Introduction

Systematic literature reviews

Methods for reviewing and evaluating the scientific literature range from highly formal, quantitative information syntheses to subjective summaries of observational data. The purpose of a systematic literature review is to evaluate and interpret all available research evidence relevant to a particular question. In this approach a concerted attempt is made to identify all relevant primary research, a standardized appraisal of study quality is made and the studies of acceptable quality are systematically (and sometimes quantitatively) synthesized. This differs from a traditional review in which previous work is described but not systematically identified, assessed for quality and synthesized.

Advantages

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

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

Thus, while some people see the mixing of 'apples and oranges' as a

problem of systematic reviews, it can be a distinct advantage because of its ability to enhance the generalizability and transferability of data.

Disadvantages

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

Method

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

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

A systematic review involves a number of discrete steps:

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

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

Question formulation

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

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

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

Finding studies

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

Appraisal and selection of studies

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

Summary and synthesis of relevant studies

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

Determining the applicability of results

Following the summary and synthesis of the studies, the next step is to ask about the overall validity, strength and applicability of any results and conclusions. How and to whom are the results of the synthesis applicable? How will the effects vary in different populations and individuals?

How much work is a systematic review?

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

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

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

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

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

About this book

The remainder of this book is divided into two parts:

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

Appendixes A and B include details of search procedures and a listing of available software.

Part 1

General methods

1 The question

1.1 What types of questions can be asked?

Clinical problems and health policies may involve many different questions which need to be informed by the best available evidence. It is useful to have a classification of the different types of health care questions that we may ask:

- Phenomena: 'What phenomena have been observed in a particular clinical problem, e.g. what problems do patients complain of after a particular procedure?'
- Frequency or rate of a condition or disease: 'How common is a particular condition or disease in a specified group?'
- Diagnostic accuracy: 'How accurate is a sign, symptom or diagnostic test in predicting the true diagnostic category of a patient?'
- Aetiology and risk factors: 'Are there known factors that increase the risk of the disease?'
- Prediction and prognosis: 'Can the risk for a patient be predicted?'
- Interventions: 'What are the effects of an intervention?'

Answering each type of question requires different study designs, and consequently different methods of systematic review. A thorough understanding of the appropriate study types for each question is therefore vital and will greatly assist the processes of finding, appraising and synthesizing studies from the literature. A summary of the appropriate study types for each question and of the issues that are important in the appraisal of the studies is also given in Table 1.1. General information on how to find and review studies is given in the remainder of Part 1 with further details for each question type in Part 2.

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

Question	Ideal study types	Major appraisal issues
1. Intervention	Randomized controlled trial	Randomization Follow-up complete Blinding of patients and clinicians
2. Frequency/rate (burden of illness)	Cross-sectional study or consecutive sample	Sample frame Case ascertainment Adequate response/follow-up achieved
3. Aetiology and risk	Cohort study	Groups only differ in exposure Outcomes measurement Reasonable evidence for causation
4. Prediction and prognosis	Cohort study	Inception cohort Sufficient follow-up
5. Diagnostic accuracy	Random or consecutive sample	Independent, blind comparison with 'gold standard' Appropriate selection of patients
6. Phenomena	Qualitative research	Appropriate subject selection and methods of observation

1.1.1 Interventions

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

To study the effects of interventions, it is necessary to compare a group of patients who have received the intervention (study group) with a comparable group who have not received the intervention (control group). A randomized controlled trial (RCT), which is a trial in which subjects are randomly allocated to the study or control groups, is usually the ideal design. A hierarchy of designs for the study of the effects of interventions is illustrated in Table 1.2.

1.1.2 Frequency or rate

How common is a particular feature or disease in a specified group in the population? This is measured as the frequency (proportion or prevalence) or rate (incidence) of the feature or disease. For example, the prevalence of osteoarthritis with ageing, or the rate of new cases of human immunodeficiency virus (HIV).

The appropriate study design in this case is a cross-sectional survey with a standardized measurement in a representative (e.g. random) sample of people; for a rate, the sample would need to be followed over time. If, instead of a single frequency, we become interested in the causes of variation of that frequency, then this becomes a question of risk factors or prediction (see below).

1.1.3 Diagnostic accuracy

How accurate is a particular diagnostic screening test? If there is good randomized trial evidence that an intervention for a particular condition works then it may be necessary to assess how accurately the condition can be diagnosed from a sign, symptom or diagnostic test. To do this, a comparison is needed between the test of interest and a 'gold standard' or reference standard. The most commonly used measures of accuracy are the sensitivity and specificity of the test.

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

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

Study design	Protocol
Systematic review	Systematic location, appraisal and synthesis of evidence from scientific studies (usually randomized controlled trials)
Experimental studies	
Randomized controlled trial	Subjects are randomly allocated to groups either for the intervention/treatment being studied or control/placebo (using a random mechanism, such as coin toss, random number table, or computer-generated random numbers) and the outcomes are compared
Pseudorandomized controlled trial	Subjects are allocated to groups for intervention/treatment or control/placebo using a nonrandom method (such as alternate allocation, allocation by days of the week or odd–even study numbers) and the outcomes are compared
Comparative (nonrandomized and observational) studies	
Concurrent control	Outcomes are compared for a group receiving the treatment/intervention being studied, concurrently with control subjects receiving the comparison treatment/intervention (e.g. usual or no care)
Historical control	Outcomes for a prospectively collected group of subjects exposed to the new treatment/intervention are compared with either a previously published series or previously treated subjects at the same institutions
Cohort	Outcomes are compared for groups of subjects who have been exposed, or not exposed, to the treatment/intervention or other factor being studied
Case-control	Subjects with the outcome or disease and an appropriate group of controls without the outcome or disease are selected and information is obtained about the previous exposure to the treatment/intervention or other factor being studied
Interrupted time series	Trends in the outcome or disease are compared over multiple time points before and after the introduction of the treatment/intervention or other factor being studied