Assessment of methods in health care

A handbook

August 2017

Preliminary version

SWEDISH AGENCY FOR HEALTH
TECHNOLOGY ASSESSMENT AND
ASSESSMENT OF SOCIAL SERVICES

Content

Chapter 1
Assessment of interventions in health care and social services – Introduction

Evidence-based care

The health care and social services has developed rapidly, becoming more and more based on scientific observations. Greater demands are placed on using interventions whose outcomes are supported by scientific evidence. EBM ("evidence-based medicine" or "evidence-based care") and EBP ("evidence-based practice") are expressions of this trend. EBM and EBP are approaches that involves critical appraisal of whether or not interventions are based on the best available scientific grounds [1].

Caregivers find it increasingly difficult to remain updated in their respective fields. The number of articles published each year grows continuously. Within health care estimates indicate that more than 1.4 million medical articles are published annually, of which approximately 10-15 percent are considered to be of practical and lasting value to the patients.

Systematic review

One way to overview the evidence is to read a *review.* The weakness of non-systematic reviews is that they are often based on studies familiar to the author. Moreover, there is a risk that authors select only those studies that supports their own views. The review may therefore give a biased picture of the actual situation.

A *systematic review* should meet great demands on reliability. A good systematic review follows certain principles that will minimize the risk that chance or arbitrariness affects the conclusions. The principles include:

- A specified question/problem
- Reproducibility: reporting selection criteria (inclusion and exclusion criteria) in order to sort out relevant literature and also strategies for searching and assessing quality
- Systematic search for all relevant literature addressing the question or problem
- Quality assessment of all studies fulfilling the inclusion criteria
- Extracting data and tabulating quality-assessed studies
- Composite appraisal of results, for example in a meta-analysis
- A judgment on how well-founded the results are (evidence-grading)

A well-executed systematic review enables readers to appraise the trustworthiness of the conclusions and to determine if the evaluation has omitted important literature.

Assessment of interventions

Health technology assessment (HTA) involves the systematic evaluation of the scientific evidence relating to the effects, risks and costs of different interventions [2]. This applies to all interventions, whether they involve prevention, diagnosis, treatment or care. SBU: s mission to systematically review effects, risks and costs should be supplemented by taking into account ethical and social aspects. The evaluation has a broader scope and therefore gives greater attention to national/local conditions than do systematic reviews.

Facts 1.1 SBU- one of the world´s oldest HTA agencies.

Since 1987, SBU is missioned by the Government to evaluate interventions used in health care. The results of these evaluations should serve as guidance both to practitioners and to political and administrative leaders at different levels. The task also includes dissemination of the results from these evaluations to the health care services in Sweden and to follow up on the effects of this effort. Since 2015, the mission also includes social services.

Selecting topics for assessment

SBU receives proposals for assessment from many sources, e.g. social workers and health care staff, specialist associations, county council leaders and other governmental agencies. Some of the assessments provide a basis for national guidelines issued by The National Board of Health and Welfare or decisions issued by The Dental and Pharmaceutical Benefits Agency (TLV).

Submitted proposals for assessment are ranked according to a number of criteria. The more criteria met, the more urgent the issue. The criteria are:

- Major importance for life and health
- affects many
- wide variation in practice
- uncertainty about the strenght of scientific evidence
- major economic consequences
- important ethical issue
- major impact on organization or staff
- controversial or popular issue

SBU: s two advisory Boards assess the scientific quality of SBU: s reports. SBU: s Committee is responsible for the conclusions of the reports. The Committee is composed of representatives of central organizations within the Swedish health care sector. The composition of the committee should guarantee that the projects have a broad support, are considered important and that the conclusions are firmly established in Swedish health care and social services.

Experts within different areas play a key role

The assessment is conducted by experts in a subject area with staff support from SBU. This is different from many other organizations producing systematic reviews. Their reports are often developed by people who are experts only on the assessment methodology itself and they have limited possibility to determine the clinical relevance of a method. The experts in SBU: s projects ensure that the assessments are based on a deep understanding of the subject area.

It is important that the composition of the project group is multidisciplinary. The questions to be answered often involve several professional categories. The group should also reflect gender and geographic distribution.

The work process

SBU:s projects normally tales 1-2 years, but can also take longer time if it involvs a complete field. At certain check points the work is presented to the scientific committees for discussion. When the manuscript is complete a comprehensive scrutiny is awaiting. The draft is scrutinized by an internal quality panel mainly evaluating the methodological quality. The draft is also sent to several external reviewers who primarily determine of the content is relevant. Once they have approved the report the manuscript is judged by the some of the internal Boards. After their approval SBU: s committee takes their views. This process can probably be regarded as more comprehensive than the process preceding publication in scientific journals.

Suggestions for reading and directions for reading

There are several Swedish [2-8] and international [9-13] publications that give basic or deeper descriptions on evidence-based medicine/care and evidence-based practice within social services.

This handbook follows the different steps in the HTA process. It may be read in succession or used as a reference by the experts in different phases of the project.

The handbook starts with an overview of the different steps for systematic assessment and evaluation (Chapter 2). This is followed by a section on formulating questions and choosing selection criteria (Chapter 3) and a section for literature search and choice of databases (Chapter 4). Chapter 5 describes the appraisal of study relevance. Chapters 6 and 7 address quality assessment of studies with different study designs. Chapter 8 discusses the assessment of qualitative studies.

The use of meta-analysis is found in Chapter 9. Chapter 10 describes how the quality of evidence should be assessed. Chapter 11 addresses health economy while ethical and social aspects are addressed in Chapter 12.

The final pages of the handbook include a glossary and appendices, including the various check lists that are used in assessing the quality of studies. Basic statistical concepts are given in appendix 10.

Referenser

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn’t. BMJ 1996;312:71-2.

2. Banta D, Jonsson E, editors. History of HTA. Int J Technol Assess Health Care 2009;25 suppl 1:1-289.

3. Brorsson B, Wall S. Värdering av medicinsk teknologi – problem och metoder. Stockholm: Medicinska forskningsrådet; 1985.

4. Nordenström J. Evidensbaserad medicin i Sherlock Holmes fotspår. 4:e upplagan. Karolinska University Press; 2007.

5. Furberg B, Furberg C. Allt är inte guld som glimmar. III Hur man värderar kliniska studier. Kungsbacka: Solutio; 2005.

6. Larsson A. Arbetsbok i evidensbaserad medicin. Södra Älvsborgs sjukhus 2006:2.2.

7. Willman A, Stoltz P, Bahtsevani C. Evidensbaserad omvårdnad. En bro mellan forskning och klinisk verksamhet. Studentlitteratur; 2006.

8. Levi R. Vettigare vård. Evidens och kritiskt tänkande i vården. Stockholm: Norstedts; 2009.

9. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.0.0 (update February 2008), Cochrane collaboration 2008. Available from www.cochrane-handbook. org.

10. Fletcher RH, Fletcher SW. Clinical epidemiology. The essentials. 4th ed. Lippincott Williams & Wilkins: Baltimore; 2005.

11. Guyatt G, Rennie D, editors. User’s guide to the medical literature. A manual for evidence-based clinical practise. JAMA & Archives Journal; 2002.

12. Egger M, Smith DG, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2001.

13. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. 2nd ed. Little, Brown and company: Boston; 1991.

Chapter 2

An overview of the steps in a systematic evaluation

Introduction

The method of evaluation used by SBU is based on a systematic assessment of the scientific literature. This means that the search for, selection and quality assessment of relevant literature are performed systematically. It is important that each step in the process is well defined and clearly evident in the report (Figure 2.1). This chapter presents a summary of what is included in the different parts of the evaluation.

Question Chapter 3 • Formulate the question in a structured manner • Establish criteria for inclusion and exclusion	**Selection of literature** Chapters 4–5 • Literature search • Initial screening of abstracts • Obtain articles in full text • Determine which articles meet inclusion-and exclusion criteria (use inclusion list)
Assessment of studies Chapters 5–8 • Determine relevance (use check list) • Quality assessment (use check list according to study design) • Data extraction for tabulation (use table template)	**Evidence-grading** Chapters 9–10 **and conclusions** • Synthesise results of separate studies • Determine the strength of evidence • Formulate evidence-graded results

Figure 2.1 Process for the systematic assessment of scientific evidence.

Formulating a project’s questions (Chapter 3)

Initially, the questions addressed by a project are usually formulated in general terms. The project group’s first task is therefore to structure the questions so that they are answerable. This work is crucial for determining which studies will be captured in the literature search and therefore has to be done with particular care.

The project group must decide which populations are of interest in relation to the question, which methods should be evaluated within the frame of the project and which outcome measures should be studied. In most cases, relevant control methods must also be defined. The question is thereafter formulated according to the PICO format (population, intervention, control, outcome) for intervention studies and according to the PIRO format (population, index test, reference test, outcome) for studies on diagnostic accuracy.

The questions are further specified by means of inclusion and exclusion criteria.

Literature search (Chapter 4)

The literature search is performed by an information specialist in consultation with the project’s experts and the project leader. The experts’ primary role is to identify relevant articles from which the information specialist then creates search strategies by analysing abstracts and how the articles are indexed.

To minimise the risk of bias, the literature search is performed in several databases. It is important to make a supplementary check of the references of identified full-text articles. The aim of the search strategies is to catch all relevant studies and minimise the number of irrelevant articles. An updated search is performed in the project’s final phase in order to catch articles that have been published in the meantime.

The search strategies and their results are stated in the report.

Assessment of a study’s relevance (Chapter 5)

Two (or more) persons working independently assess the abstracts that have been listed in the database searches. Studies that at least one of the assessors considers are relevant in the light of their titles and abstracts are obtained in full-text.

The number of articles ordered in full-text should be stated in the report.

The two assessors, working independently, then judge whether or not the included articles meet the inclusion criteria. A form for inclusion and exclusion is used as an aid. Articles that do not meet the inclusion criteria are sorted out and the reason for exclusion is noted on the form. The assessors then compare their inclusion lists; in the event of a mismatch, they discuss this and decide whether or not the article should be included.

