistically significant. This is known as 'Rosenthal's File Drawer Number' (Rosenthal 1979).

5 APPLICABILITY: RETURNING TO THE QUESTION

Having completed the systematic components of the review, it is important for the reviewer to return to the original question, and assess how well it is answered by the current evidence.

- How important are study design flaws in the interpretation of the overall results?
- Is publication bias an important issue?
- If further research is needed, then specific suggestions should be made about the necessary design features rather than a simple call for more data.

A separate handbook in this series includes detailed information on the applicability of the results of systematic reviews (*How to Use the Evidence: Assessment and Application of Scientific Evidence*, NHMRC 2000a).

Part 2

Question-specific methods

This part describes additional issues and methods specific to the different types of questions: intervention, frequency, diagnostic test accuracy, risk factors and aetiology, prognosis, and economic studies. Before reading the subsection of the specific question type you are interested in, we strongly recommend that you first read all of Part 1.

Some important resources for additional reading beyond these subsections are included in the appendixes.

Appendix C covers a search method developed by Hunt and McKibbon (1997) for current systematic reviews and randomised trials. For literature searching on specific questions, the series published in the American College of Physicians Journal Club is helpful, as well as a book by McKibbon et al (1999).

Appendix D describes some available software for performing the calculations and plots. None of the packages available are comprehensive, and they usually focus on a single question type. But even within a single question, more than one software package may be required to provide all the needed calculations and plots.

6 INTERVENTIONS

6.1 The question

There are many types of intervention that may be the subject of a systematic review, such as:

- therapy for a specific disease (eg aspirin to prevent stroke, surgery for coronary artery disease, or cognitive therapy for depression);
- a change in a risk factor (eg blood pressure lowering to prevent stroke, immunisation to prevent hepatitis, or publicity campaigns to reduce teenage smoking); or
- screening for earlier diagnosis (eg mammographic screening for breast cancer, antenatal screening for silent urinary tract infections, or screening for cholesterol).

The defining feature is that some specific activity is undertaken with the aim of improving or preventing adverse health outcomes.

6.1.1 Study design

Because of their unique ability to control for confounders, known or unknown, RCTs generally provide the best evidence of efficacy for interventions. This section therefore focuses on systematic reviews of controlled trials; other study types for intervention will be discussed in the section on aetiology and risk factors.

However, in interpreting RCTs for policy making and applying them to individuals, nontrial evidence will often be appropriate. For example, surveillance data may provide the best information on rare adverse effects; and cohort studies may provide the best information on the prognostic factors needed to predict the pretreatment risk of an individual.

6.2 Finding relevant studies

6.2.1 Finding existing systematic reviews

Appendix C gives information on finding existing systematic reviews. A check should be made of the Cochrane Database of Systematic Reviews (CDSR; Cochrane Library) and DARE for Cochrane and nonCochrane reviews

respectively. Even if the review is not considered completely appropriate, its reference list will provide a useful starting point.

6.2.2 Finding published primary studies

The best search methods are changing rapidly. The efforts of the Cochrane Collaboration have been seminal in the more systematic registering, compiling and classifying of all controlled trials in databases such as MEDLINE. Use of the Cochrane Library and contact with the Collaboration would be advisable when undertaking any new review of interventions.

An initial search should use the Cochrane Controlled Trials Registry (CCTR), which is available on the Cochrane Library CD. CCTR contains a listing of potential controlled trials. These have been identified by systematically searching databases such as MEDLINE and EMBASE, by handsearching a number of journals, and from the specialised registers of trials that are maintained by the Collaborative Review Groups.

A registry called CENTRAL has been distributed on the CD-ROM edition of the Cochrane Library since issue 4 (1997). It contains some reports of studies that are found not to be relevant for inclusion in Cochrane reviews. It is also likely to contain duplicates and errors. It includes all records in MEDLINE that contain the publication type (pt):

> **randomized controlled trial**
> OR
> **controlled clinical trial**

CCTR is the 'clean' version of CENTRAL. Controlled trials that meet the necessary quality criteria are assigned the keyword 'CCTR'.

Note that whether searching CENTRAL, MEDLINE or other databases, a carefully constructed search is still required using the structured approach described in Section 2, with synonyms and wildcards.

Does a search need to go beyond CENTRAL?

As the Cochrane Library is updated every three months, a search for more recent studies may be needed. In addition, handsearching of key journals and conference proceedings should be considered.

6.2.3 Finding unpublished primary studies

There are two approaches for searching unpublished studies. First, an appendix in the Cochrane Handbook (available on the Cochrane Library CD) contains a list of about 30 clinical trials registries with completed and ongoing studies registered in specialised areas such as AIDS and cancer. Second, it may be

helpful to contact the principal investigators of relevant studies asking whether they know of additional studies.

6.3 Appraising and selecting studies

6.3.1 Standardising the appraisal

What study features should we assess?

Numerous quality assessment methods have been used: a review in 1994 identified 25 methods (Guyatt et al 1994). The number of items ranged from 3 to 34, the times for completion per study ranged between 10 minutes and 45 minutes, and the reliability kappa (which is a measure of agreement beyond that explained by chance) ranged between 0.12 to 0.95 on a scale from 0 to 1.

As the optimal use of quality items and scales is still not clear, we recommend that items be restricted generally to those that have been shown to affect the results of trials. Empirical work by Schulz et al (1995) has shown that how well the random allocation procedure is concealed and the degree of blinding both have an important influence. These two items should be assessed in any review and are described below. A third item involving the level of patient follow-up is also important.

A. Has selection bias (including allocation bias) been minimised?

Random allocation is crucial for creating comparable groups. However, it is the allocation concealment before randomisation that is vital, rather than the 'randomness' of the random number sequence. An assessment of the allocation concealment requires an explicit statement of method, such as a central computerised randomisation system. If this is not convincing, then secondary evidence is provided by demonstration of comparability from the baseline values of the randomised groups.

B. Have adequate adjustments been made for residual confounding?

For RCTs, the elimination of bias is closely related to avoidance of selection bias (point A, above) because appropriate selection/allocation minimises bias at the sample stage. If the allocation results in important imbalances, then an estimate with statistical adjustment is desirable.

C. Was follow-up for final outcomes adequate?

Having created comparable groups through randomisation, high rates of follow-up and inclusion of all randomised patients in the analysis of outcome data ('intention-to-treat' analysis) is important. However, this control of selection bias after treatment assignment has not been empirically demonstrated to reduce bias as much as appropriate randomisation and blinding. It is still useful to extract and report data on the degree of follow-up.

D. Has measurement or misclassification bias been minimised?

Blinding of outcome measurements becomes more crucial as the measure becomes more subjective and hence more open to observer bias. This is particularly important for symptoms and other patient self-report measures. The use of adequate placebos generally provides adequate blinding of outcome measures, but blinding can also be achieved without placebos; for example, by bringing in an independent 'blinded' observer to assess the outcome measure.

Appraisal checklists

Box 6.1 is an example appraisal checklist that includes these elements, modified from a checklist developed by Iain Chalmers (Cochrane Handbook; available on the Cochrane Library CD). Other appraisal methods may be used but should always include the randomisation and blinding items. Other alternatives are given in Guyatt and Rennie (1993), Guyatt et al (1993, 1994) and Liddle et al (1996).

Should scales be generic or specific?

In addition to the generic items that have been discussed, some specific items may be useful in a particular analysis. For example, the precise methods used for the outcome measure are part of both the quality and conduct of the study and are vital for the interpretation of the results. A trial of treatment of 'glue ear' in children, for example, may have used clinical appearance of the eardrum, tympanograms, audiograms or a combination for measures of outcome.

6.4 Summarising and synthesising the studies

6.4.1 Presenting the results of the studies

Both the number of trials identified and those selected should be reported, and the reason stated for those that are not selected. For example, reviews of treatment are often limited to properly randomised trials. Hence the number of apparent trials and the number with proper randomisation would be reported.

Summary table

The generic and specific quality items should be tabulated together with the major study characteristics, such as the nature and intensity of the intervention, the outcome measures and the principal results, as illustrated in Table 6.1.

Box 6.1 Checklist for appraising the quality of studies of interventions

1. Method of treatment assignment

a. Correct, blinded randomisation method described
OR randomised, double-blind method stated
AND group similarity documented

b. Blinding and randomisation stated but method not described
OR suspect technique (eg allocation by drawing from an envelope)

c. Randomisation claimed but not described and
investigator not blinded

d. Randomisation not mentioned

2. Control of selection bias after treatment assignment

a. Intention to treat analysis AND full follow-up

b. Intention to treat analysis AND <15%
loss to follow-up

c. Analysis by treatment received only OR no mention of withdrawals

d. Analysis by treatment received
AND no mention of withdrawals
OR more than 15% withdrawals/loss-to-follow-up/post-randomisation exclusions

3. Blinding

a. Blinding of outcome assessor
AND patient and care giver

b. Blinding of outcome assessor
OR patient and care giver

c. Blinding not done

4. Outcome assessment (if blinding was not possible)

a. All patients had standardised assessment

b. No standardised assessment OR not mentioned

Source: modified from I Chalmers, Cochrane Handbook; available on the Cochrane Library CD-ROM

Table 6.1 Example summary table of quality features of a set of hypothetical intervention trials

	Trial descriptors			Quality items			
Trial	*N*	Intervention	Population and other content-specific items[a]	Random-isation procedure	Blinding	Follow-up	Results (relative risk)
1	324	20 mg daily		Central computer	Double	90% at 2 years	0.7
2	121	25 mg twice daily		Envelopes	Single	80% at 5 years	0.6
3	987	10–25 mg		Not stated	None	98% at 1 year	0.55

[a] Information relevant to particular study (eg information on participants, methods, outcomes).
Note: A Cochrane review generally has a summary table with author/reference, methods, participants (age, gender, etc), interventions, outcomes, notes (quality scores may also be included).

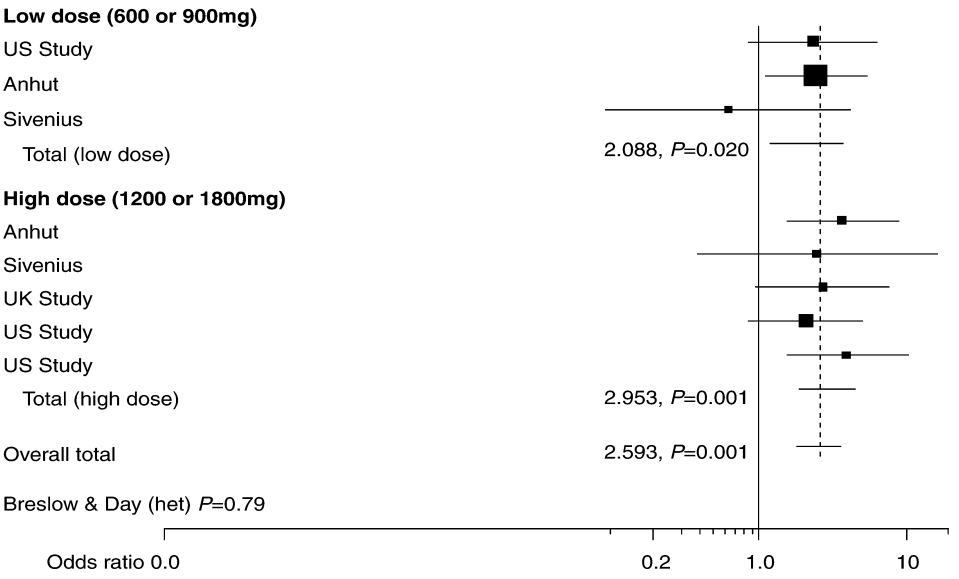
Graphical presentation

Even if studies are not to be combined, a summary plot of the results of each is invaluable. As outlined in Section 4.1, the basic plot is a summary estimate of effect together with a confidence interval (known as a 'forest plot'). The studies may be arranged in order of date, size, quality score, strength of intervention, control event rate or several other useful attributes, but should not be arranged by the size of effect. A second attribute may be indicated by subgrouping. For example, studies might be grouped by high or low intensity intervention, and then by study quality within these categories. An example of a forest plot for RCTs on treatment of epilepsy with the drug gabapentin is shown in Figure 6.1.

6.4.2 Synthesis of study results

For some systematic reviews, it may only be reasonable to present the table of study characteristics and basic data plots. However, if formal combination is considered appropriate, there are two major aims to such a meta-analytic synthesis of controlled trial data — first, to find a summary estimate of the overall effect, and second to examine whether and how this average effect is modified by other factors.

To enable a single summary outcome measure, all the trial results must be expressed in a common metric (unit). Ideally, this should be the most patient-relevant outcome and expressed in a directly interpretable manner (eg reduction in the risk of death, proportional reduction in symptoms, or days of symptoms). However, the trials will not necessarily allow this, and some pragmatic choices will need to be made.



Notes:
Dotted vertical line = the combined estimate
Total = 95% confidence interval of the combined estimate

Figure 6.1 Placebo-controlled trials of treatment of epilepsy with the drug gabapentin and the relative proportions of '50% responders' (with at least 50% reduction in seizure frequency); grouped by low (600 or 900 mg) or high (1200 or 1800 mg) doses, showing a nonsignificant trend to a greater response with the higher dosage

Outcome measures include discrete events (eg death, stroke or hospitalisation) and continuous outcomes (eg lung function, days with headache, or severity scales).

