An Introduction to Evidence Synthesis

Ashraf Nabhan

Table of contents

Preface

Systematic reviews represent a cornerstone of contemporary evidence-based practice, providing a rigorous and reproducible approach to the identification, appraisal, and synthesis of research findings. In an era characterized by the exponential growth of scientific publications, the systematic review offers a methodological framework to distill reliable evidence, reduce bias, and enhance the validity of conclusions that inform clinical practice, public health policy, and future research priorities.

The primary purpose of this volume is to provide a comprehensive and structured overview of the principles and practices that underpin systematic reviewing. It is intended to serve as both an instructional resource and a reference text for a broad readership, including graduate students, early-career investigators, experienced researchers, and practitioners who engage with evidence synthesis in their respective fields.

The book is organized to guide the reader through the entire process of conducting a systematic review. Initial chapters address the conceptual foundations and historical development of systematic reviews, followed by detailed discussions of methodological components such as protocol development, question formulation, literature searching, study selection, critical appraisal, data extraction, synthesis methods, and reporting standards. Attention is also given to methodological challenges, innovations in automation and machine learning, and the emergence of living and rapid reviews, which reflect the dynamic evolution of the field.

The contributors to this volume bring extensive expertise and diverse perspectives, ensuring that the content is both methodologically rigorous and practically relevant. While the text seeks to establish best practices, it also acknowledges that systematic reviewing is a continually developing discipline, requiring adaptability, transparency, and critical reflection from those who practice it.

It is hoped that this work will provide readers not only with technical guidance but also with an appreciation of the broader epistemological and ethical considerations inherent in evidence synthesis. By advancing the capacity to conduct systematic reviews of high methodological quality, this book aspires to contribute to the integrity of research, the advancement of knowledge, and the improvement of outcomes for individuals and societies.

1 Introduction

There are several reasons why a new review may be considered. Commissioned calls for evidence synthesis are usually on topics where a gap in knowledge has been identified, prioritized and a question posed. Alternatively, the idea for a review may be investigator led, with a topic identified from an area of practice or research interest; such approaches may or may not be funded. Whatever the motivation for undertaking a review the preparation and conduct should be rigorous.

1.1 Is a review required?

Before undertaking a systematic review, it is necessary to check whether there are already existing or ongoing reviews, and whether a new review is justified. This process should begin by searching the Database of Abstracts of Reviews of Effects (DARE), and the Cochrane Database of Systematic Reviews (CDSR). DARE contains critical appraisals of systematic reviews of the effects of health interventions. CDSR contains the full text of regularly updated systematic reviews of the effects of health care interventions carried out by the Cochrane Collaboration. Other sites to consider searching include the National Institute for Health and Clinical Excellence (NICE) and the NIHR Health Technology Assessment (NIHR HTA) program websites. The Campbell Collaboration website3 contains the Campbell Library of Systematic Reviews giving full details of completed and ongoing systematic reviews in education, crime and justice, and social welfare, and the Evidence for Policy and Practice Information (EPPI) Centre, whose review fields include education, health promotion, social care and welfare, and public health, has a database of systematic and non systematic reviews of public health interventions (DoPHER). It may also be worth looking at sites such as the National Guidelines Clearinghouse (NGC) or the Scottish Intercollegiate Guidelines Network (SIGN),6 as many guidelines are based on systematic review evidence.

Searching the previous year of MEDLINE or other appropriate bibliographic databases may be helpful in identifying recently published reviews.

If an existing review is identified which addresses the question of interest, then the review should be assessed to determine whether it is of sufficient quality to guide policy and practice. In general, a good review should focus on a well-defined question and use appropriate methods. A comprehensive search should have been carried out, clear and appropriate criteria used to select or reject studies, and the process of assessing study quality, extracting and synthesising

data should have been unbiased, reproducible and transparent. If these processes are not well-documented, confidence in results and inferences is weakened. The review should report the results of all included studies clearly, highlighting any similarities or differences between studies, and exploring the reasons for any variations.

Critical appraisal can be undertaken with the aid of a checklist such as the example outlined here. Such checklists focus on identifying flaws in reviews that might bias the results.8 Quality assessment is important because the effectiveness of interventions may be masked or exaggerated by reviews that are not rigorously conducted. Structured abstracts included in the DARE database provide worked examples of how a checklist can be used to appraise and summarize reviews.

Critically appraising review articles

- Was the review question clearly defined in terms of population, interventions, comparators, outcomes and study designs (PICOS)?
- Was the search strategy adequate and appropriate? Were there any restrictions on language, publication status or publication date?
- Were preventative steps taken to minimize bias and errors in the study selection process?
- Were appropriate criteria used to assess the quality of the primary studies, and were preventative steps taken to minimize bias and errors in the quality assessment process?
- Were preventative steps taken to minimize bias and errors in the data extraction process?
- Were adequate details presented for each of the primary studies?
- Were appropriate methods used for data synthesis? Were differences between studies assessed? Were the studies pooled, and if so was it appropriate and meaningful to do so?
- Do the authors' conclusions accurately reflect the evidence that was reviewed?

If a high-quality review is located, but was completed some time ago, then an update of the review may be justified. Current relevance will need to be assessed and is particularly important in fields where the research is rapidly evolving. Where appropriate, collaboration with the original research team may assist in the update process by providing access to the data they used. However, little research has been conducted on when and how to update systematic reviews and the feasibility and efficiency of the identified approaches is uncertain. If a review is of adequate quality and still relevant, there may be no need to undertake another systematic review.

Where a new systematic review or an update is required, the next step is to establish a review team and possibly an advisory group, to develop the review protocol.

1.2 The review team

The review team will manage and conduct the review and should have a range of skills. Ideally these should include expertise in systematic review methods, information retrieval, the relevant clinical/topic area, statistics, health economics and/or qualitative research methods where appropriate. It is good practice to have a minimum of two researchers involved so that measures to minimize bias and error can be implemented at all stages of the review. Any conflicts of interest should be explicitly noted early in the process, and steps taken to ensure that these do not impact on the review process.

1.3 The advisory group

In addition to the team who will undertake the review there may be a number of individuals or groups who are consulted at various stages, including for example health care professionals, patient representatives, service users and experts in research methods. Some funding bodies require the establishment of an advisory group who will comment on the protocol and final report and provide input to ensure that the review has practical relevance to likely end users. Even if this is not the case, and even where the review team is knowledgeable about the area, it is still valuable to have an advisory group whose members can be consulted at key stages.

Engaging with stakeholders who are likely to be involved in implementing the recommendations of the review can help to ensure that the review is relevant to their needs. The particular form of user involvement will be determined by the purpose of the consultation. For example, when considering relevant outcomes for the review, users may suggest particular aspects of quality of life which it would be appropriate to assess. An example of a review which incorporated the views of users to considerable effect is one evaluating interventions to promote smoking cessation in pregnancy, which included outcomes more relevant to users as a result of their involvement. However, consultation is time consuming, and needs to be taken into account in the project timetable. Where reviews have strict time constraints, wide consultation may not be possible.

At an early stage, members of the advisory group should discuss the audiences for whom the review findings are likely to be relevant, helping to start the planning of a dissemination strategy from the beginning of the project.

The review team may also wish to seek more informal advice from other clinical or methodological experts who are not members of the advisory group. Likewise, where an advisory group has not been established, the review team may still seek advice from relevant sources.

Summary: Getting started

- Whatever the motivation for undertaking a review the preparation and conduct should be rigorous.

- A search of resources such as the DARE database should be undertaken to check for existing or ongoing reviews, to ensure a new review is justified.
- A review team should be established to manage and conduct the review. The membership should provide a range of skills, including expertise in systematic review methods, information retrieval, the relevant clinical/topic area, statistics, health economics and/or qualitative research methods where appropriate.
- Formation of an advisory group including, for example, health care professionals, patient representatives, services users and experts in research methods may be a requirement of some funding bodies. In any event, it may be valuable to have an advisory group, whose members can be consulted at key stages.
- The review team may wish to seek advice from a variety of clinical or methodological experts, whether an advisory group is convened.

2 Systematic reviews of healthcare interventions

2.1 The protocol

The review protocol sets out the methods to be used in the review. Decisions about the review question, inclusion criteria, search strategy, study selection, data extraction, quality assessment, data synthesis and plans for dissemination should be addressed.

Specifying the methods in advance reduces the risk of introducing bias into the review. For example, clear inclusion criteria avoid selecting studies according to whether their results reflect a favoured conclusion.

If modifications to the protocol are required, these should be clearly documented and justified. Modifications may arise from a clearer understanding of the review question and should not be made because of an awareness of the results of individual studies. Further information is given in Section 1.2.4 How to deal with protocol amendments during the review.

Protocol development is often an iterative process that requires communication within the review team and advisory group and sometimes with the funder.

2.1.1 Key areas to cover in a review protocol

This section covers the development of the protocol and the information it should contain. The formulation of the review objectives from the review question and the setting of inclusion criteria are covered in detail here as these must be agreed before starting a review. The search strategy, study selection, data extraction, quality assessment, synthesis and dissemination are also mentioned briefly as they are essential parts of the review protocol. However, to avoid repetition, full details of the issues related to both protocol requirements and carrying out the review are provided in Section 1.3 Undertaking the review.

2.1.2 Background

The background section should communicate the key contextual factors and conceptual issues relevant to the review question. It should explain why the review is required and provide the rationale underpinning the inclusion criteria and the focus of the review question, for example justifying the choice of interventions to be considered in the review.

2.1.3 Review question and inclusion criteria

Systematic reviews should set clear questions, the answers to which will provide meaningful information that can be used to guide decision-making. These should be stated clearly and precisely in the protocol. Questions may be extremely specific or very broad, although if broad, it may be more appropriate to break this down into a series of related more specific questions. For example a review to 'assess the evidence on the positive and negative effects of population-wide drinking water fluoridation strategies to prevent caries',13 was undertaken by addressing five objectives:

Objective 1: What are the effects of fluoridation of drinking water supplies on the incidence of caries?

Objective 2: If water fluoridation is shown to have beneficial effects, what is the effect over and above that offered by the use of alternative interventions and strategies?

Objective 3: Does water fluoridation result in a reduction of caries across social groups and between geographical locations, bringing equity?

Objective 4: Does water fluoridation have negative effects?

Objective 5: Are there differences in the effects of natural and artificial water fluoridation?

Where there are several objectives it may be necessary to prioritise by importance and likelihood of being able to answer the question. It may even be necessary to restrict the scope of the question to a level that is manageable within set resources. For clarity, the singular term 'review question' is used throughout the guidance.

Example review objective and PICOS elements for a review protocol

Review objective

> The objective of this review is to assess the clinical effectiveness of treatments for childhood retinoblastoma.

Participants

Studies of participants diagnosed with retinoblastoma at the age of 18 years or under.

Studies of adults where childhood retinoblastoma was followed up into adulthood.

Studies of mixed diagnoses if outcomes were reported separately for children with retinoblastoma.

Interventions

Any intervention or combination of interventions given for the treatment of retinoblastoma, including (but not restricted to) enucleation, external beam radiotherapy, chemotherapy, brachytherapy, cryotherapy, thermotherapy and photocoagulation.

Outcomes

Any clinical outcome, including (but not restricted to) survival, progression-free survival, tumour response, preservation of the eye, visual acuity, disease remission and adverse effects.

Study design

Randomised controlled trials (RCTs) and controlled trials. However, it is not anticipated that many studies of these designs will be available. Therefore, if information from controlled trials is not available, cohort studies are eligible for inclusion provided that data from a comparison group are reported. Case series and case reports are excluded from the review owing to the high potential for bias in these study designs. Case–control studies (except where nested as part of a cohort study) and economic evaluations are also excluded.

2.1.4 Defining PICOS

The review question can be framed in terms of the population, intervention(s), comparator(s) and outcomes of the studies that will be included in the review. These elements of the review question, together with study design, will then be refined in order to determine the specific inclusion criteria that will be used when selecting studies for the review. Although both the acronyms PICO or PICOS are commonly used, here the term PICOS will be used throughout for consistency. In some situations, not all the elements will be relevant, for example not every review question will specify type of study design to be included. The use of PICOS in the context of reviews incorporating different study designs is discussed in the relevant chapters.

The review question may be presented in general terms, for example, 'What is the best treatment option for retinoblastoma?' More often the actual question is discussed by the review team and an objective, or series of objectives, framed by the population, the intervention and

the outcome(s) of interest agreed. For example, 'The objective of this review is to assess the clinical effectiveness of treatments for childhood retinoblastoma.'14

2.1.4.1 Population

The included population should be relevant to the population to which the review findings will be applied, and explicit inclusion criteria should be defined in terms of the disease or condition of interest. Any specified restrictions should be clinically justifiable and relevant. Eligibility must usually be applied to the whole study and consideration of how to deal with studies that include a mixed population, some of whom are relevant to the review and some of whom are not, is required. If the inclusion criteria are broad, it may be informative to investigate effectiveness across subgroups of participants.

However, in the absence of individual patient data (IPD), or very detailed reporting of data, broken down by participant characteristics, it is unlikely that inclusion can be restricted to particular types of participant or that detailed subgroup analyses will be possible. Where analysis of participant subgroups is planned, this should be specified in the protocol. Examples of factors that may be investigated include participants' gender, age, disease severity, the presence of any co-morbidities, socio-economic status, ethnicity and geographical area.

2.1.4.2 Interventions and comparators

The nature of the interventions explored in the review may be framed in very broad terms like 'psychosocial interventions' or may be more specific such as 'cognitive behavioural therapy'. Factors usually specified include the precise nature of the intervention (e.g. the method of administration of a drug), the person delivering the intervention (e.g. a community psychiatric nurse versus a non-professional carer) or setting in which the intervention is delivered (e.g. inpatient or outpatient).

Where comparative studies are to be included, the protocol should also specify which comparators are eligible. As with the interventions, comparators should be carefully defined, so that the scope of a term such as 'palliative care' or 'usual care' is clear.

The protocol should also specify whether any co-interventions carried out at the same time affect eligibility for inclusion; this applies to both the intervention(s) and the comparator(s).

2.1.4.3 Outcomes

The success or failure of a therapeutic intervention will usually be assessed in terms of differences in mortality or morbidity in the populations treated. Primary outcomes are likely to include measures of mortality and morbidity but other outcomes may also be of importance, for

example measures of quality of life and participants' subjective experiences of pain or physical functioning.

A review should explore a clearly defined set of relevant outcomes and it is important to justify each outcome included. Input from the advisory group and the findings from initial scoping searches and qualitative research may be helpful in deciding which outcomes to include.

The use of surrogate outcomes may be misleading, giving an over or underestimate of the true clinical outcome.15 Decisions about whether to consider surrogate outcomes should therefore be informed by available evidence about associations between the surrogate (e.g. blood pressure) and the outcome of interest (e.g. stroke). Often, surrogate outcomes are included only where a study also reports a relevant clinical outcome.

The review may also consider the timing of outcome assessment and possible adverse effects of the intervention. If the review is considering cost-effectiveness or economic issues as well as clinical effectiveness, the relevant economic outcomes should also be specified.

Although the review may aim to consider a series of outcomes, it is rare that inclusion would be restricted to only those studies that report all the outcomes of interest. More usually inclusion criteria will require that included studies report the main outcome.

2.1.4.4 Study design

The types of study included in the review will play a major role in determining the reliability of the results and the validity of estimates of effect is linked to the study design. While some study designs are clearly more robust than others, this should not be the only factor in determining which types of study are eligible for inclusion.16

Scoping searches may reveal that there are likely to be only a limited number of relevant randomised studies. In this case researchers have the option of justifying a decision to limit study design, bearing in mind that the identification of gaps in the current evidence base may in itself be a significant finding of the review. Alternatively, they can include quasi-experimental or observational studies. For reviews in some topic areas, these may be the only types of study available. The study design inclusion criteria given as an example in Box 1.2 have been set to take account of the paucity of experimental studies, as indicated by the scoping searches.

In some cases a range of study designs may be needed to address different questions within the same review. For example, a review seeking to include information on adverse events will often include case-control and/or case-series (see Chapter 4) whilst a review incorporating participants' experiences of an intervention is likely to include qualitative studies (see Chapter 6). The potential biases from the inclusion of a range of study designs are discussed in Section 1.3.4 Quality assessment.

2.1.5 Defining inclusion criteria

The inclusion criteria should be set out in the protocol, to ensure that the boundaries of the review question are clearly defined. In the example in Box 1.2, the population to be studied was specified in the review question as those with 'childhood' retinoblastoma. In addition to qualifying 'childhood' as under 18, appropriate timeframes for disease progression and treatment and the possibilities of concurrent disease processes have been taken into account. In reviews of interventions relating to other diseases it may be necessary to be more specific about how the disease of interest will be verified, and to specify the disease stage and severity. In the simple example given in Box 1.2 the key interventions and outcomes of interest are listed.

The nature of the intervention(s) and comparator(s) should be specified in detail. Whilst this may be more straightforward for drug interventions, more complex interventions may require detailed consideration of terms. For example, interventions such as 'stress management' or 'relaxation' may be defined differently by different study authors.

Therefore researchers need to be clear about their own definitions and what elements are acceptable. An operational definition describing the content and delivery of the intervention will usually be helpful.

The inclusion criteria should capture all studies of interest. If the criteria are too narrowly defined there is a risk of missing potentially relevant studies and the generalisability of the results may be reduced. On the other hand, if the criteria are too broad the review may contain information which is hard to compare and synthesise.17,18 Inclusion criteria also need to be practical to apply; if they are too detailed, screening may become overly complicated and time consuming.

2.1.6 Methodological quality

As previously stated, a review should be based on the best quality evidence available (see Box 1.3). Whatever the study design(s) included, it should not be assumed that all studies of the same basic design (e.g. RCT) are equally well-conducted. The quality of the included studies should be formally assessed as this will impact on the reliability of the results and therefore on the conclusions drawn. Although quality assessment can sometimes be used to exclude studies that do not meet certain criteria, this is not standard practice and differential quality is more usually assessed at the synthesis stage through sensitivity analysis. For further information see Section 1.3.4 Quality assessment and Section 1.3.5 Data synthesis.

Hierarchy of study designs to assess the effects of interventions

Randomised controlled trials

The simplest form of RCT is known as the parallel group trial which randomises eligible participants to two or more groups, treats according to assignment, and compares the groups with

respect to outcomes of interest. Participants are allocated to groups using both randomisation (allocation involves the play of chance) and concealment (ensures that the intervention that will be allocated cannot be known in advance). There are different types of randomised study designs, such as:

Randomised cross-over trials

Where all participants receive all the interventions; for example in a two arm cross-over trial, one group receives intervention A before intervention B, and the other group receive intervention B before intervention A. It is the sequence of interventions that is randomised.

Cluster randomised trials

A cluster randomised trial is a trial where clusters of people rather than single individuals are randomised to different interventions. For example, whole clinics or geographical locations may be randomised to receive particular interventions, rather than individuals.

Quasi-experimental studies

The main distinction between randomised and quasi-experimental studies is the way in which participants are allocated to the intervention and control

groups; quasi-experimental studies do not use random assignment to create the comparison groups.

Non-randomised controlled studies

Individuals are allocated to a concurrent comparison group, using methods other than randomisation. The lack of concealed randomised allocation increases the risk of selection bias.

Before-and-after study

Comparison of outcomes in study participants before and after the introduction of an intervention. The before-and-after comparisons may be in the same sample of participants or in different samples.

Interrupted time series

Interrupted time series designs are multiple observations over time that are 'interrupted', usually by an intervention or treatment.

Observational studies

A study in which natural variation in interventions or exposure among participants (i.e. not allocated by an investigator) is investigated to explore the effect of the interventions or exposure on health outcomes.

Cohort study

A defined group of participants is followed over time and comparison is made between those who did and did not receive an intervention.

Case-control study

Groups from the same population with (cases) and without (controls) a specific outcome of interest, are compared to evaluate the association between exposure to an intervention and the outcome.

Case series

Description of a number of cases of an intervention and the outcome (without comparison with a control group). These are not comparative studies.

2.1.7 Language

The ideal for most systematic reviews is to include all available relevant evidence. In principle, this includes studies written in any language to avoid the introduction of language bias into the review. Language bias arises because studies with statistically significant results that have been conducted in non-English speaking countries may be more likely to be published in English language journals than those with nonsignificant results.19 In addition, trials originating in certain countries have been found to have unusually high proportions of positive results.20

Thus, if reviews include only studies reported in English, their results and inferences may be biased.19-21 Even if language bias does not influence summary effect estimates, it is likely to affect precision, because analysis will be based on fewer data.22 Whenever feasible, all relevant studies should be included regardless of language. However, realistically this is not always possible due to a lack of time, resources and facilities

for translation. It is advisable therefore, to identify all non-English language papers, document their existence, but record 'language' as the reason for exclusion in cases where they cannot be dealt with. Although titles and abstracts are translated in many databases, full papers are usually only available in their primary language.

When a decision is made to include non-English language studies, the review question should inform the decision about which languages are chosen, as studies of particular interventions and/or settings are more likely than others to be published in certain languages. An investigation of the inclusion of non-English language reports of RCTs in systematic reviews concluded that language restrictions do not appear to bias

the estimates in reviews of conventional interventions, but may bias the results of complementary or alternative medicines.23 Researchers need to give careful thought as to whether imposing language restrictions may potentially bias the results of their individual review. When non-English language literature is included in a review, its influence on the estimation and precision of effect may be explored in a sensitivity analysis.

2.1.8 Publication type/status

Studies are not always published as full papers in peer-reviewed journals; they may be published as reports, book chapters, conference abstracts, theses or they may be informally reported or remain unpublished. Ideally a review should aim to include all relevant studies, regardless of publication status, in order to avoid publication bias.

Publication bias occurs when the publication of a study is influenced by its results, hence inclusion of only published studies may overestimate the intervention effect.24

There are practical issues that limit the inclusion of all studies regardless of publication type/status. Unpublished studies are likely to be harder to source, and more difficult to obtain, than published studies. The inclusion of conference abstracts and interim results should be considered, bearing in mind that contact with the study authors

may be required to obtain full study details.25 The effects of including any data from abstracts alone should be carefully considered, since differences often occur between data reported in conference abstracts and their corresponding full reports, although differences in results are seldom large.26, 27 Also, it can be difficult to appraise study quality from minimal details provided in an abstract. Sensitivity analyses may be carried out to examine the effect of including data from conference abstracts.28

The identification of ongoing studies is important for a number of reasons. They may provide a useful starting point for subsequent reviews and updates; they may also improve the quality of conclusions about future research by indicating where new research has already commenced. Information about ongoing studies may be available as 'partially published research' like conference abstracts – these can be classified as ongoing studies which may contribute to future reviews.29

2.1.9 Identifying research evidence

A preliminary search strategy for identifying relevant research should be included in the protocol. This should specify the databases and additional sources that will be searched, and also the likely search terms to be used. The search strategy should be constructed to take into account PICOS, although the outcome(s) of studies and/or study design are not always used. Incorporating decisions about publication status and language restrictions also needs to be made at this stage. In reviews of one year or more duration, or reviews in rapidly evolving fields, provision for repeating the searches towards the end of the review process should also be considered. In addition it may be useful to carry out current awareness searches to identify relevant papers as they are published. The approach taken will depend on the question and the topic, and also on the available time and resources. It is usual to include in the protocol details of the software that will be used to manage references. Further information is given in Section 1.3.1 Identifying research evidence for systematic reviews.

2.1.9.1 Study selection

Study selection is usually conducted in two stages: an initial screening of titles and abstracts against the inclusion criteria to identify potentially relevant papers followed by screening of the full papers identified as possibly relevant in the initial screening. The protocol should specify the process by which decisions on the selection of studies will be made. This should include the number of researchers who will screen titles and abstracts and then full papers, and the method for resolving disagreements about study eligibility. Section 1.3.2 Study selection contains more information.

2.1.9.2 Data extraction

The protocol should outline the information that will be extracted from studies identified for inclusion in the review and provide details of any software to be used for recording the data. As with study selection the protocol should state the procedure for data extraction including the number of researchers who will extract the data and how discrepancies will be resolved. The protocol should also specify whether authors of primary studies will be contacted to provide missing or additional data. If foreign language papers are to be included, it may be necessary to specify translation arrangements. Further information is given in Section 1.3.3 Data extraction.

2.1.10 Quality assessment

The protocol should provide details of the method of study appraisal to be used, including examples of the specific quality criteria. Details of how the study appraisal is to be used should be specified, for example whether the results will inform sensitivity analyses. The protocol should also specify the process for conducting the appraisal of study quality, the number of researchers involved, and how disagreements will be resolved. For a detailed discussion of these issues see Section 1.3.4 Quality assessment.

2.1.11 Data synthesis

As far as possible, the protocol should specify the strategy for data synthesis. It should state whether a meta-analysis is planned, although whether a planned meta-analysis will ultimately prove possible will depend on the studies and data that are available. As analyses will depend on what data are available, and because it is difficult to anticipate all of the statistical issues that may arise, it can be difficult to pre-specify full details of the planned synthesis. However, the protocol should outline how heterogeneity will be explored and quantified, under what circumstances a meta-analysis would be considered appropriate and whether a fixed or random-effects model or both would be used. Where appropriate, the approach to narrative synthesis should also be outlined. The protocol should also specify the outcomes of interest and what

effect measures will be used. Any planned subgroup or sensitivity analyses or investigation of publication bias should also be described. Further information is given in Section 1.3.5 Data synthesis.

2.1.12 Dissemination

Dissemination of findings is an integral part of the review process and fundamental to ensuring that the essential messages from the review reach the appropriate audiences. It is helpful to consider how the review findings will be disseminated from as early a stage as possible to allow adequate time for planning and development and to ensure that the proposed activities are properly resourced. Details are given in Section 1.3.8 Disseminating the findings of systematic reviews.

2.1.13 Approval of the draft protocol

Some commissioning or funding bodies may require that they formally approve the protocol, and will provide input to the draft protocol, in addition to the other stakeholders, such as clinical and methodological experts, patient groups and service users, who may be consulted. For commissioned reviews, even where it is not a specific requirement, it can be useful to communicate with the commissioner at the protocol development stage. This will help to ensure that the protocol meets the commissioning brief or where the review question or the scope of the project has been altered, that this is agreed before work commences.

2.1.14 How to deal with protocol amendments during the review

Sticking rigidly to a protocol when it becomes apparent that a change of direction is required, can result in a review that is not useful to end users. It is possible that consideration of the primary research may raise questions which were not anticipated at the protocol stage. Where this results from a clearer understanding of the review question, it can be appropriate to carry out documented and justified amendments to the protocol. In the report of the review findings it is helpful to distinguish between the initial review question and any subsequent amendments. It is never appropriate to modify the protocol because of awareness of the results of individual studies, as this is likely to introduce bias and affect the validity of the review's conclusions.

Many reviews undergo protocol modification.30 Where modifications are a possibility, the implications for the review process and workload should be considered carefully. In particular, the likely impact on the literature search should be assessed, as it may require modification and running again. Data extraction forms may also need to be amended, and any data that have already been extracted might require some re- working. Protocol amendments should be documented in a protocol addendum and in the final report of the review.

Summary: The review protocol

- The protocol sets out in advance the methods to be used in the review with the aim of minimizing bias.
- The background section of the protocol should communicate the key contextual and conceptual factors relevant to the review question and provide the justification for the review.
- The protocol should specify the review question.
- Study inclusion and exclusion criteria should be clearly defined using the relevant PICOS elements.
- The protocol should also specify the methods which will be used to:
  - Identify research evidence
  - Select studies for inclusion
  - Extract Data from included studies
  - Assess the risk of bias in the included studies
  - Synthesize results
  - Disseminate the review findings
- In cases when it becomes apparent that a modification to the protocol is required, protocol amendments should be clearly documented and justified.