The report should include a list of excluded articles and the reason for exclusion.

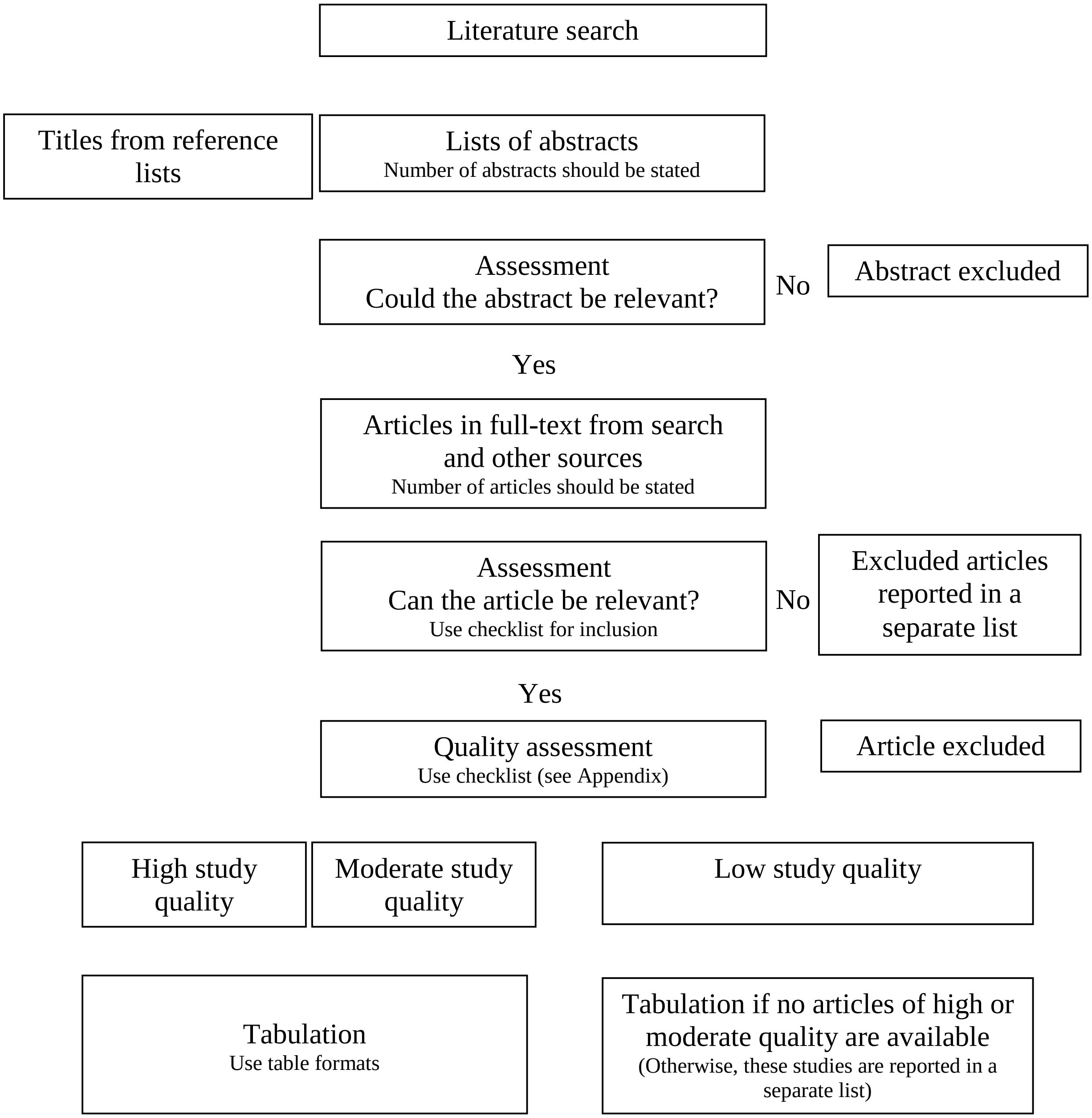
A summary of the selection process from literature search to tabulation is given in Figure 2.2.

Quality assessment and data extraction (Chapters 6–8)

In the next step the assessors independently judge the quality of the included studies. Check lists are used as an aid, one for each type of study (randomised controlled studies, observational studies, diagnostic studies, qualitative studies and systematic reviews). The judgements refer to the study's quality, which is designated as being high, moderate or low. Studies whose quality is judged to be either high or moderate form the basis for synthesising data and assessing the strength of evidence.

This more stringent assessment often leads to additional studies being judged as not meeting the criteria; these are then excluded and added to the exclusion list.

Tables of data from the studies that comprise the scientific evidence are an important part of the report. They should contain information about the authors, population, intervention, control, results and study quality. Table templates are available for questions about intervention and diagnostics, respectively. The tables are in English because, together with the English summary, they are disseminated internationally via various databases. If there are no studies of high or moderate quality, studies of low quality are tabulated. Such tables summarise the state of knowledge but the information is not sufficient to form a basis for synthesis or to determine the strength of evidence.



Literature search
Titles from reference lists
Lists of abstracts
Number of abstracts should be stated
Assessment
Could the abstract be relevant?
No
Abstract excluded
Yes
Articles in full-text from search and other sources
Number of articles should be stated
Assessment
Can the article be relevant?
Use checklist for inclusion
No
Excluded articles reported in a separate list
Yes
Quality assessment
Use checklist (see Appendix)
Article excluded
High study quality
Moderate study quality
Low study quality
Tabulation
Use table formats
Tabulation if no articles of high or moderate quality are available
(Otherwise, these studies are reported in a separate list)

Figure 2.2 The selection process.

Synthesis and weighing results (Chapter 9)

The next step in the process is to synthesise the results of the studies that have been included in the scientific base, for example by calculating an effect size. If there are several studies it is appropriate to determine whether the data can be considered for meta-analysis. If the meta-analysis shows that the studies are heterogeneous, the synthesis is reported only descriptively. The so-called forest plot of the meta-analysis is useful for studies on intervention as well as on diagnostics.

Meta-analyses can be performed using Cochrane Collaboration's programme Rev Man, which is free of charge.

Grading the strength of evidence in the results (Chapter 10)

The reliability of the composite results is expressed as the strength of evidence. SBU uses the evidence-grading system GRADE. GRADE is developed by an international group of experts and is being used increasingly by organisations and authorities such as WHO, NICE and Cochrane Collaboration.

In principal, GRADE is based on experience from previous systems but places greater emphasis on patient benefit and risks. The grading reflects the quality of the study and how its reliability is affected by factors such as the material's inconsistency, indirectness, imprecision and publication bias. The evidence is graded into one of four levels: high, moderate, low and very low. Findings rated as low or very low evidence correspond to SBU's definitions of limited and insufficient scientific basis, respectively.

Health economy (Chapter 11)

A comprehensive evaluation requires an assessment of the cost-effectiveness and economic consequences of a method being either introduced, extended, contracted or phased out.

Ethical and social aspects (Chapter 12)

Besides including the method's effects, risks and cost-effectiveness, the assessment should consider its ethical and social consequences.

Chapter 3

Structuring and defining the review questions

The first part of the project involves establishing the project's boundaries and deciding which questions to address.

The project's goal is often formulated in general terms and needs to be specified in a limited number of questions that needs to be selected. The project group may need to consult other stakeholders in order to ensure that the most essential questions are included. Such stakeholders may include those who proposed the project as well as decision-makers, experts and patients who are not participating in the assessments.

The structured question

First of all, the question must be structured. This is commonly done by breaking the question down into meaningful elements. The PICO system is applicable to most medical questions. PICO is the acronym for Patient/population/problem, Intervention/index test, Comparison/control, and Outcome [1]. Having defined which study populations, treatments, control treatments, and outcomes are relevant to the question, one can establish the main inclusion criteria. These are completed with other aspects like. treatment time/follow-up time and appropriate study design.

A well-structured PICO often leads to a greater chance of making the literature search more specific, so that less time has to be spent on excluding irrelevant article abstracts. PICO for intervention studies is summarised in Facts 3.1.

Establishing detailed inclusion criteria for a question is often beneficial because it reduces the risk of overinclusion (in that uncertainty leads to retrieval of irrelevant articles), thereby avoiding additional work at later stages.

Population

Which populations are we interested in? The degree to which populations need to be specified varies. Diagnosis by itself may suffice in some cases, while in other cases a call for a high degree of detail, eg. diagnosis by different degrees of a disease's severity, by age, gender, or ethnicity is needed. In a well-conducted study, the study population is defined in terms of clear inclusion and exclusion criteira as well as clearly reported baseline data. At times, populations are heterogeneous and encompass subgroups that are not of immediate interest. If the results for the relevant population are not reported specifically, one may need to determine how large that population's fraction of the total study population needs to be.

Facts 3.1 Components of the review question.

Population/participants

The population to be studied is defined, eg by:

- Age
- Gender
- Diagnosis
- Disease level
- Risk factors
- Other diseases

Intervention/method

Definition and description of the method

Comparison/control

Definition and description of the comparison/control method

- Other treatments
- Placebo

Outcome measures

Outcome measures of direct importance for the individual, eg survival, quality of life, morbidity, and changes in symptoms

Outcome measure can also include complications and adverse effects of the intervention

Studies in health economics often express outcomes as cost per quality-adjusted life-year (QALY)

Exclusion criteria can also apply to specific populations that meet the inclusion criteria but for various reasons are not relevant for the question, eg subgroups of the population of interest that are characterised by certain factors, such as comorbidities and medications.

Intervention

Which interventions are of interest? Here too, the level of detail varies, depending on the question and the selected population. It may be necessary to specify, eg dose, dosage form and method of administration.

For questions concerning, eg, risk factors for disease, exposure might be a more appropriate term than intervention.