Discrete events

Discrete events can be expressed as the risk difference (RD), relative risk or risk ratio (RR), odds ratio (OR), or the average time-to-event (see Table 4.1). The choice will depend on which measure is most stable and logical for that outcome. A useful initial guide is the L'Abbe plot (L'Abbe et al 1987), which graphs the event rate in the treated group against the event rate in the control group. Figure 6.2 shows the trials from a meta-analysis of six placebo-controlled trials of warfarin for nonvalvular atrial fibrillation plotted in this way.

A study of 115 meta-analyses showed that the RD varied most over different populations, whereas the RR and OR were about equally stable (Schmid et al 1995).

An ideal approach when the time-to-event varies is survival analysis based on combined data, but the necessary data for this may not be available.

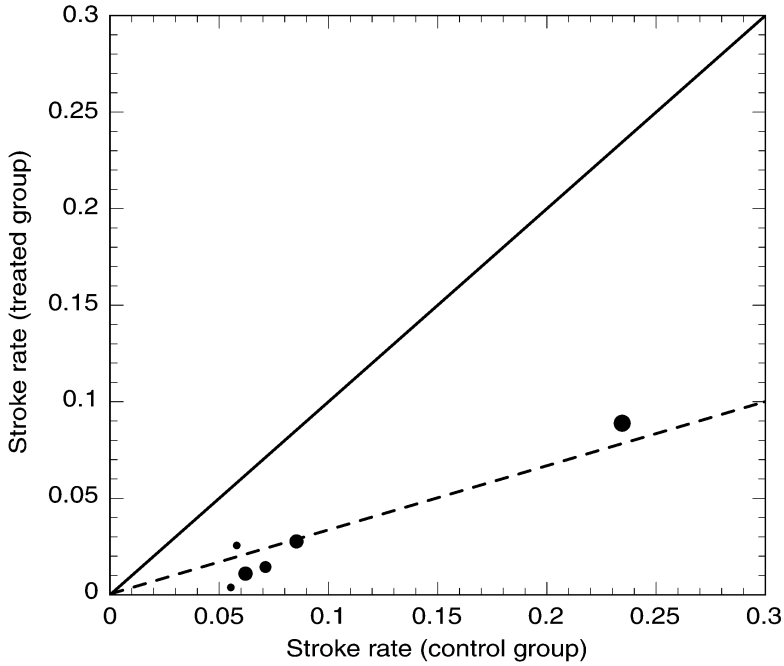


Figure 6.2 L'Abbe plot of the stroke risk in the treated group versus the stroke risk in the control group from a meta-analysis of six placebo-controlled trials of warfarin for nonvalvular atrial fibrillation. The diagonal (solid) line represents no effect. Points below this line indicate a lower rate of poor outcomes in the treated group than in the control group (ie benefit). The dashed line shows the overall (beneficial) effect of warfarin, which appears to increase with increasing risk in the control group

Continuous outcomes

Again a common interpretable outcome measure is ideal. Sometimes this is impossible, in which case a common metric is needed. Examples are:

- the proportional improvement (in say FEV1 [forced expiratory volume] or peak flow); and
- the standardised difference — the difference between the treated and control groups divided by the standard deviation in the control group (if the treatment is likely to alter the variation; otherwise the combined estimate is more stable).

6.4.3 Assessing heterogeneity

Difference in the effects seen may be caused by several factors:

- disease features, such as stage or severity;

- patient features, such as age or ethnicity;
- intensity or timing of the intervention; and, most importantly,
- study design features, such as study duration or the outcome measures used.

Even if the test for heterogeneity is nonsignificant, exploring for causes of variation is reasonable and useful (see Section 4.2.1). For example, in Figure 6.1, dealing with trials of the drug gabapentin in epilepsy, the overall test for heterogeneity was nonsignificant (P=0.79), but subgrouping appeared to show a modest dose–response relationship.

6.5 Economic evaluation

Depending on the type of economic analysis required, systematic reviews of intervention studies can also provide valuable information. The basic types of economic analysis include (Drummond et al 1997):

- cost analysis
- cost-effectiveness analysis
- cost–utility analyses
- cost–benefit analyses

The last three all contain an effectiveness component for which systematic reviews will play a role. For other components, such as the cost or pricing of resources, systematic review has a limited role. Cost-effectiveness analysis is probably the most used economic analysis in Australia. For example, submissions to the Australian Pharmaceutical Benefits Scheme require a cost-effectiveness analysis (Drummond et al 1997, Henry 1992).

A cost-effectiveness analysis has several components, including the estimate of benefit, the change in resources utilisation, and the unit costs of those resources. A general formula for cost-effectiveness ratio can be written as (Drummond et al 1997):

$$\textit{Cost-effectiveness ratio} = (\textit{Costs} - \textit{Savings}) / \textit{Effectiveness}$$

where: *Costs* = the costs of implementing the intervention

Savings = the savings from any reductions in resource use attributable to the intervention

Effectiveness = the incremental clinical benefit gained from the intervention

Savings result from changes in resource utilisation, such as hospitalisation, lengths of stay and medication usage, and are generally closely allied to clinical

outcomes. Hence, systematic reviews are relevant to the estimation of the effectiveness and, usually, to the changes in resource utilisation. However, the net cost (*Costs – Savings*) will depend on the unit costs of resources and the intervention that will vary over sites and time. Thus direct systematic review of the cost-effectiveness ratio will generally be inappropriate.

For example, for carotid endarterectomy a systematic review of the economics would first require a systematic review of the evidence of its effectiveness. This would inform us about any reductions in stroke leading to some savings. Finally, the unit costs of stroke, endarterectomy and other events would need to be established in order to calculate the net cost to be compared with the effectiveness.

Economic analysis is discussed in more detail in another handbook in this series (*How to Compare the Costs and Benefits: Evaluation of the Economic Evidence*, NHMRC 2000b).

6.6 Further information

This section provides only a brief overview of methods. A fuller description of the process for Cochrane systematic reviews is contained in the Cochrane Handbook, available as either an electronic version in the Cochrane Library or a hard copy version (Mulrow and Oxman 1996). The handbook is regularly updated and has become the principal source of guidance for systematic reviews of interventions.

7 FREQUENCY AND RATE

7.1 The question

Questions of frequency (or prevalence) arise commonly in health care. For example:

- What is the frequency of hearing problems in infants?
- What is the prevalence of Alzheimer's disease in the over 70s?
- What is the frequency of BrCa1 gene for breast cancer in women?

If the proportion changes over time, then a time period is incorporated into the definition to give a rate (or incidence). Thus, a possible question may be:

- What is the rate of incidence of influenza in different seasons and years?

Traditionally, for diseases, *prevalence* is distinguished from *incidence* as follows (Rothman and Greenland 1998):

- *prevalence* — the proportion of people who have the condition at a specific point in time (frequency of current cases);
- *incidence* — the instantaneous rate of development of new cases (also known as the incidence rate or simply the rate); and
- *incidence proportion* — the proportion of people who develop the condition within a fixed time period (also called cumulative incidence, with a specific example being the lifetime risk).

Incidence and prevalence are linked by the duration of illness, so that in a steady state population:

$$Prevalence = incidence \times duration$$

In this handbook, the terms 'frequency' and 'rate' are preferred to 'prevalence' and 'incidence' because questions do not always refer to diseases, but may refer to risk factors such as diet, or false positive rates (for diagnostic questions), and so on. The definition and calculation of frequencies and rates involves a number of subtleties, which are described by Rothman and Greenland (1998).

The apparent frequency may be greatly influenced by the case definition. For example, whether or not 'silent' myocardial infarction (incidentally detected by

later electrocardiograms) is included in estimates of myocardial infarction will change both the frequency and rate. Similarly, the precise measurements used can be influential; for example, different rates of deep venous thrombosis may be obtained from an ultrasound from those obtained from a venogram. Of particular note is that, if the true frequency is zero, the apparent frequency will consist of just the false positives, and thus be (1 – specificity). Hence it is important for any systematic review of frequencies to document both the population and the definitions and measures used.

7.1.1 Study design

The aim of a study of frequency or rate is to measure a representative sample of the target population. For frequency, this will be a random sample survey (or census) of the target population; for rate there is an additional requirement that the representative group be followed over time. Thus the major study designs are (cross-sectional) surveys for frequency, and cohort studies for rate. If the sample includes the entire population, then these become a census (for frequency) or a disease/condition registry (for rate).

7.2 Finding relevant studies

7.2.1 Finding existing systematic reviews

There have been a few systematic reviews of frequency and rates. However, it is still worth searching using the general methods: Appendix C gives information on finding existing systematic reviews. This would need to be combined with content-specific terms for the disease or risk factor being reviewed together with the terms in the next section.

7.2.2 Finding published primary studies

Unfortunately, most relevant studies are not coded as such in MEDLINE. The search requires three components:

- the alternative terms:
  incidence OR rate OR
  frequency OR proportion OR prevalence

- the condition of interest (and any synonyms), preferably using an MeSH term (exploded if possible and appropriate); and, if the number of potential studies is too large,

- a 'methodological filter' to confine this to appropriate studies of frequency, such as random or consecutive; or a filter to focus on an appropriate 'gold standard', such as audiometry for childhood hearing problems.

Various combinations of the above three components may be used. For example, a MEDLINE search for the causes of chronic cough might use:

> **chronic NEAR cough**
> where: the special search term 'NEAR' means that the 'chronic' and 'cough' need to be close together but allows for terms such as 'chronic nonproductive cough'.

This might then be restricted to references with an appropriate sample; that is, a random or consecutive set of cases, plus an adequate gold standard test or tests, and an appropriate follow-up (to catch missed or mistaken diagnoses). Together, these give the following search:

> **chronic NEAR cough AND (investigat* OR diagnos* OR cause*) AND (consecutive OR follow-up OR followup)**

7.2.3 Finding unpublished primary studies

In addition to writing to authors of published work, it is important to consider whether any government or nongovernment agencies might have relevant surveys or registries. For example, cancer registries are an obvious source for cancer incidence information; and State or Territory health departments should be contacted for information on communicable diseases. Groups and societies interested in specific diseases, such as diabetes, heart disease or cystic fibrosis, may also have done their own surveys.

7.3 Appraising and selecting studies

7.3.1 Standardising the appraisal

What study features should we assess?

There are no standard accepted quality scales for studies of proportions. However, the principal issues are similar to those described for controlled trials of interventions (see Section 6.3).

A. Has selection bias been minimised?

Random selection is important to obtain a representative sample. While simple random sampling is often appropriate, other methods include stratified random sampling and cluster sampling. The important issues are the definition and establishment of an appropriate sample frame and some form of random sampling.

B. Have adequate adjustments been made for residual confounding?

The issue of confounding is not relevant to frequency and rate studies.

C. Was follow-up for final outcomes adequate?

Having obtained a representative group by random sampling, a high response rate is needed to maintain the representativeness and avoid bias. This is particularly important if nonresponse is associated with the condition of interest. For example, if you want to know the proportion of discharged psychiatric patients who relapsed within a year, then high follow-up is important as difficult-to-follow patients often have worse outcomes.

D. Has measurement or misclassification bias been minimised?

As discussed in the introduction, a clear definition of the condition and the measurements used is crucial, as this will influence the apparent rate.

7.4 Summarising and synthesising the studies

7.4.1 Presenting the results of the studies

Summary table

A systematic description of the definitions and measurements used is critical to the comparison of studies. Hence an initial summary table is crucial, such as that shown in Table 7.1. The table should detail precise definitions of cases and the type and frequency of measurements used (eg the average of three blood pressure measurements taken in the supine position two days apart using a mercury sphygmomanometer). In addition, other potential differences between the populations should be described (eg gender mix, age range, and other inclusion and exclusion criteria).

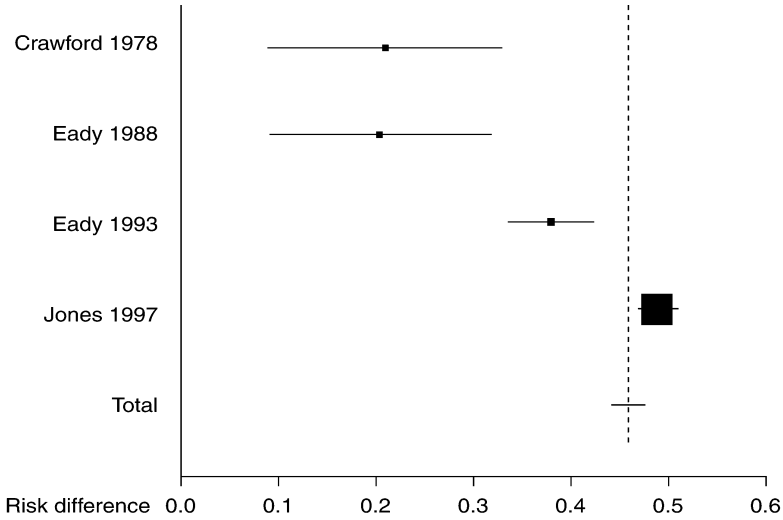
Table 7.1 Example summary table of a set of hypothetical studies of frequency

Trial	Setting	Measure	Population/ inclusion criteria	Selection	Response (%)	Results *n/N* (%)
1	Community	Single bp	Age 16–75	Random sample	70	10/105 (10%)
2	GP clinic	Average of 3 bp	All ages	Consecutive cases	80	30/240 (13%)
3	Skin clinic	Average of 2 bp on 2 occasions	Age 20–65	Consecutive cases	98	4/20 (20%)

bp = blood pressure measurements; GP = general practitioner

Graphical presentation

As with all systematic reviews, plots of the data are invaluable. For frequency and rate questions, the estimate and confidence interval should be plotted against any factors that may be predictive of the results (ie those elements provided in the descriptive table). For example, Figure 7.1 shows a plot of the rates of antibiotic resistance in *Propionibacterium acnes*, suggesting a trend with time, though other explanations, such as measurement or populations, would clearly need to be examined.