2.2 Conducting the synthesis

2.2.1 Identifying research evidence for systematic reviews

This section describes how to undertake a systematic search using a range of methods to identify studies, manage the references retrieved by the searches, obtain documents and write up the search process. Practical examples of constructing search strategies are given in Appendix 2, and Appendix 3 provides examples of how the search should be documented. Issues around the identification of research evidence that are specific to review type such as adverse effects or clinical tests are discussed in the relevant chapters.

Conducting a thorough search to identify relevant studies is a key factor in minimizing bias in the review process. The search process should be as transparent as possible and documented in a way that enables it to be evaluated and reproduced.

Studies can be located using a combination of the following approaches:

- Searching electronic databases

- Visually scanning reference lists from relevant studies
- Handsearching key journals and conference proceedings
- Contacting study authors, experts, manufacturers, and other organizations
- Searching relevant Internet resources
- Citation searching
- Using a project Internet site to canvas for studies

2.2.1.1 Minimizing publication and language biases

Decisions about where and how to search could unintentionally introduce bias into the review, so the team needs to consider, and try to minimize, the possible impact of search limitations. For example, restricting the searching to the use of electronic databases, which consist mainly of references to published journal articles, could result in the review being subject to publication bias as this approach is unlikely to identify studies that have not been published in peer reviewed journals. Wider searching is needed to identify research results circulated as reports or discussion papers. The identification of grey literature, such as unpublished papers, is difficult, but some are included on databases such as NTIS (National Technical Information Service) and HMIC (Health Management Information Consortium). Libraries of specialist research organizations and professional societies may also provide access to collections of grey literature.

Searching databases and registers that include unpublished studies, such as records of ongoing research, conference proceedings and theses, can reduce the impact of publication bias. Conference proceedings provide information on both research in progress and completed research. Conference abstracts are recorded in some major bibliographic databases such as BIOSIS Previews, as well as in dedicated databases such as Index to Scientific and Technical proceedings, ZETOC, and the Conference Papers Index.31-34 It is also worth consulting catalogues from major libraries, for example the British Library and the US National Library of Medicine. The abstracts in conference proceedings may only give limited information, and there can be differences between data presented in an abstract and that included in a final report.35, 36 For these reasons, researchers should try to acquire the full report, if there is one, before considering whether to include the results in a systematic review.

As already discussed, limiting searches to English language papers can introduce language bias. Large bibliographic databases, such as MEDLINE and EMBASE, do include a small number of non-English language journals.37 Using additional databases such as LILACS (Latin American and Caribbean Health Sciences Literature) that contain collections of non-English language research can minimize potential language bias.

2.2.1.2 Searching electronic databases

The selection of electronic databases to search will depend upon the review topic. Lists of databases are available from libraries and from database providers, such as Dialog and Wolters Kluwer, while subject experts will be familiar with the bibliographic databases in their field.

For reviews of health care interventions, MEDLINE and EMBASE are the databases most commonly used to identify studies. The Cochrane Central Register of Controlled Trials (CENTRAL) includes details of published articles taken from bibliographic databases and other published and unpublished sources.38 There are other databases with a narrower focus that could be equally appropriate. These include PsycINFO (psychology and psychiatry), AMED (complementary medicine), MANTIS (osteopathy and chiropractic) and CINAHL (nursing and allied health professions). If the topic includes social care there are a range of databases available including ASSIA (Applied Social Sciences Index and Abstracts), CSA Sociological Abstracts, and CSA Social Services Abstracts, that could be used. The databases referred to above are all subject-based but there are others, such as AgeInfo, Ageline and ChildData, that focus on a specific population group that could be relevant to the review topic.

Due to the diversity of questions addressed by systematic reviews, there can be no agreed standard for what constitutes an acceptable search in terms of the number of databases searched. For example, if the review is on a cross-cutting public health topic such as housing and health it is advisable to search a wider range of databases than if the review is of a pharmaceutical intervention for a known health condition (see Chapter 3, Section 3.3 Identifying research evidence).

2.2.1.3 Searching other sources

In addition to searching electronic databases, published and unpublished research may also be obtained by using one or more of the following methods.

Scanning reference lists of relevant studies

Browsing the reference lists of papers (both primary studies and reviews) that have been identified by the database searches may identify further studies of interest.

Handsearching key journals

Handsearching involves scanning the content of journals, conference proceedings and abstracts, page by page. It is an important way of identifying very recent publications that have not yet been included and indexed by electronic databases or of including articles from journals that are not indexed by electronic databases.39 Handsearching can also ensure complete coverage of journal issues, including letters or commentaries, which may not be indexed by databases. It can also compensate for poor or inaccurate database indexing that can result in even the most carefully constructed strategy failing to identify relevant studies. Selecting which journals

to handsearch can be done by analysing the results of the database searches to identify the journals that contain the largest number of relevant studies.

Searching trials registers

Trials can be identified by searching one or more of the many trials registers that exist. It can be a particularly useful approach to identifying unpublished or ongoing trials.

Many of the registers are available on the Internet and some of the larger ones, such as www.ClinicalTrials.gov and www.who.int/trialsearch/, include the facility to search by drug name or by condition. While some registers are disease specific, others collect together trials from a specific country or region. Pharmaceutical companies may also make information about trials they have conducted available from their websites.

Contacting experts and manufacturers

Research groups and other experts as well as manufacturers may be useful sources of research not identified by the electronic searches, and may also be able to supply information about unpublished or ongoing research. Contacting relevant research centers or specialist libraries is another way of identifying potential studies. While these methods can all be useful, they are also time consuming and offer no guarantee of obtaining relevant information.

After a thorough and systematic search has been conducted, and relevant studies have been identified, topic experts can be asked to check the list to identify any known missing studies.

Searching relevant Internet resources

Internet searching can be a useful means of retrieving grey literature, such as unpublished papers, reports and conference abstracts. Identifying and scanning specific relevant websites will usually be more practical than using a general search engine such as 'Google'.

Reviews of transport and 'welfare to work' programmes have reported how Internet searching of potentially relevant websites was effective in identifying additional studies to those retrieved from databases.40, 41 It is worth considering using the Internet when investigating a topic area where it is likely that studies have been published informally rather than in a journal indexed in a bibliographic database.

Internet searching should be carried out in as structured a way as possible and the procedure documented.

Citation searching

Citation searching involves selecting a number of key papers already identified for inclusion in the review and then searching for articles that have cited these papers. This approach should identify a cluster of related, and therefore highly relevant, papers.

As this is in effect a search forward through time, citation searching is not suitable for identifying recent papers as they cannot have been referenced by other older papers.

Citation searching used to be limited to using the indexes Science Citation Index Expanded, Social Sciences Citation Index, and Arts & Humanities Citation Index,. but other resources (including CINAHL, PsycINFO and Google Scholar) now include cited references in their records so these are also available for citation searching. Using similar services offered by journals such as the BMJ can also be helpful.

Using a project Internet site to canvas for studies

Where it has been agreed that a dedicated website should be set up for the review, for example as part of the overall dissemination strategy, this can be used to canvas for unpublished data/grey literature. Inclusion of an email contact address allows interested parties to submit information about relevant research. Posting the inclusion and exclusion criteria on the website may help to ensure submissions are appropriate. Throughout the review process the website should be continually updated with information about the studies identified. Personal responses should be sent to all respondents and where appropriate submitted material should be included in the library of references. Further details about dedicated project websites can be found in Section 1.3.8 Disseminating the findings of systematic reviews.

This approach should probably only be considered for 'high profile' reviews and then it should be as an adjunct to active canvassing for unpublished/grey literature.

2.2.1.4 Constructing the search strategy for electronic databases

Search strategies are explicitly designed to be highly sensitive so as many potentially relevant studies as possible are retrieved. Consequently the searches tend to retrieve a large number of records that do not meet the inclusion criteria. While it is possible to increase the precision of a search strategy, and so reduce the number of irrelevant papers retrieved, this may lead to relevant studies being missed.42

Constructing an effective combination of search terms involves breaking down the review question into 'concepts'. Using the Population, Intervention, Comparator, and Outcomes elements from PICOS can help to structure the search, but it is not essential that every element is used. For example it may be better not to use terms for the outcomes since inclusion might mean that the database being searched fails to show relevant studies simply because the outcome is not mentioned prominently enough in the record, even though the study measured it. For each of the elements used, it is important to consider all the possible alternative terms. For example, a drug intervention may be known by a generic name and one or more proprietary names. Advice should be sought from the topic experts on the review team and advisory group.

For a detailed discussion of how to structure a search from a review question, including the use of search filters for study design, see Appendix 2.

2.2.1.5 Text mining

Text mining is a rapidly developing approach to utilizing the large amount of published text now available. Its potential use in systematic reviews is currently being explored and it may in future be an additional useful way of identifying relevant studies.43, 44 The aim of text mining is to identify connections between seemingly unrelated facts to generate new ideas or hypotheses. A number of processes are involved in the technique: a) Information Retrieval identifies documents to match a user's query; b) Natural Language Processing provides linguistic data needed to perform c) Information Extraction, the process of automatically obtaining structured data from an unstructured natural language document; and d) Data Mining, the process of identifying patterns in large sets of data.45, 46 In future this approach may be helpful in automatically screening and ranking large numbers of potentially eligible studies prior to assessment by the researchers.

There are a variety of text mining tools available, for example TerMine and Acromine47 are tools dealing with term extraction and variation. Also of interest are KLEIO,48 which provides advanced searching facilities across MEDLINE and FACTA, which finds associated concepts using text analysis.49 Further information about text mining and the use of these tools can be found on the National Centre for Text Mining website (www.nactem.ac.uk/).)

2.2.1.6 Updating literature searches

Depending on the scope and timescale of the review, an update of the literature searches towards the end of the project may be required. If the initial searches were carried out some time before the final analysis is undertaken (e.g. six months) it may be necessary to re-run the searches to ensure that no recent papers are missed. To do this successfully the date the original search was conducted and the years covered by the search must have been recorded.

When doing update searches the update date field should be used rather than the actual date. This ensures that anything added to the database since the original search was conducted will be identified. If the database has added a lot of older material (e.g. from 1967) this will be removed by using the original date limits (e.g. 1990-2008) in combination with the update date field. For databases that do not include an update date field it may be better to run the whole search again and then use reference management software to remove those records that have already been identified and assessed.

2.2.1.7 Current awareness

If a review is covering an area where there is rapid change or if a major study is expected to report its findings soon, setting up current awareness alerts can ensure that new papers are identified as soon as they become available. Options for current awareness include e-mail alerts from journals and RSS feeds from databases or websites.

2.2.1.8 Managing references

To ensure the retrieved records are managed efficiently the team should agree working practices. For example, who will screen the references and record decisions about which documents to obtain and how to code these decisions; whether decisions about rejecting or obtaining documents should be made blind to others' decisions; and how to store documents received. In addition, one member of the team should be responsible for identifying and removing duplicate references, ordering inter-library loans, recording the receipt of documents, and following up non-arrivals.

Using bibliographic software such as EndNote, Reference Manager or ProCite to record and manage references will help in documenting the process, streamline document management and make the production of reference lists for reports and journal papers easier. EPPI-Reviewer, a web-based review management programme, also incorporates reference management functions.4, 50 Alternatively it is possible to construct a database of references using a database package such as Microsoft Access or a word processing package. By creating a 'library' (database) of references, information can be shared by the whole review team, duplicated references can be identified and deleted more easily, and customised fields can be created where ordering decisions can be recorded.42 Specialised bibliographic management software packages have the facility to import references from electronic databases into the library and interact with word processing packages so bibliographies can be created in a variety of styles.

When an electronic library of references is used, it is important to establish in advance clear rules about which team members can add or amend records in the library, and that consistent terminology is used to record decisions. It is usually preferable to have one person from the team responsible for the library of references.

2.2.1.9 Obtaining documents

Obtaining a large number of papers in a short space of time can be very labour intensive. The procedure for acquiring documents will vary according to organisational arrangements and will depend on issues such as cost, what resources are available, and whether access to an inter-library loan network is available. Most libraries in the United Kingdom will be able to obtain articles from the British Library Document Supply Centre's collection although membership is required and there is a charge per article.

Many journals are available in full text on the Internet, although a subscription may be required before articles can be downloaded. It may be cost-effective to travel to a particular library to obtain material if a large number of references are required and are available. The information specialist on the team is likely to know about networks of associated libraries and electronic resources that can be used for obtaining documents.51

2.2.1.10 Documenting the search

The search process should be reported in sufficient detail so that it could be re-run at a later date. The easiest way to document the search is to record the process and the results contemporaneously. The decisions reached during development and any changes or amendments made should be recorded and explained. It is important to record all searches, including Internet searches, handsearching and contact with experts.

Providing the full detail of searches helps future researchers to re-run or update the searches and enables readers to evaluate the thoroughness of searching. The write up of the search should include information about the databases and interfaces searched (including the dates covered), full detailed search strategies (including any justifications for date or language restrictions) and the number of records retrieved.

When systematic reviews are reported in journal articles, limits on the word count may make it impossible to provide full details of the searches. In these circumstances as much information as possible should be provided within the available space. For example, 'We searched MEDLINE, EMBASE and CINAHL' is more helpful to the reader than 'We conducted computer searches'. Many journals now have an electronic version of the publication where the full search details can be provided. Alternatively, the published report can include the review team's contact details so full details of the search strategies can be requested. If a detailed report is being written for the commissioners of the review, the full search details should be included.

Summary: Identifying research evidence for systematic reviews

- The search for studies should be comprehensive.
- The extent of searching is determined by the research question and the resources available to the research team.
- Thorough searching is best achieved by using a variety of search methods (electronic and manual) and by searching multiple, possibly overlapping resources.
- Most of the searching is likely to take place at the beginning of the review with an update search towards the end.
- Using bibliographic software to record and manage references will help in documenting the process, streamline document management and make the production of reference lists for reports and journal papers easier.
- The search process should be documented in full or details provided of where the strategy can be obtained.

2.2.2 Study selection

Literature searching may result in a large number of potentially eligible records that need to be assessed for inclusion against predetermined criteria, only a small proportion of which may eventually be included in the review. The process for selecting studies should be explicit and conducted in such a way as to minimize the risk of errors and bias. This section explains the steps involved and the issues to be considered when planning and conducting study selection.

2.2.2.1 Process for study selection

The process by which decisions on the selection of studies will be made should be specified in the protocol, including who will carry out each stage and how it will be performed. The aim of selection is to ensure that all relevant studies are included in the review.

It is important that the selection process should minimize biases, which can occur when the decision to include or exclude certain studies may be affected by pre-formed opinions.52-56 The process for study selection therefore needs to be explicit, objective and minimize the potential for errors of judgement. It should be documented clearly to ensure it is reproducible (see Figure 1.1). The selection of studies from electronic databases is usually conducted in two stages:

Stage 1: a first decision is made based on titles and, where available, abstracts. These should be assessed against the predetermined inclusion criteria. If it can be determined that an article does not meet the inclusion criteria then it can be rejected straightaway. It is important to err on the side of over-inclusion during this first stage. The review question and the subsequent specification of the inclusion and exclusion criteria are likely to determine ease of rejection in this first stage. Where the question and criteria are tightly focused then it is usually easier to be confident that the rejected studies are not relevant. Rejected citations fall into two main categories; those that are clearly not relevant and those that address the topic of interest but fail on one or more criteria such as population. For those in the first category it is usually adequate to record as an irrelevant study, without a reason why. For those in the second category it is useful to record why the study failed to meet the inclusion criteria, as this increases the transparency of the selection process. Where abstracts are available the amount and usefulness of the information to the decision-making process often varies according to database and journal. Structured abstracts such as those produced by the BMJ are particularly useful at this stage of the review process.

Stage 2: for studies that appear to meet the inclusion criteria, or in cases when a definite decision cannot be made based on the title and/or abstract alone, the full paper should be obtained for detailed assessment against the inclusion criteria.

Some searching methods provide access to full papers directly, for example handsearching journals and contact with research groups, in which case assessment for inclusion is a one stage process.

Even when explicit inclusion criteria are specified, decisions concerning the inclusion of individual studies can remain subjective. Familiarity with the topic area and an understanding of the definitions being used are usually important.

The reliability of the decision process is increased if all papers are independently assessed by more than one researcher, and the decisions shown to be reproducible. One study found that on average a single researcher is likely to miss 8% of eligible studies, whereas a pair of researchers working independently would capture all eligible studies.57 Assessment of agreement is particularly important during the pilot phase (described later in this section), when evidence of poor agreement should lead to a revision of

the selection criteria or an improvement of their coding. Agreement between assessors (inter-assessor reliability) may be formally assessed mathematically using a Kappa statistic (a measure of chance-corrected agreement).58

The process for resolving disagreements between assessors should be specified in the protocol. Many disagreements may be simple oversights, whilst others may be matters of interpretation. These disagreements should be discussed and, where possible, resolved by consensus after referring to the protocol; if necessary a third person may be consulted.

If resources and time allow, the lists of included and excluded studies may be discussed with the advisory group. In addition, these lists can be posted on a dedicated website with a request for feedback on any missing studies, an approach used in a review of water fluoridation.59 For further information see Section 1.3.8 Disseminating the findings of systematic reviews.

Piloting the study selection process

The selection process should be piloted by applying the inclusion criteria to a sample of papers in order to check that they can be reliably interpreted and that they classify

the studies appropriately. The pilot phase can be used to refine and clarify the inclusion criteria and ensure that the criteria can be applied consistently by more than one person. Piloting may also give an indication of the likely time needed for the full selection process.

Masking/blinding

Judgements about inclusion may be affected by knowledge of the authorship, institutions, journal titles and year of publication, or the results and conclusions of articles.60 Blind assessment may be possible by removing such identifying information, but the gain should be offset against the time and effort required to disguise the source of each article. Several studies have found that masking author, institution, journal name and study results is of limited value in study selection.61, 62 Therefore, the general opinion is that unmasked assessment by two independent researchers is acceptable.

Dealing with lack of information

Sometimes the amount of information reported about a study is insufficient to make a decision about inclusion, and it can be helpful to contact study authors to ask for more details. However,

this requires time and resources, and the authors may not reply, particularly if the study is old. If authors are to be contacted it may be advisable to decide in advance how much time will be given to allow them to reply. If contacting authors is not practical then the studies in question could be excluded and listed as 'potentially relevant studies'. If a decision is made to include such studies, the influence on the results of the review can be checked in a sensitivity analysis.

Dealing with duplication

It is important to look for duplicate publications of research results to ensure they are not treated as separate studies in the review. Multiple papers may be published for a number of reasons including: translations; results at different follow-up periods or reporting of different outcomes. However, it is not always easy to identify duplicates as they are often covert (i.e. not cross referenced to one another) and neither authorship nor sample size are reliable criteria for identification of duplication.63 Estimates of prevalence of duplicate publication range from 1.4% to 28%,64 and studies have been found to have up to five duplicate reports.63 Multiple reports from the same study may include identical samples with different outcomes reported or increasing samples with the same outcomes reported.

Multiple reporting can lead to biased results, as studies with significant results are more likely to be published or presented more frequently, leading to an overestimation of treatment effects when findings are combined.65 When multiple reports of a study are identified these should be treated as a single study but reference made to all the publications. It may be worthwhile comparing multiple publications for any discrepancies, which could be highlighted and the study authors contacted for clarification.

Documenting decisions

It is important to have a record of decisions made for each article. This may be in paper form, attached to paper copies of the articles, or the selection process may be partially or wholly computerised. If the search results are provided in electronic format, they can be imported into a reference management program such as EndNote, Reference Manager or ProCite which stores, displays and enables organization of the records, and allows basic inclusion decisions to be made and recorded (in custom fields). For more complex selection procedures, where several decisions and comments need to be recorded, a database program such as Microsoft Access may be of use. There are also programs specifically designed for carrying out systematic reviews which include aids for the selection process, such as TrialStat SRS and EPPI-Reviewer.

Reporting study selection

A flow chart showing the number of studies/papers remaining at each stage is a simple and useful way of documenting the study selection process. Recommendations for reporting and presentation of a flow chart when reporting systematic reviews with or without a meta-analysis have been developed by the PRISMA group, formerly the QUOROM group. Publication of these guidelines is forthcoming.66, 67 In the meantime, the existing QUOROM guidelines for the reporting meta-analysis of RCTs,9 provide guidance that is equally applicable to all

systematic reviews. Figure 1.1 is an example of a flow chart from a systematic review of treatments for childhood retinoblastoma.14

A list of studies excluded from the review should also be reported where possible, giving the reasons for exclusion. This list may be included in the report of the review as an appendix. In general, this list is most informative if it is restricted to 'near misses' (i.e. those studies that only narrowly failed to meet inclusion criteria and that readers might have expected to see included) rather than all the research evidence identified. Decisions to exclude studies may be reached at the title and abstract stage or at the full paper stage.

Summary: Study selection

- In order to minimize bias, studies should be assessed for inclusion using selection criteria that flow directly from the review question and that have been piloted to check that they can be reliably applied.
- Study selection is a staged process involving sifting through the citations located by the search, retrieving full reports of potentially relevant citations and, from their assessment, identifying those studies that fulfil the inclusion criteria.
- Parallel independent assessments should be conducted to minimize the risk of errors. If disagreements occur between assessors, they should be resolved according to a predefined strategy using consensus and arbitration as appropriate.
- The study selection process should be documented, detailing reasons for exclusion of studies that are 'near-misses'.

2.2.3 Data extraction

Data extraction is the process by which researchers obtain the necessary information about study characteristics and findings from the included studies. Data extraction requirements will vary from review to review, and the extraction forms should be tailored to the review question. The first stage of any data extraction is to plan the type of analyses and list the tables that will be included in the report. This will help to identify which data should be extracted. General guidance on the process is given here, but the specific details will clearly depend on the individual review topic.

A sample data extraction form and details of the data extraction process should be included in the review protocol. A common problem at the protocol stage is that there may be limited familiarity with the topic area. This can lead to uncertainties, for example, about comparators and outcome measures. As a result, time can be wasted extracting unnecessary data and difficulties can arise when attempting to utilise and

synthesise the data. Sufficient time early in the project should therefore be allocated to developing, piloting and refining the data extraction form.

The extraction of data is linked to assessment of study quality in that both processes are often undertaken at the same time.

Standardised data extraction forms can provide consistency in a systematic review, whilst reducing bias and improving validity and reliability.68 Use of an electronic form has the added advantage of being able to combine data extraction and data entry into one step, and to facilitate data analysis and the production of results tables for the final report.

2.2.3.1 Design

Integral to the design of the form is the category of data to be extracted. It may be numerical, fixed text such as yes/no, a 'pick list', or free text. However, the number of free text fields should be limited as much as possible to simplify the analysis of data. The form should be unambiguous and easy to use in order to minimize discrepancies. Instructions for completion should be provided, and each field should have decision rules about coding data in order to avoid ambiguity and to aid consistent completion.

Piloting the form is essential. Paper forms should only be used where access to direct completion of electronic forms is impossible, to reduce risks of error in data transcription.

2.2.3.2 Content

The nature of the data extracted will depend on the type of question being addressed and the types of study available. Box 1.4 gives an example of some of the information that might be extracted for a comparative study.

The results to be extracted from each individual study may be reported in a variety of ways, and it is often necessary for a researcher to manipulate the available data

into a common format. Manipulations of the reported findings are discussed in further detail in Section 1.3.5 Data synthesis, but can include using confidence intervals to determine standard errors or estimating the hazard ratio from a survival curve. Data can be categorised at this stage; however, it is advisable to extract as much of the reported data as is likely to be needed, and categorise at a later stage, so that detailed information is not lost during data extraction.

2.2.3.3 Software

EPPI-Reviewer is a web application that enables researchers to manage all stages of a review in a single location. RevMan and TrialStat SRS are other software packages that can be used in data extraction for systematic reviews. Other tools commonly used include general word processing packages, spreadsheets and databases. Whichever software package is used, ideally it should have the ability to provide different types of question coding. Some software will also

allow researchers to develop quality control mechanisms for minimizing data entry errors, for example, by specifying ranges of valid values.

2.2.3.4 Piloting data extraction

Data extraction forms should be piloted on a sample of included studies to ensure that all the relevant information is captured and that resources are not wasted on extracting data not required. The consistency of the data extracted should be assessed to make sure that those extracting the data are interpreting the forms, and the draft instructions and decision rules about coding data, in the same way. This will help to reduce data extraction errors. The exporting, analysis and outputs of the data extraction forms should also be pilot tested where appropriate, on a small sample of included studies.

This will ensure that the exporting of data works correctly and the outputs provide the information required for data analysis and synthesis.

When using databases, piloting is particularly important as it becomes increasingly difficult to make changes once the template has been created and information has been entered into the database. Early production of the expected output is also the best way to check that the correct data structure has been set up.

2.2.3.5 Process of data extraction

Data extraction needs to be as unbiased and reliable as possible, however it is prone to human error and often subjective decisions are required. The number of researchers that will perform data extraction is likely to be influenced by constraints on time and resources. Ideally two researchers should independently perform the data extraction (the level of inter-rater agreement is often measured using a Kappa statistic58). As an accepted minimum, one researcher can extract the data with a second researcher independently checking the data extraction forms for accuracy and completeness. This method may result in significantly more errors than two researchers independently performing data extraction but may also take significantly less time.69 Any disagreements should be noted and resolved by consensus among researchers or by arbitration by an additional independent researcher. A record of corrections or amendments to data extraction forms should be kept for future reference, particularly where there is genuine ambiguity (internal inconsistency) which cannot be resolved after discussion with the study authors. If using an electronic data extraction form that does not keep a record of amendments, completed forms can be printed and amendments recorded manually, before correcting the electronic version.

As with screening studies for inclusion, blinding researchers to the journal and author details has been recommended.70, 71 However, this is a time-consuming operation, may not alter the results of a review and is likely to be of limited value.61

Example information requirements for data extraction

General information

Researcher performing data extraction Date of data extraction

Identification features of the study:

Record number (to uniquely identify study) Author

Article title Citation

Type of publication (e.g. journal article, conference abstract) Country of origin

Source of funding

Study characteristics

Aim/objectives of the study Study design

Study inclusion and exclusion criteria

Recruitment procedures used (e.g. details of randomisation, blinding) Unit of allocation (e.g. participant, GP practice, etc.)

Participant characteristics

Characteristics of participants at the beginning of the study e.g.

Age Gender Ethnicity

Socio-economic status Disease characteristics Co-morbidities

Number of participants in each characteristic category for intervention and control group(s) or mean/median characteristic values (record whether it is the number eligible, enrolled, or randomised that is reported in the study)

Intervention and setting

Setting in which the intervention is delivered

Description of the intervention(s) and control(s) (e.g. dose, route of administration, number of cycles, duration of cycle, care provider, how the intervention was developed, theoretical basis (where relevant))

Description of co-interventions

Outcome data/results

Unit of assessment/analysis Statistical technique used

For each pre-specified outcome: Whether reported

Definition used in study Measurement tool or method used Unit of measurement (if appropriate)

Length of follow-up, number and/or times of follow-up measurements

For all intervention group(s) and control group(s): Number of participants enrolled

Number of participants included in analysis

Number of withdrawals, exclusions, lost to follow-up

Summary outcome data e.g.