If the issue involves diagnostic accuracy, the experimental method (index method) should be defined here (Chapter 7).

Comparison/control

Which intervention(s) in the comparison group can be acceptable? The choice of control intervention is often decisive for a study's relevance. For instance, drugs that have not received market authorisation in Sweden are often used as control intervention. Occasionally, such substances can still be relevant, eg if they can be seen as a representative of a group of drugs such as beta blockers or benzodiazepines. The opposite may also occur, ie a registered drug is selected as control but can not be seen as a representative of the group. One example is studies of antihypertensive drugs that use atenolol, a beta blocker, as the control intervention [2].

Sometimes the doses for intervention and control have been chosen to favort the therapeutic effects of intervention or to de-emphasize the risks of adverse effects [3]. Also, some drug studies may be designed in a way that, for pharmacokinetic reasons, it is unfavourable for the control substance. For instance, both drugs might be administered orally even though uptake of the control substance is low [4].

These factors may also be of importance outside the pharmaceutical arena. For instance, doses in psychological studies can be measured by the number, frequency, and duration of the sessions. Cognitive behavioural therapy (CBT) is administered correctly by staff with the appropriate training but in some studies CBT is provided by untrained caregivers. Placebo interventions are uncommon in psychological research. In psychological and social interventions the control group is often offered "usual (standard) care". Such care can vary widely and may be irrelevant in a Swedish context. Also, the psychological method used as control may be worse than no treatment at all, that is directly harmful, which can give the studied intervention an unfair advantage [5].

For questions concerning diagnostic accuracy, the reference standard (gold standard) should be defined here, with which the index test is compared (Chapter 7).

Outcome measures

Which outcome measures are appropriate for determining the effect of an intervention? Firstly, outcome measures should be relevant for the patient, eg mortality, morbidity, suffering, functional impairment, and quality of life. Secondly, one can choose surrogate endpoints, ie measurable factors that relate in some way to outcomes that are relevant for the patient; examples are blood lipoproteins, blood pressure, and bone density.

Composite measures are common in clinical research. The rationale behind involves combining different outcome measures to give the study a greater statistical power. However, composite measures should be used with care, particularly when surrogate endpoints are included. A statistically significant effect of a composite measure can often be explained solely by effects of a surrogate endpoint or a variable that is less relevant for the patient. It is also possible that composite measures can mask a negative treatment effect on important outcomes, eg death and cardiovascular events [6].

For questions of diagnostic accuracy, the usual outcome measures are sensitivity and specificity, which are of no direct value for the patient (Chapter 7).

Treatment duration and follow-up time

A structured review question must often include treatment duration and follow-up time. For instance, short-term studies may be considered irrelevant for treatment of chronic conditions. The same applies to questions concerning prevention.

Study design

The study design may be more or less appropriate for answering a question (Facts 3.2). For instance, randomised controlled trials are most appropriate for answering treatment questions as well as for questions on diagnostic accuracy. If numerous randomised trials are identified, this could be a reason for not including other study designs in the project. In the case of newer methods and most diagnostic studies it may be necessary to accept less reliable study designs. Questions concerning rare adverse effects, or disease risk factors, are best answered by controlled, prospective, observational studies (eg cohort studies).

Facts 3.2 Common study designs for investigating different questions. After each question, the study designs with the highest accuracy are listed first.

The question addresses:

Therapy/treatment/prophylaxis

Prognosis

Adverse effects/causal associations

Diagnosis

Screening

Economics

Aetiology

Study type

Randomised controlled trial (RCT), controlled trial, case-control

Cohort

RCT, cohort, case-control

RCT, diagnostic accuracy study,

RCT, cohort, cross-sectional,

Cost-effectiveness analysis

Cohort, case-control

Other inclusion criteria

In addition to PICO, treatment and follow-up times, and study design, it may be necessary to define other inclusion criteria.

It can be valuable to define the setting in which the study will be conducted, eg emergency departments, workplaces, or schools.

It could also be necessary to limit the assessment of literature to studies of populations of a certain minimum size. If possible, such limitations should be supported by analyses of statistical power.

A high drop-out can make the results of a study difficult to interpret since it is often uncertain why people choose to stop participating. Possible reasons can be adverse effects, or no effect. High drop-out rates are particularly common in lifestyle studies where the intervention involves more than just taking a tablet. Drop-out also increases with time, and it can be reasonable to define requirements in relation to the duration of the follow-up.

Other limitations

In practice, further limitations are often needed. The most common are language and publication date.

Language

Without language limitations, the search would encompass studies written in languages other than English. Language limitations are used partly to accommodate the expert group's language skills and partly in view of a need to consider the literature in a particular language. For instance, many studies on alternative medicine are published in Chinese, German, and Italian, and many surgical studies are published in German.

Publication date

For some projects it can be reasonable to limit the literature search to a particular time period. For example, some methods may have been modified to such an extent over time that outcome studies of older versions of the method would not be relevant. Limitations based on publication dates could also be useful when updating earlier reports.

References

1. Boudin F, Nie JY, Bartlett JC, Grad R, Pluye P, Dawes M. Combining classifiers for robust PICO element detection. BMC Med Inform Decis Mak 2010;10:29.
2. Carlberg B, Samuelsson O, Lindholm LH. Atenolol in hypertension: is it a wise choice? Lancet 2004;364:1684-9.
3. Safer DJ. Design and reporting modifications in industry-sponsored comparative psychopharmacology trials. J Nerv Ment Dis 2002;190:583-92.
4. Johansen HK, Gotzsche PC. Problems in the design and reporting of trials of antifungal agents encountered during meta-analysis. JAMA 1999;282:1752-9.
5. Moos RH. Iatrogenic effects of psychosocial interventions for substance use disorders: prevalence, predictors, prevention. Addiction 2005;100:595-604.
6. Ferreira-Gonzalez I, Permanyer- Miralda G, Busse JW, Bryant DM, Montori VM, Alonso-Coello P, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol 2007;60:651-62.

Chapter 4

Literature search

Introduction

The principles of the work on systematic reviews, which aim to minimise the risk of the conclusions being influenced by chance and arbitrariness, are outlined in Chapter 1. One of these principles is the systematic literature search, which aims in turn to capture as many studies as possible that are relevant for the questions at issue. This chapter deals with the literature search as part of the project process, focusing on searching for original articles in international databases with scientific content.

To help capture all relevant studies, complementary methods are used. The search is often made in various databases and, when necessary, in citation databases. A citation search involves starting from a researcher or an article to find out whether they are cited and in that case by whom. Examples of citation databases are Web of Science and Scopus. Another complementary method, often called chain searching, amounts to an analysis of the studies' lists of references. The experts' knowledge of the subject is, of course, also taken into account.

Many HTA organisations, such as Cochrane Collaboration [1], state that hand searches of particular journals should be performed, as well as searches of so-called grey literature,[1] as part of the work on systematic reviews [2]. In a hand search, certain journals that are considered particularly important for the question are searched page by page, which is time-consuming. SBU seldom uses either of these two methods. Some studies indicate that a search of grey literature, such as conference abstracts, does minimise publications bias (only half of the content of these abstracts results in scientific articles) but the information about their methodology is often too scanty [3].

[1] Publications that have not been quality assessed by a publisher, for example reports from public authorities.

Literature search – part of the project process

The work of designing a search strategy is done in cooperation between the information specialist and the experts involved in the review.

Some prominent features of the process are: preliminary search, test search, main search and an updated search when the project is coming to an end. The starting point for the literature search is always the project's question according to the project plan. The advantage of cooperating from the start in the work on the project plan is that a better understanding of aspects of the question makes the information specialists' work more effective. Another equally important point is that the information specialist's knowledge and experience of transforming a question into a search strategy can help to structure and formulate the question. Information about suitable databases related to the question can be presented at an early stage.

Before starting the project

Before a project is started, preparatory work should be done to check whether similar projects are in progress at other HTA organisations and whether other systematic reviews already exist. Important databases are available for this, such as Cochrane Library's databases Cochrane Review and DARE.

Facts 4.1 Databases/websites containing systematic reviews and HTA reports.

- Cochrane Library
    - Contains several databases, among others Cochrane Database of Systematic Reviews, www.thecochranelibrary.com
- PubMed Health, http://www.ncbi.nlm.nih.gov/pubmedhealth/
- The Swedish Agency on Health Technology Assessment
- National and regional HTA reports, www.sbu.se
- The Norwegian Knowledge Centre for the Health Services (NOKC)
    - National HTA organisation (Norway), www.kunnskapscenteret.no
- National Institute for Health Research Evaluation,
- Trials and Studies Coordinating Centre
- National HTA organisation (Great Britain), www.hta.ac.uk
- Canadian Agency for Drugs and Technologies in Health
- National HTA organisation (Canada), www.cadth.ca

Test search

When the decision has been taken to start a project, the information specialist and the project leader formulate search strategies for test searches and run these. Among other things, the test searches aim to investigate the following questions:

- How are relevant studies indexed and what terms occur in the titles and the abstracts?
- Is the question sufficiently well-defined, or does it need further clarification?
- How large a quantity is the search likely to yield?

The project's experts have the very important task of providing the information specialist with "core articles". Index terms and abstracts are used to develop the search strategies. The experts can also contribute terms and expressions from the subject field and judge whether the search results match the project's question(s) or whether the search strategy needs to be adjusted.

The collaboration between the information specialist and the experts can be carried out in various, sometimes complementary, ways, such as physical or online meetings, where search strategies and search results are discussed. The experts can also be given an opportunity to browse through the preliminary search results themselves. This can be done through Collections from PubMed or in the form of a library from a reference system. They can then convey their views to the information specialist. The project leader's role in this cooperation can vary, but it is important that the leader is well-informed about the progress of the work.

Main search

When the search strategy has been thoroughly prepared, the search is initiated in the database that has been chosen to be first. For projects and questions in the field of medicine, the test search is done in PubMed, which is also the first database to be used in the main search. For questions in other fields, other databases may be the first choice.