Notes:
Dotted vertical line = the combined estimate
Total = 95% confidence interval of the combined estimate

Figure 7.1 Proportion of patients with antibiotic resistance in *Propionibacterium acnes* for four studies, listed by publication date (Cooper 1998)

7.4.2 Synthesis of study results

Among a group of acceptably similar studies, the methods of quantitative synthesis are relatively straightforward. If the frequencies are considered to arise from a single common population (a 'fixed effect model'), then simple pooling will be sufficient. For example, if the prevalence of a disease was similar in all States, then an estimate of the national average prevalence would simply pool all the State disease cases, and the national prevalence (frequency) would be the total cases in the total population. The results should be reported as an overall frequency or rate and confidence interval.

However, if there is variation by case definition, measurements, population or other factors, then more complex methods are required. The first step is to look for causes of variation that may be artefacts, such as different measurements, and if possible to correct or adjust each estimate. If there appear to be true differences in the studies and populations then this should be reported. If an overall estimate is still needed, then the method depends on the aim, but may require a random effects model.

7.4.3 Assessing heterogeneity

A test for heterogeneity should be performed. Whether or not the result is significant, it is worthwhile checking whether subgroups of studies differed because of measurement method or sample frame.

8 DIAGNOSTIC TESTS

8.1 The question

Although the principles are similar across all types of study question, systematic review of diagnostic tests requires some different approaches, notably in the methods used for combining data from different studies. As with the other types of questions, the starting point for diagnostic studies is an appropriate question, including a description of:

- the disease of interest;
- the test(s) of interest;
- patient features that are likely to alter the test performance characteristics; and
- the performance characteristics of the test compared to the performance characteristics of another test(s).

If test performance characteristics vary between patient subgroups, this needs to be taken into account when applying the results of a systematic review of diagnostic tests. Common features that affect test performance characteristics include the symptoms, signs, tests and previous triage through the health care system that has got patients to the point at which you wish to evaluate the performance characteristics of a test. This issue is explored further in Section 8.3 on appraising the quality and applicability of studies.

When the performance characteristics of the test are compared to the performance characteristics of another test(s), the situation is analogous to trials in which an intervention is compared to a placebo or to another drug. For example, we may not want to know if the presence of leucocytes in an abdominal fluid aspiration has a high sensitivity and specificity for the diagnosis of appendicitis in people presenting with abdominal pain. Rather, we may want to know its incremental sensitivity and specificity compared to other features that are more easily obtained; for example, rebound tenderness in the right iliac fossa (Caldwell and Watson 1994).

8.1.1 Study design

The ideal design for studies of diagnostic test performance is usually a cross-sectional study in which the results of tests on consecutively attending patients are cross-classified against disease status determined by a reference (gold)

standard. Occasionally, the sample will be followed over time if the test is predictive of a reference standard in the future.

Most diagnostic systematic reviews have examined the test performance characteristics of individual tests. While this is useful, we are often more interested in whether a new diagnostic test is better than current alternatives. Hence there is merit in designing systematic reviews to compare tests as the many biases and heterogeneity of results in primary studies are likely to be less important if tests are compared within individuals in each study (or in individuals randomly allocated to one or other test within each study).

8.2 Finding relevant studies

8.2.1 Finding existing systematic reviews

Appendix C gives information on finding existing systematic reviews. A check should be made of DARE (available in the Cochrane Library or on the Internet) and MEDLINE. DARE compiles both intervention and diagnostic reviews, but not the other question types discussed in this guide. Even if the review is not considered completely appropriate, its reference list will provide a useful starting point.

8.2.2 Finding published primary studies

Initial searching should be done on MEDLINE, EMBASE and similar computerised databases. In MEDLINE, MeSH headings should be:

the disease of interest (all subheadings), eg
explode urinary tract infections

the name of the test (all subheadings), eg
explode reagent strips

both the disease and the test, eg
explode urinary tract infection AND explode reagent strips

Ideally, no further keywords should be used to restrict the search.

Only if inspection of the abstracts suggests that this initial approach is unmanageably nonspecific should the search be restricted. If you really need to restrict the search, try linking the disease and/or test (all subheadings) with the following:

sensitivity AND specificity (exploded MeSH heading, which includes 'predictive value of tests' and 'ROC curve')
OR
sensitivit* (textword)
OR
specificit* (textword)
OR
predictive value (textword)

(Note: sensitivit* is a shorthand which allows for both 'sensitivity' and

This method of restricting the search while minimising loss of sensitivity is based on evidence from a set of journals on general and internal medicine with high impact factors in 1986 and 1991 (Haynes et al 1994). It may not be applicable now or to a wider range of journals. If it does not capture articles of known relevance, reconsider the feasibility of manually searching the abstracts of the unrestricted search based only on the disease and test. If you still consider that is not feasible, check how missed articles have been indexed to get ideas on additional restriction terms. Having two searchers develop strategies independently may be helpful. Some additional MeSH headings that may help generate relevant articles are:

diagnostic errors (exploded heading, which includes 'false negative reactions', 'false positive reactions' and 'observer variation')

diagnosis, differential

reproducibility of results

Some additional textwords that may help are:

accuracy, ROC, likelihood ratio

You may also find more articles by clicking on 'related articles' in relevant articles identified in PubMed.[5]

An alternative but perhaps less successful method of restricting is to search for the disease and/or test of interest, including only those subheadings concerned with diagnosis, for example:

diagnosis, pathology, radiography, radionuclide imaging, ultrasonography and diagnostic use

[5] www.ncbi.nlm.nih.gov/PubMed/

Avoid using the MeSH heading 'diagnosis' because it differs from diagnosis as a subheading of a disease and is not designed to capture articles on diagnostic tests.

Articles on diagnostic tests may not be indexed as well as articles on intervention studies. Therefore, as demonstrated by the example of near-patient testing described in Section 2.3, it is more important to search the references of studies, handsearch relevant journals and conference proceedings, and examine articles suggested by experts in the relevant field (McManus et al 1998). It is helpful to record and report the details of your search strategy for future reference.

8.2.3 Finding unpublished primary studies

Publication bias is probably as much of a problem for systematic reviews of diagnostic tests as it is for observational studies in general. This is because reviews are often produced using available data sets and only those that show features of interest may reach publication.

Methods for detecting and dealing with publication bias for diagnostic test studies are not well developed. We are not aware of any attempt to develop registries of studies at the design stage, as has been done for RCTs.

8.3 Appraising and selecting studies

8.3.1 Standardising the appraisal

The quality of diagnostic studies is determined by the extent to which biases have been avoided. However, a high quality study (sometimes referred to as internally valid) may not be applicable in your setting (ie externally valid) if the exact test used differs from that to which you have local access or the test has been evaluated in a tertiary care setting, while you are interested in using it in primary care. The applicability of high quality studies is determined by whether the test methods and population accord with your area of interest.

Information about the characteristics that define the quality and applicability of studies may be used to decide the 'boundaries' of the question to be answered by the systematic review, when reviewing abstracts or after having reviewed full papers. Alternatively, a more informative approach is to explore the extent to which some or all of the characteristics affect estimates of test performance when combining studies, as outlined in Section 8.4. For example, if the primary studies choose two different reference standards, it is possible to explore whether the estimated test performance characteristics vary with the choice of reference standard.

What study features should we assess?

Several checklists for quality and applicability of primary studies of diagnostic tests have been developed (Bruns 1997, Irwig et al 1994, Jaeschke et al 1994ab, Reid et al 1995, Liddle et al 1996). The most comprehensive checklist has been developed by the Cochrane Methods Working Group on Screening and Diagnostic Tests.[6] A shortened and updated version of this checklist is shown in Box 8.1. However, only a few studies are known that have given empirical evidence about the effect of quality on estimated test performance characteristics (Lijmer et al 1999, Fahey et al 1995). Nevertheless, any checklist should include the elements of quality and applicability outlined below.

Quality

A. Has selection bias been minimised?

Consecutive patients with the features of interest should be enrolled. Some studies, however, do not use this method and instead estimate test performance based on people who have been diagnosed with the disease and those without the disease. These studies tend to include the more severe or 'definite' end of the disease spectrum and the nondiseased group tends to be people without a clinical problem. Such 'case-control' studies are likely to overestimate both sensitivity and specificity (Lijmer et al 1999).

B. Have adequate adjustments been made for residual confounding?

For diagnostic tests, the issue of 'confounding' can generally be considered as the incremental value of the new test over other tests that have been done (and which may be cheaper, less invasive, etc). In this instance, this is an issue of applicability rather than quality and is discussed in more detail under Applicability, below. Another context in which confounding arises is if the reference standard is a later event that the test aims to predict. In this case, any interventions should be blind to the test result, to avoid the 'treatment paradox': a test may appear to be poorly predictive because effective treatment in the test-positives has prevented the poor outcomes that the test would otherwise predict.

C. Was follow-up for final outcomes adequate?

To maintain the sample, all those enrolled should be verified by the reference standard and included in the analysis. Verification bias occurs when the reference standard is applied differently to the test-positives and the test-negatives. This is most likely when the reference standard is an invasive procedure, in which case the test-negatives are less likely to be subjected to it.

[6] www.cochrane.org/cochrane/sadt.htm

Box 8.1 Checklist for appraising the quality of studies of diagnostic accuracy

Descriptive information about the study

- Study identification
- What is the study type?
- What tests are being evaluated?
- What are the characteristics of the population and study setting?
- Is the incremental value of the test being compared to other routine tests?

Has selection bias been minimised?

- Were patients selected consecutively?

Was follow-up for final outcomes adequate?

- Is the decision to perform the reference standard independent of the test results (ie avoidance of verification bias)?
- If not, what per cent were not verified?

Has measurement bias been minimised?

- Was there a valid reference standard?
- Are the test and reference standards measured independently (ie blind to each other)?
- Are tests measured independently of other clinical and test information?
- If tests are being compared, have they been assessed independently (blind to each other) in the same patients or done in randomly allocated patients?

Has confounding been avoided?

- If the reference standard is a later event that the test aims to predict, is any intervention decision blind to the test result?

Source: modified from Cochrane Methods Working Group on Diagnostic and Screening Tests

Likewise, the proportion of the study group with unobtainable test results should be reported; for example, the number of needle biopsies that provided an inadequate sample. It is inappropriate to omit from analysis those test results that are uncertain; for example, some, but not full-colour, development on a reagent strip. The test performance characteristics of uncertain test results should be obtained or uncertain results combined with positives or negatives.

D. Has measurement or misclassification bias been minimised?

A validated reference standard should be used and the test and reference standard should be measured independently of (blind to) each other. The tests should also be measured independently of other clinical and test information. Although independent assessment is generally desirable, there are some situations where prior information is needed; for example, in identifying the exact site of an abnormality for which a radiograph is being viewed.

If tests are being compared, have they been assessed independently?

If tests are being compared, a systematic review based on studies in which the tests are being compared is a much stronger design than if performance characteristics of the tests come from different studies. The strongest within-study design is when both tests are done on the same individuals or individuals are randomly allocated to each test. It is especially important that two or more tests whose performance characteristics are being compared are assessed independently in each individual. For example, if mammography and ultrasound are being compared as a diagnostic aid in young women presenting with breast lumps, the two techniques should be assessed without knowledge of the results of the other imaging technique.

Applicability

Estimated test performance characteristics may depend heavily on details of how the test was performed and the population tested. This information should be collected and presented so that readers can judge applicability by the extent to which the clinical problem is being addressed and if the exact test used is similar to those in the setting in which they practice.

A. About the test(s)

- How were tests performed (eg kits from different manufacturers)?

- What threshold was used to differentiate 'positive' from 'negative' tests? Ideally, tests will be looked at using several categories of test result (or even as a continuum), and this should be noted when it is done. Because data are usually dichotomised around a single threshold in primary studies published to date, and accessible meta-analysis methods are best developed for dichotomised data, this will be the only approach considered further.

B. About the population

- Presenting clinical problem — the condition that defined entry into the study.

- Disease spectrum — the spectrum of disease in the diseased group (those with the disease of interest) is described directly by the stage or severity of disease. Spectrum in the so-called nondiseased group (those without the disease of interest) is described by the final diagnoses in that group. Indirect

measures of spectrum include the setting (eg primary or tertiary care), previous tests and the referral filter through which people had to pass to get to the point where they were eligible for the study.

- Incremental value of tests — although a test may appear to give good results, it may not provide any more information than simpler (eg less invasive or cheaper) tests that are usually done in a particular setting. This is like thinking of these other tests as 'confounders' that must be taken into account when assessing the test performance characteristics of the test of interest (eg by restriction, stratification or modelling).

Indirect measures

The above features may not capture all aspects of quality and applicability, as the information you want is often not provided in the primary studies. Therefore, it is worth looking at some additional measures.

- Prevalence of the condition — this may be a proxy for the 'setting' in which the test is being assessed. More importantly, it has been shown that error in the reference standard is an important cause of sensitivity and specificity variation (nonlinear) with the observed prevalence of the condition (Brenner and Savitz 1990, Valenstein 1990).

- Year of the study — the quality of studies, the way tests have been done and the populations on which the tests are being performed may have altered over time.