Dichotomous: number of events, number of participants

Continuous: mean and standard deviation

Type of analysis used in study (e.g. intention to treat, per protocol)

Results of study analysis e.g.

Dichotomous: odds ratio, risk ratio and confidence intervals, p-value Continuous: mean difference, confidence intervals

If subgroup analysis is planned the above information on outcome data or results will need to be extracted for each patient subgroup

Additional outcomes

Record details of any additional relevant outcomes reported Costs

Resource use Adverse events

NB: Notes fields can be useful for occasional pieces of additional information or important comments that do not easily fit into the format of other fields.

Reviews that include only published studies may be at risk of overestimating the treatment effect. Including data from unpublished studies (or unpublished outcomes) is therefore important in minimizing bias. However, this can be time-consuming, and the original data may no longer be available. Although those performing IPD meta- analyses,72 have generally been successful in obtaining data from the authors of unpublished studies, the same may not be true of other types of review. The practical difficulties of locating and obtaining information from unpublished studies may, for example, make the ideal of including relevant unpublished studies unachievable in the timescales available for many commissioned reviews. When information from unpublished studies is obtained, the published and unpublished material should be subjected to the same methodological evaluation.

Summary: Data extraction

- Standardized data extraction forms provide consistency in a systematic review, thereby potentially reducing bias, improving validity and reliability.
- Data extraction forms should be designed and developed with both the review question and subsequent analysis in mind. Sufficient time should be allocated early in the project for developing and piloting the data extraction forms.
- The data extraction forms should contain only information required for descriptive purposes or for analyses later in the systematic review.

  Information on study characteristics should be sufficiently detailed to allow readers to assess the applicability of the findings to their area of interest.
- Data extraction needs to be unbiased and reliable; however it is prone to human error and often subjective decisions are required. Clear instructions and decision rules about coding data should be used.
- As a minimum, one researcher should extract the data with a second researcher independently checking the data extraction forms for accuracy and detail. If disagreements occur between assessors, they should be resolved according to a predefined strategy using consensus and arbitration as appropriate.

2.2.4 Risk of bias assessment

2.2.4.1 Introduction

Research can vary considerably in methodological rigour. Flaws in the design or conduct of a study can result in bias, and in some cases this can have as much influence on observed effects as that of treatment. Important intervention effects, or lack of effect, can therefore be obscured by bias.

Recording the strengths and weaknesses of included studies provides an indication of whether the results have been unduly influenced by aspects of study design or conduct (essentially the extent to which the study results can be 'believed'). Assessment of study quality gives an indication of the strength of evidence provided by the review and can also inform the standards required for future research. Ultimately, quality assessment helps answer the question of whether the studies are robust enough to guide treatment, prevention, diagnostic or policy decisions.

Many useful books discuss the sources of bias in different study designs in detail, or provide an in-depth guide to critical appraisal.73-75 No single approach to assessing methodological quality is appropriate to all systematic reviews. The best approach will be determined by contextual, pragmatic and methodological considerations. However, the following sections describe the underlying principles of quality assessment and the key issues to consider.

2.2.4.2 Defining quality

Quality is a complex concept and the term is used in different ways. For example, a project using the Delphi consensus method with experts in the field of quality assessment of RCTs was unable to generate a definition of quality acceptable to all participants.76

Taking a broad view, the aim of assessing study quality is to establish how near the 'truth' its findings are likely to be and whether the findings are of relevance in the particular setting or patient group of interest. Quality assessment of any study is likely to consider the following:

- Appropriateness of study design to the research objective
- Risk of bias
- Other issues related to study quality
  - Choice of outcome measure
  - Statistical issues
  - Quality of reporting
  - Quality of the intervention
  - Generalisability

The importance of each of these aspects of quality will depend on the focus and nature of the review. For example, issues around statistical analysis are less important if the study data are to be re-analysed in a meta-analysis, and the quality of reporting is irrelevant where data (either individual patient or aggregate) and information are obtained directly from those responsible for the study.

Appropriateness of study design

As discussed previously, types of study used to assess the effects of interventions can be arranged into a hierarchy, based broadly on their susceptibility to bias (Box 1.3). Although the RCT is considered the best study design to evaluate the effect of an intervention, in cases where it is unworkable or unethical to randomise participants (e.g. when evaluating the effects of smoking on health), researchers may instead have to use a quasi-experimental or an observational design. Simply grading studies using this hierarchy does not provide an adequate assessment of study quality, because it does not take into account variations in quality among studies of the same design. Even RCTs can be implemented in such a way that findings are likely to be seriously biased and therefore of little value in decision-making.

It should be noted that the terminology used to describe study designs (e.g. cohort, prospective, retrospective, historical controls, etc.) can be ambiguous and used in different ways by different researchers. Therefore it is important to consider the individual aspects of the study design

that may introduce bias rather than focussing on the descriptive label used. This is particularly important for the description of non- randomised studies.

Risk of bias

Bias refers to systematic deviations from the true underlying effect brought about by poor study design or conduct in the collection, analysis, interpretation, publication or review of data. Bias can easily obscure intervention effects, and differences in the risk of bias between studies can help explain differences in findings.

Internal validity is the extent to which an observed effect can be truly attributed to the intervention being evaluated, rather than to flaws in the design or conduct of the study. Any such flaws can increase the risk of bias.

The types of bias, and the ways in which they can be minimized by each type of study design, are described below.

Randomized controlled trials

The RCT is generally considered to be the most appropriate study design for evaluating the effects of an intervention. This is because, when properly conducted, it limits

the risk of bias. The simplest form of RCT is known as the parallel group trial which randomises eligible participants to two or more groups, treats according to assignment, and compares the groups with respect to outcomes of interest.

Participants are allocated to groups using both randomisation (allocation involves the play of chance) and concealment (ensures that the intervention that will be allocated cannot be known in advance of assignment). When appropriately implemented, these aspects of design should ensure that the groups being compared are similar in all respects other than the intervention. The groups should be balanced for both known and unknown factors that might influence outcome, such that any observed differences should be attributable to the effect of the intervention rather than to intrinsic differences between the groups.

Allocation in this way avoids the influence of confounding, where an additional factor is associated both with receiving the intervention and with the outcome of interest. For example, babies who are breast fed are less likely to have gastrointestinal illnesses than those who are bottle fed. Though this might suggest evidence for the protective effect of breastfeeding, mothers who breast feed also tend to be of higher socio-economic status, which in itself is associated with a range of health benefits to the baby. Therefore,

when evaluating any possible protective effects of breastfeeding socio-economic status should be considered as a potential confounding factor. In some cases, the possible confounding factor(s) may not be known or measurable. In an RCT, so long as a sufficient number of participants are assigned then the groups should be balanced with respect to both known and unknown potential confounding factors.

Selection bias or allocation bias occurs where there are systematic differences between comparison groups in terms of prognosis or responsiveness to treatment. Concealed assignment prevents investigators being able to predict which intervention will be allocated next and using that information to select which participant receives which treatment. For example, clinicians may want to 'try out' the new intervention in patients with a poorer prognosis. If they succeed in doing this by knowing or correctly 'guessing' the order of allocation, the intervention group will eventually contain more seriously ill participants than the comparison group, such that the intervention will probably appear less effective than if the two groups had been properly balanced.

The most robust method for concealing the sequence of treatment allocation is a central telephone randomisation service, in which the care provider calls an independent trial service, registers the participant's details and then discovers which intervention they are to be given. Similarly, an on-site computer-based randomisation system that is not readable until the time of allocation might be used. Envelope methods of randomisation, where allocation details are stored in pre-prepared envelopes, are less robust and

more easily subverted than centralised methods. Where this method is adopted, sealed opaque sequentially numbered envelopes that are only opened in front of the participant being randomised should be used. Unfortunately, the methods which are used to ensure that the randomisation sequence remains concealed during implementation (frequently referred to as concealment of allocation) are often poorly reported making it difficult to discern whether the methods were susceptible to bias.

Some studies, which may describe themselves as randomised, may allocate participants to groups on an alternating basis, or based on whether their date of birth is an odd or even number. Allocation in these studies is neither random nor concealed.

Performance bias refers to systematic differences (apart from the intervention of interest) in the treatment or care given to comparison groups during the study and detection bias refers to systematic differences between groups in the way that outcomes are ascertained. The risk of these biases can be minimized by ensuring that people involved in the study are unaware of which groups participants have been assigned to (i.e. they are blinded or masked). Ideally, the participants, those administering the intervention, those assessing outcomes and those analysing the data should all be blinded. If not, the knowledge of which comparison group is which may consciously or unconsciously influence the behaviour of any of these people. The feasibility and/ or success of blinding will partly depend on the intervention in question. There are situations where blinding is not possible owing to the nature of the intervention, for example where a particular intervention has an obvious physiological effect whereas the comparator does not, and others where it may be unethical (e.g. sham surgery carries risks with no intended benefit). Methods of blinding for studies of drugs involve the use of pills and containers of identical size, shape and number (placebos). Sham devices can be used for many device interventions and for some procedural interventions sham procedures can be used (e.g. sham acupuncture). Blinding of outcome assessors is particularly important for more subjective outcome measures such as pain, but less important for objective measures such as

mortality. Implementation of a blinding process does not however guarantee successful blinding in practice. In study reports, terms such as double-blind, triple-blind or single-blind can be used inconsistently77 and explicit reporting of blinding is often missing.78 It is important to clarify the exact details of the blinding process.

A well-conducted RCT should have processes in place to achieve complete and good quality data,79 in order to avoid attrition bias. Attrition bias refers to systematic differences between the comparison groups in terms of participants withdrawing or being excluded from the study. Participants may withdraw or drop-out from a study because the treatment has intolerable adverse effects, or on the other hand, they may recover and leave for that reason. They may simply be lost to follow-up, or they may be withdrawn due to a lack of data on outcome measures. Other reasons that participants may be excluded include mistaken randomisation of participants who, on review, did not meet the study inclusion criteria, and participants receiving the wrong intervention due to protocol violation. The likely impact of such withdrawals and exclusions needs to be considered carefully; if the exclusion is related to the intervention and outcome then it can bias the results (for example, not accounting for high numbers of withdrawals due to adverse effects in one intervention arm will unduly favour that intervention). Serious bias can arise as a result of participants being withdrawn for apparently ad hoc reasons that are related to the success or failure of an intervention. There is evidence from the field of cancer research that exclusion of patients from the analysis may bias results,80 though how this may apply to other fields is unclear. An intention to treat (ITT) analysis is generally recommended in order to reduce the risk of bias.

An ITT analysis includes outcome data on all trial participants and analyses them according to the intervention to which they were randomised, regardless of the intervention(s) they actually received. Complete outcome data are often unavailable for participants who drop-out before the end of the trial, so in order to include all participants, assumptions need to be made about their missing outcome data (for example by imputation of missing values). ITT analysis generally provides a more

conservative, and potentially less biased, estimate in trials of effectiveness (see Section 1.3.5.2 Quantitative synthesis of comparative studies). However, ITT analyses are often poorly described and applied81 and if assessing methodological quality associated with statistical analysis, care needs to be taken in judging whether the use of ITT analysis has minimized the risk of attrition bias and whether it was appropriately applied. If

an ITT analysis is not used, then the study should at least report the proportion of participants excluded from the analysis to allow a researcher to judge whether this is likely to have led to bias.

The minimum criteria for assessment of risk of bias in RCTs are set out in Box 1.5. While all these criteria are relevant to assessing risk of bias, their relative importance can

be context specific. For example, the importance of blinded outcome assessment will vary depending on whether the outcomes involve subjective judgement (this may vary between different outcomes measured within the same trial). Therefore, when planning which criteria

to use it is important to think carefully about what characteristics would realistically be considered ideal. The Cochrane handbook provides a detailed assessment tool for use when assessing risk of bias in an RCT.82

Criteria for assessment of risk of bias in RCTs

- Was the method used to generate random allocations adequate?
- Was the allocation adequately concealed?
- Were the groups similar at the outset of the study in terms of prognostic factors, e.g. severity of disease?
- Were the care providers, participants and outcome assessors blind to treatment allocation? If any of these people were not blinded, what might be the likely impact on the risk of bias (for each outcome)?
- Were there any unexpected imbalances in drop-outs between groups? If so, were they explained or adjusted for?
- Is there any evidence to suggest that the authors measured more outcomes than they reported?
- Did the analysis include an intention to treat analysis? If so, was this appropriate and were appropriate methods used to account for missing data?

Other randomised study designs

In addition to parallel group RCTs, there are other randomised designs where further quality criteria may need to be considered. These are described below.

Randomised cross-over trials

In randomised cross-over trials all participants receive all the interventions. For example in a two arm cross-over trial, one group receives intervention A before intervention B, and the other group receives intervention B before intervention A. It is the sequence

of interventions that is randomised. The advantage of cross-over trials is that they are potentially more efficient than parallel trials of a similar size, in which each participant receives only one of the interventions. The criteria for assessing risk of bias in RCTs also apply to cross-over trials, but there are some additional factors that need to be taken into consideration.

The cross-over design is inappropriate for conditions where the intervention may provide a cure or remission, where there is a risk of spontaneous improvement or resolution of the condition, where there is a risk of deterioration over the period of the

trial (e.g. degenerative conditions) or where there is a risk that patients may die.83 This is because these outcomes lead either to the participant being unable to enter the second period or, on entering the second period, their condition is systematically different from that in the first period.

The possibility of a 'carryover' of the effect of the intervention provided in the first period into the second intervention period is an important concern in this study design.83 This risk is dealt with by building in a treatment-free or placebo 'washout period' between the intervention periods.83 The adequacy of the washout period length will need to be considered as part of the assessment of risk of bias.

The statistical analysis appropriate to cross-over trials are discussed in the synthesis section and statistical advice is likely to be required (see Section 1.3.5 Data synthesis).

Cluster randomised trials

A cluster randomised trial is a trial where clusters of people rather than single individuals are randomised to different interventions.84 For example, whole clinics or geographical locations may be randomised to receive particular interventions, rather than individuals.

The distinctive feature of cluster trials is that the outcome for each participant within a cluster may not be independent, since individuals within the cluster are likely to respond in a similar way to the intervention. Underlying reasons for this intra-cluster correlation include individuals in a cluster being affected in a similar manner due to shared exposure to a common environment such as specific hospital policies on discharge times; or personal interactions between cluster members and sharing of attitudes, behaviours and norms that may lead to similar responses.84 This has implications for estimating the sample size required (i.e. the sample needs to be larger than for an individually randomised trial) and the statistical analysis.

When assessing the risk of selection bias in cluster randomised trials there are two factors that need to be considered: the randomisation of the clusters and how

participants within clusters are selected into the study.85 The first can be dealt with by using an appropriate randomisation method with concealed allocation (clusters are often allocated at the outset). However, where the trial design then requires selection

of participants from within a cluster, the risk of selection bias should also be assessed.

There is a clear risk of selection bias when the person recruiting participants knows in advance the clinical characteristics of a participant and which intervention they will receive. Also, potential participants may know in advance which intervention their

cluster will receive, leading to different participation rates in the comparison groups.85 Two key methods for reducing bias in the selection of individuals within clusters have been identified: recruitment of individuals prior to the random allocation of clusters and, where this is not possible, use of an impartial individual to recruit participants following randomisation of the clusters.86

The statistical analyses appropriate to cluster randomised trials are discussed in Section

Data synthesis and statistical advice is likely to be required.

Wider reading is recommended prior to conducting a quality assessment of cluster randomised trials. Several texts discuss the design, analysis and reporting of this trial design.75, 84, 87, 88

Quasi-experimental studies

The main distinction between randomised and quasi-experimental studies is the way in which participants are allocated to the intervention and control groups; quasi- experimental studies do not use random assignment to create the comparison groups.

In non-randomised controlled studies, individuals are allocated to concurrent comparison groups, using methods other than randomisation. The lack of concealed randomised allocation increases the risk of selection bias.

Before-and-after studies evaluate participants before and after the introduction of an intervention. The comparison is usually made in the same group of participants, thus avoiding selection bias, although a different group can be used. In this type of design however, it can be difficult to account for confounding factors, secular trends, regression to the mean, and differences in the care of the participants apart from the intervention of interest.

An alternative to this is a 'time series' design. Interrupted time series studies are multiple observations over time that are 'interrupted', usually by an intervention or treatment and thus permit separating real intervention effects from other long-term trends. It is a study design used where others, such as RCTs, are not feasible, for example in the evaluation of a screening service or a mass media campaign. It is also frequently used in policy evaluation, for example to measure the effect of a smoking ban.

The circumstances in which, and extent to which, studies without randomisation are at risk of bias are not fully understood.89 A key influencing factor may be the extent to which prognosis influences selection for a particular intervention as well as eventual outcome.89 Because of the risk of bias, careful consideration should be given to the inclusion of quasi-experimental studies in a review to assess the effectiveness of an intervention. If included, researchers should think carefully about the strength of this evidence and how it should be interpreted.

A review of quality assessment tools designed for or used to assess studies without randomisation identified key aspects of quality as being particularly pertinent:89

- How the treatment groups were created (how allocation occurred; and whether the study was designed to generate groups that are comparable on key prognostic factors e.g. by 'matching' participants in each group).
- The comparability of intervention and comparison groups at the analysis stage. For example, whether prognostic factors were identified; and whether case-mix adjustment was used to account for any between group differences.

Other quality issues identified were similar to those for assessing performance, detection and attrition bias in RCTs: blinding of participants and investigators; the level of confidence that the participants received the intervention to which they were assigned and experienced the reported outcome as a result of that intervention; the adequacy of the follow-up; and appropriateness of the analysis.

Observational studies

In observational studies the intervention(s) that individuals receive are determined by usual practice or 'real-world' choices, as opposed to being actively allocated as part of the study protocol.

Observational studies are usually more susceptible to bias than experimental studies, and the conclusions that can be drawn from them are necessarily more tentative and are often hypothesis generating, highlighting areas for further research.

Observational designs such as cohort studies, case-control studies and case series are often considered to form a hierarchy of increasing risk of bias. However, such a hierarchy is not always helpful because, as noted before, the same label can be used

to describe studies with different design features and there is not always agreement on the definitions of such studies. Attention should focus on specific features of the studies (e.g. participant allocation, outcome assessment) and the extent to which they are susceptible to bias.

In a cohort study design, a defined group of participants is followed over time and comparison is made between those who did and did not receive an intervention (e.g. a study may follow a cohort of women who choose to use oral contraceptives and compare them over time with women who choose other forms of contraception).

Prospective cohort studies are planned in advance and define their participants before the intervention of interest and follow them into the future. These are less likely to be susceptible to bias than retrospective cohort studies, which identify participants from past records and follow them from the time of that record.

Case-control studies compare groups from the same population with (cases) and without (controls) a specific outcome of interest, to evaluate the association between exposure to an intervention and the outcome. The risk of selection bias in such studies will be dependent on how the control group was selected. Groups may be matched

to make them comparable for potential confounding factors. However, since analysis cannot be performed on matched variables, the matching criteria must be selected carefully, as this can give rise to 'over-matching' when the factors are related to allocation to the intervention.

Case series are observations made on a number of individuals (with no control group) and are not comparative. They can, however, provide useful information, for example about the unintentional effects of an intervention (see Chapter 4) and in such situations it is important to assess their quality.

Other issues related to study quality

Choice of outcome measure

As well as using blinding to minimize bias when assessing outcomes, it is usually necessary to consider the reliability or validity of the actual outcome measure being used (e.g. several different scales can be used to measure quality of life or psychological outcomes). It is important that the scales are fully understood to enable comparison, (e.g. a high score implies a favourable outcome in some scales and an unfavourable one in others).

The outcome should also be relevant and meaningful to both the intervention and the evaluation (i.e. a treatment intended to reduce mortality should measure mortality, not merely a range of biochemical indicators).

Statistical issues

Although issues around statistical analysis are less important if the study data are to be combined in a meta-analysis, when studies are not being quantitatively pooled it is also important to assess statistical issues around design and analysis. For

example, assessing whether a study is adequately powered to detect an effect of the intervention.90 The assessment of statistical power may involve relying on the sample size calculation in the primary study, where reported. However, defining population parameters for sample size calculations is a subjective judgement which may vary between investigators;91 for some review topics it may be appropriate to define a priori an adequate sample size for the purposes of the review.

Quality of reporting

Inadequate reporting of important aspects of methodological quality such as allocation concealment, blinding and statistical analysis is common,92 as is failure to report detail about the intervention and its implementation. Quality of reporting does not necessarily reflect the quality of the underlying methods or data, but when planning quality assessment it is important to decide how to deal with poor reporting. One approach is to assume that if an item is not reported then the criterion has not been met. While this may often be justifiable,93, 94 there is evidence to suggest that failure to report a method does not necessarily mean it has not been used.95-97 Therefore it is important to be accurate and distinguish between failure to report a criterion and failure to meet a criterion. For example, a criterion can be described as being met, not met, or unclear due to inadequate reporting.

There have been a number of initiatives aimed at improving the quality of reporting of primary research. The CONSORT statement contains a set of recommendations for the reporting of RCTs,98 the TREND statement provides guidelines for the reporting of non-randomised evaluations of behavioural and public health interventions,99 and the

STROBE statement is an initiative to improve reporting of observational studies.100 The EQUATOR network promotes the transparent and accurate reporting of health research in a number of ways, including the use of these consensus reporting guidelines.101 It is anticipated

that implementation of these guidelines will help improve the standard of reporting, which should make quality assessment more straightforward.

Quality of the intervention

In addition to study design, it is often helpful to assess the quality of the intervention and its implementation. At its most simplistic, the quality of an intervention refers to whether it has been used appropriately. This is a fairly straightforward assessment where, for example drug titration studies have been conducted. It is more problematic where there is no preliminary research suggesting that an intervention should be administered in a particular way,102 or where the intervention requires a technical skill such as surgery or

physiotherapy.103 It is important to establish to what extent these are standardised, as this will affect how the results should be interpreted.

The quality of the intervention is particularly relevant to complex interventions made up from a number of components, which act independently and inter-dependently.104, 203, 204 These include clinical interventions such as physiotherapy as well as public health

interventions such as community-based programmes. The quality of an intervention can be conceptualised as having two main aspects: (i) whether the intervention has been appropriately defined and (ii) whether it has been delivered as planned (the integrity or fidelity of the intervention).

If the quality of the intervention is relevant, the review should assess whether the intervention was implemented as planned in the individual studies (i.e. how many participants received the intervention as planned, whether consistency of implementation was measured, and whether it is likely that participants received an unintended intervention/contamination of the intervention that may influence the results). In some topic areas, for example when a sham device or procedure is being used, it may also be relevant to assess the quality of the comparator. When an intervention relies on the skill of the care provider it may be useful to assess whether the performance of those providing the intervention was measured. For more detailed information on complex interventions see Chapter 3.

Generalisability

Generalisability, also known as applicability or external validity, is not considered in detail in this section. In addition to assessing the risk of bias (internal validity),

researchers may also consider how closely a study reflects routine practice or the usual setting where the intervention would be implemented. However, this is not an inherent characteristic of a study as the extent to which a study is 'generalisable' depends

also on the situation to which the findings are being applied.105 Therefore the issue of generalisability is also raised in Section 1.2 The review protocol in the context of defining inclusion criteria for the review, Section 1.3.3 Data extraction and in Section 1.3.6 Report writing.

2.2.4.3 The impact of study quality on the estimate of effect

Several empirical studies have explored how quality can influence the results of clinical trials (and therefore the results of reviews of trials). Trials with double-blinding and adequate concealment of allocation have been found to indicate less beneficial treatment effects than trials without these features.106 Similarly, exclusion of lower quality studies has led to less beneficial effects in meta-analyses.106 In meta-analyses of subjectively assessed outcomes (e.g. patient reported outcomes), inadequate allocation concealment and lack of blinding have been associated with substantially more beneficial treatment effects, whereas for objective outcomes (e.g. mortality) there was a modest effect of inadequate allocation concealment and no effect of lack of blinding.107 There is some evidence about the relationship between study quality and the estimate of effect that is contradictory to the above,108, 109 though this may be due to the data sets used and how specific quality criteria were defined.

2.2.4.4 The process of risk of bias assessment

There are two main approaches towards assessing quality. One involves the use of checklists of quality items and the other of scales which provide an overall numerical quality score for each study.

Tools for assessing quality

Checklists can be a reliable means of ensuring that all the studies assessed are critically appraised in a standardised way. There are many different checklists and scales readily available, 75, 111-116 which can be modified to meet the requirements of the review, or a new detailed checklist, specific to the review, may be developed.

Because some items included may require a degree of subjective judgement, it is important to pilot the use of the checklist and to ensure that the quality assessment is undertaken independently by two researchers.

The use of scales with summary scores to distinguish high and low quality studies is questionable and not recommended.117, 118 Very few scales have been developed using standard techniques to establish their validity and reliability.113 The weighting assigned to methodological items varies considerably between scales,117 and does not usually take into account the direction of bias.119 An investigation comparing low-molecular-weight heparin (LMWH) with standard heparin for thromboprophylaxis in general surgery found that trials identified as 'high quality' by some of the 25 scales investigated indicated that LMWH was not superior to standard heparin, whereas trials identified as 'high quality' by other scales led to the opposite conclusion, that LMWH was beneficial.117 It is therefore preferable that aspects of quality such as blinding and treatment allocation (and their potential impact on study results) should be considered individually.117

Checklists by type of study design

In general checklists tend to be specific to particular study designs, and where reviews include more than one type of study design, separate lists can be used or a combined list selected or developed. Checklists have also been developed for use with both randomised and non-randomised studies such as that by Downs and Black.

There are multiple systems available for the evaluation of RCTs, in addition to the Cochrane handbook assessment tool for assessing risk of bias. In a review of checklists for the assessment of non-randomised studies, nearly 200 tools were identified. From these, six were recommended as being suitable for use in systematic reviews including non-randomised studies.89 The Cochrane Effective Practice and Organisation of Care Group (EPOC) have developed guidelines to assist researchers in making decisions about when to include studies that use interrupted time series designs and how to assess their methodological quality.115, 116 A useful checklist for observational studies was published as part of the US Agency for Healthcare Research and Quality's (AHRQ) 'Systems to Rate the Strength of Scientific Evidence'.112 The most recent version of the Cochrane Handbook also contains guidance on dealing with non-randomised studies in systematic reviews of interventions, from the protocol to synthesis stages.75

How will the quality assessment information be used?

Simply reporting which quality criteria were met by studies included in a systematic review is not sufficient. The implications of the quality assessment for interpreting results need to be explicitly considered.

Study quality can be incorporated into the synthesis either quantitatively through subgroup or sensitivity analyses (see Section 1.3.5.2: Quantitative synthesis), or in a narrative synthesis. In the latter, the quality assessment can be used to help interpret and explain differences in results across studies (e.g. unblinded studies with subjective outcomes may have consistently larger effects than blinded studies) and inform a qualitative interpretation of the risk of bias (see Section 1.3.5.1 Narrative synthesis).

Summary: Quality assessment

- An important part of the systematic review process is to assess the risk of bias in included studies caused by inadequacies in study design, conduct or analysis that may have led to the treatment effect being over or underestimated.
- Various tools are available but there is no single tool that is suitable for use in all reviews. Choice should be guided by:
  - Study design
  - The level of detail required in the assessment
  - The ability to distinguish between internal validity (risk of bias) and external validity (generalisability)

- Using quality scores is problematic; it is preferable to consider individual aspects of methodological quality in the quality assessment and synthesis.
- Where appropriate, the potential impact that methodological quality had on the findings of the included studies should be considered.
- Detailed quality assessment can be time consuming if a review includes a large number of studies and may require considerable expertise in critical appraisal. If resources are limited, priority should be given to assessment of the key sources of bias.