The next step is to adjust the search strategy to the remaining databases. Search strategies and searches in health economics and ethical issues are then formulated and performed. All search strategies and search results are carefully documented as an important part of SBU's demand for reliability and transparency.

The search results are imported to a reference management programme where checking for duplicates is done. When all searches have been performed and all duplicates removed, it is time for the manual assessment of retrieved abstracts. The search identifies references but the task of deciding their relevance in relation to the question has to be performed manually.

Updated search

In the event of a long interval between the main search and the publication of the report, the search should be updated. This is done to identify the very latest published studies and have them included in the report.

Designing the search strategy

As described in previous chapters, a properly structured and defined question is crucial for an effective literature search. Structuring a question is simply a matter of breaking it down into components and analysing each component separately, after which the decisions that have been made should be documented in the project plan.

From PICO to search

As an aid to structuring the question, the acronym PICO (population, intervention, comparison/control, outcome) is used for studies on intervention and diagnosis. For studies based on qualitative data, the project's questions can be structured using the acronym SPICE (setting, perspective, intervention, comparison, evaluation).

Structuring the questions involves formulating criteria for inclusion and exclusion. Any limitations are also decided on, such as confining the search to a certain period, language or study design. All this has a bearing on how the search strategy is formed. While the search strategy is based on the PICO of the question, it is important to note that this does not mean that every part of a PICO/SPICE should always be included in the search strategy.

Building-block strategy

The so-called building-block strategy is commonly used when formulating a search strategy. Each part of the PICO that has been selected for use in the search strategy usually corresponds to a block of search terms and search phrases. Parts of the questions sometimes correspond to two blocks in the search. For instance, if the question concerns the population "elderly persons" with urinary incontinence, this will correspond to two blocks: one for elderly persons and one for urinary incontinence. A separate search is made for each block and they are then combined to form a final search result.

Boolean operators for combining search words

The separate blocks of terms and expressions that are included in the search are created by combining concepts and terms with a boolean operator. The boolean operators AND, OR, NOT are programmed to give the database specific instructions and should not be confused with the ordinary meanings of these words.

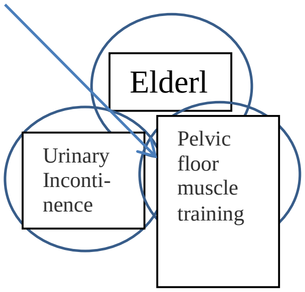
Within each block of search terms, synonyms and other closely related terms are combined with the boolean operator OR. Inserting the operator OR between each search term in a block instructs the database to search either one or the other search term or all terms that occur in the block. Using OR is a safeguard for the multitude of expressions that may be used for one and the same disease, intervention, etc. in an abstract. This extends the search compared with using a single search word.

When each block of search words is searched for, the blocks are combined with the boolean AND between them. Here the instruction to the database is that at least one word from each block must be found in each reference in the search results. This specifies the search and narrows its results.

One way of combining blocks is to use the respective database's search history function.

The boolean operator NOT is used to instruct the database that something should not be included in the search result. This operator is usually used very cautiously.

Search results



Search results

Elderly

Urinary incontinence Pelvic floor muscle training

Study design

Figure 4.1 Search result with the boolean operator AND.

Parenthesis search

Parentheses are used in a search strategy that includes different boolean operators to decide the order in which the database is to search for terms and operators.

> Example: incontinence AND (urine OR urinary OR stress OR urge)

The parentheses instruct the database to start the search inside the parentheses. The results are then combined with the search term “incontinence” and a boolean AND.

Different types of search words – indexing words

In order to capture as many relevant studies as possible, a search strategy for a systematic review consists of a mixture of indexing words and free text words.

The indexing words are obtained from the specific alphabetical, hierarchically formed glossary, thesaurus, that is a part of every large international subject database. Medline’s (PubMed) thesaurus is called MeSH and PsycInfo’s is called Thesaurus of Psychological Index Terms. Since different databases use different concepts and expressions, indexing words and controlled search words, every search strategy has to be reformulated and adapted to each specific database.

A majority of the articles in a database are indexed, that is, an indexer adds a number of terms from the thesaurus to each article. These indexing words aim to describe the content of the article and sometimes also descriptions such as the study design or publication type. A thesaurus is designed to provide a uniform way of naming the contents of a database and at the same time create relations between the concepts in the hierarchical system.

Free text words

The other type of search word, free text words, is chosen to match words that occur in the database’s description of each specific study. Here one can decide where in the description the word is relevant. The occurrence of free text words is commonly limited to the references’ titles and abstracts.

Pros and cons of indexing words and free text words

The advantage of using the databases’ indexing words is that they are uniform, that is, each reference obtains a number of distinct terms aimed at describing the article’s contents. An abstract can also describe an article’s contents but searching for words in descriptive texts can result in irrelevant hits. Using indexing words does away with the need, when using free text words, to consider synonyms and spelling variations. A disadvantage can be that indexing words are sometimes too general for the question at hand.

It is also important to consider how the indexing of an article is affected by the choice of title and how the abstract is written. Another matter that has to be considered in connection with incorrect indexations is the human factor.

The advantage of free text words is that studies which have not yet been indexed can be caught. Using only indexing words is not sufficient for capturing the most recently published articles in, for example, important PubMed. A combination with free text words is necessary for this. Free text terms can also be helpful when the database's indexing words are too general to fit the question at issue.

Limitations

The work on a question's PICO also involves deciding limits to the question and whether they should be included in the search strategy or applied manually when assessing the abstracts.

Limits may concern the population's age, gender, language, time, study design, etc.

International databases have built-in functions, Limits, for this. In some databases, such as PubMed, the use of certain Limits is equivalent to searching with MeSH terms, which means that new articles which have not yet been indexed with these terms are missed. This applies to functions such as Ages, Article Type and Species. Other limits, such as language and time, are not linked to MeSH, so in such cases even unindexed articles will be caught.

Decisions about limits are a matter for the project group

The experts' knowledge of developments in the research field is of great importance for setting suitable limitations. The decisions are made in the light of the SBU's staff's knowledge and experience of creating systematic reviews, particularly the information specialists' knowledge about databases and information retrieval. The time frame and the resources for the project also have to be considered.

Language: limitations can be made to certain languages in the databases. The project group must decide whether there is a need to obtain an overview of studies written in other languages than English (the abstract is always in English but not the study itself) or whether limits should be imposed.

Time period: There may be reasons to limit the search to a particular time period. When a search is updated, the earlier search is supplemented.

Study design: The project group must also decide whether to include the study design in the search strategy rather than as inclusion criteria that feature in the assessment of abstracts.

Search filters, or hedges, are an aid in the search for a certain type of study, such as study design. Search filters are validated, that is, they are controlled to find as many relevant studies as possible while minimising the capture of irrelevant studies. The search filter is adapted to different versions of a database, for example Medline, as well as to different databases. The search filter is combined with the other blocks in the search strategy.

Search filters are collected, assessed and published by the InterTASC Information Specialists' Sub-Group (ISSG)[4].

Here is a summary of some points to consider when creating a search strategy for a systematic review:

- Create search blocks consisting of both indexing words and free text words
- Search as few parts of PICO as possible, sift out the rest when screening abstracts
- For some questions, *one* part of PICO corresponds to several blocks
- It is usually appropriate to search for the population and the intervention

The extent of the literature search – striking a balance

Systematic reviews are hopefully based on all the relevant existing literature. The optimal literature search for such a project would be one that captures all relevant studies and nothing else, that is, a search precision of 100 percent. In practice, this is of course not possible. There are various approaches to a literature search, which can be anything from comprehensive/sensitive to narrow. The extensiveness of a search is often determined in practice by how many hits the search generates, since the results of the search (in the form of listed abstracts) are always gone through manually.

Facts 4.2 Concepts for describing search results [5].
Precision = The number of relevant articles as a proportion of the total number of retrieved articles.
Recall= The number of relevant hits as a proportion of the total number of relevant articles.

Narrow search

A literature search that, for example, searches for two words in the article's title and combines these with a boolean AND, will of course yield just a few hits and most of those found will probably be relevant. At the same time, the search will probably miss much of the relevant literature because it did not take variations in the terminology into account. A narrow search is therefore often high on precision but may, of course, be "off target". A narrow search is usually inadequate for a systematic review but can serve its purpose for other literature searches.

Comprehensive search

A systematic review should aim for a comprehensive search: a search strategy that considers varying indexations, insufficient indexations, and non-indexation of some studies. A sensitive search is intended to achieve a high recall, to find as many as possible of the existing studies that answer the question. Of course, the search has to be made without knowing how large a proportion of the relevant studies it will capture. Nevertheless, a comprehensive search increases the chances of finding most of them. The disadvantage is that the number of irrelevant hits increases with the breadth of the search.

There are various ways of widening or narrowing the literature search; some suggestions to make the search more sensitive are given in Facts 4.3.

Facts 4.3

Suggestions for comprehensive searches, suitable for systematic reviews

- Search with both indexing words and free text words.
- Consider that a thesaurus changes over time.
- There may be different ways of indexing the related or kindred phenomena.
- Search in a larger number of relevant databases
- Search with few blocks (often blocks for population AND intervention).
- Add alternative spellings and inflected forms for free text words-.
- When suitable, truncate the free text words, that is, search for word stems that end with a truncation sign (usually *). But check and refrain from truncation if it results in too many hits.

The difference between comprehensive and narrow searches is illustrated in Figure 4.2.

All literature
Comprehensive search
Narrow search
Relevant literature

Comprehensive search
+ Captures most of the relevant literature
– Can result in rather a lot of irrelevant hits

Narrow search
+ Often high precision
– Loss of relevant literature

Figure 4.2 The difference between comprehensive and narrow searches.