8.4 Summarising and synthesising the studies

8.4.1 Presenting the results of the studies

Summary table

Studies should be listed, tabulating the extent to which they fulfil each criterion for quality and applicability. Studies can be categorised by the most important quality and applicability criteria for the topic being addressed. If the number of studies is large or many criteria are considered equally important, provide a summary table showing the proportion of papers that fall into each category (or important combinations of criteria). Table 8.1 shows an example summary table.

Table 8.1 Example summary table of quality features of a set of hypothetical diagnostic accuracy trials

Study descriptors			Quality			
Study	***N***	**Setting**	**Consecutive attenders**	**Verification bias avoided**	**Test and reference standard measured independently**	**Tests being compared assessed independently**
1	300	Hospital	Yes	Yes	Yes	Yes
2	800	Primary care	Yes	No	Yes	No
3	1000	Specialist clinic	No	Yes	No	Yes

Graphical presentation

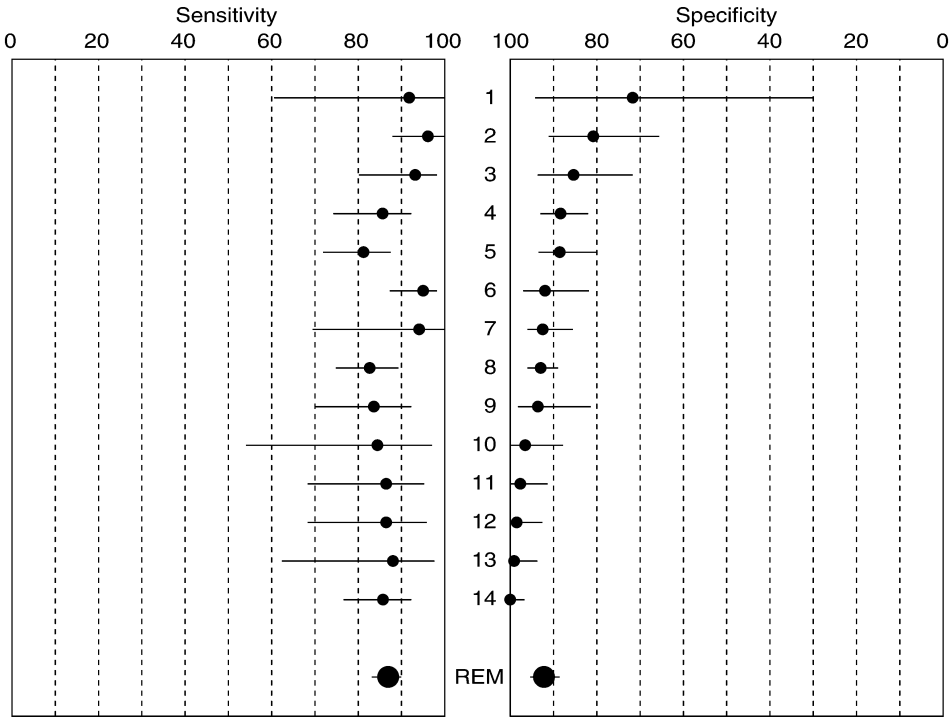
Simple plot of sensitivity and specificity

Show the sensitivity and specificity of each study with its confidence intervals. This is best done graphically, with the specificity for a particular study shown alongside the sensitivity for that study (as shown in Figure 8.1). Ordering the studies by some relevant characteristic helps interpretation. For example, test threshold may differ between studies, so that those studies with lowest sensitivity may have the highest specificity and vice versa. If studies are ranked by their sensitivity, the visual display is the first step towards understanding the magnitude of this phenomenon.

The plot and all the following steps can be done using Metatest software (see Appendix D). However, the currently available version of the software (Metatest 0.5) does not test statistical significance. The Internet website will be updated as the software is developed further (see Appendix D). Statistical modelling and significance testing can be done in any statistical package, but requires expertise in applying the transformations outlined below and its back-transformation.

Plot sensitivity against specificity

The next step is to plot sensitivity against specificity in ROC space, ideally showing the points as ovoids with an area proportional to the square root of the number of people on whom sensitivity and specificity have been calculated (see Figure 8.2). As in the last step, this may display the trade-off between sensitivity and specificity because studies have different thresholds.



REM = pooled estimate using the random effects model
Note that as specificity improves, sensitivity appears to decrease

Figure 8.1 Plot of sensitivity versus specificity (with 95% confidence intervals) for 14 studies of carotid ultrasound for carotid stenosis (graph prepared with Metatest software) (Hasselblad and Hedges 1995)

8.4.2 Synthesis of study results

Fit a summary ROC (SROC)

A good method of combining data, which takes account of the interdependence of sensitivity and specificity, is the SROC (Moses et al 1993, Irwig et al 1994, 1995). This is difficult to do directly and is therefore done in three steps:

1. the true positive rate (TPR, or sensitivity) and the false positive rate (FPR, or 1 – specificity) are first transformed through the logarithm of their odds

2. the regression analysis is done; and

3. the results are back-transformed and plotted in the standard ROC format.

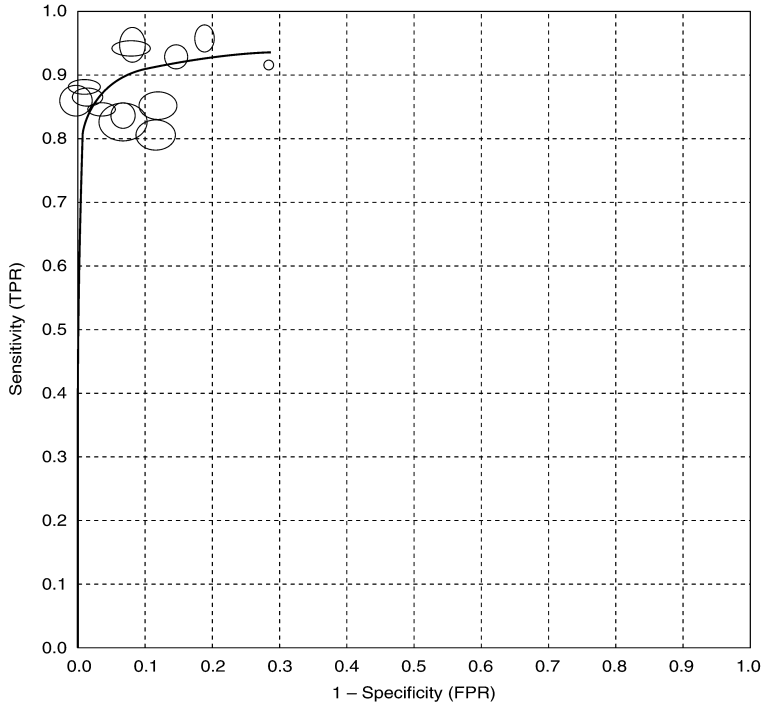


Figure 8.2 Receiver–operator curve (ROC) plotting true positive rate (sensitivity) against false positive rate (1 – specificity) for a meta-analysis of carotid ultrasound accuracy showing the individual study points and the fitted summary ROC (SROC) (Hasselblad and Hedges 1995)

The transformation of the data in the second step examines the linear relationship:

$$D = a + bS$$

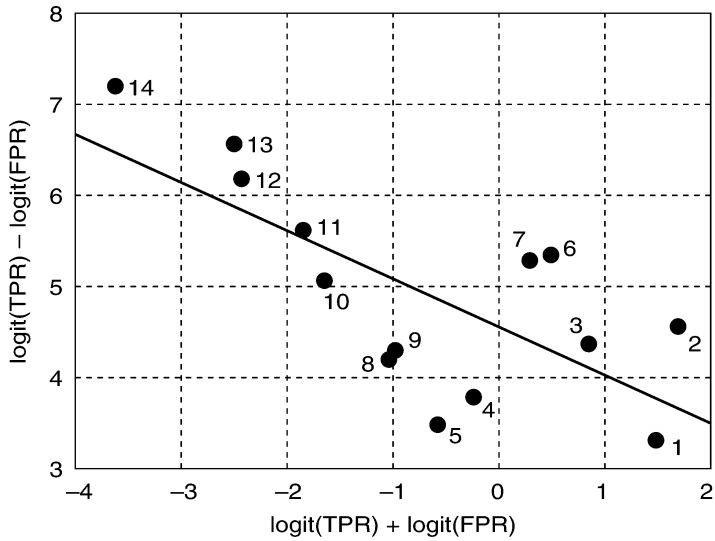
where: D = (logit TPR) – (logit FPR) = log (odds ratio)

S = (logit TPR) + (logit FPR) = log (odds product), which is a proxy for the threshold

a = estimated linear intercept

b = estimated regression coefficient (slope)

This relationship can be plotted as a regression of D on S as shown in Figure 8.3 and provides the estimates of a and b needed for the SROC (see Figure 8.2). If the slope, b, is nonsignificant and close to 0, then we can focus on the intercept, a, back-transforming it to the odds ratio. A single constant odds ratio defines a single symmetric SROC. If the slope, b, is significant, the situation is more complex — the odds ratio changes with the threshold, resulting in an asymmetric SROC.



Note: Solid line shows unweighted best fit with intercept $a = 4.55$; slope $b = -0.53$

Figure 8.3 Plot of *D* versus *S* for a meta-analysis of carotid ultrasound accuracy showing the individual study points and the fitted line (Hasselblad and Hedges 1995)

One difficulty with SROCs is that they do not give a particular set of sensitivity and specificity values as the summary estimate. We therefore suggest using the sensitivity at average specificity. For screening tests, read off the sensitivity at a false positive rate (1 – specificity) equivalent to the rate of positive results in your population. The sensitivity and specificity obtained can be used to generate post-test probabilities for a range of pretest values.

Compare tests

If the objective is to compare tests, use only those studies that do both tests and plot them using different symbols against the 'common' SROC (Loy et al 1996). Testing can be done by adding test type to the regression of the *D* (log odds ratio) on *S* (log odds product) mentioned in the previous section.

8.4.3 Assessing heterogeneity

Assess whether the test performance characteristics vary by study quality or population and test characteristics (Moons et al 1997). Start by plotting the data for subgroups defined by each important criterion for study quality given in Section 3.1 and examine how they fall around the common regression. To test significance, add each feature individually in the SROC model. If there are sufficient studies, this can be extended to include several variables simultaneously using conventional approaches to modelling.

9 AETIOLOGY AND RISK FACTORS

9.1 The question

Questions of aetiology and risk factors commonly arise in relation to public health. For example:

- Does the evidence support a likely causal effect of a factor (eg obesity) on a particular disease (eg breast cancer)?

Clearly, in public health terms, you may want to know the whole array of health effects of an exposure, but the evidence for each causal or preventive influence has to be first assessed separately, along the lines we suggest here. Largely, such evidence will come from case-control and cohort studies, although in some instances RCTs provide critical tests of causal hypotheses.

In a striking recent example, RCTs showed beta-carotene to be an ineffective preventive of lung cancer, contrary to deductions made from a large number of observational studies that evaluated diet and laboratory data. This role of RCTs needs to be borne in mind when constructing a search strategy, as indicated below.

Getting the right balance in the question being addressed may be straightforward (eg 'Do oral contraceptives cause breast cancer?'), especially when it derives from a clear clinical or public health question. But should a review of body size and breast cancer include:

- all measures of body size (height, weight, skinfolds, circumferences, derived variables);

- only those that are modifiable (removing height, which is of biological interest); or

- only the most direct estimates of adiposity, such as skinfolds?

Such issues usually make systematic reviews of aetiology and risk factors more complex than systematic reviews of interventions.

9.1.1 Study design

Epidemiology studies of aetiology (often called observational studies) relate individual characteristics, personal behaviours, environmental conditions, and treatments as 'exposures' that may modify risk of disease. In contrast to randomised trials, most epidemiology studies relate naturally occurring

exposures to the onset of disease. These studies are often called 'observational studies' and may be cross-sectional, prospective or retrospective. Cohort (prospective) studies relate exposure to subsequent onset of disease, comparing the rates among the exposed to those in the unexposed. Case-control studies (retrospective) compare the exposure histories of a group of cases to those among controls (disease free).

For the study of aetiology, prospective studies usually provide stronger evidence than case-control studies. Rarely are cross-sectional studies of importance, although in the case of obesity and breast cancer they may shed light on the relation between adiposity and hormone levels, giving support for a biological mechanism for the relation under study.

9.2 Finding relevant studies

9.2.1 Finding existing systematic reviews

You should first check whether an appropriate systematic review already exists (see Appendix C). If no such review is found or if there is none that directly matches your needs and is up to date, then you face the challenge of constructing your own.

9.2.2 Finding published primary studies

The initial approach parallels that for searching for diagnostic studies, essentially searching MEDLINE for combinations of the disease and the exposure of interest, and, if the resulting set is too large, adding a methodological filter. For example:

> **obesity AND breast cancer**
>
> will be quite specific, especially if combined with
>
> **human**
> OR
> **epidemiology**
>
> but much relevant research will be excluded, whereas
>
> **obesity OR any of its alternatives OR breast cancer**
>
> will spread a wider (more sensitive) net, but at the expense of retrieving a mass of unwanted material to sift through.

The amount of unwanted material can be reduced substantially by using a methodological filter; for example, focusing on study types most likely to yield sound data relevant to causality, and/or pertinent measures of association (odds ratio, relative risk, or hazard ratio).

It needs to be noted that (as with our suggestion above) the 'aetiological filter' excludes RCTs. In light of the beta-carotene example (see Section 9.1), if such methodological filters are to be used, then RCTs should be included as a specific design option (as outlined in Section 6.2).