2.2.5 Synthesis

Synthesis involves the collation, combination and summary of the findings of individual studies included in the systematic review. Synthesis can be done quantitatively using formal statistical techniques such as meta-analysis, or if formal pooling of results is inappropriate, through a narrative approach. As well as drawing results together, synthesis should consider the strength of evidence, explore whether any observed effects are consistent across studies, and investigate possible reasons for any inconsistencies. This enables reliable conclusions to be drawn from the assembled body of evidence.

Deciding what type of synthesis is appropriate

Many systematic reviews evaluating the effects of health interventions focus on evidence from RCTs, the results of which, generally, can be combined quantitatively. However, not all health care questions can be addressed by RCTs, and systematic reviews do not automatically involve statistical pooling. Meta-analysis is not always possible or sensible. For example, pooling results obtained from diverse non-randomised study types is not recommended.120 Similarly, meta-analysis of poor quality studies could be seriously misleading as errors or biases in individual studies would be compounded and the very act of synthesis may give credence to poor quality studies.

However, when used appropriately, meta-analysis has the advantage of being explicit in the way that data from individual studies are combined, and is a powerful tool for combining study findings, helping avoid misinterpretation and allowing meaningful conclusions to be drawn across studies.

The planned approach should be decided at the outset of the review, depending on the type of question posed and the type of studies that are likely to be available. There may be topics where it can be decided a priori that a narrative approach is appropriate. For example, in a systematic review of interventions for people bereaved by suicide, it was anticipated there would be such diversity in the included studies, in terms of settings, interventions and outcome measures, that a narrative synthesis alone was proposed in the protocol.121

Narrative and quantitative approaches are not mutually exclusive. Components of narrative synthesis can be usefully incorporated into a review that is primarily

quantitative in focus and those that take a primarily narrative approach can incorporate some statistical analyses such as calculating a common outcome statistic for each study.

Initial descriptive synthesis

Both quantitative and narrative synthesis should begin by constructing a clear descriptive summary of the included studies. This is usually done by tabulating details about study type, interventions, numbers of participants, a summary of participant characteristics, outcomes and outcome measures. An indication of study quality or risk of bias may also be given in this or a separate table (see Section 1.3.2 Study selection and Section 1.3.4 Quality assessment). An example is given in Table 1.1. If the review will not involve re-calculating summary statistics, but will rather rely on the reported results of the author's analyses, these may also be included in the table. The descriptive process should be both explicit and rigorous and decisions about how to group and tabulate data should be based on the review question and what has been planned in the protocol. This initial phase will also be helpful in confirming that studies are similar and reliable enough to synthesise, and that it is appropriate to pool results.

2.2.5.1 Narrative synthesis

All systematic reviews should contain text and tables to provide an initial descriptive summary and explanation of the characteristics and findings of the included studies. However simply describing the studies is not sufficient for a synthesis. The defining characteristic of narrative synthesis is the adoption of a textual approach that provides an analysis of the relationships within and between studies and an overall assessment of the robustness of the evidence.

A narrative synthesis of studies may be undertaken where studies are too diverse (either clinically or methodologically) to combine in a meta-analysis, but even where a meta-analysis is possible, aspects of narrative synthesis will usually be required in order to fully interpret the collected evidence.

Narrative synthesis is inherently a more subjective process than meta-analysis; therefore, the approach used should be rigorous and transparent to reduce the potential for bias. The idea of narrative synthesis within a systematic review should not be confused with broader terms like 'narrative review', which are sometimes used to describe reviews that are not systematic.

A general framework for narrative synthesis

How narrative syntheses are carried out varies widely, and historically there has been a lack of consensus as to the constituent elements of the approach or the conditions for establishing credibility. A project for the Economic and Social Research Council (ESRC) Methods Programme has developed guidance on the conduct of narrative synthesis in systematic reviews.123-126 The guidance offers both a general framework and specific tools and techniques that help to increase the transparency and trustworthiness of narrative synthesis.

The general framework consists of four elements:

- Developing a theory of how the intervention works, why and for whom
- Developing a preliminary synthesis of findings of included studies
- Exploring relationships within and between studies
- Assessing the robustness of the synthesis

Though the framework is divided into these four elements, the elements themselves do not have to be undertaken in a strictly sequential manner, nor are they totally independent of one another. A researcher is likely to move iteratively among the activities that make up these four elements.

For each element of the framework, this guidance presents a range of practical tools and techniques. It is not mandatory (or indeed appropriate) to employ each one of these

for every narrative synthesis, but the appropriate tools/techniques should be selected depending upon the nature of the evidence being synthesised. The reason for the choice of tool or technique should be specified in the methods section of the review.

A fuller description of these tools and techniques and narrative synthesis in general can be found in the ESRC guidance report.125, 126 It should be noted that the list given here is not comprehensive and other tools and techniques may be appropriate in certain circumstances.

Developing a theory of how the intervention works, why and for whom The extent to which theory will play a role will partly depend upon the type of intervention(s) being evaluated. For example, theory may only play a minor role in a systematic review looking at the effects of a single therapeutic drug on patient

outcomes because many aspects of the 'mechanism of action' will have been established in early studies investigating pharmacodynamics, dose-finding etc. Alternatively, in a systematic review evaluating the effects of a psychosocial or educational programme, theories about the causal chain linking the intervention to the outcomes of interest will be of crucial importance and might be presented descriptively or in diagrammatic form.

Developing a preliminary synthesis of findings of included studies

Once the relevant studies have been data extracted, the first step is to bring together, organise and describe their findings. The direction and size of the reported effects may be the starting point. Or, for example, a collection of studies evaluating one kind of intervention might be divided into subgroups of studies with distinct populations, such as children and adults. It is important to remember that this is only the first step of the synthesis. The remaining elements of the framework need to be taken into account before it can be considered adequate as a narrative synthesis.

Table describes a range of tools and techniques that might be employed at this stage of the synthesis.

Developing a preliminary synthesis of findings of included studies

Textual descriptions of studies	A descriptive paragraph on each included study. These descriptions should be produced in a systematic way, including the same type of information for all studies if possible and in the same order. It may be useful for recording purposes to do this for all excluded studies as well.
Groupings and clusters	The included studies might be grouped at an early stage of the review, though it may be necessary to refine these initial groups as the synthesis develops. This can also be a useful way of aiding the process of description and analysis and looking for patterns within and across groups. It is important to use the review question(s) to inform decisions about how to group the included studies.
Tabulation	A common approach, used to represent data visually. The way in which data are tabulated may affect readers' impressions of the relationships between studies, emphasising the importance of a narrative interpretation to supplement the tabulated data.
Transforming data into a common measure	In both narrative and quantitative synthesis it is important to ensure that data are presented in a common measure to allow an accurate description of the range of effects.
Vote-counting as a descriptive tool	Simple vote-counting might involve the tabulation of findings according to direction of effect. More complex approaches can be developed both in terms of the categories used and by assigning different weights or scores to different categories. However, vote-counting can disregard sample size and be misleading. So, the interpretation of the results must be approached with caution and subjected to further scrutiny.
Translating data: thematic analysis	A technique used in the analysis of qualitative data in primary research can be used to systematically identify the main, recurrent and/or most important (based on the review question) themes and/or concepts across multiple studies.127
Translating data: content analysis	A technique for compressing many words of text into fewer content categories based on explicit rules of coding.128 Unlike thematic analysis, it is essentially a quantitative method, since all the data are eventually converted into frequencies.

Exploring relationships within and between studies

Patterns emerging from the data during the preliminary synthesis need to be rigorously scrutinised in order to identify factors that might explain variations in the size/direction of effects.

At this stage there is a clear attempt to explore relationships between: (a) characteristics of individual studies and their reported findings; and (b) the findings of different studies.

However, when exploring heterogeneity in this way, it is necessary to be wary of uncovering associations between characteristics and results that are based on comparisons of many subgroups – some of these may simply have occurred by chance.

Subgroup comparisons which are specified in advance (i.e. as part of the review protocol) are more likely to be plausible than those which are not.129, 130

The extent to which these factors can be explored in the review depends on how clearly they are reported in the primary research studies. The amount of detail may depend on the type of publication and the nature of the intervention being reviewed (e.g. highly standardised interventions may not be described as fully as more unusual ones).

Tools and techniques that might be employed at this stage of the synthesis are described in Table 1.3.

Exploring relationships within and between studies

Graphs, frequency distributions, funnel plots, forest plots and L'Abbe plots	There are several visual or graphical tools that can help reviewers explore relationships within and between studies. These include: presenting results in graphical form; plotting findings (e.g. effect size) against study quality; plotting confidence intervals; and/or plotting outcome measures.
Moderator variables and subgroup analyses	This refers to the analysis of variables which can be expected to moderate the main effects being examined in the review. This can be done at the study level, by examining characteristics that vary between studies (such as study quality, study design or study setting) or by analysing characteristics of the sample (such as subgroups of participants).
Idea webbing and conceptual mapping	Involves using visual methods to help to construct groupings and relationships. The basic idea underpinning these approaches is (i) to group findings that are empirically and/or conceptually similar and (ii) to identify (again on the basis of empirical evidence and/or conceptual/theoretical arguments) relationships between these groupings.

Graphs, frequency distributions, funnel plots, forest plots and L'Abbe plots	There are several visual or graphical tools that can help reviewers explore relationships within and between studies. These include: presenting results in graphical form; plotting findings (e.g. effect size) against study quality; plotting confidence intervals; and/or plotting outcome measures.
Qualitative case descriptions	Any process in which descriptive data from studies included in the systematic review are used to try to explain differences in statistical findings. For example why one intervention outperforms another apparently similar intervention or why some studies are statistical outliers.
Investigator/methodological/conceptual triangulation	Triangulation makes use of a combination of different perspectives and/or assessment methods to study a particular phenomenon. This could apply to the methodological and theoretical approaches adopted by the researchers undertaking primary studies included in a systematic review, e.g. investigator triangulation explores the extent to which heterogeneity in study results may be attributable to the diverse approaches taken by different researchers. Triangulation involves analysing the data in relation to the context in which they were produced, notably the disciplinary perspectives and expertise of the researchers producing the data.

Assessing the robustness of the synthesis

Towards the end of the synthesis process, the analysis of relationships as described above should lead into an overall assessment of the strength of the evidence. This is essential when drawing conclusions based on the narrative synthesis.

Robustness can relate to the methodological quality of the included studies (such as risk of bias), and/or the credibility of the product of the synthesis process. Obviously, these are related. The credibility of a synthesis will depend on both the quality and the quantity of the evidence base it is built on, and the method of synthesis and the

clarity/transparency of its description. If primary studies of poor methodological quality are included in the review in an uncritical manner then this will affect the integrity of the synthesis. Attempts to minimize the introduction of bias might include 'weighting' the findings of studies according to technical quality (i.e. giving greater credence to the findings of more methodologically sound studies) and providing a clear justification for this. Similarly, a clear description of the potential sources of bias within the synthesis process itself helps establish

credibility with the reader. Table 1.4 describes the tools and techniques that might be employed at this stage of the synthesis.

Assessing the robustness of the synthesis

Use of validity assessment	Use of specific rules to define weak, moderate or good evidence. An example is the approach used by the US Centers for Disease Control and Prevention131 although there are many other evidence grading systems available. Decisions about the strength of evidence are explicit although the criteria used are often debated.
Reflecting critically on the synthesis process	Use of a critical discussion to address methodology of the synthesis used132 (especially focusing on its limitations and their potential influence on the results); evidence used (quality, validity, generalisability) – with emphasis on the possible sources of bias and their potential influence on results of the synthesis; assumptions made; discrepancies and uncertainties identified; expected changes in technology or evidence (e.g. identified ongoing studies); aspects that may have an influence on implementation and effectiveness in real settings. Such a discussion would provide information on both the robustness and generalisability of the synthesis.
Checking the synthesis with authors of primary studies	It is possible to consult with the authors of included primary studies in order to test the validity of the interpretations developed during the synthesis and the extent to which they are supported by the primary data.133 The authors of the primary studies may have useful insights into the possible accuracy and generalisability of the synthesis; this is most likely to be useful when the number of primary studies is small. This is a technique that has been used with qualitative evidence.

2.2.5.2 Quantitative synthesis of comparative studies

As with narrative synthesis, quantitative synthesis should be embedded in a review framework that is based on a clear hypothesis, should consider the direction and size of any observed intervention effects in relation to the strength of evidence, and should explore relationships within and between studies. The requirements for a careful and thoughtful approach, the need to assess the robustness of syntheses, and to reflect critically on the synthesis process, apply equally but are not repeated here.

This section aims to outline the rationale for quantitative synthesis of comparative studies and to focus on describing commonly used methods of combining study results and exploring heterogeneity. A more detailed overview of quantitative synthesis for systematic review is

given in the Cochrane Handbook.75 Comprehensive accounts are also given by Whitehead134 and Cooper and Hedges,135 and a discussion of recent developments and more experimental approaches is given in a paper by Sutton and Higgins.136

Decisions about which comparisons to make, and which outcomes and summary effect measures to use, should have been addressed as part of the protocol development.

However, as synthesis depends partly on what results are actually reported, some planned analyses may not be possible, and others may have to be adapted or developed. Any departures from the analyses planned in the protocol should be clearly justified and reported.

Decisions about what studies should and should not be combined are inevitably subjective and require careful discussion and judgement. As far as possible a priori consideration at the time of writing the protocol is desirable. There will always be differences between studies that address a common question. Reserving meta- analyses for only those studies that evaluate exactly the same interventions in near identical participant populations would be severely limiting and seldom achievable in practice. For example, whilst it may not be sensible to average the results of studies using different classes of experimental drugs or comparators, it may be reasonable to combine results of studies that use analogues or drugs with similar mechanisms of

action. Likewise, it will often be reasonable to combine results of studies that have used similar but not identical comparators (e.g. placebo and no treatment). Where there are substantial differences between studies addressing a broadly similar question, although combining their results to give an estimate of an average effect may be meaningless,

a test of whether an overall effect is present might be informative. It can be useful to calculate summary statistics for each individual study to show the variability in results across studies. It may also be helpful to use meta-analysis methods to quantify this heterogeneity, even when combined estimates of effect are not produced.

Reasons for meta-analysis

Combining the results of individual studies in a meta-analysis increases power and precision in estimating intervention effects. In most areas of health care, 'breakthroughs' are rare and we may reasonably expect that new interventions will lead to only modest improvements in outcome; such improvements can of course be extremely important

to individuals and of significant benefit in terms of population health. Large numbers of events are required to detect modest effects, which are easily obscured by the play

of chance, and studies are often too small to do so reliably. Thus, in any group of small trials addressing similar questions, although a few may have demonstrated statistically significant results by chance alone, most are likely to be inconclusive. However, combining the results of studies in a meta-analysis provides increased numbers of participants, reduces random error, narrows confidence intervals, and provides a greater chance of detecting a real effect as

statistically significant (i.e. increases statistical power). Meta-analysis also allows observation and statistical exploration of the pattern of results across studies and quantification and exploration of any differences.

Combining comparative study results in a meta-analysis

Most meta-analyses take a two-step approach in that they first analyse the outcome of interest and calculate summary statistics for each individual study. In the second stage, these individual study statistics are combined to give an overall summary estimate.

This is usually calculated as a weighted average of the individual study estimates. The greater the weight awarded to a study, the more it influences the overall estimate.

Studies are usually, at least in part, weighted in inverse proportion to their variance (or standard error squared), a method which essentially gives more weight to larger studies and less weight to smaller studies. It is also possible to weight studies according to other factors such as trial quality, but such methods are very seldom implemented and not recommended.

Two main statistical models are used. Fixed-effect models weight the contribution of each study proportional to the amount of information observed in the study. This

considers only variability in results within studies and no allowance is made for variation between studies. Random-effects models allow for between-study variability in results by weighting studies using a combination of their own variance and the between-study variance. Where there is little between-study variability, the within-study variance will dominate and the random-effects weighting will tend towards that of the fixed-effect weighting. If there is substantial between-study variability, this dominates the weighting factor and within-study variability contributes little to the analysis. In this way, all trials will tend towards contributing equally towards the overall estimate and it can be argued that small studies will unduly influence the estimate. Those in favour of random-effects argue that it formally allows for between-study variability and that the fixed-effect approach unrealistically assumes a single effect across trials and gives over-precise estimates. In practice, with well-defined questions, the results of both approaches

are often very similar and it is common to run both to test robustness of the choice of statistical model.

Generic inverse variance method of combining study results

The generic inverse variance method is a widely used and easy to implement method of combining study results that underlies many of the approaches that are described

later. It is very flexible and can be used to combine any type of effect measure provided that an effect estimate and its standard error is available from each study. Effect estimates may include adjusted estimates, estimates corrected for clustering and repeat measurements, or other summaries derived from more complex statistical methods.

A fixed-effect meta-analysis using the generic inverse variance method calculates a weighted average of study effect estimates (EEIV) by summing individual effect

estimates (EEi), for example, the log odds ratio or the mean difference, and weighting these by the reciprocal of their squared standard errors (SEi) as follows:

A random-effects approach involves adjusting the study specific standard errors to incorporate between-study variation, which can be estimated from the effects and standard errors associated with the included studies.138

Types of data

Other ways to combine studies of effectiveness are available, some of which are specific to the nature of the data that have been collected, analysed and presented in the included studies.

Dichotomous/binary outcomes

Dichotomous outcomes are those that either happen or do not happen and an individual can be in one of only two states, for example having an acute myocardial infarction or not having an infarction. Dichotomous outcomes are most commonly expressed in terms of risks or odds. Although, in everyday use, the terms risk and odds are often used to mean the same thing, in the context of statistical evaluation they have quite specific meanings.

Risk describes the probability with which a health outcome will occur and is often expressed as a decimal number between 0.0 and 1.0, where 0.0 indicates that there is no risk of the event occurring, and 1.0 indicating certainty that the event will take place. A risk of 0.4 indicates that about four in ten people will experience the event. Odds describe the ratio of the probability that an event will happen to the probability that it

will not happen and can take any value between zero and infinity. Odds are sometimes expressed as the ratio of two integers such that 0.001 can be written 1:1000 indicating that for every one individual who will experience the event, one thousand will not.

Risk ratios (RR), also known as relative risks, indicate the change in risk brought about by an intervention and are calculated as the probability of an event in the intervention group divided by the probability of an event in the control group (where the probability of an event is estimated by the total number of events observed in the group divided by the total number of individuals in that group). A risk ratio of 2.0 indicates that the intervention leads to the risk becoming twice that of the comparator. A risk ratio of 0.75 indicates that the risk has been reduced to three quarters of that of the comparator.

This can also be expressed in terms of a reduction in risk whereby the relative risk reduction (RRR) is given as one minus the risk ratio multiplied by 100. For example, a risk ratio of 2.0 corresponds to a relative risk reduction of –100% (a 100% increase), while a risk ratio of 0.75 corresponds to a relative risk reduction of 25%. Box 1.6 illustrates the calculation of these measures and further details of the formulae can be found elsewhere.137

Risk ratios can be combined using the generic inverse variance method applied to the log risk ratio and its standard error (either in a fixed effect or a random-effects model). Odds ratios (OR) describe the ratio of the odds of events occurring on treatment to the odds of events occurring on control, and therefore describes the multiplication of the odds of the outcome that occur with use of the intervention. Box 1.6 illustrates how

to calculate the odds ratio for a single study. Odds ratios can be combined using the generic inverse variance method applied to the log odds ratio and its standard error as described above.

The Mantel-Haenszel method for combining risk ratios or odds ratios, which uses a different weighting scheme, is more robust when data are sparse, but assumes a fixed effect model.137

The Peto odds ratio139 (ORPeto) is an alternative estimate of a combined odds ratio in a fixed effect model, and is based on the difference between the observed number of events and the number of events that would be expected (O – E) if there was no difference between experimental and control interventions (see Box 1.6). Combining studies using the Peto method is straightforward, and it may be particularly useful for meta-analysis of dichotomous data when event rates are very low, and where other methods fail.

This approach works well when the effect is small (that is when the odds ratio is close to 1.0), events are relatively uncommon, and there are similar numbers in the

experimental and control groups. The approach is commonly used to combine data from cancer trials which generally conform to these expectations. Correction for zero cells is not necessary (see below) and the method appears to perform better than alternative approaches when events are very rare. It can also be used to combine time-to-event data by pooling log rank observed minus expected (O – E) events and associated variance. However, the Peto method does give biased answers in some circumstances,

especially when treatment effects are very large, or where there is a lack of balance in treatment allocation within the individual studies.140 Such conditions will not usually apply to RCTs but may be particularly important when combining the results of observational studies which are often unbalanced.

Although both risk ratios and odds ratios are perfectly valid ways of describing a treatment effect, it is important to note that they are not the same measure, cannot be used interchangeably and should not be confused. When events are relatively rare, say less than 10%,141 differences between the two will be small, but where the event

rate is high, differences will be large. For treatments that increase the chance of events, the odds ratio will be larger than the risk ratio and for interventions that reduce the chance of events, the odds ratio will be smaller than the risk ratio. Thus if an odds

ratio is misinterpreted as a risk ratio it will lead to an overestimation of the effect of intervention. Unfortunately, this error in interpretation is quite common in published reports of individual studies and systematic reviews. Although some statisticians prefer odds ratios owing to their

mathematical properties (they do not have inherent range limitations associated with high baseline rates and naturally arise as the antilog of coefficients in mathematical modelling, making them more suitable for statistical

manipulation), they have been criticised for not being well understood by clinicians and patients.142, 143 It may therefore be preferable, even when calculations have been based on odds ratios, to transform the findings to describe results as changes in the more intuitively understandable concept of risk.

Neither the risk ratio nor the odds ratio can be calculated for a trial if there are no events in the control group (as calculation would involve division by zero), and so in this situation it is customary to add 0.5 to each cell of the 2x2 table.137 If there are no events (or all participants experience the event) in both groups, then the trial provides no information about relative probability and so it is omitted from the meta-analysis. These situations are likely to occur when the event of interest is rare, and in such situations the choice of effect measure requires careful thought. A simulation study has shown that when events are rare, most meta-analysis methods give biased estimates of effect,144 and that the Peto odds ratio (which does not require a 0.5 correction) may be the least biased.

Continuous outcomes

Continuous outcomes are those that take any value in a specified range and can theoretically be measured to many decimal places of accuracy, for example, blood pressure or weight. Many other quantitative outcomes are typically treated as continuous data in meta-analysis, including measurement scales. Continuous data are usually summarized as means and presented with an indication of the variation around the mean using the standard deviation (SD) or standard error (SE). The effect of an intervention on a continuous outcome is measured by the absolute difference between the mean outcome observed for the experimental intervention and control, termed the mean difference (MD). This estimates the amount by which the treatment changes the outcome on average and is expressed:

Study mean differences and their associated standard errors can be combined using the generic inverse variance method.

Where studies assess the same outcome but measure it using different scales (for example, different quality of life scales), the individual study results must be standardised before they can be combined. This is done using the standardised mean difference (SMD), which considers the effect size in each study relative to the variability in the study and is calculated as the mean difference divided by the standard deviation among all participants. Where scales differ in direction of effect (i.e. some increase with increasing severity of outcome whilst others decrease with increasing severity), this needs to be accounted for by assigning negative values to the mean of one set of studies thereby giving all scales the same direction of measurement. There are three commonly used methods of recording the effect size in the standardised mean difference method, Cohen's d,145 Hedges adjusted g,145 and Glass' delta.146 The first two differ in whether the standard deviation is adjusted for small sample bias. The third differs from the other two by standardizing by the control group standard deviation rather than an

average standard deviation across both groups. The standardised mean difference assumes that differences in the standard deviation between studies reflect differences in the measurement scale and not differences between the study populations. The summary intervention effect can

be difficult to interpret as it is presented in abstract units of standard deviation rather than any particular scale.

Note that in social science meta-analyses, the term 'effect size' usually refers to versions of the standardised mean difference.

Time-to-event outcomes

Time-to-event analysis takes account not only of whether an event happens but when it happens. This is especially important in chronic diseases where even although we may not be able to ultimately stop an event from happening, slowing its occurrence can be beneficial. For example, in cancer studies in adult patients we rarely anticipate cure, but hope that we can significantly prolong survival. Time-to-event data are often referred to as 'survival' data since death is often the event of interest, but can be used for many different types of event such as time free of seizures, time to healing or time to conception. Each study participant has data capturing the event status and the time

of that status. An individual may be recorded with a particular elapsed time-to-event, or they may be recorded as not having experienced the event by a particular elapsed time or period of follow-up. When the event has not (yet) been observed, the individual is described as censored, and their event-free time contributes information to the analysis up until the point of censoring.

The most appropriate way to analyse time-to-event data is usually to use Kaplan Meier analysis and express results as a hazard ratio (HR). The HR summarises the entire

survival experience and describes the overall likelihood of a participant experiencing an event on the experimental intervention compared to control. Meta-analyses that collect individual participant data are able to carry out such analysis for each included study and then pool these using a variant of the Peto method described above. Alternatively a modelling approach can be used.

Meta-analyses of aggregate data often treat time-to-event data as dichotomous and carry out analyses using the numbers of individuals who did or did not experience an event by a particular point in time. However, using such dichotomous measures in a meta-analysis of time-to-event outcomes is discarding information and can pose additional problems. If the total number of events reported for each study is used to calculate an odds ratio or risk ratio, this can involve combining studies reported at

different stages of maturity, with variable follow-up, resulting in an estimate that is both unreliable and difficult to interpret. This approach is not recommended. Alternatively, ORs or RRs can be calculated at specific points in time. Although this makes estimates comparable,

interpretation can still be difficult, particularly if individual studies contribute data at different time points. In this case it is unclear whether any observed difference in effect between time points is attributable to the timing or to the analyses being based on different sets of contributing studies. Furthermore, bias could arise if the time points are subjectively chosen by the researcher or selectively reported by the study author at times of maximal or minimal difference between intervention groups.

A preferable approach is to estimate HRs by using and manipulating published or other summary statistical data or survival curves.147, 148 This approach has also been described in non-technical step-by-step terms.149 Currently, such methods are under-used in

meta-analyses,149 which may reflect unfamiliarity with the methods and that study reports do not always include the necessary statistical information150, 151 to allow the methods to be used.

Ordinal outcomes

Outcomes may be presented as ordinal scales, such as pain scales (where individuals' rate their pain as none, mild moderate or severe). These are sometimes analysed as continuous data, with each category being assigned a numerical value (for example, 0 for none, 1 for mild, 2 for moderate and 3 for severe). This is usual when there are many categories, as is the case for many psychometric scales such as the Hamilton depression scale or the Mini-Mental State Examination for measuring cognition.

However, a mean value may not be meaningful. Thus, an alternative way to analyse ordinal data is to dichotomise them (e.g. none or mild versus moderate or severe) to produce a standard 2 2 table. Methods are available for analysing ordinal data directly, but these typically require expert input.

Counts and rates

When outcomes can be experienced repeatedly they are usually expressed as event counts, for example, the number of asthma attacks. When these represent common events, they are often treated and analysed as continuous data (for example, number of days in hospital) and where they represent uncommon events they are often dichotomised (for example, whether or not each individual had at least one stroke).

When events are rare, analyses usually focus on rates expressed at the group level, such as the number of asthma attacks per person, per month. Although these can be combined as rate ratios using the generic inverse variance method, this is not always appropriate

as it assumes a constant risk over time and over individuals, and is not often done in practice. It is important not to treat rate data as dichotomous data because more than one event may have arisen from the same individual.