Number needed to read

The precision of a search can also be expressed as the “number needed to read” (NNR), that is, the number of abstracts that need to be read in order to find a relevant article (NNR=1/precision). NNR is determined not only by how comprehensive/narrow the search is but also by the extent of the research field and by how well the question has been limited. If the aim of the project is to answer a question involving few published studies, a comprehensive search tends to be fairly straightforward. It does not risk missing relevant articles, neither does it require a great deal of work for those screening the captured references.

If, on the other hand, the project’s question/questions makes a large number of published studies relevant, the extent of the search becomes an even more important issue. How many abstracts are the rewievers willing to read to ensure that nothing is missed?

The balance between a narrow and a sensitive search is thus an issue about not missing relevant studies, time, how many people are involved in the project and where the work load lies. It may sometimes be quicker and easier to screen a large number of references compared with the time it takes to narrow the search in a way that does not lead to too many relevant studies being missed. At the same time, problems also arise with the alternative where too many hits have a high NNR (that is, a large number of irrelevant articles have to be read in order to find one that is relevant). It can be hard for people to remain concentrated when assessing a large number of abstracts, so even relevant studies may be sifted out by mistake. However, screening a list of abstracts may not take all that long even though it seems cumbersome at first sight.

“At a conservatively-estimated reading rate of two abstracts per minute, the results of a database search can be scan-read at the rate of 120 per hour (or approximately 1000 over an 8-hour period)” [6].

Choosing databases

If the aim is to find as many relevant studies as possible that matches the question, a number of reports indicate that the search needs to include more than one database [7]. Which and how many databases should be used depends on the subject of the question.

According to the checklist used to assess the quality of systematic reviews, AMSTAR (Appendix 6), to be considered sufficient, a thorough search needs to include at least two databases. At SBU at least three databases are searched; for questions in the field of medicine, it is usually sufficient to search in

PubMed, Embase and the Cochrane Library. For questions of a more multidisciplinary nature, the choice of databases should be considered in relation to this.

Note that even if a reference is included in a database, it may not be easy to find. Complementary searches in other databases can be worthwhile since the same reference may be indexed differently in different databases.

Facts 4.4 Examples of bibliographic databases of importance for systematic literature reviews in the health care field.

PubMed (www.ncbi.nlm.nih.gov/pubmed)

PubMed is a general medical and biosciences database. Its main component, MEDLINE, contains approximately 25 million references to articles and a selection of full-text articles from more than 5,000 biomedical journals (2016). The database gives broad coverage in the field of heath care. The articles in Medline are indexed according to the database's thesaurus MeSH (Medical Subject Heading). In addition, a growing number of articles awaiting indexation in PubMed are recognised through the comments "PubMed-in process" alternatively "Supplied by publisher". The database is produced by the US National Library of Medicine and is free of charge via the internet.

Embase (www.embase.com)

Embase is the other large database in the field of medicine and contains about 30 million references from 8,500 journals (2016). With Embase the search can be integrated with the Medline database but Embase does not have either the rest of PubMed's contents or the MeSH-database. Embase has a developed thesaurus, Emtree, considered particularly suitable for pharmaceuticals. As in PubMed, articles "In process" can be searched for, but also "Article in press", that is, articles not yet published. Embase is produced by the scientific company Elsevier and covers proceedings and more European journals than the American PubMed. The database charges a fee.

Cochrane Library (www.thecochranelibrary.com)

The Cochrane Library contains several databases. Besides the Cochrane Database of Systematic Reviews, containing Cochrane's own systematic reviews, the other databases include the Cochrane Central Register of Controlled Trials (Central), the Cochrane Methodology Register and the NHS Economic Evaluation Database (NHS EED). NHS EED and DARE (another part database in the Cochrane Library)[1] are both produced by CRD, the Centre for Reviews and Dissemination. The Health Technology Assessment Database assembles new and ongoing projects outside the Cochrane Collaboration. The database charges a fee.

Facts continue on the next page

[1] NHS EED and DARE are not updated in Cochrane Library since April 2015.

Facts 4.4 continued

CINAHL (www. ebscohost.com/biomedical-libraries/the-cinahl-database)
CINAHL (Cumulative Index to Nursing and Allied Health Literature) is a database for articles, books and dissertations about nursing, physiotherapy, occupational therapy, etc. It contains about 2.9 million references from about 3,000 journals. The database is distributed by EBSCO and charges a fee .

PsycInfo (www.apa.org/psycinfo)
PsycInfo, a database in psychology, behavioural sciences and related fields, provides references to about 3 million references from about 2,500 journals, books and dissertations. PsycInfo is produced by the American Psychological Association (APA) and charges a fee.

Documentation

Documenting a database search is important in order to be able to repeat the procedure. The documented search strategy should be available to readers of the systematic review. There is no common standard for how to document a search strategy but the following information should be reported:

- The databases's names
- The databases's producers
- Date of the search
- Exact search terms and the types of term, that is, indexing words and free text words
- Any limitations in the search
- How the terms were combined.

SBU's checklist for the documentation of searches is presented in Example 4.1.

Any complementary search methods that have been used for identifying relevant literature, such as hand search and chain searching, should also be reported.

Managing references

A powerful reference management programme is necessary to cope with the large numbers of references related to work with systematic reviews. Such programmes are EndNote and Zotero amongst others. The programme makes it possible to import all references from the database search to a library that is specific for the question or the whole project.

Example 4.1 Documenting a search.

Pubmed via NLM 17 November 2011

Title: Pelvic floor muscle training as an intervention for elderly with urinary incontinence

Search terms		Items found
Population: aged		
	"Aged"[Mesh:NoExp] OR "Aged, 80 and over"[Mesh] OR "Frail Elderly"[Mesh] OR Geriatrics[MeSH] OR Homes for the Aged[MeSH]	2038796
	(older patient*[ti] OR older adult[ti] OR older adults[ti] OR older women[ti] OR older men[ti] OR geriatric[ti] OR geriatrics[ti] OR elderly[ti] OR elders[ti] OR Vulnerable elder[ti] OR Vulnerable elders[ti] OR senior[ti] OR seniors[ti] OR community-dwelling[tiab] OR nursing home[ti] Or nursing homes[ti] OR care home[ti] OR care homes[ti] OR oldest old[ti] OR frail[ti]) NOT medline[SB])	7972
	1 OR 2	*2046528*
Population: urinary incontinence		
	Urinary Incontinence[MeSH:NoExp] OR Urinary Incontinence, Stress[MeSH] OR Urinary Incontinence, Urge[MeSH] OR Nocturia[MeSH] OR Urinary Bladder, Overactive[MeSH] OR "Diurnal Enuresis"[Mesh] OR overactive bladder[tiab]	25556
	(Mixed incontinence[tiab] OR Stress incontinence[tiab] OR Stress urinary[tiab] OR overactive bladder[tiab] OR bladder overactivity[tiab] OR bladder control[tiab] OR urge to void[tiab] OR (Incontinence[ti] AND (urine[ti] OR urinary[ti] OR stress[ti] OR urge[ti])) NOT medline[SB]	1146
	4 OR 5	*26393*
Intervention: pelvic floor muscle training		
	(Pelvis[MeSH:NoExp] OR Pelvic Floor[MeSH]) AND (Muscle Contraction[MeSH] OR Exercise Therapy[MeSH:NoExp] OR Physical Therapy Modalities[MeSH])	1407
	pelvic muscles exercise*[tiab] OR Pelvic muscle exercise*[tiab] OR Bladder and pelvic muscle training[tiab] OR pelvic floor muscle training[tiab] OR pelvic floor re-education[tiab] OR pelvic exercise*[tiab] OR pelvic floor training[tiab] OR pelvic muscle precontraction[tiab] OR pelvic floor exercise*[tiab] OR pelvic muscle re-education[tiab] OR (pelvic floor[ti] AND (training[ti] OR exercise*[ti] OR education[ti]))	1040
	7 OR 8	*1972*
Combined sets		
1.	**3 AND 6 AND 9**	**350**

The search result, usually found at the end of the documentation, forms the list of abstracts.

[MeSH] = Term from the Medline controlled vocabulary, including terms found below this term in the MeSH hierarchy
[MeSH:NoExp] = Does not include terms found below this term in the MeSH hierarchy
[MAJR] = MeSH Major Topic
[TIAB] = Title or abstract

[TI] = Title
[AU] = Author
[TW] = Text Word
Systematic[SB] = Filter for retrieving systematic reviews
* = Truncation

References

1. Cochrane Collaboration. [2012; cited 14th April 2016]. Available from: http://www.cochrane.org/

2. Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. [Updated March 2011, cited 14th April 2016]. Available from: http://handbook.cochrane.org/

3. Dundar Y, Dodd S, Dickson R, Walley T, Haycox A, Williamson PR. Comparison of conference abstracts and presentations with full-text articles in the health technology assessments of rapidly evolving technologies. Health Technol Assess 2006;10:1-145.

4. Centre for Reviews and Dissemination (CRD): The InterTASC Information Specialists' Sub-Group Search Filter Resource [cited 14th April 2016]. Available from: http://www.york.ac.uk/inst/crd/intertasc/index.htm

5. Shariff SZ, Cuerden MS, Haynes RB, McKibbon KA, Wilczynski NL, Iansavichus AV, et al. Evaluating the impact of MEDLINE filters on evidence retrieval: study protocol. Implementation Sci 2010;5:58.

6. Lefebvre C, Manheimer E, Glanville J. Chapter 6.4.4: Sensitivity versus precision. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. [Updated March 2011, cited 14th April 2016]. Available from: http://handbook.cochrane.org/

7. Lefebvre C, Manheimer E, Glanville J. Chapter 6.1.1.2: Minimizing bias. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. [Updated March 2011, cited 14th April 2016]. Available from: http://handbook.cochrane.org/

Chapter 5

Assessing a study's relevance

The selection of studies for inclusion in a review begins with an assessment of the studies' relevance to the question of interest, that is, how well they meet the pre-specified inclusion criteria. The aim is to identify all studies relevant to the review question. Only studies that are considered relevant are retained for the assessment of quality.