Your choice of a sensitive or a specific initial strategy will depend on your purpose — a fully comprehensive review requires the former as a starting point — and perhaps the size of the available literature (if it is small, you should probably scan it all anyway). But if you simply want a reasonable array of sound, relevant studies, you should pursue a more restrictive search strategy from the start.

Finally, a comprehensive review should include handsearching of current relevant journals and scanning bibliographies of retrieved articles. A useful source is the IARC Directory of Ongoing Research in Cancer Prevention (Sankaranarayanan et al 1996). This directory is available on the Internet from the IARC site.[7]

9.2.3 Finding unpublished primary studies

Relevant databases should be searched for possible unpublished work, including the database of dissertation abstracts. A number of services provide access to this and similar databases of unpublished thesis work.

9.3 Appraising and selecting studies

9.3.1 Standardising the appraisal

The question at hand is whether any selection, measurement bias or confounding is great enough to seriously distort the size (and qualitative interpretation) of the effect estimate. The importance of these errors will vary with study type and problems specific to the question at hand. For example, exposure measurement error will be minimal for a biochemical value studied prospectively, but may be more important for self-reported exercise habits in a case-control study.

What study features should we assess?

Individual studies can be reviewed against a set of methodological criteria, systematically applied within study types. There are many different checklists that give different weight to elements of the design and conduct of

[7] www-dep.iarc.fr/prevent.htm

observational studies. Box 9.1 gives an example derived from Liddle et al (1996), which is a thoughtful publication in this area.

The problems of bias and their solutions are qualitatively similar for case-control studies and for RCTs, although exposure measurement is usually far more challenging, and accessing a sufficient proportion of an appropriate control group often provides difficulties for case-control studies.

A reviewer must give greatest weight to factors that are most problematic for the issue under scrutiny, and then consistently assess their likely role in each study. It is also essential to consider the practical consequences of error. If, for example, despite the gross misclassification of physical activity due to poor patient recall, a study shows an association with, say, lower risks of osteoporosis, then the true effect of exercise is actually likely to be much larger than that observed.

A. Has selection bias been minimised?

Sampling (selection and allocation) bias arises when noncomparable criteria have been used to enrol participants in a retrospective, or case-control, investigation.

B. Have adequate adjustments been made for residual confounding?

In an RCT, given a reasonable randomisation process (and large enough sample size), confounding should not be an issue, whereas for observational research, it is always a possible explanation for an effect. Exactly how to proceed is not clear and attempts to deal with confounding in case-control and cohort studies deserve close attention (Colditz et al 1995). Studies that do not control for known strong confounders (eg cigarette smoking in an analysis of diet and throat cancer) are likely to be of little value. One approach that has been used is to document the range of covariates considered in each of the studies identified and use this information to qualitatively assess the magnitude of confounding observed across the studies at hand.

Better quality studies always address confounding openly and thoughtfully, although this could lead to different actions — for example, including all possible confounders in a model, or excluding from a model variables that are shown not to confound, either practically (no/small change in estimate) or theoretically (not associated with exposure or not a risk indicator for disease).

There will also be areas where knowledge of risk factors (and hence confounders) is limited, leading some authors to 'control' for confounding, while others do not. Here, considering whether adjusted and crude estimates differ may help judge whether confounder control implies higher quality. And it should always be borne in mind that even careful statistical control of scores of the known possible confounders may be inadequate to deal with unknown or

unmeasured confounding. This seems especially likely to occur with self-selection of life-habits (eg exercise, long-term use of pharmaceutical preventives).

Box 9.1 Checklist for appraising the quality of studies of aetiology and risk factors

This set of criteria should be used for appraising studies of the extent to which the characteristics or behaviour of a person, an environmental exposure or the characteristics of a disease alter the risk of an outcome.

Information about the study

- Study identification.
- What is the study type?
- What risk factors are considered?
- What outcomes are considered?
- What other factors could affect the outcome(s)?
- What are the characteristics of the population and study setting?

Evaluation criteria for the study

- Are study participants well-defined in terms of time, place and personal characteristics?
- What percentage of individuals or clusters refused to participate?
- Are outcomes measured in a standard, valid and reliable way?
- Are risk factors and outcomes measured independently (blind) of each other?
- Are all important risk factors included in the analysis?
- What percentage of individuals or clusters recruited into the study are not included in the analysis (ie loss to follow-up)?

Overall assessment of the study

- How well does the study minimise bias? What is the likely direction in which bias might affect the study results?
- Include other comments concerning areas for further research, applicability of evidence to target population, importance of study to policy development.

Source: modified from Liddle et al (1996)

C. Was follow-up for final outcomes adequate?

In a prospective study, ascertainment bias arises when the intensity of surveillance and follow-up varies according to the exposure status of study participants. Documenting participation rates and methods of surveillance and diagnosis of endpoints is essential to assess sampling bias. Ascertainment bias also may arise when the actual diagnosis of interest is not independent of the exposure. This can arise in either prospective or retrospective studies.

D. Has measurement or misclassification bias been minimised?

Collection of noncomparable information from cases and non-cases accounts for measurement bias. This bias may arise when interviewers elicit information differentially between different study groups. Alternatively, participants may recall information with different levels of accuracy depending on their past disease experience. In retrospective or case-control studies this is referred to as recall bias.

Finally, measurement error due to general inaccuracy in the assessment of exposure leads to bias in the measure of association between exposure and outcome. In any study if such error in exposure assessment is random, it will lead to underestimates of the association between exposure and disease.

9.4 Summarising and synthesising the studies

9.4.1 Presenting the results of the studies

Summary table

A summary table is essential to show individual study characteristics. You should prepare a common data abstraction form on which to summarise the main elements of each study:

- descriptive data on numbers, demographic and other characteristics; and
- relevant outcome measures — frequencies, effect estimates (simple, adjusted, ordered) and confidence intervals or *P*-values, in particular.

And be warned — this abstract form should be extensively piloted before starting the summary table itself. It can be hard to imagine the array of different ways results of observational studies are presented. Early consultation with a statistician may also be helpful in deciding how to deal with (and record) an absence of point estimates and/or confidence intervals; and factors that are sometimes treated as a cause (eg overweight) and elsewhere as preventive (eg low/normal weight). Analyses over a range of doses will also often be based on different categories in different settings.

Here and elsewhere in the presentation of results, it is best to begin by considering studies according to study design. That is, evaluate the prospective studies as a group and compare the results to those reported from the retrospective or case-control studies, and from any RCTs. Such presentation sets the stage for combining data within each study design type as a first step to data summarisation.

Graphical presentation

The best way to show the pattern of effects is to plot point estimates (generally shown with 95% confidence intervals). There should be some sensible order to the data; for example, ranking or grouping by quality score or study type, and examining them for consistency (or lack thereof) first within and then across groups. Whether or not it is sensible or useful to combine the estimates across all studies, or subgroups, is a decision to be made, as noted previously, according to subjective judgments (perhaps aided by a formal test of heterogeneity) on the data pattern and the similarity of the studies.

9.4.2 Synthesis of study results

The estimates in observational studies will usually need to be adjusted for the major confounding factors such as age, gender, etc. A quantitative synthesis will thus aim at combining these *adjusted* estimates. Hence, the methods described for interventions in Section 6.4 will only occasionally be applicable. However, the general principle will still be to obtain a weighted combination where the weights are the inverse of the variance of the study estimates. Thus the standard error of each adjusted estimate will be required. If this is not given directly, it may need to be inferred from the confidence interval (the width of which will be about 4 standard errors of the log RR or OR on a log scale) or the exact *P*-value. Several software packages now automate these calculations for use in meta-analysis.

Ideally, all studies should have been adjusted for all major confounders. If some have not, then these would need to be grouped separately, or 'external' adjustments made. A good general discussion of such methods for synthesis of observational studies is available in Rothman and Greenland (1998; chapter on meta-analysis).

9.4.3 Assessing heterogeneity

Heterogeneity arises when the results vary among the studies more than can be attributed to chance (see Section 4.3). The 'homogeneity assumption' is that the results of all studies estimate the same true effect and that the observed variation is due to within-study sampling error. However, in practical research applications it is impossible to know whether this is true, and it is most likely that it is not.

Many investigators use statistical tests of heterogeneity (lack of homogeneity) to see whether the assumption is correct, and to evaluate the test at the $P=0.05$ level (see Section 4.3). However, because tests for homogeneity have low power, homogeneity should not be assumed uncritically. That is, the purpose of the test is not to determine whether heterogeneity exists at all, but to get an idea of how much heterogeneity exists.

For observational studies there are more sources of heterogeneity than for RCTs. For example, where a series of RCTs may have all used the same doses of a drug and a common protocol for definition of the endpoint of interest, there is relatively little potential for variation among the studies. For observational studies, however, the approach to measuring the exposure may vary among the studies, and the criteria used for diagnosis or the class of endpoints studied often also differ. When the intensity of surveillance for endpoints differs across studies, there is a range of sources of heterogeneity.

Therefore, for observational studies, in particular, it is better to act as if there is heterogeneity among study results when the chi-square goodness-of-fit test statistic is greater than the number of studies minus 1 (which is the mean value when there is no heterogeneity).

If the purpose of meta-analysis is to study a broad issue then the true values can be expected to vary from study to study, and both an estimate of this variability and the mean of the true values are important and should be reported (Colditz et al 1995). The contribution of factors such as study design and methods to the observed heterogeneity can then be evaluated.

9.5 Judging causality

Once the most important features of the data and the heterogeneity between studies have been explored with respect to study-specific flaws of sample (size and/or selection bias), measurement or confounding, the reviewer is positioned to explore formally whether the observed effects allow a causal interpretation.

- Is there a clear (reasonably strong) and important effect, which is fairly consistently seen, at least among the better studies?
- Has the effect been demonstrated in human experiments?
- Is this effect greater at higher exposures?
- Does it have an accepted or probable biological basis?
- Is it evident that the time-direction of the association is clear (cause always precedes effect)?

Other elements may be added, such as those suggested by Bradford-Hill (1965). Studies that provide a critical test of a causal hypothesis are particularly important. These will often be experimental studies, as with the beta-carotene example described in Section 9.1. However you approach it, there is no doubt that a careful and logical assessment of the broad causal picture is extremely helpful to a reader struggling to understand the mass of facts a review may contain.

10 PREDICTION AND PROGNOSIS

10.1 The question

Prognostic questions generally contain two parts:

- the definition of the patient population of interest (eg recent onset diabetes, newly detected colorectal cancer); and
- the outcomes of interest, such as morbidity and mortality.

The implicit third part of the usual three-part question is the set of risk factors that have been used for the prediction of prognosis. Section 9 looked at a single risk factor, with a particular focus on whether that risk factor was causally associated with the outcome. In this section this idea is extended but with a focus on prediction or prognosis for individuals. This section should therefore be read in conjunction with Section 9 on risk factors but differs in two ways.

- First, the principal aim is prediction of outcomes, whether or not the factors are causal. For example, an earlobe crease might be considered a valid marker of cardiovascular disease risk and form a useful part of a risk prediction model, though clearly it is a marker rather than being causal.
- Second, the combination of multiple factors for prediction will often give better prediction than the single factors considered in Section 9 (eg Framingham cohort study risk equations for heart disease).

10.1.1 Why should we be interested in prediction?

There are two principal reasons for investigating questions about prediction. First, patients are intrinsically interested in their prognosis, so they can adapt and plan for their future. Second, separation of individuals with the same disease into those at high and low risk may be extremely valuable in appropriately targeting therapy. Generally, those with high risk have more to gain, and hence benefits are more likely to outweigh disadvantages, and also to be more cost-effective — the importance of prognostic information in applying the results of systematic reviews and clinical trials is more fully specified in an accompanying handbook in this series (*How to Use the Evidence: Assessment and Application of Scientific Evidence*, NHMRC 2000a).

In using prognostic information to help decide on treatment, it is generally important to know the 'natural history' of the condition. That is, what would happen without any effective therapy. It is important to realise that this is often impossible information to collect, as some therapy will often have been started.

Hence it is important to consider that prognosis is conditional. For example, you may ask about the prognosis of noninsulin dependent diabetic patients conditional on antihypertensive treatment being given for high blood pressure and antihypercholesterolaemic agents being given for high cholesterol levels.

The placebo groups of trials may be considered a reasonable source of information for natural history but even these groups will often be undergoing some forms of treatment and the new treatment that is being compared with placebo is an add-on to these other therapies.

10.1.2 Study design

The ideal study design for prognostic studies should focus on cohort studies with an 'inception' cohort of patients with a condition followed for a sufficiently long period of time for the major outcomes to have occurred.

10.2 Finding relevant studies

10.2.1 Finding existing systematic reviews

Systematic reviews for the influence of single factors on prognosis are becoming more common (eg the effect of blood pressure level on stroke risk) and it is clearly worth trying to find them using the methods described in Part 1 and Appendix C. However, there are methodological problems for systematic reviews that look at several prognostic factors simultaneously and hence only a few have been undertaken.

10.2.2 Finding published primary studies

The search strategies given in Appendix C (Haynes et al 1994) focus on identifying longitudinal studies that potentially have such predictive information. An important alternative to consider is the use of the control groups in RCTs, as this is often the only place where sufficient investment in follow-up has been made to be able to provide adequate information on prognosis.

10.2.3 Finding unpublished primary studies

Since there is no registry of prognostic studies, these will be particularly difficult to track down. Some exceptions occur when routine data are kept on a group of patients. For example, cancer registries provide locally-relevant survival data, although the degree of detail on prognostic factors varies considerably.