Presentation of quantitative results

Results should be expressed in formats that are easily understood, and in both relative and absolute terms.

Where possible, results should be shown graphically. The most commonly used graphic is the forest plot (see Box 1.7), which illustrates the effect estimates from individual studies and the overall summary estimate. It also gives a good visual summary of the review findings, allowing researchers and readers to get a sense of the data. Forest plots provide a simple representation of the precision of individual and overall results and of the variation between-study results. They give an 'at a glance' identification of any studies with outlying or unusual results and can also indicate whether particular studies are driving the overall results. Forest plots can be used to illustrate results for dichotomous, continuous and time-to-event outcomes.152

Individual study results are shown as boxes centred on their estimate of effect, with extending horizontal lines indicating their confidence intervals. The confidence interval expresses the uncertainty around the point estimate, describing a range of values within which it is reasonably certain that the true effect lies; wider confidence intervals reflect greater uncertainty. Although intervals can be reported for any level of confidence,

in most systematic reviews of health interventions, the 95% confidence interval is used. Thus, on the forest plot, studies with wide horizontal lines represent studies with more uncertain results. Different sized boxes may be plotted for each of the individual studies, the area of the box representing the weight that the study takes in the analysis providing a visual representation of the relative contribution that each study makes to the overall effect.

The plot shows a vertical line of equivalence indicating the value where there is no difference between groups. For odds ratios, risk ratios or hazard ratios this line will be drawn at an odds ratio/risk ratio/hazard ratio value of 1.0, while for risk difference and mean difference it will be drawn through zero. Studies reach conventional levels

of statistical significance where their confidence intervals do not cross the vertical line. Summary (meta-analytic) results are usually presented as diamonds whose extremities show the confidence interval for the summary estimate. A summary estimate reaches conventional levels of statistical significance if these extremities do not cross the line

of no effect. If individual studies are too dissimilar to calculate an overall summary estimate of effect, a forest plot that omits the summary value and diamond can be produced.

Odds ratios, risk ratios and hazard ratios can be plotted on a log-scale to introduce symmetry to the plot. The plot should also incorporate the extracted numerical data for the groups for each study, e.g. the number of events and number of individuals for odds ratios, the mean and standard deviation for continuous outcomes. Other forms of graphical displays have also been proposed.153

Relative and absolute effects

Risk ratios, odds ratios and hazard ratios describe relative effects of one intervention versus another, providing a measure of the overall chance of the event occurring on the experimental

intervention compared to control. These relative effects do not provide information on what this comparison means in absolute terms. Although there may

be a large relative effect of an intervention, if the absolute risk is small, it may not be clinically significant because the change in absolute terms is minimal (a big percentage of a small amount may still be a small amount). For example, a risk ratio of 0.8 may represent a 20% relative reduction in events from 50% to 40% or it could represent

a 20% relative reduction from 5% to 4% corresponding to absolute differences of 10% and 1% respectively. There may be situations where the former is judged to be

clinically significant whilst the latter is not. Meta-analysis should use ratio measures; for example, dichotomous data should be combined as risk ratios or odds ratios and pooling risk differences should be avoided. However, when reporting results it is generally useful to convert relative effects to absolute effects. This can be expressed as either an absolute difference or as a number needed to treat (NNT). Absolute change is usually

expressed as an absolute risk reduction which can be calculated from the underlying risk of experiencing an event if no intervention were given and the observed relative effect as shown in Box 1.8.

Consideration of absolute effects is particularly important when considering how results apply to different types of individuals who may have different underlying prognoses and associated risks. Even if there is no evidence that the relative effects of an intervention vary across different types of individual (see Subgroup analyses and Meta-regression below), if the underlying risks for different categories of individual differ, then the

effect of intervention in absolute terms will be different. It is therefore important when reporting results to consider how the absolute effect of an intervention varies for different types of individual and a table expressing results in this way, as shown in Table 1.5, can be useful. The underlying risk for different types of individual can be estimated from the studies included in the meta-analysis, or generally accepted standard estimates can be used. Confidence intervals should be calculated around absolute effects.

The NNT, which is derived from the absolute risk reduction as shown in Box 1.8, also depends on both relative effect and the underlying risk. The NNT represents the number of individuals who need to be treated to prevent one event that would be experienced on the control intervention. The lower the number needed to treat, the fewer the patients that need to be treated to prevent one event, and the greater the efficacy of the treatment. For example a meta-analysis of antiplatelet agents for the prevention of pre-eclampsia found an RR of 0.90 (0.84 – 0.97) for pre-eclampsia.158 Plausible underlying risks of 2%, 6% and 18% had associated NNTs of 500 (313-1667), 167 (104-556) and 56 (35-185) respectively.

Sensitivity analyses

Sensitivity analyses explore the robustness of the main meta-analysis results by repeating the analyses having made some changes to the data or methods.159 Analyses run with and without

the inclusion of certain trials will assess the degree to which studies (perhaps those with poorer methodology) affect the results. For example, analyses might be carried out on all eligible trials and a sensitivity analysis restricted to only those that used a placebo in the control group. If results differ substantially, the final results will require careful interpretation. However, care must be taken in attributing reasons for differences, especially when a single or small numbers of trials are included/excluded in the sensitivity analysis, as a study may differ in additional ways to the issue being explored in the sensitivity analysis. Some sensitivity analyses should be proposed in the protocol, but as many issues suitable for exploration in sensitivity analyses only come to light whilst the review is being done, and in response to decisions made or difficulties encountered, these may have to change and/ or be supplemented.

Exploring heterogeneity

There will inevitably be variation in the observed estimates of effect from the studies included in a meta-analysis. Some of this variation arises by chance alone, reflecting the fact that no study is so large that random error can be removed entirely. Statistical heterogeneity refers to variation other than that which arises by chance. It reflects methodological or clinical differences between studies. Exploring

statistical heterogeneity in a meta-analysis aims to tease out the factors contributing to differences, such that sources of heterogeneity can be accounted for and taken into consideration when interpreting results and drawing conclusions.

There is inevitably a degree of clinical diversity between the studies included in a review,160 for example because of differing patient characteristics and differences in interventions. If these factors influence the estimated intervention effect, then there will be some statistical heterogeneity between studies. Methodological differences that influence the observed intervention effect will also lead to statistical heterogeneity. For example, combining results from blinded and unblinded studies may lead to statistical heterogeneity, indicating that they might best be analysed separately rather than in combination. Although it manifests itself in the same way, heterogeneity arising from clinical differences is likely to be because of differences in the true intervention effect, whereas heterogeneity arising from differences in methodology is more likely to be because of bias.

An idea of heterogeneity can be obtained straightforwardly by visually examining forest plots for variations in effects. If there is poor overlap between the study confidence intervals, then this generally indicates statistical heterogeneity.

More formally a ^{2} (chi-squared) test (see Box 1.9), often also referred to as Q- statistic, can assess whether differences between results are compatible with chance alone. However, care must be taken in interpreting the chi-squared test as it has low power, consequently a larger P value ($P<0.1$) is sometimes used to designate statistical significance. Although a statistically significant test result may point to a problem with heterogeneity, a nonsignificant test result does not preclude important between-study differences, and cannot be taken as evidence of no heterogeneity. Conversely, if there are many studies in a meta-analysis, the test has high

power to detect a small amount of heterogeneity that, although statistically significant, may not be clinically important.

Accepting that diversity is likely to be inherent in any review, methods have also been developed to quantify the degree of inconsistency across studies, shifting the focus from significance testing to quantifying heterogeneity. The I2 statistic160, 161 describes the percentage of variability in the effect estimates that can be attributed to heterogeneity rather than chance (see Box 1.9).

Although the I2 statistic often has wide confidence intervals and it is difficult to provide hard and fast rules on what level of inconsistency is reasonable in a meta-analysis, as a rough guide it has been suggested that I2 values of up to 40% might be unimportant, 30% to 60% might be moderate, 50 to 90% may be substantial and 75% to 100% considerable.75

If statistical heterogeneity is observed, then the possible reasons for differences should be explored162 and a decision made about if and how it is appropriate to combine studies. A systematic review does not always need to include a meta-analysis and, if there are substantial differences between study estimates of effect, particularly if they are in opposing directions, combining results in a meta-analysis can be misleading.

One way of addressing this is to split studies into less heterogeneous groups according to particular study level characteristics (e.g. by type of drug), and perform separate analyses for each group. Forest plots can be produced to show subsets of studies on the same plot. Each subset of studies can have its own summary estimate, and if appropriate an overall estimate combined across all studies can also be shown. Showing these groupings alongside each other in this way provides a good visual summary of how they compare. This approach allows the consistency and inconsistency between subsets of studies to be examined. Differences can be summarised narratively, but where possible they should also be evaluated formally. A ^{2} test for differences across subgroups can be carried out (see Box 1.9).

The influence of patient-level characteristics (e.g. age, gender) or issues related to equity (e.g. ethnicity, socioeconomic group) can also be explored through subgroup analyses, meta-regression or other modelling approaches. However, there is generally insufficient information in published study reports to allow full exploration of heterogeneity in this way and this can usually only be addressed satisfactorily when IPD are available. Such exploration of heterogeneity may enable additional questions to be addressed, such as which particular treatments perform best or which types of patient will benefit most, but is unlikely to be helpful when there are few studies. Wherever possible, potential sources of heterogeneity should be considered when writing the review protocol and possible subgroup analyses pre-specified rather than trying to explain statistical heterogeneity after the fact.

Subgroup analyses

Subgroup analyses divide studies (for study level characteristics) or participant data (for participant level characteristics) into subgroups and make indirect comparisons between them.

These analyses may be carried out to explore heterogeneity (see above) as well as to try to answer particular questions about patient or study factors. For example a subgroup analysis for study level characteristics might examine whether the results of trials carried out in primary health care settings are the same as trials carried out in a hospital setting. A participant level subgroup analysis might examine whether the effect of the intervention is the same in men as in women.

In individual studies it is unusual to have sufficient numbers and statistical power to permit reliable subgroup analyses of patient characteristics. However, provided that such data have been collected uniformly across studies, a meta-analysis may achieve sufficient power in each subgroup to permit a more reliable exploration of whether the effect of an intervention is larger (or smaller) for any particular type of individual. Although, owing to the multiplicity of testing, these analyses are still potentially misleading, subgroup analysis within the context of a large meta-analysis may be the only reasonable way of performing such exploratory investigations. Not only do the greater numbers give increased statistical power, but consistency across trials can be

investigated. Indeed, the possibility of undertaking such analyses is a major attraction of IPD meta-analyses as dividing participant data into groups for subgroup analysis is seldom possible in standard reviews of aggregate data.163 Subgroup analyses in most (non IPD) systematic reviews focus on grouping according to trial attributes.

The interpretation of the results of subgroup analyses must be treated with some caution. Even where the original data have come from RCTs, the investigation of between-study differences is indirect and equivalent to an observational study.164, 165 There may be explanations for the observed differences between groups, other than the attributes chosen to categorise groupings. Comparisons which are planned in advance on the basis of a plausible hypothesis and written into the protocol are more credible than findings that are found through post hoc exploratory analyses. Furthermore, the likelihood of finding false negative and false positive significance tests rises rapidly as more subgroup analyses are done. Subgroups should therefore be restricted to a few potentially important characteristics where it is reasonable to suspect that the characteristic will interact with or modify the effect of the intervention. Note that there is often confusion between prognostic factors and potential effect modifiers; just because a characteristic is

prognostic does not mean that it will modify the effect of an intervention. For example, whilst gender is prognostic for survival (women live longer than men) it does not necessarily mean that women will benefit more than men will from a drug to treat lung cancer.

Meta-regression

Meta-regression can be used to investigate the effects of differences in study characteristics on the estimates of the treatment effect,140 and can explore continuous as well as categorical characteristics. In principle it can allow for the simultaneous exploration of several characteristics and their interactions, though in practice this is seldom possible because of small numbers of studies.166 As in any simple regression analysis, meta- regression aims to predict

outcome according to explanatory variables or covariates of interest. The covariates may be constant for the entire trial, for example, the protocol dose of a drug, or a summary measure of attributes describing the patient population, for example, mean age or percentage of males. The regression is weighted by precision of study estimates such that larger studies have more influence than smaller studies.

The regression coefficient is tested to establish whether there is an association between the intervention effect and the covariate of interest. Provided that enough data are available (at least 10 studies),82 the technique may be a useful exploratory tool.

However, there are limitations. Not all publications will report on all the covariates of interest (and there could be potential bias associated with selective presentation of data that have shown a positive association within a primary study). If a study is missing

a covariate it drops out of the regression, limiting the power and usefulness of the analysis, which is already likely to be based on relatively few data points.

Meta-regression is not a good way to explore differences in treatment effects between different types of individual as summary data may misrepresent individual participants.167 What is true of a study with a median participant age of 60 may not necessarily be true for a 60-year-old patient. Potentially all the benefit could have been shown in the 50-year-olds and none in the 60 and 70-year-olds. Comparison of treatment effects between different types of individual, for example between men and women, should be done using subgroup analyses and not by using meta-regression

incorporating the proportion of women in each trial. It should always be borne in mind that finding a significant association in a meta-regression does not prove causality and should rather be regarded as hypothesis generating.

Assessing the possibility of publication bias

Although thorough searches should ensure that a systematic review captures as many relevant studies as possible, they cannot eliminate the risk of publication bias. As publication and associated biases can potentially influence profoundly the findings of

a review, the risk of such bias should be considered in the review's conclusions and inferences.24 The book by Rothstein et al provides a comprehensive discussion of publication bias and associated issues.168

The obvious way to test for publication bias is to compare formally the results of published and unpublished studies. However, more often than not unpublished studies are hidden from the reviewer, and more ad hoc methods are required. A common technique to help assess potential publication bias is the funnel plot.

This is a scatter plot based on the fact that precision in estimating effect increases with increasing sample size. Effect size is plotted against some measure of study precision – of which standard error is likely to be the best choice.169 A wide scatter in results of

small studies, with the spread narrowing as the trial size increases, is expected. If there is no difference between the results of small and large studies, the shape of the plot should resemble an inverted funnel (see Box 1.10). If there are differences, the plot will be skewed and a gap where the small unfavourable studies ought to be is often cited

as evidence of publication bias. However, the shape of a funnel plot can also depend on the measures selected for estimating effect and precision169, 170 and could be attributable to differences between small and large studies other than publication bias. These differences could be a result of other types of methodological bias, or genuine clinical differences. For example, small studies may have a more selected participant population where a larger treatment effect might be expected. Funnel plots are therefore more accurately described as a tool for investigating small study effects.

Although visual inspection of funnel plots has been shown to be unreliable,170, 171 this might be improved if contour zones illustrating conventional levels of significance are overlaid on the plot to illustrate whether 'missing' studies are from zones of statistical significance or not. If the 'missing' studies are from nonsignificant zones, this may support a publication bias. On the other hand if 'missing' studies are from statistically significant zones, the asymmetry may be more likely to be attributable

to other causes.172 Over time a range of statistical and modelling methods have been developed to test for asymmetry, the most frequently used of which are those based on rank correlation173 or linear regression174, 175 and complex modelling176 methods.

Some methods (for example, the trim and fill method177, 178) attempt to adjust for any publication bias detected.176 However, all methods are by nature indirect and the appropriateness of many methods is based on some strict assumptions that can be difficult to justify in practice.

Although frequently used to help assess possible publication bias, funnel plots and associated statistical tests are often used and interpreted inappropriately,179, 180 potentially giving false assurance where a symmetrical plot overlooks important bias or undermining important valid evidence because of an asymmetric plot.179 The methods are inappropriate where there is statistical heterogeneity; have low power

and are of little use where there are few studies; and are meaningless where studies are of similar size. Consequently, situations where they are helpful are few and their use is not generally a good way of dealing with publication bias.181 Therefore use of these methods to identify or adjust for publication bias in a meta-analysis should be considered carefully and generally be restricted to sensitivity analyses. Results should be interpreted with caution. Statistical tests will not resolve bias and avoidance of publication bias is preferable. In time this may become easier with more widespread registration of clinical trials and other studies at inception.182

Dealing with special study designs and analysis issues

Intention to treat analyses

ITT is usually the preferred type of analysis as it is less likely to introduce bias than alternative approaches. True intention to treat analysis captures two criteria: (i) participants should be analysed irrespective of whether or not they received their allocated intervention and irrespective of what occurred subsequently, for example, participants with protocol violations or those subsequently judged ineligible should be included in the analysis; (ii) all participants should be included irrespective of whether outcomes were collected. Although the first criterion is generally accepted, there is no clear consensus on the second81 as it involves including participants in the analyses whose outcomes are unknown, and therefore requires imputation of data. Many authors describe their analyses as ITT when only the first criterion has been met. Alternative analysis of all participants for whom outcome data are available is termed available case analysis. Some studies present analysis of all participants who completed their allocated treatment, this is per protocol or treatment received analysis which may be seriously biased.

Imputing missing data

Although statistical techniques are available to impute missing data, this cannot reliably compensate for missing data184 and in most situations imputation of data is not recommended. It is reasonable for most systematic reviews to aim for an available case analysis and include data from only those participants whose outcome is known.

Achieving this may require making contact with the study author if individuals for whom outcome data were recorded have been excluded from the published analyses. The extent and implications of missing data should always be reported and discussed in

the review. If the number of participants missing from the final analysis is large it will be helpful to detail the reasons for their exclusion.

In some circumstances, it might be informative to impute data in sensitivity analyses to explore the impact of missing data.185 For missing dichotomous data the analysis can assume that either all participants with missing data experienced the event, or that they all did not experience the event. This generates the theoretical extremes of possible effect. Data could also be imputed using the rate of events observed in the control group, however this does not add information, gives inflated precision and is not recommended. Where missing data are few, imputation will have little impact on the results. Where missing data are substantial, analysis of worst/best case scenarios will give a wide range of possible effect sizes and may not be particularly helpful.

Approaches to imputing missing continuous data have received little attention. In some cases it may be possible to use last observation carried forward, or to assume that no change took place. However, this cannot be done from aggregate data and the value of such analysis is unclear. Any researcher contemplating imputing missing data should consult with an experienced statistician.

Cluster randomised trials

In cluster randomised trials, groups rather than individuals are randomised, for example clinical practices or geographical areas. Reasons for allocating interventions in this

way include evaluating policy interventions or group effects such as in immunisation programmes, and avoiding cross-contamination, for example, health promotion information may be easily shared by members of the same clinic or community. In many instances clustering will be obvious, for example where primary care practices are allocated to receive a particular intervention. In other situations the clustering may be less obvious, for example where multiple body parts on the same individual are allocated treatments or where a pregnant woman has more than one fetus. It is important that any cluster randomised trials are identified as such in the review.

As participants within any one cluster are likely to respond in a manner more similar to each other than to other individuals (owing to shared environmental exposure or personal interactions), their data cannot be assumed to be independent. It is therefore inappropriate to ignore the clustering and analyse as though allocation had been at the individual level. This unit of analysis error would result in overly narrow confidence intervals and straightforward inclusion of trials analysed in this way would give undue

weight to that study in a meta-analysis. Unfortunately, many primary studies have ignored clustering and analysed results as though from an individual randomized trial.186, 187 One way to avoid the problem of inappropriately analysed cluster trials is to carry out meta-analyses using a summary measure for each cluster as a single observation. The sample size becomes the number of clusters (not the number of individuals) and the analysis then proceeds as normal. However, depending on the size and number of clusters, this will reduce the statistical power of the analysis considerably and unnecessarily. Indeed, the information required to do this is unlikely to be available in many study publications.

A better approach is to adjust the results of inappropriately analysed primary studies to take account of the clustering, by increasing the standard error of the estimate of effect.75 This may be achieved by multiplying the original standard error by the square root of the 'design effect'. The design effect can be calculated from the intracluster correlation coefficient, which, although seldom reported, can use external values from similar studies such as those available from the University of Aberdeen Health Services Research Unit (www.abdn.ac.uk/hsru/epp/iccs-web.xls).) A common design effect is usually adopted across the intervention groups.

These values can then be used in a generic inverse variance meta-analysis alongside unadjusted values from appropriately analysed trials.

Cross-over trials

Cross-over trials allocate each individual to a sequence of interventions, for example one group may be allocated to receive treatment A followed by treatment B, and the other group allocated to receive B followed by A. This type of trial has the advantage that each participant acts as their own control, eliminating between participant variability such that fewer participants are required to obtain the same statistical power. They are suitable for evaluating interventions

that have temporary effects in treating stable conditions. They are not appropriate where an intervention can have a lasting effect that compromises treatment in subsequent periods of the trial, or where a condition has rapid evolution, or the primary outcome is irreversible. The first task of

the researcher is to decide whether the cross-over design is appropriate in assessing the review question.

Appropriate analysis of cross-over trials involves paired analysis, for example using a paired t-test to analyse a study with two interventions and two periods (using experimental measurement – control measurement) for each participant, with standard errors calculated for these paired measurements. These values can then be combined in a generic inverse variance meta-analysis. Unfortunately, cross-over trials are frequently inappropriately analysed and reported.

A common naive analysis of cross-over data is to treat all measurements on experimental and control interventions as if they were from a standard parallel group trial. This results in confidence intervals that are too wide and the trial receives too little weight in the meta-analysis. However, as this is a conservative approach, it might not be unreasonable in some circumstances. Where the effect of the first intervention is thought to have influenced the outcome in subsequent periods (carry-over), a common approach is to use only the data from the first period for each individual. However, this will be biased if the decision to analyse in this way is based on a test of carry-over and studies analysed in this way may differ from those using paired analyses. One approach to combining studies with differing types of reported analyses is to carry out an analysis grouped by type of study i.e. grouped by cross-over trial paired analysis, cross-over trial with first period analysis, parallel group trial, and explore whether their results differ (see Subgroup analyses above).

Alternatively, the researcher can carry out their own paired analysis for each trial if (i) the mean and standard deviation or standard error of participant differences are available; (ii) the mean difference plus a t-statistic, p-value or confidence interval from a paired analysis is available; (iii) a graph from which individual matched measurements can be extracted; or (iv) if individual participant data are available.188 Another approach is to attempt to approximate a paired analysis by imputing missing standard errors by 'borrowing' from other studies that have used the same measurement scale or by a correlation coefficient obtained from other studies or external sources.75 Researchers

will need to decide whether excluding trials is preferable to inferring data. If imputation is thought to be reasonable, advice should be sought from an experienced statistician. Authors should state explicitly where studies have used a cross-over design and how this has been handled in the meta-analysis.

Mixed treatment comparisons

Mixed treatment comparisons (MTC), or network meta-analyses, are used to analyse studies with multiple intervention groups and to synthesise evidence across a series of studies in

which different interventions were compared. These are used to rank or identify the optimal intervention. They build a network of evidence that includes both direct evidence from head-to-head studies and indirect comparisons whereby interventions that have not been compared directly are linked through common comparators. A framework has been described that outlines some of the circumstances in which such syntheses might be considered.189 Methods for conducting indirect comparisons190, 191 and more complex mixed treatment methods192, 193 require expert advice. Researchers wishing to undertake such analyses should consult with an appropriately experienced statistician.

Bayesian methods

Unlike standard analysis techniques, Bayesian analyses allow for the combination of existing information with new evidence using established rules of probability.194 A simple Bayesian analysis model includes three key elements:

1. Existing knowledge on the effect of an intervention can be retrieved from a variety of sources and summarised as a prior distribution
2. The data from the studies are used to form the likelihood function
3. The prior distribution and the likelihood function are formally combined to provide a posterior distribution which represents the updated knowledge about the effect of the intervention

Bayesian approaches to meta-analysis may be useful when evidence comes from a diverse range of sources particularly when few data from RCTs exist.195, 196 They can also be used to account for the uncertainty introduced by estimating the between-study variance in the random-effects model, which can lead to reliable estimates and predictions of treatment effects.197 While there are several good texts available,198-200 if a Bayesian approach is to be used, the advice of a statistical expert is strongly recommended.

Describing results

When describing review findings, the results of all analyses should be considered as a whole, and overall coherence discussed. Consistency across studies should be considered and results interpreted in relation to biological and clinical plausibility.

Where there have been many analyses and tests, care should be taken in interpreting unexpected or implausible findings as among a large number of tests the play of chance alone is likely to generate spurious statistically significant results.

Quantitative results of meta-analyses should be expressed as point estimates together with associated confidence intervals and exact p-values. They should not be presented or discussed only in terms of statistical significance. This is particularly important where results are not statistically significant as nonsignificance can arise both when estimates are close to no effect with narrow confidence intervals, or when estimates of effect

are large with wide confidence intervals. Whilst in the former, we can be confident that there is little difference between the interventions compared, in the latter there is

insufficient evidence to draw conclusions. Researchers should be aware that describing a result as 'there is no statistical (or statistically significant) difference between the two interventions' can be (mis)read as there being no difference between interventions.

It is important that inconclusive results are not interpreted as indicating that an intervention is ineffective and estimates with wide confidence intervals that span no effect should be described as showing no clear evidence of a benefit or harm rather than as there being no difference between interventions. Demonstrating lack of sufficient evidence to reach a clear conclusion is an important finding in its own right.

Similarly, care should be taken not to overplay results that are statistically significant, as with large enough numbers, even very modest differences between interventions can be statistically significant. The size of the estimated effect, and its confidence intervals, should be considered in view of how this relates to current or future practice (see Section 1.3.6 Report writing).

It is usually helpful to present findings in both relative and absolute terms and in particular to consider how relative effects may translate into different absolute effects for people with differing underlying prognoses (see Relative and absolute effects section above). Where a number of outcomes or subgroup analyses are included in a review it can be helpful to tabulate the main findings in terms of effect, confidence intervals and inconsistency or heterogeneity statistics.

Summary: Data synthesis

- Synthesis involves bringing the results of individual studies together and summarising their findings.
- This may be done quantitatively or, if formal pooling of results is inappropriate, through a narrative approach.
- Synthesis should also explore whether observed intervention effects are consistent across studies, and investigate possible reasons for any inconsistencies.

2.2.5.3 Initial descriptive synthesis

All syntheses should begin by constructing a clear descriptive summary of the included studies.

Narrative synthesis is frequently an essential part of a systematic review, and as with every other stage of the process, bias must be minimized.

Narrative synthesis has typically not followed a strict set of rules. However, a general framework can be applied in order to help maintain transparency and add credibility to the process. The four elements of this framework are:

- Developing a theory of how the intervention works, why and for whom
- Developing a preliminary synthesis of findings of included studies
- Exploring relationships within and between studies
- Assessing the robustness of the synthesis

  Each element contains a range of tools and techniques that can be applied. A researcher is likely to move iteratively among the four elements, choosing those tools and techniques that are appropriate to the data being synthesised and providing justifications for these choices.

2.2.5.4 Quantitative synthesis

- Meta-analysis increases power and precision in estimating intervention effects.
- Results of individual studies are combined statistically to give a pooled estimate of the 'average' intervention effect.
- Most meta-analysis methods are based on calculating a weighted average of the effect estimates from each study.
- The methods used to combine results will depend on the type of outcome assessed.
- Quantitative results should be expressed as point estimates together with associated confidence intervals and exact p-values.
- Variation in results across studies should be investigated.
- Sensitivity analyses give an indication of the robustness of results to the type of study included and the methods used.