The relevance of studies, unlike their quality, is not graded. In other words, a study is either relevant or irrelevant for the question.

The assessment of the quality of relevant studies includes a judgment of their external validity or generalisability. This should not be confused with the assessment of relevance.

Two steps for assessing relevance

Assessing relevance is done in two steps (Figure 2.2). Each step should be made by two assessors working independently. Each step should be carefully documented. Step one consists of a screening of titles and abstracts based on the pre-specified selection criteria. In the event of uncertainty at this stage it is often better to give the article the benefit of the doubt. If at least one of the assessors judges an article to be of possible relevance, it is obtained in full-text.

Step two involves assessing the full-text articles for their relevance (see the check list, Appendix 1). Studies judged to be relevant are retained for the assessment of their quality (Chapters 6–7). Studies that are not considered relevant at this stage are excluded. Each excluded study and the primary reason for exclusion must be documented. In the event of disagreement, the study is discussed in the project group.

Documentation of excluded studies

The selection process is reported in a flow chart where the number of abstracts and full-text articles should be traceable to the literature search (Figure 5.1). The high demands for transparency in the production of a systematic review makes it important to document the reason why a full-text article is not included in the review. The most common reasons can sometimes also be stated in the flow chart. References of excluded full-text articles are usually reported in an appendix together with the reason for exclusion. Some examples are given in Facts 5.1.

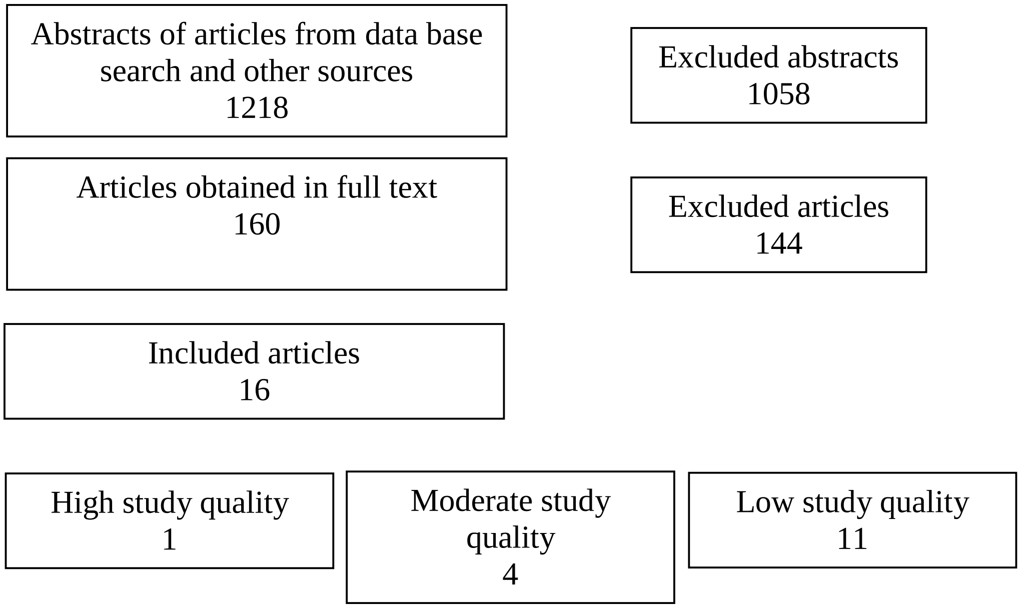


Figure 5.1 Example of a flow chart showing the number of included/excluded articles.

Facts 5.1 Examples of reasons for excluding studies:

- Irrelevant study design
- Background material
- Irrelevant population
- Irrelevant intervention
- Irrelevant control intervention
- Irrelevant outcome
- Too short follow-up time
- Duplicate publication
- Baseline data not fully reported

Chapter 6

Assessing study quality

This chapter describes different types of intervention studies, the methodological problems associated with each type and how study quality is assessed. The chapter is ended with methodological problems associated with systematic reviews and how these reviews are appraised.

Randomised controlled trials (RCT) are generally considered to provide the most reliable evidence on the effectiveness of interventions. Compared with other study designs, they offer better possibilities to control for baseline factors that are not involved in the intervention itself. When it comes to appraising risks, observational studies including case-control studies may be preferable. Cross-sectional studies or case series without controls are less reliable and are seldom included in SBU's projects.

Checklists

In order to assess the credibility of intervention research, SBU use checklists that appraise the presence or absence of several types of bias found in reported research (Appendices 2-3). Bias refers to systematic error, meaning that multiple replications of the same study would reach the wrong answer on average. These refer to selection, detection, and performance; attrition; reporting; publication; and conflict of interest.

In practice the questions are often difficult to answer, eg the reporting may be incomplete. Hence it is important for the project group to practice together in using the checklists. This provides opportunities to address uncertainties and how to interpret the questions. To ensure that the assessments are conducted uniformly, kappa-calculations can be made at some points during the project.

Before starting the assessment process, the project group should determine whether certain aspects are particularly important for quality and whether other aspects are irrelevant.

Studies seldom meet all the quality criteria and the picture will differ from case to case. The checklist reveals a study's shortcomings and the reviewers then must determine the extent to which these shortcomings affect the study's validity.

The checklists are used as a *support* for assessing a study's quality. They are not intended to summarize the number of "yes" answers as an indication of the study's quality. The completed checklists are to be kept including notes of the specific reasons connected with the assessment. This facilitates writing the report at a later stage.

Part A of the checklist (A1-A6) is used to assess the risk of bias in a study. To be able to summarise the results and rate the outcomes in terms of quality according to the GRADE approach, additional information is needed. This concerns the compilation of the following aspects of all the included studies: insufficient agreement between studies (B), the studies' applicability (C), the assessment of the studies' precision (D), publication bias (E), effect size (F), dose-response relation (G) and the possibility that the effect is underestimated (H). The quality according to GRADE is assessed in a later stage of the process but it can be appropriate to comment on these factors already when reading the individual studies (B-H).

Randomised controlled studies (RCT)

The CONSORT statement, is an international agreement on the reporting of RCTs (1). Thus, the CONSORT statement is a guideline for the aspects of a study that it is important to assess. SBU's checklist (Appendix 2) is therefore based on the CONSORT statement.

Selection bias

A major threat to the internal validity of a trial is that experimental and control groups differ at start, that is selection bias. The best way to refute this is by randomizing the participants to the conditions. With a proper randomization, risk and protective factors will be equally distributed among groups pending the study is large enough, and the only systematic difference will be the treatment.

The randomisation procedure should therefore be described in detail, which is often not the case in old studies where the randomization procedure may be lacking. The extent to which the lack of this information affects a study's quality must be decided from case to case. The study groups' comparability at baseline should however always be evaluated.(Example 6.1). In small studies, randomization may lead to imbalances at baseline. The risk of selection bias is dealt with in the checklist (Appendix 2, A1 Selection bias).

Example 6.1 An older control group

In a study on whether treatment with the alpha-blocker alfuzosin increased the chance of successful removal of a urinary catheter, the patients in the control group happened to be five years older than the treatment group (2). Since age appeared to be the strongest factor for predicting unsuccessful use of a catheter, irrespective of treatment, the result was difficult to appraise.

Performance bias

All interventions need to be described in detail, not only what is planned but also what is implemented. This is especially important with complex interventions; examples are frequent of professionals that drop parts that are considered less relevant. Lack of blinding may lead to risk of performance bias if researchers and participants are aware of the treatment. An example could be that the control group and intervention group to a different extent are exposed to another treatment that may influence the study result. Other supportive treatment could for example affect study results for hospital inpatients if the intervention is unblended.

Questions related to performance bias, blinding and compliance are included in the checklist (Appendix 2.A2 Treatment bias).

Detection bias

Expectations about the treatment effects can also influence the outcome measure. This is particularly important for outcomes like quality of life or estimation of the burden of symptoms. Outcomes like survival or fractures are less likely to be affected. As many as possible of the parties involved in a study should therefore be blinded, that is, they should be unaware of which treatment is being given (experiment or control). The ideal situation is that all parties (caregiver, patient, the one who measures the effect and the one who evaluates the results) are blinded, so called triple blinding.

Blinding can be difficult in practice. This applies, for example, in surgery, physiotherapy, psychotherapy, and social service interventions. Provided it is considered ethical, so-called sham surgery (simulated surgery) can be used to minimise systematic errors due to the patient's expectations. Even if it is not possible to blind the caregiver/administrator and the patient/client, those who register and evaluate the results can be blinded. This type of blinding can be considered as a minimum requirement. Interventions in the social service field are most often not possible to blind but the outcome assessor may be blinded.

(Appendix 2, A3 Detection bias)

Attrition bias

An important aspect of study quality is attrition (Appendix2, A4 drop-out bias). The credibility is affected by the extent to which those included in a study are followed-up all the time and can be included in the analysis. The results from a study with a large drop-out rate is not trustworthy. It is possible that individuals experiencing no symptom relief are more prone to leave the study than those with an improvement or individuals with side effects may also drop out. A drop-out rate that differs between the experimental and control groups is particularly problematic. For studies on pharmaceuticals, SBU uses the following approximate limits:

- <10%: drop outs hardly affect credibility and does not affect the study's quality
- >30%: drop outs may seriously affect credibility to such an extent that information from the study is of no value. The study is excluded.

The limits of acceptance of drop out and to what extent it will influence the study quality should be defined in advance by the study group. The assessment can sometimes be moderated if the authors can argue, for instance from an analysis of the drop-out rate, that this did not affect the results. The dropout rate should always be evaluated in relation to the event rate in the study. In studies using continuous variables or scales, the calculation *LOCF* (last observation carried forward) is sometimes used to compensate for attrition. The latest measured observation is then considered valid also at later points in time for which data are missing. Other methods are also available to compensate for attrition. Sensitivity analyses can indicate whether the effect remains under the worst possible conditions. One such condition is for instance that none of the drop-outs showed an improvement. Did the effect remain even in the worst possible scenario?