10.3 Appraising and selecting studies

10.3.1 Standardising the appraisal

What study features should we assess?

The requirements for good quality information on prognosis are similar to those of an RCT (see Section 6.3). The two principal differences are that randomisation is unnecessary, and that good baseline measurements of the potential prognostic factors have been included. The following information is critical to appraise.

A. Has selection bias been minimised?

A consecutive or random sample of patients should have been selected at a similar time point in their disease. If this is at the beginning of the disease, this is known as an 'inception cohort'.

B. Have adequate adjustments been made for residual confounding?

Confounding plays a very different role when we are focused on risk prediction rather than aetiology. Even if 'confounders' are not causal, but merely good markers of risk, they may be useful in a prognostic or prediction model. Hence they do not need to be used for 'adjustment' but may be a useful part of the model.

C. Was follow-up for final outcomes adequate?

Having obtained a representative group through consecutive or random selection, high rates of follow-up and inclusion of all patients are important. Hence, it is useful to extract and report data on the level of follow-up.

D. Has measurement or misclassification bias been minimised?

Outcomes should preferably be measured blind to the prognostic factors being considered. For example, knowing about factors such as cholesterol or smoking may influence the decision about whether a person has ischaemic heart disease. This becomes more common if the outcome to be assessed is more subjective and hence more open to observer bias.

10.4 Summarising and synthesising the studies

10.4.1 Presenting the results of the studies

A summary table of the identified studies should be included with the characteristics of the study population, the follow-up and outcomes measured, and the quality features mentioned in the previous section.

In this case the data must be emphasised because multiple factors and synthesis across several studies are usually much more difficult or even impossible to deal with. Hence, it may be necessary to forgo any synthesis, and choose the 'largest' acceptable quality study available. It should be noted, however, that the study with the largest number of patients will not necessarily be the most statistically reliable. It is the outcomes that really provide the information, and this will depend on three factors:

- the number of patients;
- the length of follow-up; and
- the level of risk of the patients.

The number of patients and the length of follow-up combined represent the total 'person time' of the study, which is commonly used as a measure of the relative power of studies.

10.4.2 Synthesis of study results

Combining studies is occasionally possible, although it is uncommon. Ideally, this would involve the use of individual patient data pooled from the studies. One example of this is the INDANA project, which has pooled the prognostic information from several trials of hypertension (Gueffier et al 1995). Another example is the Atrial Fibrillation Trialists Collaborative, which has combined prognostic information on patients with atrial fibrillation (Atrial Fibrillation Investigators 1994).

However, as discussed in Part 1, combining studies is sometimes impossible and, even if it is possible, requires considerable effort and cooperation. It also relies on common prognostic factors having been measured at a baseline, and in a similar way. For example, the atrial fibrillation collaboration could not include the predictive value of echocardiograms because most studies had not included this as a prognostic measure. Hence, such pooling will usually be confined to the prognostic factors common to all studies.

(Statistical note: it may not be necessary to combine all of the individual data in a single file but may be sufficient to pool the variance–covariance matrices – see the *Handbook of Research Synthesis*, Cooper and Hedges 1994.)

10.4.3 Assessing heterogeneity

When considering single prognostic factors, the issues and methods are similar to those for intervention studies (see Section 6.4.3). That is, is there heterogeneity of the size of the effect, and is it explained by study design or patient factors? However, for multiple factor prognostic studies, appropriate methods have not been described. If there is a pooling of individual study data then each factor within the multivariate model could be examined to check whether there were interactions with other factors.

APPENDIX A

MEMBERSHIP OF PRODUCTION TEAM FOR HANDBOOK

NHMRC Assessment Panel

Professor Paul O'Brien (Chair)	Department of Surgery, Monash Medical School Member of HAC
Professor Chris Silagy	Monash Institute of Public Health and Health Services Research Member of HAC
Professor John McCallum	Faculty of Health, University of Western Sydney Member of HAC

Consultant authors *

Associate Professor Paul Glasziou	Department of Social and Preventive Medicine, University of Queensland
Professor Les Irwig	Department of Public Health and Community Medicine, University of Sydney
Associate Professor Chris Bain	Department of Social and Preventive Medicine, University of Queensland
Professor Graham Colditz	Harvard School of Public Health, United States

Technical writer / editor

Dr Janet Salisbury	Biotext, Canberra

Secretariat

Ms Roz Lucas, Ms Janine Keough, Ms Monica Johns	Health Advisory Unit, Office of NHMRC

*The authors would also like to acknowledge the assistance of Dr Dianne O'Connell for provision of the Glossary and Table 1.2, Dr Joseph Lau for making Metatest available and Ms Ann McKibbon for help with some of the search methods.

APPENDIX B

PROCESS REPORT

During the 1997–99 NHMRC triennium the Health Advisory Committee focused its work on the areas of coordination and support rather than on collating and reviewing scientific evidence. However, the committee recognised that a key part of its coordination and support function was to provide a methodology on how to develop evidence-based guidelines.

The NHMRC publication *A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines* (NHMRC 1999), which had been produced by the health Advisory Committee as a resource for people wishing to develop clinical practice guidelines to a standard acceptable to the NHMRC, was revised during 1998. Early in the revision process, the committee realised that there was a need for a number of complementary handbooks to expand on the principles outlined in the document. This complementary series would cover other aspects of the identification, collation and application of scientific evidence. It was envisaged that these handbooks would be of invaluable assistance to agencies wishing to develop clinical practice guidelines of a high standard either independently, or on behalf of the NHMRC.

It was agreed that there would initially be five handbooks in the series:

- how to review the evidence;
- how to use the evidence;
- how to put the evidence into practice;
- how to present the evidence for consumers; and
- how to compare the costs and benefits.

They would be published individually to allow flexibility in their production and revision, as well as to allow any later additions to the series.

Recognising the need for a transparent and competitive process for contracting the services of an expert(s), tenders were sought for the preparation of each handbook. A selection committee was then appointed by the Health Advisory Committee to consider the tenders.

Once the successful tenderers had been contracted to prepare the handbooks, an assessment panel, composed of Health Advisory Committee members, was formed to manage the progress of each project (see Appendix A).

When first drafts of each handbook were received, they were distributed to a small number of experts in that particular field for peer review. The documents were subsequently revised in the light of these comments. A technical writer was employed to ensure consistency in content and style within and between the handbooks.

The finalised documents were referred, in turn, to the Health Advisory Committee for approval before being forwarded to the NHMRC for endorsement.

APPENDIX C

LITERATURE SEARCHING METHODS

Finding existing systematic reviews

As discussed in Section 1, it is always worth checking to see whether a previous systematic review or meta-analysis has already been done. Even if it needs modification or updating, it will provide a useful base of studies and issues. For interventions, Cochrane reviews are available in the Cochrane Database of Systematic Reviews (CDSR) in the Cochrane Library, which also contains the Database of Abstracts and Reviews (DARE). DARE compiles and appraises many nonCochrane reviews of both interventions and diagnostic accuracy.

Hunt and McKibbon (1997) have developed simple and complex search strategies for MEDLINE for finding systematic reviews and meta-analyses, which are applicable to all question types.

The simple search consists of the following steps:

1. **meta-analysis (pt)**
2. **meta-anal: (tw)**
3. **review (pt) AND medline (tw)**
4. **1 OR 2 OR 3**

[pt = publication type; tw = textword; : = wildcard symbol]

The comprehensive search consists of the following steps:

1. **meta-analysis (pt)**
2. **meta-anal: (tw)**
3. **metaanal: (tw)**
4. **quantitativ: review: OR quantitative: overview: (tw)**
5. **systematic: review: OR systematic: overview: (tw)**
6. **methodologic: review: OR methodologic: overview: (tw)**
7. **review (pt) AND medline (tw)**
8. **1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7**

An alternative strategy has recently been developed by the York Centre for Reviews and Dissemination, United Kingdom.[8]

Finding randomised trials

The following search strategy (developed by Kay Dickersin) is used by the Cochrane Collaboration to identify randomised trials in MEDLINE. This search is run regularly and forms part of the process of identifying trials for the Cochrane Controlled Trials Registry.

#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=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#9	#7 not #8
#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=ANIMAL not (TG=HUMAN and TG=ANIMAL)

[8] www.york.ac.uk/inst/crd/search.htm

#24	#22 not #23
#25	#24 not #9
#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=ANIMAL not (TG=HUMAN and TG=ANIMAL)
#34	#32 not #33
#35	#34 not (#9 or #25)
#36	#9 or #25 or #35

PubMed clinical queries using research methodology filters

A free MEDLINE facility is available from the National Library of Medicine.[9] A section of this is the PubMed Clinical Queries, which uses methodological filters developed by Haynes et al (1994) for many of the question types discussed in Part 2 of this handbook. These searches, which are shown below, are less extensive than the methods discussed in Part 2, but may be useful as a quick initial search.

[9] www.nlm.nih.gov/

Clinical queries using research methodology filters

Category	Optimised for:	ELHILL terms[a]	Sensitivity / specificity[b]	PubMed equivalent[c]
Therapy	sensitivity	randomized controlled trial (pt) or drug therapy (sh) or therapeutic use (sh) or all random: (tw)	99%/74%	'randomized controlled trial' [PTYP] | 'drug therapy' [SH] | 'therapeutic use' [SH:NOEXP] | 'random*' [WORD]
	specificity	all double and all blind: (tw) or all placebo: (tw)	57%/97%	(double [WORD] & blind* [WORD]) | placebo [WORD]
Diagnosis	sensitivity	exp sensitivity a#d specificity or all sensitivity (tw) or diagnosis & (px) or diagnostic use (sh) or all specificity (tw)	92%/73%	'sensitivity and specificity' [MESH] | 'sensitivity' [WORD] | ('diagnosis' [SH] | 'diagnostic use' [SH] | 'specificity' [WORD]
	specificity	exp sensitivity a#d specificity or all predictive and all value: (tw)	55%/98%	'sensitivity and specificity' [MESH] | ('predictive' [WORD] & 'value*' [WORD])
Aetiology	sensitivity	exp cohort studies or exp risk or all odds and all ratio: (tw) or all relative and all risk (tw) or all case and all control: (tw)	82%/70%	'cohort studies' [MESH] | 'risk' [MESH] | ('odds' [WORD] & 'ratio*' [WORD]) | ('relative' [WORD] & 'risk' [WORD]) | ('case' [WORD] & 'control*' [WORD])
	specificity	case-control studies or cohort studies	40%/98%	'case-control studies' [MH:NOEXP] | 'cohort studies' [MH:NOEXP]
Prognosis	sensitivity	incidence or exp mortality or follow-up studies or mortality (sh) or all prognos: (tw) or all predict: (tw) or all course (tw)	92%/73%	'incidence' [MESH] | 'mortality' [MESH] | 'follow-up studies' [MESH] | 'mortality' [SH] | prognos* [WORD] | predict* [WORD] | course [WORD]
	specificity	prognosis or survival analysis	49%/97%	prognosis [MH:NOEXP] | 'survival analysis' [MH:NOEXP]

[a] MEDLINE terms used by ELHILL (search engine for MEDLARS and Grateful MED)

[b] Sensitivity = the proportion of high quality studies in MEDLINE that are detected
Specificity = the proportion of irrelevant or poorer quality studies detected

[c] Approximate equivalent in PubMed query language, as used on the Clinical Queries Using Research Methodology Filters page (www.ncbi.nlm.nuh.gov/entrez/query.fcgi)

APPENDIX D

SOFTWARE FOR META-ANALYSIS

Many standard statistical software packages provide facilities that would enable meta-analysis (eg by the use of logistic regression). Some packages have had routines or macros specifically developed to allow meta-analysis. For example, STATA has a macro for the analysis of multiple 2 × 2 table data (this is available in the STATA Technical Bulletin Reprints, Volume 7, sbe16: Meta-analysis. S. Sharp and J. Sterne.[10]

In addition to standard statistical software, over the last several years a number of programs specifically designed to perform meta-analysis have been developed, some of which are described below. None of these is comprehensive or capable of performing all of the types of analysis discussed in Section 2. In particular, there is only one program available for performing adequate meta-analytic studies. Even for a single question type, however, it may be necessary to use more than one package to get an adequate range of analyses done. Some of the packages are freely available and others are commercial packages costing a few hundred dollars. This list is not comprehensive; another listing is also available on the Internet.[11]

Meta-analysis of intervention study

RevMan

This is the standard systematic review software for the Cochrane Collaboration.[12] This is a comprehensive package for managing the process of systematic reviews of intervention studies. RevMan is used for the writing of protocols, keeping the list of included and excluded publications, writing the Cochrane Review Text and performing the statistical meta-analytic functions. The latter are done through a separate program — metaview — which then enables the final reviews published in the Cochrane Library to be analysed by the same piece of software. Because of its comprehensive nature and structured processes, the package can be a little more difficult to learn than most others.

[10] www.stata.com/

[11] www.bmj.com/cgi/content/full/316/7126/221

[12] hiru.mcmaster.ca/cochrane/resource.htm

MetaAnalyst

This is an MS-DOS program developed by Dr Joseph Lau at the New England Medical Center and New England Cochrane Center in the United States. MetaAnalyst is designed to analyse a set of trials, each of whose data can be presented as a 2×2 table. The program performs a wide variety of meta-analytic statistics on trial data, providing both fixed and random fixed models for relative risk, odds ratio, and risk differences. It also provides heterogeneity statistics, cumulative meta-analytic plots and regressions against control rate, all with publication quality plots — however, the types of print are limited.