2.2.6 Report writing

Report writing is an integral part of the systematic review process. This section deals with the primary scientific report of the review which often takes the form of a comprehensive report to the commissioning body. Many commissioners have their own guidance for production and submission of the report. Alternatively the primary report may take the form of a journal article, where space limitations may mean that important details of the review methods have to be omitted. These can be made available through the journal's or the review team's website. Whatever the format, it is important to take as much care over report preparation as over the review itself. The report should describe the review methods clearly and in sufficient detail that others could, if they wished, repeat them. There is evidence that the quality of reporting in reports of primary studies may affect the readers' interpretation of the results, and the same is likely to be true of systematic reviews.201 It has also been argued that trials and reviews

often provide incomplete or omit the crucial 'how to' details about interventions, limiting a clinicians' ability to implement findings in practice.202-204

The QUOROM statement9 has set standards for how reviews incorporating meta-analysis should be reported, and many journals require articles submitted to adhere to these standards. The QUOROM checklist and flow chart are useful resources for all authors of systematic review reports. However, recognising that the quality of reporting of many systematic reviews is disappointing,205 the QUOROM group have broadened their remit, been renamed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta- Analyses),206 and developed a flow chart and checklist for the reporting of systematic reviews with or without a meta-analysis.66, 67

1. General considerations

Resources for writers

There are many resources for writers available in both printed and electronic form. These include guides to technical writing and publishing,207-209 style manuals210, 211 and guides to use of English.212 The EQUATOR Network is an initiative that seeks to improve the quality of scientific publications by promoting transparent and accurate reporting of health research.101 It provides an introduction to reporting guidelines, and information for authors of research reports, editors and peer reviewers as well as those developing reporting guidelines.

Style and structure

Commissioning bodies and journals usually have specific requirements regarding presentation and layout that should be followed when preparing a report or article. Some organisations offer detailed guidance while others are less specific. In the absence of guidance, a layered approach such as a one page summary of the research 'actionable messages', three-page executive summary and a 25-page report is advocated as the optimal way to present research evidence to health service managers and policy-makers.213 Box 1.11 presents a suggested outline structure for a typical report of a systematic review.

Many journals publish papers electronically ahead of print publication and electronic publishing often allows additional material, such as large tables, or search strategies to be made available through the journal's website. There is no specific word limit for reports published in electronic format only, for example in the Cochrane Library, although Cochrane reviews 'should be as succinct as possible'.75

Suggested structure of a systematic review report

Title

Contents list Abbreviations/glossary

Executive summary or structured abstract

Background Objectives

Methods (data sources, study selection, data extraction, quality assessment, data synthesis)

Results Conclusions

Main text Background/introduction Review question(s) Review methods

Identification of studies

Study selection (inclusion and exclusion criteria; methods) Data extraction

Quality assessment Data synthesis

Results of the review

Details of included and excluded studies Findings of the review

Secondary analyses (sensitivity analyses etc.) Discussion (interpretation of the results) Conclusions

Recommendations/implications for practice/policy Recommendations/implications for further research

Acknowledgements or list of contributors and contributions Funding

Conflicts of interest References Appendices

Researchers should familiarise themselves with the conventions favoured by their commissioning body or 'target' journal. Many journals now prefer a clear and active style that is understandable to a general audience. Weaknesses in the use of grammar and spelling constitute obstacles to clear communication and should be eliminated as far as possible. The field of scientific and technical communication predominantly uses English as its common language, so those who are unsure of their ability in written English may find it helpful to have their report checked by an accomplished speaker/ writer who is familiar with the subject matter before submission.

Contents lists and headings are essential for guiding the reader through longer documents. Inclusion of an index may also be helpful. It is particularly important to adopt a consistent style (e.g. font, point size, font style) for different levels of main headings and sub-headings.

Planning

Time spent preparing a brief outline covering the main points to be included in the report can save time overall. The outline should focus on who the intended audience is and what they need to know. The review team will need to agree the outline and, if the report is to be written by multiple authors, allocate writers for each section. Dividing the work amongst a number of people reduces the burden on each individual but there is a risk of loss of consistency in style and terminology. In addition, completion of the report relies on all the team members working to the agreed schedule. It is essential for the lead author (corresponding author for journal articles) to monitor progress and take responsibility for accuracy and consistency.

Authorship and contributorship

The report of a systematic review will usually have a number of authors. According to the International Committee of Medical Journal Editors (ICMJE),214 authorship credit should be based on:

1. Substantial contributions to conception and design, or acquisition of data, or analysis and interpretation of data
2. Drafting the article or revising it critically for important intellectual content; and
3. Final approval of the version to be published

All authors should meet all of these conditions. The review team should agree amongst themselves who will be authors and the order of authorship. Order of authorship is often taken to reflect an individual's contribution to the report and methods are available for scoring contributions to determine authorship.215 Alternatively authors can simply be listed alphabetically. Contributions that do not meet the criteria for authorship (for example, data extraction or membership of an advisory group) should be included in the acknowledgements.

Some journals, for example the BMJ, favour a system of contributorship.216 In addition to the standard list of authors, there is a list of all those who contributed to the paper with details of their contributions. One contributor (occasionally more than one) is listed as guarantor and accepts overall responsibility for the work. This system gives some credit to those who do not meet the ICMJE criteria for authorship and provides accountability for each stage of the review.

Peer review and feedback

Most systematic reviews have an expert advisory group assembled at the beginning of the project and members of this group should be asked to review the draft report and comment on its scientific quality and completeness. The commissioning body may also organise its own independent peer review of the draft report before publication.

Medical journals almost invariably seek external peer review of manuscripts submitted for publication. Draft manuscripts may also be posted on institutional websites or electronic preprint servers, allowing an opportunity for feedback from a wide range of interested parties, although for reports intended for journals it is important to ensure that such posting will not be considered as prior publication.

In addition to scientific peer review, end users may also be asked to assess the relevance and potential usefulness of the review. They may recommend changes that would help in identifying the main messages for dissemination and important target audiences as well as possible formats and approaches.

When feedback from external reviewers has been received, a final report can be prepared. A record of the comments and the way in which they were dealt with should be kept with the archive of the review.

Conflict of interests

The ICMJE state that a conflict of interests exists if 'an author (or the author's institution), reviewer, or editor has financial or personal relationships that inappropriately influence (bias) his or her actions'.214 Relationships that might constitute a conflict of interests are common and there is nothing wrong with having such relationships. However, it is important that they are declared so that readers are aware of the possibility that authors' judgements may have been influenced by other factors. Review authors need to be explicit about any potential conflict of interests because such transparency is important in maintaining the readers' confidence.

2.2.6.1 Executive summary or abstract

The executive summary (for full-length reports) or abstract (for journal articles) is the most important part of the report because potentially it is the only section that many readers will actually read (perhaps in conjunction with the discussion and conclusions). It should present the findings of the review clearly and concisely and allow readers to quickly judge the quality of the review and the generalisability of its findings. Providing a good balance between detail of the intervention and how the review was conducted, and the results and conclusions is always a challenge, and may require several iterations across the whole review team. The summary is usually the last section to be written so that full consideration can be given to all relevant aspects of the project. However, the process of summary writing may help in the further development of the recommendations by forcing review teams to identify the one or two most important findings and the conclusions which flow from them. It should be remembered that revisions to the report or article following peer review may also need to be reflected in the summary. Assistance from outside parties and medical writers may be helpful in developing a good summary.

2.2.6.2 Formulating the discussion

The purpose of the discussion section of a report is to help readers to interpret the results of the review. This should be done by presenting an analysis of the findings and outlining the strengths and weaknesses of the review. The discussion should also place the findings in the context of the existing evidence base, particularly in relation to any existing relevant reviews. It has been suggested that more could and should be done in discussion sections to contextualise both the nature of the research and the findings to the existing evidence base.217 There should be a balance between objectively describing the results, and subjectively speculating on their meaning.218 It is important to present a clear and logical train of thought and reasoning, supported by the findings of the review and other existing knowledge. For example although statistically significant results and clear evidence of effectiveness may have been demonstrated, without an exploration of the impact on clinical practice it may not be clear whether they are clinically significant. Information on the interpretation of the analysis is given throughout Section 1.3.5 Data synthesis.

Some commissioners and most journals have a set format or structure for the report. This may require the discussion section to incorporate the conclusions and any implications or recommendations, or may require these as separate sections. In the absence of a structured format for the discussion section, the framework given in Box 1.12 may be helpful.

Framework for the discussion section of a review

> Statement of principal findings
>
> Strengths and weaknesses of the review
>
> Appraisal of methodological quality of the review
>
> Relation to other reviews, in particular considering any differences
>
> Meaning of the review's findings
>
> Strengths and weaknesses of the evidence included in the review Direction and magnitude of effects observed in the included studies Applicability of the findings of the review
>
> Implications
>
> Practical implications for clinicians and policy-makers Unanswered questions and implications for further research

2.2.6.3 Conclusions, implications, recommendations

Faced with the need to make decisions and limited time to read the whole report, many readers may go directly to the conclusions. Therefore, whether incorporated in the discussion section or presented separately, it is essential that the conclusions be clearly worded and based solely on the evidence reviewed. The conclusions should summarise the evidence and draw out the implications for health care, and preferably be worded to show how they have been derived from the evidence.

Conclusions are generally a standard requirement, however, many commissioners and journals have their own conventions about implications and recommendations. For example, the NIHR HTA programme require the conclusions section of reports to include the implications for health care and specify recommendations for future research,

in order of priority. They specifically exclude making recommendations for policy or clinical practice.220 Authors' conclusions from Cochrane reviews are presented as the implications for practice and research; recommendations are not made.130

In the absence of guidance from the commissioner, it is generally advisable to avoid making recommendations about policy or practice, unless this is the focus of the review. The nature of the review question should therefore guide whether it is appropriate to include recommendations or focus on the implications for policy, practice and/or further research, and how these

are best presented. Whether recommendations are made or implications drawn, it is important to ensure that these are supported by the evidence and to avoid making any statements that are outside the defined scope of the review.

The way in which a recommendation or implication is phrased can considerably influence the way in which it is interpreted and implemented (or ignored). Hence, it is important to make all statements as precise as possible.221-223

Recommendations for practice are usually only made in guidelines, and are formulated from a variety of sources of information in addition to review findings. There are a number of schemes available for grading practice recommendations according to the strength of the evidence that supports them.224-230 Systematic review reports should aim to provide the information required to implement any of these systems if used. It should be noted that not all the schemes take into account the generalisability of the findings of the review to routine clinical practice. This should always be a consideration when drawing up the implications or if making recommendations.

A clear statement of the implications or recommendations for future research should be made; vague statements along the lines of 'more research is needed' are not helpful and should be avoided. Specific gaps in the evidence should be highlighted to identify the research questions that need answering. Where methodological issues have been identified in existing studies, suggestions for future approaches may be made. Where possible, research recommendations should be listed in order of priority, and an indication of how rapidly the knowledge base in the area is developing should be included. This can assist in planning an update of the review and help guide commissioners when allocating funding.

The DUETs initiative (Database of Uncertainties about the Effects of Treatments; (www.duets.nhs.uk), recommends the presentation of research recommendations in a structured format represented by the acronym EPICOT (Evidence, Population(s), Intervention(s), Comparison(s), Outcome(s), Time stamp). Timeliness (duration of intervention/follow-up), disease burden and suggested study design are considered as optional additional elements of a structured research recommendation. Further details and an example of how to formulate research recommendations using the EPICOT format can be found in an article published by the DUETS Working Group.231 It is worth noting that there is some debate about the applicability of the EPICOT format for some reviews, particularly those of complex interventions.

Summary: Report writing

- Report writing is an integral part of the systematic review process.
- Reviews may be published as a report for the commissioner, as a journal article or both. Researchers should be aware of the requirements of commissioning bodies and journals and adhere to them.

- Readability is a key aspect of reporting; a review's findings will not be acted on if they are not clearly presented and understood.
- The executive summary (for full-length reports) or abstract (for journal articles) is the most important part of the report, because it is potentially the only section that many readers will actually read (perhaps in conjunction with the discussion and conclusions).
- A structured framework can be helpful for preparing the discussion section of the report.
- Implications for practice or policy and recommendations for further research should be based solely on the evidence contained in the review.
- The findings from systematic reviews are frequently used to inform guideline development. Guideline recommendations are often formulated using a grading scheme. Systematic review reports should therefore aim to provide the information required for such grading schemes.
- A structured format for the presentation of research recommendations has been developed as a result of the DUETS initiative.

2.2.7 Archiving the review

There are published guidelines relating to the retention of primary research data.233 While these do not currently relate to systematic reviews, they do represent appropriate good practice. Where policies on retention, storage and protection are not specified

by a commissioner, researchers might consider including this information in research proposals so that it is clear from the outset what will be kept and for how long.

Decisions need to be made about which documents are vital to keep and which can be safely disposed of. Extracted data and quality assessment information should be preserved. In addition, records of decisions made during protocol development, inclusion screening and data extraction, are unique and should be kept. Minutes of meetings, correspondence as well as peer review comments and responses might also be held for a specific period of time as further records of the decision-making process. It is always advisable to permanently store a copy of the final report, particularly if the only other copy in existence is the one submitted to the commissioners.

Some information used in the review such as conference abstracts, additional information from authors, and unpublished material may be particularly difficult to obtain at a later stage so hard copies should be archived. This also applies to material retrieved from the Internet, which should be printed for the archive, as links to web pages are not permanent.

Whilst it may be easy and space saving to archive material electronically, paper records are often preferable as the equipment used to access documents stored in electronic formats can become obsolete after a relatively short period of time.

2.2.8 Disseminating the findings of systematic reviews

In recent years, there has been substantial investment in the commissioning of systematic reviews assessing the effects of a range of different health care interventions. To improve the quality of health care, and ultimately health outcomes, the review findings need to be effectively communicated to practitioners and policy- makers. The transfer of knowledge obtained through research into practice has long been acknowledged as a complex process234-238 that is highly dependent on context and the interaction of a multitude of interconnected factors operating at the level of the individual, group, organisation and wider health system.

A number of conceptual frameworks have attempted to represent the complexity of knowledge translation processes.234, 236, 238-244 One recent framework,244 whilst recognising the importance of non-linear diffusion, highlights a pivotal role for the direct or planned dissemination of contextualised, actionable messages derived from systematic reviews to inform practice and policy decision-making processes.

2.2.8.1 What is dissemination?

As interest in enhancing the impact of health research has increased, so too has the terminology used to describe the approaches employed.241, 245 Terms like dissemination, diffusion, implementation, knowledge transfer, knowledge mobilisation, linkage and exchange and research into practice are all being used to describe overlapping and interrelated concepts and practices. Given this, it is helpful to explain how the term dissemination is used here.

Dissemination is a planned and active process that seeks to ensure that those who need to know about a piece of research get to know about it and can make sense of the findings. As such it involves more than making research accessible through the traditional mediums of academic journals and conference presentations. It requires forethought about the groups who need to know the answer to the question a review is addressing, the best way of getting the message directly to that audience, and doing so by design rather than chance. Hence an active rather than passive process.

The term dissemination is often used interchangeably with implementation but it is more appropriate to see the terms as complementary. Dissemination and implementation are

part of a continuum.239, 246, 247 At one end are activities that focus on making research accessible, raising awareness of new findings and encouraging consideration of practice alternatives and policy options. At the other end of the continuum are activities that seek to increase the adoption of research findings into practice and policy and that facilitate, reinforce and maintain changes in practice.

Dissemination should not be viewed as an adjunct to the review process or as something to be considered at the end when thoughts turn to publication. Nor should it be seen as separate from the wider social context in which the review findings are expected to be used. It is

an integral part of the review process and should be considered from an early stage to allow adequate time for planning and development, for the allocation of responsibilities and to ensure that the proposed activities are properly resourced.

2.2.8.2 A suggested approach to dissemination

Traditionally, research on dissemination and implementation has tended to focus on the use of research knowledge, rather than on the effects of dissemination activities.

However, several conceptual frameworks have been put forward which consistently suggest that the effectiveness of dissemination activities is determined by careful consideration of a number of key attributes.234, 237, 254-258 These are:

- The characteristics of the research message
- The setting in which the message is received
- The characteristics of the target audience(s)
- The source of the research message
- The presentation of the research message
- The communication channel(s) used

Assuming that all research has an audience (but not that all research should be widely disseminated), whether the message provides an unequivocal answer or simply highlights the need for further research, our approach is structured around six key attributes which are interlinked and difficult to consider in isolation. The key messages from the review are the starting point for determining the audience to be targeted.

Characteristics of the research message and the setting in which it will be received

The literature on communication259 and diffusion239 (i.e. how, why, and at what rate ideas/innovations spread through social systems) highlights three types of messages that can impact on the knowledge and attitudes of target audiences: awareness, instruction ('how to') and persuasion (information that reduces uncertainty about expected consequences). Message characteristics to consider include the nature of the intervention, the strength of the evidence, its transferability, the degree of uncertainty and whether the findings confirm or reject existing predispositions or practices.

Messages also have to be perceived as relevant and meaningful by the audiences being targeted. Knowledge about both the wider setting (economic, social, organisational and political environments) within which a target audience resides and the context (hostile or receptive) in which a message is to be received, should be used to inform the development of appropriate dissemination strategies.

Characteristics of the target audience(s)

Deciding who to target usually involves an element of prioritisation (segmentation) as resource constraints can make it difficult to reach all possible audiences. In prioritising, relevance (who needs to know about this research) and receptivity (who is most likely to be influenced and to influence others) need to be considered. The question of how best to reach target audiences can in part be answered by drawing upon the theoretical literature on research utilization (the ways by which different audiences become aware of, access, read and make use of research findings).260, 261 This literature helps to inform the selection of the most appropriate or feasible communication channels for the audiences being targeted. Channels frequently used to promote review findings include paper and electronic publishing, email alerting services, direct and relationship marketing, mass media campaigns as well as engaging directly with target audiences.

Presentation of the research message and communication channel(s) used The literature on diffusion239 makes a distinction between mass media channels and interpersonal (face to face) channels. The former are generally regarded as being more important for dissemination purposes whereas interpersonal channels are more important for activity at the implementation end of the continuum. A combination of communication channels is helpful in increasing the likelihood that target audiences will encounter the review messages being promoted.

The selection of communication channels may also inform the presentation (tailoring) of the research message itself. When tailoring messages, consideration is given to the target audience, language used, the format, structure and style of presentation, the types of appeal and the amount of repetition. It is generally appropriate to try to write for an educated but non-research specialist health professional or decision-maker. Lay terms are used rather than technical language and statistics presented in as simple

a form as possible. The aim is to make information accessible to a broad range of readers and anyone who would like more details can access the full report. It has been advocated that a layered structure such as the '1:3:25' format (i.e. one page of the research 'bottom lines' or 'actionable messages', three-page executive summary and a 25-page report) is the optimal way to present research evidence to health service managers and policy-makers.213 This type of structuring involving a front page of key messages has become common place and reflects documented audience preferences for the 'bottom line' up front. There is some evidence that this order of presentation can increase overall understanding of the research findings but may also in some instances alienate those who are less receptive to or in disagreement with the conclusions presented.262, 263

Dissemination strategies

It has been proposed that there are four dissemination models that can be employed to link 'research to action'.262, 263 These are:

- Push strategies which are largely associated with supply (researcher) led distribution of new research findings

- Pull strategies which facilitate demand (audience) led access to research
- Linkage and exchange264 efforts which involve two-way communications and partnerships between producers and users of research
- Integrated approaches that incorporate aspects of all three

In reality, push, pull and exchange strategies are not mutually exclusive; facilitating user pull often requires the application of a promotional push strategy (e.g. utilizing email alerting services or RSS feeds) to inform and remind target audiences about review findings that are forthcoming or have been made available online for example. The integrated approach that incorporates elements of all three strategies is recommended, but where the emphasis shifts according to the topic and the audiences to be targeted.

Evaluation of impact

There is an increasing requirement, particularly from funders, for the impact of research to be predicted in advance of the work and then assessed after completion.265, 266 There are a number of specialised research impact assessment approaches, but these usually require specialist skills and additional resources.267, 268 Taking the issue of whether academic quality or practical use and impact of research is most important, a pragmatic framework has been proposed which addresses both points.269 The framework is based on the assessment criteria used in UK universities. It provides a structure for a narrative description of the impact of the findings from why the research question was first posed and funded, to where the results were sent, discussed, and put into policy and/or practice.

3 Systematic reviews of clinical tests

Clinical tests are routinely used for diagnosis, confirming or excluding the presence of a disease or condition (such as pregnancy). They are also used to monitor disease progression, assess prognosis, and screen asymptomatic populations for disease. Any process that yields information used to inform patient management can be regarded as a clinical test.1 This includes a wide range of processes from history taking and physical examination to complex imaging techniques. The test itself is an intervention and forms part of the continuum of patient care. New tests are adopted into clinical practice for a number of reasons, including replacement of an existing test (where the new test is expected to reduce the negative impact on the patient, provide better information, or equivalent information for less cost), triage (to decide whether a more expensive or invasive test is necessary), or as an addition to the existing testing protocol.

The ultimate aim of any research on clinical tests should be to determine impact upon patient management and outcome. An RCT comparing the effect of different diagnostic strategies on one or more clinical outcomes could be considered ideal, as it provides direct information on the benefit to patients and can be modified to address various types of diagnostic question.2 However, RCTs may not be appropriate for addressing all diagnostic questions3, 4 and to date much of the research on diagnostic tests is in the form of test accuracy studies. The basic aim of test accuracy studies is to assess how well a test can distinguish between people with and without the disease/condition of interest. The outcome measures used describe the probabilistic relationships between positive and negative test results, and the presence or absence of disease, as compared with the best currently available method (i.e. the clinical reference standard). As such, test accuracy studies do not directly measure the relative benefits and harms to patients of testing. Evidence on the accuracy of a test, combined with evidence of a prognostic link between the target condition and preventable morbidity/mortality, may be considered indicative of the likely effectiveness of the test.5 Where a new test is being evaluated, evidence for a prognostic link between the target disease/condition and long-term morbidity or mortality should be available as should an effective intervention. However, this is not always the case as tests can be established in clinical practice with limited supporting evidence.

When considering a systematic review of test accuracy studies, it is important to assess whether review findings will be able to provide the information necessary to inform clinical practice. Any review of test accuracy is likely to be of limited value where evidence is lacking that the disease/condition is associated with long-term morbidity or mortality, or where no effective intervention is available. This is illustrated by the following examples:

- Magnetic Resonance Angiography (MRA) versus intra-arterial Digital Subtraction Angiography (DSA) for the detection of carotid artery stenosis.6 There is evidence from RCTs that carotid endarterectomy is an effective treatment for symptomatic carotid artery stenosis at thresholds defined by DSA. MRA is a less invasive test option. A review of test accuracy is therefore likely to be informative.
- Ultrasound versus Micturating Cystourethrography (MCUG) for the detection of vesicoureteric reflux (VUR) in children with urinary tract infection (UTI).7 There is conflicting evidence of a link between VUR and long-term renal damage and the effectiveness of treatment options, such as prophylactic antibiotics, is also uncertain. A review of test accuracy alone is therefore unlikely to be informative.

Although some study designs, such as those based upon multivariable prediction modelling, may better reflect the true nature of the diagnostic workup and are potentially more informative than test accuracy studies,8, 9 they are rare. Consequently, systematic review methods for assessing clinical tests have largely focused upon test accuracy studies and this chapter discusses methods developed specifically to deal with such studies. Section 2.2 focuses on diagnostic accuracy studies, but the methods described also apply to test accuracy studies used to assess the performance of new screening tests, within established screening programmes. The clinical effectiveness

of screening programmes is best evaluated using RCTs and systematic reviews of such studies should follow the principles outlined in Chapter 1. Section 2.3 describes methods for reviewing prognostic studies.

In light of the limitations described in relation to test accuracy studies, careful consideration should always be given to the likely informative value and any additional data requirements before undertaking a systematic review of test accuracy.

3.1 Diagnostic tests

3.1.1 The review question

As with all systematic reviews, the development of a clear, well-defined question is essential to maintaining transparency of the review process and to the quality and relevance of the findings. Some aspects of the question require consideration when planning a review of test accuracy.

3.1.1.1 Population

Diagnostic tests perform differently in different populations,10, 11 for example it would generally be inappropriate to evaluate the performance of a test in a secondary care population when

the test is mainly used in primary care. Both frequency and severity of the target condition would be expected to be greater in secondary care. It is therefore important to clearly define the population of interest. The ideal study sample for a test accuracy study is a consecutive or randomly selected series of patients in whom the target condition is suspected, or for screening studies, the target population. Because participant sampling methods are often poorly reported in test accuracy studies,12 using the sampling method as an inclusion/exclusion criterion is likely to result in a substantial reduction in available data. It is likely to be more useful to consider the sampling method and/or its reporting as an aspect of study quality (see Section 2.2.5 Quality assessment) and to base the inclusion criteria relating to the population upon participant characteristics. For example in a review comparing the accuracy of different imaging techniques, the inclusion criteria might state that only patients with a specified level of symptoms, representative of those in whom the test would be used for intervention planning, are eligible.

3.1.1.2 Intervention (index test)

In reviews of test accuracy the 'index test' (the test whose performance is being evaluated) can be viewed as the intervention. As with any review, the scope of the question can be broad such as 'what is the optimum testing pathway for the diagnosis and follow-up investigation of childhood urinary tract infection (UTI)?'13 or it can be narrow; for example 'what is the diagnostic accuracy of magnetic resonance angiography (MRA) when compared with intra-arterial x-ray angiography, for the detection of carotid artery stenosis?'6 The former is likely to include a number of different technologies, addressing multiple target conditions, whereas the latter compares the performance of an alternative (replacement), less invasive or less costly diagnostic technology with that of the reference standard for the detection of a specified target condition. The rate of technological development may be an important consideration; in this latter example inclusion of MRA techniques that are already obsolete in clinical practice, is unlikely to be useful.

Careful consideration should always be given to the equivalence of different analytical techniques when setting inclusion criteria. For example, a systematic review of faecal occult blood tests to screen for colorectal cancer14, 15 evaluated both immunochemical and colourimetric methods for detecting blood in the faeces; though both methods target blood, they cannot be considered equivalent tests.

The traditional concept of test accuracy often implies the dichotomisation of data into test results which are classified as positive (target condition present) or negative (target condition absent). Any systematic review of test accuracy will therefore need to consider diagnostic thresholds (points at which results are classified as positive or negative) for each included index test.

3.1.1.3 Reference standard/comparator

The reference standard is usually the best test currently available, and is the standard against which the index test is compared. It need not be the test used routinely in practice (although it can be), and may include information which is not known for some time after the tests have been done (e.g. follow-up of test negatives in cancer).

The test accuracy study is based upon a one-sided comparison between the results of the index test and those of the reference standard. Any discrepancy is assumed to arise from error in the index test. Selection of the reference standard is therefore critical to the validity of a test accuracy study and the definition of the diagnostic threshold forms part of that reference standard.

It is important to note that the assumption of 100% accuracy for the reference standard rarely holds true in practice. This represents a fundamental flaw in the test accuracy study design, since the index test can never be deemed to perform better than the reference standard, and its value may therefore be underestimated.16

Where several tests are available to diagnose the target condition, there is often no consensus about which test constitutes the reference standard. In such cases a

composite reference standard, which combines the results of several available tests to produce a better indicator of true disease status may be used.17 A number of statistical methods have been proposed to estimate the performance of tests in the absence of a single accepted reference standard.18, 19

There may be instances when it is deemed unethical to use an invasive procedure as a reference standard in a study.20 In such cases, clinical follow-up and final diagnosis

may sometimes be used as a surrogate reference standard. There will also be occasions when clinical follow-up and final diagnosis provides the most appropriate reference standard. The length of follow-up should ideally be defined in advance. Studies using follow-up and clinical outcome in this way may be viewed as prognostic studies in that they are measuring the accuracy with which the test is able to predict a future event, rather than the accuracy with which it is able to determine current status. Where such studies are included in a systematic review, it is important to define, in advance, what constitutes appropriate follow-up and hence an adequate reference standard.