Reporting bias (selective outcome reporting)

It should be assured that the main result as pointed out by the authors is based on the primary outcome measure. If there is no statistically significant result for the primary outcome, the authors may report another endpoint that turned out significant. The authors may also perform various types of ad-hoc subgroup analysis to find significant results. (Appendix 2, A5 selective outcome reporting)

Conflicts of interest

The studies are also assessed for financial or other types of conflicting interests that might have biased the results (Appendix 2, A6 Conflicts of Interest). Such conflicts of interest are documented for methods sponsored by the industry (pharmaceuticals, medical devices) as well as when researchers investigate a method which they themselves developed.

Observational studies

Observational studies (quasi experimental studies) can be planned and conducted in various ways; methodologically they can be divided into non-randomised comparative studies, cohort studies, case-control studies and cross-sectional studies. Cohort studies that are included in SBU's assessments should be controlled, that is, they should have a comparison group.

Guidelines and advice on how to rate cohort and other observational studies have been developed internationally, for example STROBE [3]. The checklist for observational studies used by SBU is presented in Appendix 3. In principal, risk of bias is the same as for randomised studies but risk of selection bias is more pronounced (see below).

For observational studies, problems with quality tend to be more serious for small studies, studies with historical controls and studies that have not adjusted for confounders.

Cohort studies

Controlled cohort studies compare one group receiving an intervention or being exposed to a risk with another group receiving an alternative or no intervention or not exposed to the risk. In a cohort study a group of individuals is followed prospectively to see what happens to them. A great deal of information about the individuals in the two groups is usually collected at the start of the study. Examples of important basic information are age, gender, socio-economic situation, living habits, and diseases. The relation between smoking and several severe diseases was demonstrated in the 1950s in a cohort study that followed what happened to British doctors who either smoked or did not smoke [4]. Considering how smoking habits have decreased in the light of this knowledge, this is probably the study that has saved most lives. Our knowledge of risk factors for cardiovascular diseases, such as high blood pressure, high cholesterol-levels, smoking and obesity, comes from several large cohort studies such as the Framingham study in the USA and "Men born in 1913 in Göteborg".

Drawbacks with cohort studies are that they can be costly and difficult to perform in the case of rare diseases or when a long time must pass before an outcome can be measured. Such studies require large populations and long follow-up times. A suitable and cost-effective alternative is case-control studies (see below). Well-developed health data and quality registries are already available for large study populations in Sweden and other Nordic countries. This enables cohort studies to be used for rare diseases and when long follow-up times are needed (Example 6.2).

Example 6.2 Register-based cohort study

A Danish register-based cohort study analysed the risk of autism among children who have been vaccinated for measles, mumps, and rubella. More than 537,000 children were followed for eight years. No increased risk of autism was found for vaccinated children compared with children who had not been vaccinated [5].

Case-control studies

Case-control studies are backward-looking (retrospective). Here, the individuals (cases) with a particular outcome (eg disease or death) are identified and compared with a control group of individuals who do not have that outcome. It is then possible to study whether the two groups differ regarding exposure for risk factors or the type of intervention they have received. Cases and controls must represent the same study base, and the controls must be selected entirely independently of possible exposure to the intervention.

The data usually must be collected retrospectively. This can be done by interviewing cases and controls or by using patients' records or register data (Example 6.3). The problem with interviews is that people do not always remember everything that happened earlier in life and tend to recall and report memories differently depending on whether they belong to the patient or the control group.

Example 6.3 Case-control study

A Swedish case-control study investigated whether aspirin and NSAID preparations can reduce the risk of stomach cancer [6]: 567 persons with stomach cancer and 1,165 persons without stomach cancer (controls) were interviewed about, among other things, their use of analgesics. After controlling for gender, age and socioeconomic status, the study showed that aspirin-users had a reduced risk of developing stomach cancer (odds ratio=0.7). The risk decreased with an increased use of aspirin. No association was found between stomach cancer and other pain-killers.

The choice between a case-control and a cohort study depends on practical and financial circumstances. Given adequate data collection from the start of the study, cohort studies can often elucidate many different hypotheses. When it comes to rare outcomes, case-control studies tend to be more effective than other study designs, but are less effective for demonstrating associations of rare exposures.

Cross-sectional studies

Cross-sectional studies collect information about interventions (past or present) and current health outcomes on a particular point in time. They are a good way of estimating the prevalence of diseases and exploring any co-variation between exposures and diseases. An example of a cross-sectional study is Statistics Sweden's (SBC's) investigations of living conditions, where a random sample of Swedes are interviewed annually about several conditions. These data have been used to estimate disease prevalence that are not easy to estimate from register data. SBU's projects often rule out cross-sectional studies because the difficulty in determining the temporal and causal association between an intervention/exposure and outcome.

Assessing the risk of bias in observational studies (see Appendix 3, checklist)

One main methodological problem with observational studies, compared with randomised controlled studies, is that the former should control for *all* potentially important factors that might influence the results. In observational studies, study groups may differ initially in important respects that can influence the outcome (selection bias). If these differences are measured reliably at the individual level, they can be controlled for statistically; if they are not measured, one cannot claim that any observed effect depends on inter-group differences rather than on the intervention in question.

Selection bias

The allocation of individuals to either the experimental or the control group can be influenced by both the professional (eg, doctor or social worker) and the individual. The professional can select individuals who are considered to benefit most from the intervention; these individuals can be either healthier (less problems) or sicker (more problems) than those who will not receive the intervention. Well-informed individuals, who tend to be healthier or have less problems, may know more about alternative interventions and therefore demand specific interventions of which other individuals are unaware. The quality assessment of observational studies must therefore focus on whether the groups are initially comparable (Appendix 3, A1 Selection bias). It is often particularly important to adjust the analyses about the study participants health and socio-economic status.

The risk of selection bias in observational studies can never be ruled out entirely. It is, however, less likely to occur when it comes to unknown or unexpected adverse effects [8]. Among other things, this might be because in such cases the professional is less aware of risks and is consequently less prone to assign participants to the experimental or the control group based on risk. Example 6.4 illustrates how the results can be influenced by selection bias. A systematic meta-analysis on adverse effects showed that there was no difference in the effect size between randomised and observational studies [9]. The authors concluded that systematic reviews on adverse effects should not be restricted to study design.

Confounders

Un-measured variables may cause effects rather than the focused intervention. Thus, those confounding variables, may be a major problem in observational studies.

Confounding variables can be accounted for in various ways. The study can be designed so that the material is limited in a way that rules out problems with confounding. This can be done by excluding persons with one of the two factors that co-vary. Other ways of controlling for confounders are to stratify or match the material or to use different types of multivariate statistical methods. Stratifying the material involves dividing it into subgroups with different exposures to reduce the risk of confounding; examples of such variables are gender, age and smoking. Matching involves choosing persons who are the same as regards the confounders that need to be controlled for. Statistical methods, for example regression modelling, can be used to keep confounding variables constant. Another method for controlling for confounders is propensity score [10]. "Residual confounding" may always exist even if attempts have been made to control for several known confounders.

Methodological reasons indicate that, if anything, associations in observational studies are underestimated. If the misclassification of exposure, for example levels of blood pressure, is independent of the outcome, the strength of association will always be underestimated, so called regression dilution bias [11]. This applies to cohort studies where information about exposure is collected before the start of the observation period. Moreover, data are probably classified more correctly in randomised controlled studies than in observational studies.

Example 6.4 Risk of selection bias and the need to control for confounders.
The risk of selection bias in observational studies can be illustrated by a much-cited example concerning oestrogen treatment and the risk of cardiovascular disease. Several observational studies had shown that oestrogen treatment reduced the risk of cardiovascular disease; this applied also when the authors of some studies tried to adjust for some risk factors. These findings were contradicted in a large randomised study: when the risk was adjusted by considering socio-economic differences, it was no longer reduced [12]. This was not surprising considering that several other studies had shown that well-educated women received oestrogen treatment to a greater extent than other women. The spurious conclusions could probably have been avoided if socio-economy had been included in the observational studies' analyses. Similarly, when controlled for socio-economy, observational studies were unable to support the hope that vitamins can protect against cardiovascular disease and lung cancer [13].

Studies on adverse effects

Observational studies are valuable for demonstrating negative effects and risks, that is, adverse effects. Such effects are often unexpected and some serious adverse effects are rare but of such dignity that even a low incidence is unacceptable (Example 6.5). Randomised controlled studies are seldom dimensioned to capture such adverse effects, either in terms of the material's size, time of follow-up or reporting routines. Register-based cohort studies and case-control studies with large populations are often good alternatives in these situations.

Example 6.5 The value of observational studies for analysing adverse effects/risks

The drug aprotinin has been used around the world since 1993 to reduce bleeding, for instance in connection with by-pass- surgery. Many small randomised controlled studies did not observe any risks with the use of aprotinin, whereas large cohort- and case-control studies showed an increased risk of renal failure and death [14,15]. More recently, an increased mortality rate within 30 days led to the suspension of a large randomised controlled study and the drug was withdrawn. A few deaths could probably have been avoided if the observational studies had been taken more seriously and if meta-analyses of RCT had been scrutinised more carefully [16]. The European Medicines Agency (EMA) does, however, allow the use of aprotinin on certain indications.

Facts 6.1 The Swedish Medical Products Agency assesses reports on adverse effects and grades the causal relationship of reported adverse effects as a) certain or probable; b) possible; c) not probable, and d) not possible to judge. The grading is based on biological likelihood, possible mechanisms of action, time associations, possible re-exposure and, more seldom, the results of provoking studies where the subjects of an experiment are either exposed or not exposed. The assessments do not attempt to rate the quality of individual studies or the overall quality of evidence.

Tabulating studies

Relevant information is extracted from the studies of high or moderate quality and compiled in tables. The purpose is to give readers of the report an overview of included studies and how they have been assessed. Another purpose is to structure the data to facilitate subsequent work.