EasyMA

EasyMA is an MS-DOS program with a user-friendly interface developed to help physicians and medical researchers to synthesise evidence in clinical or therapeutic research. The program was developed by the Michel Cucherat, a teaching hospital in Lyon, France. The latest version (99) is available on the Internet.[13]

SCHARP

The Survival Curve and Hazard Ratio Program (SCHARP) for meta-analysis of individual patient data was developed by the Medical Research Council Cancer Trials Office, Cambridge, United Kingdom and the Instituto 'Maria Negri', Milan, Italy. This is a windows-based menu-driven program aimed at producing summary survival plots and 'forest plots' of the hazard ratios.

Comprehensive Meta-analysis

This is a versatile Windows 95 program. Like RevMan it creates a database of studies, with full citations that can be imported from MEDLINE. Data may be entered in a wide variety of formats, with conversions performed automatically. The available effect size indices include mean difference, correlations, rate difference, relative risk and odds ratio. The graphs allow subgroup and cumulative meta-analyses.[14]

[13] www.spc.univ-lyon1.fr/~mcu/easyma/

[14] www.Meta-analysis.com

Meta-analysis of diagnostic tests

Metatest

This is an MS-DOS program developed by Dr Joseph Lau at the New England Medical Center and New England Cochrane Center in the United States. This is a software package that produces a meta-analysis of diagnostic test accuracies. The plots include simple plots of sensitivity and specificity with their confidence intervals, summary receiver–operator curve (ROC) plots giving the data points of the individual studies, and a summary ROC (see Section 8). It is available on the Internet.[15]

[15] www.cochrane.org/cochrane/sadt.htm

Glossary

ATTENTION ALL READERS

Due to the 'cutting-edge' nature of the information contained in this handbook '*How to Review the Evidence: Systematic Identification and Review of the Scientific Literature*', the definitions contained in the Glossary (on page 95) relating to magnitude of treatment effect and strength of evidence have been modified. For these terms, readers should refer to strength of evidence and size of effect in the Glossary of the accompanying handbook '*How to Use the Evidence: Assessment and Application of Scientific Evidence*' (NHMRC 2000a).

GLOSSARY

Absolute risk reduction

The effect of a treatment can be expressed as the difference between relevant outcomes in the treatment and control groups by subtracting one rate (given by the proportion who experienced the event of interest) from the other. The reciprocal is the number needed to treat (NNT).

Accuracy (*see also* validity)

The degree to which a measurement represents the true value of the variable which is being measured.

Adverse event

A nonbeneficial outcome measured in a study of an intervention that may or may not have been caused by the intervention.

Allocation (or assignment to groups in a study)

The way that subjects are assigned to the different groups in a study (eg drug treatment/placebo; usual treatment/no treatment). This may be by a random method (*see* randomised controlled trial) or a nonrandom method (*see* pseudorandomised controlled study).

Applicability (*see also* external validity, generalisability)

The application of results to both individual patients and groups of patients. This term is preferred to generalisability as it includes the idea of particularising or individualising treatment and is closest to the general aim of clinical practice. It addresses whether a particular treatment that showed an overall benefit in a study can be expected to convey the same benefit to an individual patient.

Before-and-after study (*see also* pretest–post-test study)

A study design where a group of subjects is studied before and after an intervention. Interpretation of the result is problematic as it is difficult to separate the effect of the intervention from the effect of other factors.

Bias

Bias is a systematic deviation of a measurement from the 'true' value leading to either an over- or underestimation of the treatment effect. Bias can originate from many different sources, such as allocation of patients, measurement, interpretation, publication and review of data.

Blinding

Blinding or masking is the process used in epidemiological studies and clinical trials in which the observers and the subjects have no knowledge as to which treatments subjects are assigned to. This is done in order to minimise bias occurring in patient response and outcome measurement. In single-blind studies only the subjects are blind to their allocations, whilst in double-blind studies both observers and subjects are ignorant of the treatment allocations.

Bradford-Hill criteria (*see* causality)

Case-control study

Patients with a certain outcome or disease and an appropriate group of controls without the outcome or disease are selected (usually with careful consideration of appropriate choice of controls, matching, etc) and then information is obtained on whether the subjects have been exposed to the factor under investigation.

Case series

The intervention has been used in a series of patients (may or may not be consecutive series) and the results reported. There is no separate control group for comparison.

Causality

The relating of causes to the effects they produce. The Bradford–Hill criteria for causal association are: temporal relationship (exposure always precedes the outcome — the only essential criterion), consistency, strength, specificity, dose–response relationship, biological plausibility, coherence and experiment.

Clinical outcome

An outcome for a study that is defined on the basis of the disease being studied (eg fracture in osteoporosis, peptic ulcer healing and relapse rates).

Clinically important effect (*see also* statistically significant effect)

An outcome that improves the clinical outlook for the patient. The recommendations made in clinical practice guidelines should be both highly statistically significant *and* clinically important (so that the 95% confidence interval includes clinically important effects).

Cochrane Collaboration

The Cochrane Collaboration is an international network that aims to prepare, maintain and disseminate high quality systematic reviews based on randomised controlled trials (RCTs) and, when RCTs are not available, the best available evidence from other sources. It promotes the use of explicit methods to minimise bias, and rigorous peer review.

Cohort study

Data are obtained from groups who have been exposed, or not exposed, to the new technology or factor of interest (eg from databases). Careful consideration is usually given to patient selection, choice of outcomes, appropriate controls, matching, etc. However, data on outcomes may be limited.

Comparative study

A study including a comparison or control group.

Concurrent controls

Controls receive the alternative intervention and undergo assessment concurrently with the group receiving the new intervention. Allocation to the intervention or control is not random.

Confidence interval (CI)

An interval within which the population parameter (the 'true' value) is expected to lie with a given degree of certainty (eg 95%).

Confounding

The measure of a treatment effect is distorted because of differences in variables between the treatment and control groups that are also related to the outcome. For example, if the treatment (or new intervention) is trialled in younger patients then it may appear to be more effective than the comparator, not because it is better, but because the younger patients had better outcomes.

Cross-sectional study

A study that examines the relationship between diseases (or other health-related characteristics) and other variables of interest as they exist in a defined population at one particular time (ie exposure and outcomes are both measured at the same time).

Cumulative meta-analysis

In a systematic review, the results of the relevant studies are ordered by some characteristic, and sequential pooling of the trials is undertaken in increasing or decreasing order.

Degrees of freedom (df)

The number of independent comparisons that can be made between the members of a sample.

Discounting

The process by which benefits and costs are adjusted to net present values to take account of differential timing.

Double-blind study (*see* blinding)

Ecological fallacy

The bias that may occur because an association observed between variables on an aggregate (eg study or country) level does not necessarily represent the association that exists at an individual (subject) level.

Effect modification, effect modifier (*see also* interaction)

The relationship between a single variable (or covariate) and the treatment effect. Significant interaction between the treatment and such a variable indicates that the treatment effect varies across levels of this variable.

Effectiveness

The extent to which an intervention produces favourable outcomes under usual or everyday conditions.

Efficacy

The extent to which an intervention produces favourable outcomes under ideally controlled conditions such as in a randomised controlled trial.

Efficiency (technical and allocative)

The extent to which the maximum possible benefit is achieved out of available resources.

Evidence

Data about the effectiveness of a new treatment or intervention derived from studies comparing it with an appropriate alternative. Preferably the evidence is derived from a good quality randomised controlled trial, but it may not be.

Evidence-based medicine/health care

The process of finding relevant information in the medical literature to address a specific clinical problem. Patient care based on evidence derived from the best available studies.

External validity (*see also* generalisability, applicability)

Also called generalisability or applicability, is the degree to which the results of a study can be applied to situations other than those under consideration by the study, for example, for routine clinical practice.

Extrapolation

Refers to the application of results to a wider population and means to infer, predict, extend or project the results beyond that which was recorded, observed or experienced.

Generalisability (*see also* external validity, applicability)

Refers to the extent to which a study's results provide a correct basis for generalisation beyond the setting of the study and the particular people studied. It implies the application of the results of a study to another group or population.

Gold standard

A method, procedure or measurement that is widely regarded or accepted as being the best available. Often used to compare with new methods.

Hazard ratio (HR)

When time to the outcome of interest is known, this is the ratio of the hazards in the treatment and control groups where the hazard is the probability of having the outcome at time t, given that the outcome has not occurred up to time t.

Heterogeneity

Refers to the differences in treatment effect between studies contributing to a meta-analysis. If there is significant heterogeneity, this suggests that the trials are not estimating a single common treatment effect.

Historical controls

Data from either a previously published series or previously treated patients at an institution that are used for comparison with a prospectively collected group of patients exposed to the technology or intervention of interest at the same institution.

Incidence

The number of new events (new cases of a disease) in a defined population, within a specified period of time.

Intention to treat

An analysis of a clinical trial where participants are analysed according to the group to which they were initially randomly allocated, regardless of whether or not they dropped out, fully complied with the treatment, or crossed over and received the other treatment. By preserving the original groups one can be more confident that they are comparable.

Interaction

The relationship between a single variable (or covariate) and the treatment effect.

Interrupted time series

Treatment effect is assessed by comparing the pattern of (multiple) pretest scores and (multiple) post-test scores (after the introduction of the intervention) in a group of patients. This design can be strengthened by the addition of a control group which is observed at the same points in time but the intervention is not introduced to that group. This type of study can also use multiple time series with staggered introduction of the intervention.

Intervention

An intervention will generally be a therapeutic procedure such as treatment with a pharmaceutical agent, surgery, a dietary supplement, a dietary change or psychotherapy. Some other interventions are less obvious, such as early detection (screening), patient educational materials, or legislation. The key characteristic is that a person or their environment is manipulated in the hope of benefiting that person.

Level of evidence

Study designs are often grouped into a hierarchy according to their validity, or degree to which they are not susceptible to bias. The hierarchy indicates which studies should be given most weight in an evaluation.

Magnitude of treatment effect

Refers to the size (or the distance from the null value indicating no treatment effect) of the summary measure (or point estimate) of the treatment effect and the values included in the corresponding 95% confidence interval.

Meta-analysis

Results from several studies, identified in a systematic review, are combined and summarised quantitatively.

Meta-regression

The fitting of a linear regression model with an estimate of the treatment effect as the dependent variable and study level descriptors as the independent variables.

Nonrandomised cross-over design

Participants in a trial are measured before and after introduction or withdrawal of the intervention and the order of introduction and withdrawal is not randomised.

Null hypothesis

The hypothesis that states that there is no difference between two or more interventions or two or more groups (eg males and females). The null hypothesis states that the results observed in a study (eg the apparent beneficial effects of the intervention) are no different from what might have occurred as a result of the operation of chance alone.

Number needed to harm (NNH) (*see also* number needed to treat)

When the treatment increases the risk of the outcome, then the inverse of the absolute risk reduction is called the number needed to harm.

Number needed to treat (NNT) (*see also* number needed to harm)

When the treatment reduces the risk of specified adverse outcomes of a condition, NNT is the number of patients with a particular condition who must receive a treatment for a prescribed period in order to prevent the occurrence of the adverse outcomes. This number is the inverse of the absolute risk reduction.

Observational studies

Also known as epidemiological studies. These are usually undertaken by investigators who are not involved in the clinical care of the patients being studied, and who are not using the technology under investigation in this group of patients.

Odds ratio (OR)

Ratio of the odds of the outcome in the treatment group to the corresponding odds in the control group.

Patient expected event rate (PEER)

The probability that a patient will experience a particular event (eg a stroke or myocardial infarction) if left untreated. Also known as baseline risk.

Patient-relevant outcome

Any health outcome that is meaningful to the patient. It can be the best surrogate outcome, resources provided as part of treatment, impact on productivity (indirect) or one that cannot be measured accurately (eg pain, suffering). Common examples include: primary clinical outcomes, quality-of-life and economic outcomes.

Post-test only study

Patients undergo the intervention being studied and outcomes are described. This does not allow any comparisons.

Pretest–post-test study

Outcomes (pain, symptoms, etc) are measured in study participants before receiving the intervention being studied and the same outcomes are measured after. 'Improvement' in the outcome is reported. Often referred to as before-and-after studies.

Precision

Statistical precision indicates how close the estimate is to the true value. It is defined as the inverse of the variance of a measurement or estimate.

Prevalence

The measure of the proportion of people in a population who have some attribute or disease at a given point in time or during some time period.

Prognostic model

A statistical model that estimates a person's probability of developing the disease or outcome of interest from the values of various characteristics (such as age, gender, risk factors).

Pseudorandomised controlled study

An experimental comparison study in which subjects are allocated to treatment/intervention or control/placebo groups in a nonrandom way (such as alternate allocation, allocation by day of week, odd–even study numbers, etc). These groups may therefore differ from each other in ways other than the presence of the intervention being tested. This contrasts to 'true' experiments (RCTs) where the outcomes are compared for groups formed by random assignment (and are therefore equivalent to each other in all respects except for the intervention).

Publication bias

Bias caused by the results of a trial being more likely to be published if a statistically significant benefit of treatment is found.

***P*-value** (*see also* confidence interval, statistically significant effect)

The probability (obtained from a statistical test) that the null hypothesis (that there is no treatment effect) is incorrectly rejected.

NOTE: The *P*-value is often misunderstood. It does not, as commonly believed, represent the probability that the null hypothesis (that there is no treatment effect) is true (a small *P*-value therefore being desirable). The *P*-value obtained from a statistical test corresponds to the probability of claiming that there is a treatment effect when in fact there is no real effect.