The comparator is an alternative test, usually that which is used in current practice, against which the index test must be evaluated in order to assess its potential role. Ideally, this should be done by comparing index test and comparator to the reference standard in the same population.

3.1.1.4 Outcome measures

The primary outcome of interest for any systematic review of test accuracy is the data required to populate 2 x 2 contingency tables. These describe the relationship between the results of the index test and the reference standard at a given diagnostic threshold (point at which results are classified as positive or negative). The table includes the number of true positives (TP: those that have the disease and test positive), false positives (FP: those that do not have the disease and test positive), false negatives (FN: those that do have the disease and test negative) and true negatives (TN: those that do not have the disease and test negative).

		Reference standard	
		Disease	No disease
Index test	Positive	TP	FP
	Negative	FN	TN

From the 2 x 2 contingency table, the following commonly used measures of test performance can be calculated:

Sensitivity =vThe proportion of people with the target condition who have a positive test result.

Specificity =vThe proportion of people without the target condition who have a negative test result.

Overall accuracy =vThe proportion of people correctly classified by the test.

Positive predictive value =vThe probability of disease among persons with a positive test result.

Negative predictive value =vThe probability of non-disease among persons with a negative test result.

Positive likelihood ratio and Negative likelihood ratio.

Likelihood ratios (LR) describe how many times more likely it is that a person with the target condition will receive a particular test result than a person without. Positive likelihood ratios greater than 10 or negative likelihood ratios less than 0.1 are sometimes judged to provide convincing diagnostic evidence.21

Diagnostic odds ratio = Used as an overall indicator of diagnostic performance and calculated as the odds of a positive test result among those with the target condition, divided by the odds of a positive test result among those without the condition.

In primary studies, a receiver operating characteristic (ROC) curve describes the relationship between 'true positive fraction' (sensitivity) and 'false positive fraction' (1– specificity) for

different positivity thresholds. It is used to display the trade-offs between sensitivity and specificity as a result of varying the diagnostic threshold.

Below is an example ROC analysis for serum thyroid stimulating hormone (TSH) as a diagnostic test for primary hypothyroidism:

Test results (Serum TSH) vs. reference standard (thyroid status)

Serum TSH (mIU/L)	Number with primary hypothyroidism	Number without primary hypothyroidism
<6	17	325
6-12	42	158
13-15	46	48
16-20	66	33
>20	284	5

2 x 2 contingency data for serum TSH diagnostic threshold (derived by summing the numbers of participants, with and without primary hypothyroidism, on either side of the diagnostic threshold)

Diagnostic threshold for a positive test result (mIU/L)	TP	FP	FN	TN
6	438	244	17	325
>12	396	86	59	483
>15	350	38	105	531
>20	284	5	171	564

Sensitivity and specificity values for each diagnostic threshold (derived from the 2 x 2 contingency data and expressed as percentages)

Diagnostic threshold for a positive test result (mIU/L)	Sensitivity	Specificity
6	96.2%	57.1%
>12	87.0%	84.9%
>15	76.9%	93.3%
>20	62.4%	99.1%

'Q*', or maximal joint sensitivity and specificity, is the point on the ROC curve that intersects with the line of symmetry. It is sometimes used as an indicator of overall test performance where there is no clinical preference for maximising either sensitivity (minimizing false negatives) or specificity (minimizing false positives). However Q* is not useful if the thresholds at which tests have been evaluated do not lie close to the line of symmetry and can then give misleading results if used to compare performance between tests.

In some scenarios (e.g. tests used in population screening) a threshold which skews diagnostic performance may be preferable (e.g. minimizing the number of false negatives at the expense of some increase in the number of false positive results, in conditions/diseases where missing the presence of disease will lead to serious consequences). Overall diagnostic accuracy is summarised by the area under the curve (AUC); the closer the curve is to the upper left hand corner the better the diagnostic performance.22 The AUC ranges from 0 to 1, with 0.5 indicating a poor test where the accuracy is equivalent to chance.

As with other types of intervention, when assessing the clinical effectiveness of a diagnostic test, it is important to consider all outcome measures which may be relevant to the use of the test in practice. These might include adverse events (see Chapter 4) and the preferences of patients, although inclusion of such information is rare.

3.1.1.4.0.1 Study design

There are two basic types of test accuracy study: 'single-gate' which are similar to consecutive series (and previously sometimes called diagnostic cohort studies) and 'two-gate' which are similar to case-control studies. The term 'two-gate' being used

where two sets of inclusion criteria or 'gates' are applied, one for participants who have the target condition and one for those who do not. These designs differ in structure from other cohort and case-control studies in that both are generally cross-sectional in nature.23

- The single-gate design includes participants in whom the disease status is unknown, and compares the results of the index test with those of the reference standard used to confirm diagnosis, i.e. it is broadly representative of the scenario in which the test would be used in practice.
- The two-gate design compares the results of the index test in patients with an established diagnosis of the target condition with its results in healthy controls or controls with another diagnosis (known status, with respect to the target condition, is therefore treated as the reference standard); i.e. it is

  unrepresentative of practice and is unlikely to contain the full spectrum of health and disease over which the test would be used.

There are inherent problems with the two-gate design that may lead to bias. The selective inclusion of cases with more advanced disease is likely to lead to over estimations of sensitivity and inclusion of healthy controls is likely to lead to over estimations of specificity. The recruitment of healthy controls from the general population has been associated with two- to three-fold increases in measures of test performance time-to-events derived from a diagnostic cohort design.11, 24, 25 This over estimation can be increased further when cases of severe disease are used alongside healthy controls.26 By contrast, where cases are derived from individuals with mild disease, underestimations of sensitivity can result.27 Where the control group is derived from patients with alternative diagnoses, specificity may be under or over-estimated, depending upon the alternative diagnosis.23 In theory, the two-gate study design could produce a valid estimate of test performance if the cases were sampled to match the reference standard positive patients in a single-gate study (in terms of the spectrum of disease severity) and controls were matched to the reference standard negative patients (in terms of the spectrum of alternative conditions). In practice however, this is difficult to achieve.23 Whilst two-gate studies are therefore of limited use in assessing how a test is likely to perform in clinical practice, they can be useful in the earlier phases of test development.28

Where systematic reviews include both single and two-gate study designs, careful consideration should be given to the methods of analysis and the impact of study design should be assessed in any meta-analyses.29

3.1.2 Identifying research evidence

3.1.2.1 Sources

The importance of searching a wide range of databases to avoid missing relevant diagnostic test accuracy studies has been demonstrated, with MEDLINE, EMBASE, BIOSIS, LILACS, Pascal and Science Citation Index all providing unique records.30 The reference lists of included studies can also be a useful resource.

The Cochrane Diagnostic Test Accuracy Working Group31 is creating a database of test accuracy studies,32 similar to the non-topic specific Cochrane Central Register of Controlled Trials (CENTRAL) which includes details of published articles taken from bibliographic databases and other published and unpublished sources.33

3.1.2.2 Database searching

Many electronic databases do not have appropriate indexing terms to label test accuracy studies, and those that do tend not to apply them consistently.30, 34-36 They also vary in their design which adds to the difficulty in identification.34 The problem is compounded by the fact that the original authors are often poor at identifying their studies as being test accuracy.30

It has been reported that the use of filters to identify reports of diagnostic test accuracy studies in electronic databases may miss a considerable number of relevant articles and is therefore not generally considered appropriate.34, 36, 37 Database searching should concentrate on terms for index tests and target conditions. If further restriction is required, it can be achieved by means of topic specific terms, rather than using a filter.36, 38 It is hoped, however, that in time, as the issues of reporting and indexing diagnostic, screening and prognostic studies are more widely realised, the situation will improve allowing the development of more accurate filters.

3.1.2.3 Publication bias

As the data used in studies of test accuracy are often collected as part of routine clinical practice (and in the past have tended not to require formal registration) it has been argued that test accuracy studies are more easily conducted and abandoned than RCTs. They may therefore be particularly susceptible to publication bias.39 Simulation studies have, however, indicated that the effect of publication bias on meta-analytic estimates of the Diagnostic Odds Ratio (DOR) is not likely to be large.40

It has been demonstrated that the unique features of the test accuracy study make the application of the Begg, Egger, and Macaskill tests of funnel plot asymmetry potentially misleading.40 An alternative approach uses funnel plots of (natural logarithm (ln) DOR) vs. ($1/\sqrt{}$effective sample size) and tests for asymmetry using related regression or rank correlation tests.40 It should be noted that the power of all statistical tests for funnel plot asymmetry decreases with increasing heterogeneity of DOR. It should also be noted that factors other than publication bias, for example aspects of study quality and population characteristics, may be associated with sample size.

Given the limitations of current knowledge, to ignore the possibility of publication bias would seem unwise, however, its assessment in reviews of test accuracy is complex.

3.1.3 Data extraction

The same precautions against reviewer bias and error should be employed whilst extracting data from test accuracy studies as would be applied in any other type of review. Independent checking of 2x2 data is particularly important, as test accuracy studies are often poorly reported,12, 41 and the production of a 2x2 table from these studies can be far from straightforward.

Some studies may provide the actual results for each test for individual patients. In this case the researcher may need to classify each patient according to the diagnostic thresholds defined in the review protocol.

Studies may provide categorical data, which may represent multiple categories or stages of disease. In this case data will need to be extracted for the numbers of index test positive and negative participants (using the threshold(s) defined in the review protocol, which may include all thresholds reported) with and without the target condition (as defined by the reference standard, using the threshold(s) defined in the review protocol).

There may be instances when the raw data are not reported, but 2x2 data can be calculated from reported accuracy measures and total numbers of diseased or non- diseased patients.

Somewhat more problematic are cases when the data do not 'fit' the 2x2 contingency table model. 'Forcing' data into a 2x2 contingency table, for example by classifying uncertain index test results as FP or FN, may be inappropriate. The contingency table can be extended to form a six cell table, which accommodates uncertain or indeterminate index test results.

The informative value of an indeterminate test result can be assessed using an indeterminate likelihood ratio (or LR+/-), defined as the probability of an indeterminate test result in the presence of disease divided by the probability of an indeterminate test result in the absence of disease.

When index test and reference standard give clear results (ie considered determinate), but there is incomplete concordance, the 2x2 table may be expanded to accommodate a more complete clinical picture.

3.1.4 Risk of bias assessment

Structured appraisal of methodological quality is key to assessing the reliability of test accuracy studies included in a systematic review.44 Quality assessment should consider the association of individual elements of methodological quality with test accuracy; generating overall 'quality scores' is not recommended.45

There are many differences in the design and conduct of diagnostic accuracy studies that can affect the interpretation of their results. Some differences lead to systematic bias such that estimates of diagnostic performance will differ from their true values,

others give rise to variation in results between studies, which can limit applicability. The distinction between bias and variation is not always clear, and quality assessment checklists have tended to include items that are pertinent to both.46, 47 Sources of

variation and bias that are potentially relevant when considering studies of test accuracy are described in Table 2.1. Whilst it is clear that variation (e.g. in the demographic characteristics or severity of disease in the study population) can affect the applicability of the results of both individual studies and systematic reviews, there is limited evidence on the effects of design-related biases in primary studies on the results of systematic reviews.11, 24, 26, 48 Research on the impact of design-related biases is largely a work in progress, being dependent upon the availability of adequate data sets and consistent methods of quality assessment.

Guidelines for assessing the methodological quality of test accuracy studies were first developed in the 1980s.16, 46 A large number of quality assessment tools and checklists have since been published, often as part of individual systematic reviews. Methodological work has identified 67 tools designed to assess the quality of test accuracy studies and 24 guides to the interpretation, conduct or reporting of test

accuracy studies.49 Only six of the quality assessment tools specified which aspects of quality they aimed to cover.50-55 One quality assessment tool46 and one guide to the reporting of diagnostic accuracy studies56 provided detailed information of how items had been selected for inclusion in the tool, and none reported systematic evaluation of the tool.

QUADAS was the first attempt to develop an evidence-based, validated, quality assessment tool specifically for use in systematic reviews of test accuracy studies.47 The items included in QUADAS were derived by combining empirical evidence from three systematic reviews, reported in two publications11, 49 with expert opinion, using a formal consensus method.47 The QUADAS criteria and the sources of bias and variation

to which they relate are given in Table 2.2. Each item is scored as 'Yes', 'No' or 'Unclear' and generic guidance on scoring has been published.47, 57 It is, however, impossible

to provide a universally applicable description of how some QUADAS items should be scored, e.g. the definition of 'an appropriate patient spectrum', or 'a reference standard likely to correctly classify the target condition.' It is therefore important that guidance on scoring be refined for individual reviews, with the definition of what should be scored as 'Yes', 'No' and 'Unclear' being specified for each QUADAS item and agreed by the whole review team at the start of the review; this should be done in close consultation with clinical experts.57 Piloting of the quality assessment process on a small sample

of included studies should be done in an attempt to eliminate any discrepancies in understanding between reviewers.

Table 2.1: Sources of bias and variation in test accuracy studies11

Population

Demographic characteristics	Variation	Test may perform differently in different populations.
Disease severity	Variation	Differences in disease severity may lead to different estimates of diagnostic performance.
Disease prevalence	Variation	The prevalence of the target condition varies with the setting and may affect estimates of diagnostic performance. In
	Bias	settings of higher prevalence, interpreters are more prone to classify test results as abnormal (context bias).

Demographic characteristics	Variation	Test may perform differently in different populations.
Participant selection	Variation	A selection process that may not include a spectrum of patients similar to that in which the test will be used in practice may limit the applicability of study findings.

Test methods

Test execution	Variation	Differences in the execution of the index test and/or reference standard can result in different estimates of diagnostic performance; clear reporting of the methods used is therefore important.
Technological	Variation development	Diagnostic performance of tests can change over time due to technological improvements.
Treatment paradox	Bias	Occurs when treatment is started, based upon the results of one test prior to undertaking the other; thus disease state is potentially altered between tests.
Disease progression	Bias	Occurs when there is sufficient time delay between the application of the index test and the reference standard to allow change in the disease state.

Application of the reference standard

Use of an inappropriate reference standard	Bias	The error in diagnoses derived from an imperfect reference standard can result in underestimation of the performance of the index test.
Differential verification	Bias	Occurs when the diagnosis is verified using different reference standards, depending upon the result of the index test.
Partial verification	Bias	Occurs where only a selected sample of participants undergoing the index test also receive the reference standard.

Test or diagnostic review	Bias	Where interpretation of either the index test or reference standard may be influenced by knowledge of the results of the other test. Diagnostic review bias may be present when the results of the index test are known to those interpreting the reference standard. Test review bias may be present when the results of the reference standard are known to those interpreting the index test.
Clinical review	Bias	The availability of other relevant clinical information (e.g. symptoms, co-morbidities) may also affect estimates of test performance.
Incorporation	Bias	Occurs when the result of the index test is used in establishing the final diagnosis (i.e. it forms part of the reference standard).
Observer	Variation	The interpretation placed upon a test result may vary between observers and this can affect estimates of test accuracy. The reproducibility of a test within (intra) and between (inter) observers affects its applicability in practice.

Analysis

Handling of uninterpretable results	Bias	Diagnostic tests fail or produce un-interpretable results with varying frequency. Study participants for whom a test result could not be obtained are often removed from reported analyses. This may lead to a biased assessment of test performance.
Arbitrary choice of threshold value (the diagnostic threshold is derived from the same data set in which test performance is evaluated)	Variation	The choice of a threshold value based upon that which maximises sensitivity and specificity for the study data may result in exaggerated estimates of test performance. The test may perform less well at the chosen threshold when evaluated in a new independent patient set.

QUADAS is a generic tool, which may be adapted to optimize its usefulness for specific topic areas. Researchers should, therefore, also consider in advance whether all QUADAS items are

relevant to their topic area, and whether there are any additional items that are not included in QUADAS.57 For example, disease progression bias may not be a relevant issue where the clinical course of the target condition is slow; when comparing the performance of imaging tests, or other tests which require subjective interpretation by the operator, the impact of observer variation may need to be considered as variation in test performance with individual operators of the same test (e.g. different individuals conducting and/or interpreting an ultrasound examination) can exceed, and therefore mask, a difference in performance between two different tests (e.g. ultrasound and magnetic resonance imaging).58, 59

Table 3.10: The QUADAS items

QUADAS criterion	Bias/variation assessed
Was the spectrum of patients representative of the patients who will receive the test in practice?	Population characteristics (demographic, severity and prevalence of disease)
Were the selection criteria clearly described?	Participant selection
Is the reference standard likely to correctly classify the target condition?	Use of an inappropriate reference standard
Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?	Disease progression
Did the whole sample or random selection of the sample receive verification using a reference standard of diagnosis?	Partial verification
Did the patients receive that same reference standard regardless of the index test results?	Differential verification
Was the reference standard independent of the index test?	Incorporation
Was the execution of the index test described in sufficient detail to permit replication?	Test execution
Was the execution of the reference standard described in sufficient detail to permit replication?	
Were the index test results interpreted without knowledge of the results of the reference standard?	Test review
Were the reference standard results interpreted without knowledge of the results of the index test?	Diagnostic review
Were the same clinical data available when the test results were interpreted as would be available when the test is used in practice?	Clinical review

QUADAS criterion	Bias/variation assessed
Were un-interpretable/intermediate test results reported?	Handling of un-interpretable or missing results
Were withdrawals from the study explained?	

It is worth noting that the information that can be derived from the quality assessment of test accuracy studies is often limited by poor reporting. Where QUADAS items are scored 'unclear' the researcher cannot be certain whether this indicates poor methods with the attendant consequences for bias/variation, or simply poor reporting of a methodologically sound study. The STARD initiative60 has proposed standards for the reporting of diagnostic accuracy studies. If these standards are widely adopted and lead to a general improvement in the reporting of test accuracy studies, reviewers will increasingly be able to assess methodological quality rather than the quality of reporting.

3.1.5 Synthesis

A thorough investigation of heterogeneity should be undertaken before deciding if studies are suitable for combining in a meta-analysis and if so what method to use. Clinical and methodological differences such as patient populations, tests, study design and study conduct, should be considered in addition to statistical variation in the accuracy measures reported by studies. Where a meta-analysis is not considered clinically or statistically meaningful, a structured narrative synthesis can be carried out which can include the presentation of results in one or more graphical formats.61 For example the results of individual studies can be plotted in ROC space, as in Figure 2.4, whether or not a summary curve is included. As well as stratification by index test characteristics, reviews which focus on determining the optimal diagnostic pathway for a condition, rather than the diagnostic performance of a single test, should consider structuring narrative reports to represent the order in which tests would be applied in clinical practice. Reviews which consider differential diagnosis from a common presenting symptom, such as a review of the performance tests to determine the cause of haematuria, should consider stratifying the narrative by target condition with the most common diagnosis addressed first. These approaches aim to increase readability for practitioners and can equally be applied to the structure of reports which include meta-analyses.

Assessment of statistical heterogeneity

Threshold effect

A source of heterogeneity unique to test accuracy studies, which requires careful assessment, arises from the choice of the threshold used to define a positive result.62 Even when different thresholds are not explicitly defined, variation in interpretation by observers may result in

implicit variation in threshold. This can be assessed visually using a ROC space plot and statistically by measuring the correlation between sensitivity and specificity. However, statistical tests may be unreliable where studies in a systematic review have small sample sizes; threshold effect may be present but undetected by statistical tests. A ROC space plot is a plot of the 'true positive rate' (sensitivity) from each study against the 'false positive rate' (1 - specificity). If a threshold effect exists then the plot will show a curve (as the threshold decreases the sensitivity will increase and the specificity will decrease). This curve follows the operating characteristics of the test at varying thresholds.

Figure 2.4 clearly shows a curve in the top left hand corner of the plot, indicating the presence of a threshold effect. The presence of a threshold effect can also be investigated using a regression62 or a hierarchical summary ROC (HSROC) model63 which are described in more detail in the meta-analysis section below.

Heterogeneity of individual diagnostic accuracy measures

Variability amongst each of the individual measurements (sensitivity, specificity, positive and negative likelihood ratio, and DOR) can be assessed using the same methods as for other study types. Forest plots can be used to visually assess differences between studies, although these will not show any threshold effects. Paired forest plots should be used when illustrating paired outcome measures such as sensitivity and specificity. Use of statistical tests of heterogeneity does not reliably indicate absence of heterogeneity and it is generally advisable to assume the presence of heterogeneity and to fit models which aim to describe and account for it.

3.1.5.1 Meta-analysis

The meta-analysis of diagnostic accuracy studies requires the use of some specific statistical methods which differ from standard methods. Meta-analysis has two main aims: to obtain a pooled measure of diagnostic accuracy and in the case of summary ROC (SROC) models, to explore the heterogeneity amongst studies. Diagnostic accuracy is usually represented by a pair of related measurements, for example: sensitivity and specificity; positive and negative likelihood ratio; and this relationship needs to be incorporated into the analysis methods.

Pooling individual diagnostic accuracy measures

A robust approach to combining data and estimating the underlying relationship between sensitivity and specificity is the construction of an SROC curve. Methods that involve pooling sensitivities and specificities from individual studies, or combining positive and negative likelihood ratios fail to account for the paired nature of the parameters, and should generally be avoided. However, where only one parameter (e.g. sensitivity, but not specificity) is presented, simple pooling of proportions is the only option. Assessment of single parameters is usually inappropriate, but is sometimes used when there is a specific clinical reason why only one parameter should be the focus of interest.

Diagnostic odds ratios can be pooled using standard fixed or random-effects methods for pooling odds ratios. However, these methods do not help estimate average sensitivity and specificity and may produce erroneous results where there is a relationship between DOR and threshold.64

Predictive values should not be pooled in meta-analyses as they are affected by the prevalence of disease in the populations of the studies. Overall predictive values are sometimes calculated using estimates of prevalence from the included studies and pooled estimates of likelihood ratios. However, the potentially misleading nature of such estimates should be considered carefully.

Simple methods of estimating summary ROC curves

The Moses-Littenburg regression based method,62 has been used as a simple method of pooling study results in the presence of a suspected threshold effect. It can be used in preliminary exploratory analyses and is helpful in understanding the data.65 However, it has limitations and should not be used to obtain summary estimates of sensitivity and specificity. The usual regression model assumptions are not met.66, 67 It also assumes that there is only one result per study and so cannot deal adequately with studies which have multiple data sets per test (e.g. data for a number of different thresholds).

It is possible to pool ROC curves, or the AUC from individual studies although this is not recommended and would not be practical in the case where some studies reported data for a single threshold and others presented data (or a ROC curve) for a number of thresholds.21

Optimal methods of modelling SROC curves

Statistical models, including hierarchical and bivariate models, have been developed for the estimation of SROC curves in the meta-analysis of test accuracy results. The HSROC model63 accounts for both within- and between-study variation in true positive and false positive rates. The model estimates parameters for the threshold, log DOR and the shape of the underlying ROC curve. It has been shown that it is possible to fit this model using statistical package SAS, and that this method provides results that agree with the more complex Bayesian methods.68 The HSROC model can be extended to deal with studies that provide results for more than one threshold, but programming

is challenging. The bivariate model67 analyses sensitivity and specificity jointly, therefore retaining the paired nature of the original data (a STATA command function has recently been produced for the bivariate model). The HSROC and bivariate models have been

shown to produce equivalent results in the absence of other study-level covariates.69 It is recommended that meta-analyses using these models should be undertaken with the assistance of a statistician.

Exploring heterogeneity

Sources of methodological and/or clinical heterogeneity can be explored using subgroup analyses. Ideally subgroups should be planned at the protocol stage. However, where this is dependent upon what data are available, and an adaptive process is needed,

this should be stated clearly in the protocol. Results from different groups, for example different tests, or study designs, can be visually assessed by using a ROC space plot with different symbols. Figure 2.5 illustrates the divergent accuracy results between different study designs from a systematic review of faecal occult blood tests used in screening for colorectal cancer,15 which indicates that two-gate studies (white circles) overestimate test performance compared with single-gate studies (black circles).

HSROC and bivariate models can be used to assess heterogeneity by including covariates. These models allow investigation of the effect of covariates on sensitivity and specificity separately, rather than just the DOR (although this can still be obtained). Further research is needed to determine which SROC models are the most appropriate for the exploration of heterogeneity as the choice of model may depend on which accuracy measure (DOR, sensitivity, specificity) is most affected.69 An overview of the different methods used to explore heterogeneity in systematic reviews of diagnostic test accuracy is available.70 It should be noted that, as for meta-regression analyses of other study designs, these analyses are exploratory, can only include covariates reported by

the studies and should not be conducted if there are only a small number of studies (a minimum of 10 studies per covariate is needed). Regardless of the approach used, study-level factors to be examined should be defined in the protocol and aspects of methodological quality, (e.g. QUADAS items) should be considered individually, rather than as overall quality scores.45, 48

3.1.5.2 Software

Methods for calculating outcome measures, assessing heterogeneity, producing plots (both with and without summary estimates) and undertaking exploratory analyses using the Moses model are available in a user-friendly form in the Meta-DiSc software (www.hrc.es/investigacion/metadisc_en.html).71.71) Systematic reviews of diagnostic accuracy studies have been incorporated in version 5.0 of the Cochrane Review Manager software. More specialist statistical software packages, such as STATA, SAS or WINBUGS, are needed to fit HSROC/bivariate models and the support of a statistician with knowledge of the field is generally recommended.

3.1.6 Presentation of results

When presenting the results of a systematic review of clinical tests it is important to consider how these results will be understood by clinicians and applied in practice. The understanding of and preferences for measures of test performance by clinicians has been the subject of much

research and comment.72-74 The 'best' method remains elusive but some general points, which may improve clarity and aid interpretation, are given below.

The presentation of diagnostic measures should be similar for both narrative and meta- analytic approaches, with graphical representation and/or tabulation of individual study results and additional results presented if meta-analysis was performed. Sufficient detail of the tests, participants, study design and conduct should be presented in tables.75

The 2 x 2 table results of TP, FP, FN and TN together with sensitivity and specificity, as a minimum should be presented for each study. The choice of accuracy measures presented depends on the aims and anticipated users of the review. Sensitivity and

specificity and likelihood ratios are measures of test performance; likelihood ratios may be more useful in a clinical setting as they can be used to calculate the probability of disease given a particular test result, whereas DORs are difficult to interpret clinically.22 Forest plots or ROC space plots provide useful visual summaries and can be easier

to interpret than large tables of numbers. The ranges should be presented when summarising results which have not been subject to meta-analytic pooling. For paired results it may be useful to also present the corresponding measure for the studies at each end of the range, e.g. 'sensitivity ranged from 48% (at a specificity of 80%) to 92% (at a specificity of 70%)'.