Quality of evidence

Degree to which bias has been prevented through the design and conduct of research from which evidence is derived.

Quality of life

The degree to which a person perceives themselves able to function physically, emotionally and socially. In a more 'quantitative' sense, an estimate of remaining life free of impairment, disability or handicap as captured by the concept of quality-adjusted life-years (QALYs).

Random error

The portion of variation in a measurement that has no apparent connection to any other measurement or variable, generally regarded as due to chance.

Randomisation

A process of allocating participants to treatment or control groups within a controlled trial by using a random mechanism, such as coin toss, random number table, or computer-generated random numbers.

Randomised controlled trial

An experimental comparison study in which participants are allocated to treatment/intervention or control/placebo groups using a random mechanism (*see* randomisation). Participants have an equal chance of being allocated to an intervention or control group and therefore allocation bias is eliminated.

Randomised cross-over trial

Patients are measured before and after exposure to different interventions (or placebo) which are administered in a random order (and usually blinded).

Relative risk or risk ratio (RR)

Ratio of the proportions in the treatment and control groups with the outcome. This expresses the risk of the outcome in the treatment group relative to that in the control group.

Relative risk reduction (RRR)

The relative reduction in risk associated with an intervention. This measure is used when the outcome of interest is an adverse event and the intervention reduces the risk. It is calculated as one minus the relative risk, or:

RRR = 1 – (event rate in treatment group / event rate in control group)

Reliability

Also called consistency or reproducibility. The degree of stability that exists when a measurement is repeatedly made under different conditions or by different observers.

Risk difference (RD)

The difference (absolute) in the proportions with the outcome between the treatment and control groups. If the outcome represents an adverse event (such as death) and the risk difference is negative (below 0) this suggests that the treatment reduces the risk — referred to as the absolute risk reduction.

Selection bias

Error due to systematic differences in characteristics between those who are selected for study and those who are not. It invalidates conclusions and generalisations that might otherwise be drawn from such studies.

Statistically significant effect (*see also* clinically important effect)

An outcome for which the difference between the intervention and control groups is statistically significant (ie the *P*-value is 0.05). A statistically significant effect is not necessarily clinically important.

Strength of evidence

Magnitude, precision and reproducibility of the intervention effect (includes magnitude of the effect size, confidence interval width, *P*-value, and the exclusion of clinically unimportant effects). In the case of nonrandomised studies, additional factors such as biological plausibility, biological gradient and temporality of associations may be considered.

Surrogate outcome

Physiological or biochemical markers that can be relatively quickly and easily measured and that are taken as predictive of important clinical outcomes. They are often used when observation of clinical outcomes requires longer follow-up. Also called intermediate outcome.

Systematic review

The process of systematically locating, appraising and synthesising evidence from scientific studies in order to obtain a reliable overview.

Time series

A set of measurements taken over time. An interrupted time series is generated when a set of measurements is taken before the introduction of an intervention (or some other change in the system), followed by another set of measurements taken over time after the change.

Validity

- Of measurement: an expression of the degree to which a measurement measures what it purports to measure; it includes construct and content validity.

- Of study: the degree to which the inferences drawn from the study are warranted when account is taken of the study methods, the representativeness of the study sample, and the nature of the population from which it is drawn (internal and external validity, applicability, generalisability).

Variance

A measure of the variation shown by a set of observations, defined by the sum of the squares of deviation from the mean, divided by the number of degrees of freedom in the set of observations.

ACRONYMS AND ABBREVIATIONS

AIDS	acquired immune deficiency syndrome
CCTR	Cochrane Controlled Trials Registry
CD-ROM	Compact disk-read only memory
CDSR	Cochrane Database of Systematic Reviews
CI	confidence interval
Cochran Q	Cochran chi-square
DARE	Database of Abstracts and Reviews, Cochrane Library
df	degrees of freedom
exp	explode
FPR	false positive rate
HR	hazard ratio
IARC	International Agency for Research on Cancer
MeSH	Medical Subject Heading
NHMRC	National Health and Medical Research Council
NLM	National Library of Medicine (United States)
NNH	number needed to harm
NNT	number needed to treat
OR	odds ratio
P-value	probability
RCT	randomised controlled trial
RD	risk difference
ROC	receiver–operator curve
RR	relative risk/risk ratio
SCHARP	Survival Curve and Hazard Ratio Program developed by MRC Cancer Trials Office, Cambridge, United Kingdom
SROC	summary receiver–operator curve
TPR	true positive rate

REFERENCES

Allen IE and Olkin I (1999). Estimating time to conduct a meta-analysis from number of citations retrieved. Journal of the American Medical Association 282:634–635.

Atrial Fibrillation Investigators (1994). Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation: pooled data from five randomised controlled trials. Archives of Internal Medicine 28:957–960.

Begg CB and Mazumdar M (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics 50:1088–1101.

Berlin J (1997). Does blinding of readers affect the results of meta-analyses? Lancet 350:185–186.

Bradford-Hill A (1965). The environment and disease: association or causation. Proceedings of the Royal Society of Medicine 58:295–300.

Brenner HS and Savitz DA (1990). The effects of sensitivity and specificity of case selection on validity, sample size, precision, and power in hospital-based case-control studies. American Journal of Epidemiology 132:181–192.

Breslow NE and Day NE (1987). Statistical methods in cancer research 2: The design and analysis of cohort studies. Lyon: IARC.

Bruns DE (1997). Reporting diagnostic accuracy. Clinical Chemistry 43:2211.

Caldwell M and Watson R (1994). Peritoneal aspiration cytology as a diagnostic aid in acute appendicitis. British Journal of Surgery 81:276–278.

Colditz GA, Burdick E and Mosteller F (1995). Heterogeneity in meta-analysis of data from epidemiologic studies: a commentary. American Journal of Epidemiology 142:371–382.

Cooper AJ (1998). Systematic review of *Propionibacterium acnes* resistance to systemic antibiotics. Medical Journal of Australia 169:259–261.

Cooper H and Hedges LV (eds) (1994). Handbook of Research Synthesis. New York: Russell Sage Foundation.

Detsky A, Naylor CD, O'Rourke K, McGreer AJ and L'Abbe KA (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. Journal of Clinical Epidemiology 45:255–265.

Dickersin K, Min YI and Meinert CL (1992). Factors influencing publication of research results: followup of applications submitted to two institutional review boards. Journal of the American Medical Association 267:374–378.

Stoddart GL and Torrance GW (1997). Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press.

EBCTCG (Early Breast Cancer Trialists' Collaborative Group) (1992). 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 339:71–85.

Egger M, Smith GD, Schneider M and Minder C (1997). Bias in meta-analysis detected by a simple, graphical test. British Medical Journal 315:629–634.

Fahey M, Irwig L and Macaskill P (1995). Meta analysis of pap test accuracy. American Journal of Epidemiology 141:680–689.

Gelber R and Goldhirsch A (1987). Interpretation of results from subset analyses within overviews of randomised clinical trials. Statistics in Medicine 6:371–378.

Gueffier F, Boutitie F, Boissel JP et al (1995). INDANA: a meta-analysis on individual patient data in hypertension. Protocol and preliminary results. Therapie 50:353–362.

Guyatt GH and Rennie D (1993). User's guide to medical literature. Journal of the American Medical Association 270:2096–2097.

Guyatt GH, Sackett DL and Cook DJ (1993). Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. Journal of the American Medical Association 270:2598–2601.

Guyatt GH, Sackett DL and Cook DJ (1994). 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? Journal of the American Medical Association 271:59–63.

Hasselblad V and Hedges LV (1995). Meta-analysis of screening and diagnostic tests. Psychological Bulletin 117:167–178.

Haynes RB, Wilczynski N, McKibbon KA, Walker CJ and Sinclair JC (1994). Developing optimal search strategies for detecting clinically sound studies in Medline. Journal of the American Medical Informatics Association 1:447–458.

Henry D (1992). Economic analysis as an aid to subsidisation decisions: the development of Australian guidelines for pharmaceuticals. PharmacoEconomics 1:54–67.

Hunt DL and McKibbon KA (1997). Locating and appraising systematic reviews. Annals of Internal Medicine 126:532–38. (www.acponline.org/journals/annals/01apr97/ systemat.htm)

Irwig L, Tosteson ANA, Gastonis C et al (1994). Guidelines for meta-analyses evaluation of diagnostic tests. Annals of Internal Medicine 120:667–676.

Irwig L, Macaskill P, Glasziou P et al (1995). Meta-analytic methods for diagnostic test accuracy. Journal of Clinical Epidemiology 48:119–130.

Jaeschke R, Guyatt G and Sackett DL (1994a). Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results valid? Journal of the American Medical Association 271:389–391.

Jaeschke R, Guyatt GH and Sackett DL (1994b). Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? Journal of the American Medical Association 271:703–707.

Juni P, Witschi A, Bloch R and Egger M (1999). The hazards of scoring the quality of clinical trials for meta-analysis. Journal of the American Medical Association 208:1054–1060.

L'Abbe KA, Detsky AS and O'Rourke K (1987). Meta-analysis in clinical research. Annals of Internal Medicine 107:224–233.

Liddle J, Williamson M and Irwig L (1996). Method for Evaluating Research and Guideline Evidence. Sydney: NSW Health Department.

Lijmer J, Mol B, Heisterkamp S, Bonsel GJ and Bossuyt P (1999). Empirical evidence of bias in the evaluation of diagnostic tests. Journal of the American Medical Association 282:1061–1066.

Linde K, Ramirez G, Mulrow CD, Pauls A, Weidenhammer W and Melchart D (1996). St John's wort for depression—an overview and meta-analysis of randomised clinical trials. British Medical Journal 313:253–258.

Loy CT, Irwig LM, Katelaris PH and Talley NJ (1996). Do commercial serological kits for *Helicobacter pylori* infection differ in accuracy? A meta-analysis. American Journal of Gastroenterology 91:1138–1144.

McKibbon A, Eady A and Marks S (1999). PDQ Evidence-Based Principles and Practice. Hamilton: BC Becker Inc.

McManus RJ, Wilson S, Delaney BC et al (1998). Review of the usefulness of contacting other experts when conducting a literature search for systematic reviews. British Medical Journal 317:1562–-1563.

Mahoney MJ (1977). Publications prejudices: an experimental study of confirmatory bias in the peer review system. Cognitive Therapy and Research 1:161–175.

Moons KG, van Es GA, Deckers JW, Habbema JD and Grobbee DE (1997). Limitations of sensitivity, specificity, likelihood ratio and Bayes' Theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 8:12–17.

Moses L, Shapiro D and Littenberg B (1993). Combining independent studies of a diagnostic test into a summary ROC curve: data analytic approaches and some additional considerations. Statistics in Medicine 12:1293–1316.

Mulrow CD and Oxman A (1996). How to conduct a Cochrane Systematic Review. 3. Oxford: The Cochrane Collaboration.

Mulrow CD and Oxman A (1997). Cochrane Collaboration Handbook [updated Sept 1997]. Oxford: The Cochrane Collaboration.

NHMRC (National Health and Medical Research Council) (1999). A Guide to the Development, Implementation and Evaluation of Clinical Practice Guidelines. Canberra: NHMRC.

NHMRC (2000a). How to Use the Evidence: Assessment and Application of Scientific Evidence. Canberra: NHMRC.

NHMRC (2000b). How to Compare the Costs and Benefits: Evaluation of the Economic Evidence. Canberra: NHMRC.

Reid MC, Lachs MS and Feinstein AR (1995). Use of methodological standards in diagnostic test research. Getting better but still not good. Journal of the American Medical Association 274:645–651.

Rosenthal A (1979). The file drawer problem and tolerance for null results. Psychological Bulletin 86:638–641.

Rothman KJ and Greenland S (1998). Modern Epidemiology. Philadelphia: Lippincott-Raven.

Sankaranarayanan R, Becker N and Demaret E (1996). Directory of Ongoing Research in Cancer Epidemiology. International Agency for Cancer Reseach (IARC) Scientific Publication No 137. (see www-dep.iarc.fr/prevent.htm)

Schmid CH, McIntosh M, Cappelleri JC, Lau J and Chalmers TC (1995). Measuring the impact of the control rate in meta-analysis of clinical trials. Controlled Clinical Trials 16:66s.

Schmid CH, Lau J, McIntosh M and Cappelleri JC (1998). An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Statistics in Medicine 17:1923–1942.

Schulz KF, Chalmers I, Hayes RJ et al (1995). Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal of the American Medical Association 273:408–412.

Simes RJ (1987). Confronting publication bias: a cohort design for meta-analysis. Statistics in Medicine 6:11–29.

Stern J and Simes RJ (1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects. British Medical Journal 315:640–645.

Thompson SG (1995). Why sources of heterogeneity in meta-analysis should be investigated. British Medical Journal 309:1351–1355.

Towler B, Irwig L, Glasziou P, Kewenter J, Weller D and Silagy C (1998). A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, Hemoccult. British Medical Journal 317:559–565.

Tramer MR, Reynolds DJM, Moore RA and McQuay HJ (1997). Impact of covert duplicate publication on meta-analysis: a case study. British Medical Journal 315:635–640.

Valenstein PN (1990). Evaluating diagnostic tests with imperfect standards. American Journal of Clinical Pathology 93:252–258.

Whitehead A and Whitehead J (1991). A general parametric approach to the meta-analysis of randomized clinical trials. Lancet 10:1665–1677.