If a meta-analysis was undertaken then the presentation of results depends on the methods used. If sensitivity or specificity have been pooled as individual measures then the summary estimate together with the 95% confidence intervals should be presented. If an SROC model has been used then the relevant SROC curve(s) should be presented. Where the performance of a number of index tests is being compared it may

be useful to present multiple SROC curves (or un-pooled data sets) on the same plot. Summary measures of overall diagnostic accuracy, such as AUC or the Q* point (the point on the curve where sensitivity and specificity are equal) may also be presented. However, the relevance of the Q* point is debatable, as its use may lead to summary estimates of sensitivity and specificity outside the values in the original studies.67 Pairs of sensitivity and specificity values can also be read from the SROC curve and presented as a number of summary points in order to provide an overall description of the curve. The estimated SROC curves should also be presented if HSROC or bivariate models have been used. These models enable the calculation of summary estimates of sensitivity

and specificity, which should be reported along with their 95% confidence intervals. Although the use of HSROC or bivariate models to generate summary likelihood ratios is not recommended,76 where likelihood ratios are considered helpful to interpretation, summary likelihood ratios can be calculated from the pooled estimates of sensitivity and specificity generated by these models. For results from a HSROC or bivariate model, as these retain the paired nature of sensitivity and specificity, a region can be plotted around the summary operating point which represents the 95% confidence intervals of both measures.67 Confidence interval regions can also be plotted for the results of individual studies, but care is required to ensure that

these are not mistakenly interpreted as representations of study weighting. Both models can also be used to plot a prediction region; this is the region which has a particular probability of including the true sensitivity and specificity of a future study.69

Summary: Diagnostic studies

- Researchers planning systematic reviews of test accuracy should give careful consideration to context (e.g. is there evidence of a prognostic link between the target condition and preventable morbidity/mortality).
- Diagnostic tests should be evaluated in patients who are representative of those in whom the test will be used in practice; ideally a consecutive or

  randomly selected series whose diagnosis is unknown at the time of testing.
- Careful consideration should be given to what is the appropriate reference standard to establish diagnosis.
- Difficulties in searching bibliographic databases for test accuracy studies and the lack of suitable methodological search filters mean that more specific searches carry a risk of missing studies. Searches based upon index test and target condition, which are designed to maximise sensitivity, are therefore recommended.
- Test accuracy studies are often poorly reported, hampering data extraction, quality assessment and synthesis.
- Though often unable to provide a definitive estimate of test accuracy, systematic reviews can highlight important gaps in the evidence base and aid in the design of future studies.

3.2 Prognostic tests

Prognostic markers (biomarkers) are characteristics that help to identify or categorise people with different risks of specific future outcomes. They may be simple clinical measures such as body mass index, but are more often pathological, biochemical, molecular or genetic measures or attributes. Identifying those who are or who are not at risk can facilitate intervention choice, and aid patient counselling.

Prognostic research has to date received much less attention than research into therapeutic or diagnostic areas, and an evidence-based approach to the design, conduct and reporting of primary studies of prognostic markers is needed.77 Reviews have shown that primary prognostic studies are often of poor quality.78

Synthesis of prognostic studies is a relatively new and evolving area in which the methods are less well developed than for reviews of therapeutic interventions or of diagnostic accuracy, and available reviews have often been of poor quality.79-82

Although numbers of completed prognostic reviews are relatively few,83 they are becoming more common. Of 294 reviews of prognostic studies published since 1966, almost all have appeared since 1996, occurring most commonly in cancer (15%), musculoskeletal disorders and rheumatology (13%), cardiology (10%), neurology (10%), and obstetrics (10%).79 Available reviews often include large numbers of studies and patients. For example, some reviews in cancer and cardiovascular disease have reported data on over 10,000 patients for a single marker.84-87

This section focuses mainly on reviews of studies of potential prognostic markers and builds on previous work.88 Given that this is a developing area where methods and approaches will undoubtedly change rapidly, this section presents a discussion rather than firm guidance. Systematic reviews of studies which develop a prognostic model (risk score) are not considered here.

3.2.1 Defining the review question: setting inclusion criteria

Defining the review question and setting inclusion criteria should be approached in the same way as set out in Chapter 1, Section 1.2 The review protocol. However, some aspects of methodology require particular attention when planning a systematic review of prognostic studies, and should be considered at an early stage.

3.2.1.1 Population/study design

Patients included in a prognostic study are usually selected as an 'inception' cohort of patients identified very early in the course of their disease, perhaps at diagnosis. Even if the cohort is identified retrospectively, it should be followed forwards in time from a particular point, such as diagnosis or (if relevant) randomisation. The case-control design is liable to bias.89 Careful thought as to what study designs will be included in the review is needed.

3.2.1.2 Intervention

Although often ignored in prognostic studies, if the intervention that patients receive varies on account of perceived prognosis, this precludes an unbiased assessment of the prognostic ability of a marker (unless alternative interventions are equally effective).

Although the intervention effect may be small compared to the effect of important prognostic variables and consequently will have little impact on findings, ideally, prognostic variables should be evaluated in a cohort of patients treated the same way, or that have been included in an RCT.90, 91 The intervention received is rarely reported in primary studies.

3.2.2 Defining the review question: other considerations

3.2.2.1 Publication bias and sample size

Evidence of publication and associated reporting biases is accumulating for prognostic studies.92, 93 For example, in a systematic review of studies of a marker Bcl2 in non- small cell lung cancer, almost all the smaller studies showed a statistically significant relationship between Bcl2 and risk of dying, with large hazard ratios, whereas the three large studies were all non-significant and showed a much smaller effect.94 A recent review of the prognostic importance of TP53 status in head and neck cancer showed clearly that published studies had larger effects than unpublished studies.80 This is in keeping with the belief that epidemiological studies are more prone to publication bias than randomised trials.80, 95 Publication bias may indeed be worse as many studies are based on retrospective analysis of existing clinical databases, and so in essence they do not really exist until published.

Adequate sample size is equally as important for prognostic studies as for clinical trials, but has received little attention. For example, three quarters of 47 papers reporting prognostic studies in osteosarcoma had fewer than 100 cases.96 The likely presence of publication bias means that small studies are unreliable and for prognostic reviews there is a good argument for omitting small studies from meta-analysis, for example those with fewer than 100 patients or even 100 events.

Selective reporting of outcomes is also a concern in prognostic studies. For example, in cancer studies the two principal outcomes are time to death (overall survival) and time to recurrence of disease ('disease-free survival'). Many studies, such as in the case- study in Section 2.3.7, report only one of these outcomes, which may have been chosen in relation to the findings.

3.2.2.2 Cutpoints

Most markers are continuous measurements. However, it is very common in cancer, and occasionally in other fields, for continuous marker values to be converted to binary variables whereby each patient is characterised as having a high or low value. Dichotomisation is statistically inefficient,97, 98 but in some fields, notably cancer, it is ubiquitous. Dichotomising does not introduce bias if the split is made at the median or some other pre-specified percentile. However, if the cutpoint is chosen based on analysis of the data, by splitting at the value which produced the largest difference in outcome between categories, then severe bias will be introduced.99 Significant findings associated with a data-derived cutpoint will be overoptimistic, perhaps by a large amount. Such studies may best be excluded from any meta-analysis.

Many reports do not state how cutpoints were chosen. When the numbers above and below the cutpoint differ or are not stated, and when the chosen cutpoint is unique to that study, it may be unwise to assume that the choice was made in a valid way.

3.2.2.3 IPD vs summary data

Several authors have noted the considerable advantages of obtaining individual patient data (IPD),100, 101 and it is clear that IPD could be especially valuable for systematic reviews of prognostic markers. In addition to the usual advantages of IPD over published summary statistics100 (see Appendix 1), there are some specific advantages. Firstly, it may allow inclusion of more studies as not all studies provide the necessary outcome data. Secondly, it allows all data sets to be analysed in a consistent way, which in this case means adjusting for the same variables and using the same analysis method. Thirdly, the marker values can be kept continuous, increasing statistical power and informativeness. Finally, it is possible to conduct analyses restricted to clinical subgroups, for example by stage of disease.

The natural extension of standard systematic reviews would be to try to collect IPD from all identified studies, whether published or not. Although this has been attempted for prognostic studies it has been found to be very time consuming.102, 103 Concerns about publication bias and the overhead attached to identifying, obtaining and processing each data set have led to the suggestion that for a prognostic meta-analysis of IPD, restriction to only the larger studies or perhaps those carried out in one region104 would be preferable to one based on summary published data that included every published study.77

3.2.3 Identifying research evidence

Identifying prognostic studies is hampered by an absence of standard descriptors and indexing terms. In recent years search strategies have been developed to identify prognostic studies in MEDLINE105 (see Box 2.1) and EMBASE.106 An improved search strategy for MEDLINE, CINAHL and HealthStar has recently been presented107 but is as yet unpublished.

3.2.4 Data extraction

Aspects of particular relevance in prognostic studies include recording how the measurements were made (e.g. equipment or assay used), length of follow-up, distribution of the marker, any cutpoints used (with rationale), amount of missing data, methods of statistical analysis, including variables adjusted for, and the number of participants included in the final model.

A prognostic study with a dichotomous endpoint, such as 30 day mortality after surgery, is statistically no different from a diagnostic accuracy study and poses no additional difficulties for extraction of results. Random-effects endpoints are desirable but there are often difficulties in extracting the log hazard ratio and its standard error from published reports. Guidance on how to estimate these quantities when they are not given explicitly is available.108

3.2.5 Risk of bias assessment

The assessment of the appropriateness of the methodology used in the primary studies is a key element of any systematic review, but has been performed in a minority of cases in prognostic systematic reviews.79, 109 This may reflect the absence of widely agreed criteria for assessing the quality of prognostic studies. Although it is not good practice to use quality as an inclusion criterion, an evaluation of reviews79 found that this was done in 55/163 (34%) reviews.

Reviews of prognostic studies have demonstrated that generally the methodological quality of included studies is poor. For example, one review which assessed 104 prognostic studies in kidney disease against eight criteria, found that three-quarters of the studies satisfied four or fewer of the eight criteria.78

As with other study designs, quality scores are problematic.48, 110, 111 For example, a quality score was developed which evaluated aspects of study methodology grouped into four main categories: the scientific design; laboratory methodology; the generalisability of the results; and the analysis of the study data.112 No details were provided of the development of this scoring system, and as it includes elements of both methodology and reporting it is hard to interpret. Further, for many of the items (e.g. 'source of samples') there is no explanation of the coding scheme. It is preferable to consider specific aspects of methodology related to the risk of bias.

Despite the lack of empirical evidence to support the importance of particular study features affecting the reliability of study findings, especially the risk of bias, theoretical considerations and common sense point to several methodological aspects that are likely to be important.

3.2.5.1 Generic criteria

Table 2.3 lists methodological features that are likely to be important for the internal validity of prognostic studies.88 The items are not phrased as questions but rather as domains of likely importance. Most authors have presented their checklists as questions. For example, 'Was there a representative and well-defined sample of patients at a similar point in the course of the disease?', taken from a checklist produced by the Evidence-Based Medicine Working Group,113 is a question that includes three elements from Table 2.3. This checklist is widely quoted, for example in a guide for clinicians,114 but it omits several of the items in Table 2.3.

It is generally agreed that to be reliable (and clinically interpretable) a prognostic study requires a well-defined ('inception') cohort of patients at the same stage of their disease, preferably at diagnosis.115 This also illustrates the more general requirement that the cohort can be clearly described, which is necessary for the study to have external validity.

3.2.5.2 Context-specific criteria

There may also be context-related quality aspects that should be considered in individual reviews. For example, some studies may have used inferior laboratory methods to measure the marker. However, it is important to distinguish aspects of a study that might be a cause of bias, and hence be genuinely a matter of quality, and those that just reflect variation in study conduct but where no bias is likely. Examples of the latter are patient inclusion criteria, length of follow-up, and choice of measuring device or assay kit. Such factors may well be a cause of heterogeneity and it may be prudent to perform separate (subgroup) analyses to investigate whether they are in fact of importance. There are several published checklists for assessing prognostic studies in cancer.116-118

3.2.5.3 Implementing quality assessment

Quality assessment in prognostic systematic reviews is often incomplete and there is wide variation in current practice. A review of reviews identified 14 methodological domains grouped within six dimensions relating to the risk of bias of prognostic studies.

A framework for assessing the internal validity of articles describing prognostic factor studies88

Study feature	Qualities sought
Sample of patients	Inclusion criteria defined Sample selection explained
	Adequate description of diagnostic criteria
	Clinical and demographic characteristics fully described Representative
	Assembled at a common (usually early) point in the course of their disease
	Complete
Follow-up of patients	Sufficiently long
Outcome	Objective
	Unbiased (e.g. assessment blinded to prognostic information) Fully defined
	Appropriate
	Known for all or a high proportion of patients
Prognostic variable	Fully defined, including details of method of measurement if relevant Precisely measured
	Available for all or a high proportion of patients If relevant, cutpoint(s) defined and justified

Study feature	Qualities sought
Analysis	Continuous predictor variable analysed appropriately Statistical adjustment for all important prognostic factors
Intervention subsequent to inclusion in cohort	Fully described
	Intervention standardised or randomised

3.2.5.4 Quality of reporting

Assessment of study quality is often seriously hampered by poor reporting of methodological details, as is well known for other types of research. The REporting recommendations for tumour MARKer prognostic studies (REMARK) initiative has proposed guidelines for reporting prognostic studies in cancer, most of which apply to any medical context.121 Adoption of the REMARK guidelines should lead to improved reporting of prognostic studies.

System for assessing quality of prognostic factor studies, with proportion of 153 prognostic systematic reviews meeting each item79

Potential bias	% reviews adequately assessing bias	Domains addressed	% reviews assessing domain
1. Study participation The study sample represents the population of interest on key characteristics, sufficient to limit potential bias to the results	55	1. Source population clearly defined	50 21
		2. Study population described	50
		3. Study population represents source population or population of interest	
2. Study attrition Loss to follow-up (from sample to study population) is not associated with key characteristics, sufficient to limit potential bias (i.e., the study data adequately represent the sample)	42	4. Completeness of follow-up described	19 42
		5. Completeness of follow-up adequate	

Potential bias	% reviews adequately assessing bias	Domains addressed	% reviews assessing domain
3. Prognostic factor measurement The prognostic factor of interest is adequately measured	59	6. Prognostic factors defined in study participants to sufficiently limit potential bias	31 59
		7. Prognostic factors measured appropriately	
4. Outcome measurement The outcomes of interest are adequately measured in study participants to sufficiently limit potential bias	51	8. Outcome defined 9. Outcome measured appropriately	42 51

Potential bias	% reviews adequately assessing bias	Domains addressed	% reviews assessing domain
5. Confounding measurement and account Important potential confounders are appropriately accounted for, limiting potential bias with respect to the prognostic factor of interest	13	10. Confounders defined and measured	21 53
		11. Confounding accounted for	
6. Analysis The statistical analysis is appropriate for the design of the study, limiting potential for presentation of invalid results	33	12. Analysis described	8
		13. Analysis appropriate	33 32
		14. Analysis provides sufficient presentation of data	

3.2.6 Synthesis

3.2.6.1 Outcome measures

In prognostic studies the focus of interest is what may happen in the future. It is natural, therefore, that most prognostic studies have outcomes that are the time to a specific event, such as death. However, some prognostic studies with dichotomous outcomes may inappropriately ignore the time element. For example, a study looking at death within three years may classify all patients as dead or alive, but those patients who are lost to follow-up before three years (i.e. have censored survival times) cannot be so classified and may be excluded. One exception is studies of prognosis in pregnancy where outcomes often relate to the birth of the baby (e.g. predicting caesarean section or pre-term birth). Such outcomes are genuinely dichotomous and can be analysed in the same way as a study of diagnostic accuracy.

Meta-analysis of time-to event outcomes of aggregate data derived from publications is usually done using the generic inverse-variance approach and may use a fixed effect or random-effects model (see Chapter 1, Section 1.3.5 Data synthesis). This type of analysis and extensions have been discussed, as has investigation of heterogeneity in such studies.122, 123 Although the preferred statistical summary is the hazard ratio (HR) (see Chapter 1, Section 1.3.5 Data synthesis) many publications do not report the HR or the information needed to calculate it. Consequently, some of the identified studies cannot be included in the synthesis. Furthermore, non-reporting of appropriate statistical summary measures may be more likely if the marker was found not to be statistically significantly related to outcome, leading to bias. Statistical methods for analysing IPD time-to-event data have been compared,124 and methods have been published for combining IPD with published summary data.125

When all studies have reported data as dichotomous or continuous, meta-analysis may be relatively straightforward. However, if there is a mixture of binary, multi-category, and continuous representation of the same marker, meta-analysis will be problematic and expert input will be advisable. Similar problems have been reported in meta- analysis of epidemiological studies.126

In principle researchers may need to combine estimates of a marker that is kept continuous in some studies and dichotomised in others. It is important to note that the hazard ratios for those two cases are not comparable so they should not be combined. There is a related literature on combining data on dose-response relationships in epidemiology.127-129

3.2.6.2 Adjustment for other variables

In RCTs the groups being compared are expected to be very similar with regard to prognostic factors (baseline characteristics) through the use of a random sequence of intervention assignment. In non-randomised studies there is no such safeguard and we should expect the groups being compared to differ in various ways. In prognostic studies we are comparing individuals

with different levels of a marker, whether binary or continuous. That comparison could easily be biased by other variables that are associated with both the marker and patient prognosis – in other words the comparison may be 'confounded'.

Furthermore, while it may be of interest to know if a marker considered alone is prognostic, in most cases the real aim of a prognostic marker study should be to ascertain if the marker adds useful clinical information to what is already known. In many clinical contexts much is already known about prognosis, and it is important to know whether the new marker offers additional prognostic value over and above that achieved with previously identified prognostic variables. As an example, a study examined the 'incremental usefulness' of 10 biomarkers for predicting the risk of cardiovascular events, adjusted for age, sex, and conventional risk factors.130 That approach implies the addition of the marker to a statistical model that includes other known prognostic variables. As well as addressing the most sensible clinical question, adjustment should greatly reduce the risk of confounding.

Dealing with adjustment presents a problem for synthesis, as individual studies are likely to have used different statistical approaches for adjustment and adjusted for different selections of variables. Some syntheses avoid this methodological variation by using unadjusted estimates.131 While this approach is standard in systematic reviews of RCTs, in prognostic studies it replaces one problem with a worse one; unadjusted analyses are likely to be biased. Although the unadjusted estimate provides the maximum opportunity for comparison of consistent estimates across studies,131 it is important to adjust for other prognostic variables to get a valid picture of the relative prognosis for different values of the marker. Prognostic studies thus generally require analysis using multiple regression analysis, although stratification may be useful in simpler situations. For outcomes which are dichotomous or time to a specific event, logistic or Cox proportional hazards regression models respectively are appropriate for examining the influence of several prognostic factors simultaneously. For this purpose, known prognostic factors should preferably not be subjected to a variable selection process. Even though such variables may not reach specified levels of significance in a particular study, they should be included in the models generated in order to compare results to other reported studies. Comparison of models with and without the marker of interest provides an estimate of its independent effect and a test of statistical significance of whether the new marker contains additional prognostic information.

In practice, researchers will often find a mixture of adjusted and unadjusted results. Only 47/129 (36%) of prognostic marker studies in cancer used multivariate modelling in which the marker was added to standard clinical variables.132 A recent review presented separate meta-analyses of adjusted and unadjusted results of BCL-2 as a protective prognostic marker in breast cancer.133 It demonstrated, as expected, that the adjusted hazard ratio was lower than the unadjusted value but these differences were small (disease free survival (DFS) HR 1.58 vs HR 1.66). This approach reduces the need for speculation about the value of adjustment, which seems a good strategy even if all studies are then combined.

3.2.6.3 Sensitivity analyses

General considerations of investigating the sensitivity of the review findings to various choices apply equally to reviews of prognostic studies. In the specific context of prognosis, given the evidence about publication bias, it may be advisable to conduct a sensitivity analysis in which smaller studies are excluded.

3.2.6.4 Case study

An example of a systematic review addressing a prognostic question.

Objective

This systematic review of aggregate data obtained from study publications aimed to obtain better quantification of the prognostic importance of Ki-67/MIB- 1 expression as a marker of cell proliferation in early breast cancer. Ki-67 is present in all proliferating cells and there is great interest in its role as a marker of proliferation. MIB-1 is a monoclonal antibody against recombinant parts of the Ki-67 antigen.

Inclusion criteria

The review included studies evaluating the relationship between Ki-67/MIB-1 status and prognosis in early breast cancer published by May 2006. Studies had to have been published as a full paper in English. No minimal sample size or minimal median duration of follow-up was defined.

Searching

PubMed was searched using the following keywords: 'breast cancer', 'Ki-67', 'MIB-1', 'proliferative index', 'proliferative marker', 'survival' and 'prognostic'. The authors also screened references from the relevant literature, including all the

identified studies and reviews. When the same patient population was reported in more than one publication, only the most recent or complete study was included.

Data extraction

The methods of Parmar et al134 were used to extract log HR and SE(log HR). Three people independently extracted information from survival curves.

Data availability

Sixty-eight eligible studies were identified of which 46 studies (including 12,155 patients) could be included in meta-analyses; 38 studies for disease free survival and 35 studies for overall survival.

Study characteristics

Table 2.5 shows that there was considerable variation in study characteristics, for example in patient characteristics, cutpoint used to define high Ki-67, and prevalence of raised levels of the marker. All studies dichotomised Ki-67 values. Even studies with the same threshold had prevalence of high values ranging from 11% to 88%. The studies also varied considerably in the interventions patients had received and in the antibody used in laboratory evaluations of Ki-67.

Systematic review of Ki-67 as a prognostic marker in early breast cancer: excerpt from table of study characteristics and results for disease-free survival (hazard ratios and 95% confidence intervals) 87

Study	N	Follow- up (median months)	Threshold	Prevalence	How chosen	HR	95% CI
Bevilacqua, 1996	107	74	10%	88%	arbitrary	2.75	1.02 7.39
Bos, 2003	150	106 (mean)	10%	42%	arbitrary	2.47	1.08 5.65
Brown, 1996	674	72	5%	25%	optimal cut-off	1.19	0.79 1.80
Caly, 2004	244	72 (min)	32%	50%	unclear	1.95	0.92 4.14
Domagala (N0), 1996	111	88	10%	60%	median	3.04	1.03 8.99
Domagala (N+), 1996	75	88	10%	53%	median	1.38	0.66 2.86
Erdem, 2005	47	73	10%	28%	median	17.23	2.42 122.4
Fresno, 1997	146	75	10%	58%	arbitrary	1.81	0.71 4.59
Gasparini, 1994	165	60	7.5%	50%	mean	2.58	1.21 5.49

Study	N	Follow- up (median months)	Threshold	Prevalence	How chosen	HR	95% CI
Gonzalez, 2003	221	103	30%	NR	arbitrary	3.18	1.52 6.65
Goodson, 2000	112	61	24%	50%	mean	2.90	1.18 7.15
Heatley, 2002	59	60	10%	44%	mean	0.81	0.36 1.81
Hlupic (N+), 2004	192	180	10%	61%	arbitrary	1.30	0.80 2.11
Jacquemier, 1998	152	60	3.5%	49%	median	3.29	1.49 7.22
Jansen, 1998	321	128	7%	48%	median	1.35	1.01 1.80
Jensen, 1995	118	104	17%	46%	median	3.41	1.44 8.06
Liu, 2001	773	196	17.8%	50%	median	1.76	1.41 2.20
Locker, 1992	67	27	9%	34%	tertile	4.19	1.19 14.7
Mottolese, 2000	157	60	10%	55%	arbitrary	1.82	0.90 3.67
Pellikainen, 2003	414	57	20%	44%	arbitrary	2.56	1.46 4.50
Pierga, 1996	136	70	8%	49%	median	1.37	0.64 2.91
Pietilainen, 1996	188	103 (mean)	20%	53%	arbitrary	1.88	1.16 3.05

Study	N	Follow- up (median months)	Threshold	Prevalence	How chosen	HR	95% CI
Pinder, 1995	177	NR	34%	42%	tertile	1.66	1.09 – 2.52
Pinto, 2001	295	39.6	10%	46%	arbitrary	1.46	0.74 – 2.87
Querzoli, 1996	170	66.5	13%	25%	tertile	2.05	1.11 – 3.77
Railo, 1993	326	32.4 (mean)	10%	11%	unclear	2.39	0.77 – 7.38

N: Number of participants; HR: Hazard ratio; CI: Confidence interval

Meta-analysis

Study results were combined using the Peto-Yusuf method. No studies were excluded because of methodological quality but some studies were excluded because suitable data were not available – those included studies which did not provide unadjusted results. Random-effects meta-analyses were used because there was considerable heterogeneity. Separate meta-analyses were performed for overall (OS) and DFS. Both showed a significant association between raised Ki-67 and worse survival: HR 1.93 (95% CI: 1.74 – 2.14) and 1.95 (1.70 – 2.24)

respectively. Table 2.5 shows the reported characteristics and the results (HR) for DFS for a subset of the studies.

The 17 omitted studies were included in a sensitivity analysis with no appreciable change to the findings. The authors did not consider possible publication bias.

Conclusions

The authors concluded that 'Despite some limitations, this meta-analysis supports the prognostic role of Ki-67 in early breast cancer, by showing a significant association between its expression and the risk of recurrence and death in all populations considered and for both outcomes, DFS and OS.' They also noted that the reporting of the individual studies was suboptimal and that they had assessed only the univariate prognostic value of Ki-67. They suggested that a prospective study to examine whether Ki-67 was of prognostic importance over and above known factors. Thus, in common with many reviewers of such studies, these authors did not feel that the existing literature was strong enough on which to base clinical decisions.

3.2.7 Systematic review as a driver for improved study quality

Systematic reviews can play a valuable role not just in summarising the findings of published studies but also in drawing attention to the poor and inconsistent methods used. Good systematic reviews are needed to highlight the weaknesses of the evidence base behind prognostic markers and to provide guidance on how better-quality studies can be carried out in the future. This is true of prognostic studies and it has been commented that 'one has to question why it is acceptable for tumor marker studies to be performed with less scientific rigor than studies of new pharmaceutical agents.

As an example, a review of 26 published systematic reviews of prognostic markers in cancer found common deficiencies in both conduct and reporting.109 Less than 75% of the systematic reviews stated clearly their aims and objectives, the literature search strategy, and the study eligibility criteria. Only 20% reported the final number of primary studies used. Less than 50% of the systematic reviews reported elements of primary study description and analysis, such as sampling methods, cancer stage, cutpoint, and numeric results including CIs and P-values. The exception was the sample size, which was reported in 73% of the systematic reviews. About half of the systematic reviews had carried out a meta-analysis. Of those, some did not include a forest plot or numerical summary with confidence intervals. Most had explored heterogeneity,

4 Summary

This volume offers a comprehensive examination of systematic reviews as a methodological framework for evidence synthesis. It provides a structured account of the rationale, principles, and procedures underlying systematic reviewing, with emphasis on methodological rigor, transparency, and reproducibility.

The book addresses the full continuum of the review process: formulation of research questions, development of protocols, systematic searching, study selection, critical appraisal, data extraction, synthesis of evidence (qualitative and quantitative), and standards for reporting. Consideration is also given to methodological challenges, risk of bias, and the application of established guidelines such as PRISMA.

In addition to foundational methods, the volume discusses emerging approaches, including automation, living systematic reviews, and rapid review techniques, reflecting the evolving landscape of evidence synthesis. Ethical and epistemological issues, as well as the implications of systematic reviews for clinical practice, health policy, and research agendas, are also explored.

Intended for students, researchers, and practitioners, this book functions both as an instructional resource and as a reference text. Its aim is to strengthen the capacity to conduct and critically appraise systematic reviews, thereby contributing to the advancement of evidence-based practice and the integrity of research synthesis.

References