The table provides information on each study's reference, questions, methods, selection, execution, results and methodological quality. The tables are in English so non-Swedish speaking individuals can benefit from SBU' s review work; also, as most of the articles are in English, terms used in the articles do not have to be translated. Moreover, the tables are often consulted by persons with a scientific education and experience of reading scientific literature in English. However, any tables in the summary should always be in Swedish. Table 6.1 is an example of a table design which ensures that relevant information is included.

Excluded trials, are tabulated in an appendix.

The table can be used to make a qualitative assessment regarding heterogeneity or large variations between studies. It also gives a summary picture of current knowledge of a specific question.

Table 6.1 Example of a structure for tables.

Author **Year** **Reference** **Country**	**Study design**	**Population characteristics**	**Inter-vention**	**Follow-up period** **Drop out rate**	**Results**	**Study quality and relevance** **Comments**
	(RCT, CT, cohort, case control etc)	Inclusion/ exclusion criteria Setting No at baseline Male/ female	Interven-tion (I) (dose, interval, duration) Control (C) (active, placebo, usual care, etc)	(From baseline to follow-up, or from end of intervention to follow-up) Drop out (%)	Results (I, C) (Absolute difference, HR, RR, OR, p-value, confidence interval for the difference, sensitivity, specificity, observer reliability, cost-effectiveness, etc)	High, moderate or low study quality if appropriate

C = Control; CT = Controlled trial; HR = Hazard ratio; I = Intervention; OR = Odds ratio; RCT = Randomised controlled trial; RR = Risk ratio

The basis for an overall assessment according to GRADE

An overall assessment (Chapter 10) considers any lack of agreement between the studies (B), their applicability (C), their precision (D), publication bias (E), effect size (F), dose-response relations (G) and the probability of an underestimated effect (H).

The assessment of applicability should focus on whether there is sufficient similarity between the populations in the studies and the population on which the report is focussed.

The precision of the data mainly depends on the size of the study, the prevalence of harmful outcomes and the size of an effect. Small studies can be problematic for several reasons. One is the greater difficulty in demonstrating statistically significant results. Another is the risk of such studies being less well planned than larger ones. There is also a greater risk of imbalances in known *and* unknown background factors in small studies. There is also a risk of type-2 error (the null-hypothesis is accepted although false). This means that the authors have not managed to demonstrate a true effect of an intervention since the study population is too small to reveal a statistically significant effect.

It is important that the authors have specified "points in time" for the final analysis and any interim analyses and also how these were handled statistically. Otherwise there is the possibility that the authors added participants successively to the study until they achieved statistical significance.

There are two ways of calculating the effect. In a *per protocol analysis* (also known as the Treatment of the treated), the results include only those persons who completed the study protocol (completers). An *ITT analysis* (intention to treat) is a more conservative method, aimed at reducing the risk of overestimating the results of treatment. It implies that all those who were randomised are followed in their respective treatment group irrespective of whether they received the intended treatment. ITT is the best method for estimating the effect of an intervention. A per protocol analysis is often preferable for measuring adverse effects, since these are measured only for those who received the treatment or were exposed to a risk factor. Otherwise dilution effects can result in possible risks being overlooked. In the case of non-inferiority studies, both per protocol and ITT analysis should be reported.

Effect size (F) and any dose-response relations (G) clearly affect confidence in the results.

Including previously published systematic reviews

A Research question can be answered by make use of a previously published systematic review. Prerequisites are that:

- the research question is entirely compatible with the project’s question
- the systematic review is conducted so to decrease the risk of bias
- the systematic review it is not older than two years.

Each review is carefully scrutinised since systematic reviews often have methodological shortcomings. For instance, a cross-sectional study on the quality of systematic reviews in 2007 found that about 30 percent of the reviews did not include a quality assessment of individual studies [17].

In order to improve the quality of systematic reviews and the reporting of meta-analyses, an international group published a report, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses, www.prisma-statement.org). The first step in the assessment is to decide whether the review is relevant. Do the review’s questions agree with the project’s and are the inclusion and exclusion criteria the same as the project’s? If the review has a different purpose or other criteria it is excluded. The next step is to assess the quality of the systematic review. Among other things, the review should include:

- Clearly formulated questions
- A clear description of the methods used for literature search and selection of articles
- An assessment of the quality of appropriate studies according to inclusion and exclusion criteria
- Tables showing data of included studies
- Compilation and weighting of the studies’ results, using appropriate methods such as meta-analysis
- Formulation of conclusions, indicating that the authors have considered the scientific quality of included studies

Systematic reviews are assessed and valued with the aid of a checklist (Appendix 6), which is based on an internationally developed checklist, AMSTAR [18, 19]. Like other checklists, AMSTAR consists of several questions with alternative answers: “yes”, “no”, “cannot answer” and “not applicable”. The standards for approval of a review’s quality, such as which items must have a “yes” (or “not applicable”) answer, should be decided in advance.

It is important to consider whether the review has captured all relevant articles. A missing published study, indicates an inadequate search strategy. Conflicts of interest can also influence which studies are included, something that is difficult to control for.

An additional measure to ensure quality is to control for facts and interpretations by reading some of the studies that are included in the review. This is because some studies may have been wrongly assessed. Studies that are coded as randomised may on closer investigation be observational studies.

Table 6.2 summarises SBU's requirements for accepting a systematic review to answer a question.

Table 6.2 Assessing the usefulness of systematic reviews and HTA-reports.

Alternatives	Quality assessment (according to Appendix 6)	Conclusion and evidence grading	Action
Questions and criteria for inclusion and exclusion agree with the project's	Accepted	Conclusions are accepted if not contradicted by more recent literature	The review is included and is complemented by any later articles. The review is evidence rated using GRADE
	Not accepted	Conclusions not accepted	The review is excluded. The reference list and other information are used in the project's work
The questions agree but criteria for inclusion and exclusion do not		Conclusions not accepted	The review is excluded. The reference list and other information are used in the project's work

References

1. Schulz KF, Altman DG, Moher D; CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010;8:18. http://www.biomedcentral.com/1741-7015/8/18

2. McNeill SA, Daruwala PD, Mitchell ID, Shearer MG, Hargreave TB. Sustained-release alfuzosin and trial without catheter after acute urinary retention: a prospective placebo-controlled. BJU Int 1999;84:622-7.

3. von Elm E, Egger M, Altman DG, Pocock SJ, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806-8.

4. Doll R, Hill AB. The mortality of doctors in relation to their smoking habits: a preliminary report. BMJ 1954;1:1451-5.

5. Madsen KM, Hviid A, Vestergaard M, Schendel D, Wohlfahrt J, Thorsen P, et al. A population-based study of measles, mumps, and rubella vaccination and autism. N Engl J Med 2002;347:1477-82.
6. Akre K, Ekström AM, Signorello LB, Hansson L-E, Nyrén O. Aspirin and risk for gastric cancer: a population-based case-control study in Sweden. Br J Cancer 2001; 84:965-8.
7. Rosén M, Axelsson S, Lindblom J. Släng inte ut observationsstudier med badvattnet. Bedöm deras kvalitet istället. Läkartidningen 2008;105:3191-4.
8. Vandenbroucke JP. When are observational studies as credible as randomised trials? Lancet 2004;363:1728-31.
9. Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomized controlled trials as compared to observational studies: methodological overview. PLoS Medicine 2011;8:1-13.
10. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265-81.
11. MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, et al. Blood pressure, stroke and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet 1990;335:765-74.
12. Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. Ann Intern Med 2002;137:273-84.
13. Lawlor DA, Davey Smith G, Brucksdorfer KR, Kundu D, Ebrahim S. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? Lancet 2004;363:1724-7.
14. Mangano DT, Tudor JC, Dietzel C. The risk associated with aprotonin in cardiac surgery. N Engl J Med 2006;354:353-65.
15. Schneeweiss S, Seeger JD, Landon J, Walker AM. Aprotinin during coronary-artery bypass grafting and risk of death. N Engl J Med 2008;358:771-83.
16. Rosén M. The aprotinin saga and the risks of conducting meta-analysis on small randomised controlled trials – a critique of a Cochrane review. BMC Health Serv Res 2009;9:34.
17. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemi-ology and reporting characteristics of systematic reviews. PLoS Med 2007;4:e78.
18. Shea BJ, Grimshaw JM, Wells GA, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.
19. Shea BJ, Hamel C, Wells GA, Bouter LM, Kristjansson E, Grimshaw J, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. J Clin Epidemiol 2009;62:1013-20.

Chapter 7

Assessing the Evidence for Diagnostic Tests

Background

Tests, including eg clinical findings, imaging, psychological tests and biochemical analyses, can be used for a variety of purposes. Some examples are to determine prognosis and risks and to monitor the effects of treatment. This chapter focuses on tests for diagnostics.

Diagnostic tests aim to identify individuals with diseases or health problems. A special case of diagnostics is screening. Screening is population-based and the prevalence of the disease usually is much lower than among people who seek health care for a suspected problem. The potential to predict which individuals are sick or healthy therefore is lower in screening and the likelihood of false positives is higher. According to WHO [1] there are requirements for screening that must be fulfilled eg the availability of effective treatments for the disease and that benefits of the screening outweigh the estimated risks and costs.

A systematic review of studies of diagnostic tests differs in several ways from a systematic review of intervention studies:

- Literature searches are often a major challenge since the indexing of diagnostic studies is less developed than for intervention studies.
- Although there are randomised studies of diagnostic tests most studies use a cross-sectional or observational (cohort) study design.
- Criteria for study quality are different. The QUADAS checklist (Appendix 4) is used to assess the quality of the included studies [2].
- Descriptions of the study population and the tests used as well as the reporting of the results are often poor, which can affect the statistical analysis [3, 4].
- Studies of diagnostic tests are heterogeneous to a higher degree than intervention studies. Common sources of heterogeneity are differences in the spectrum of patients included and different threshold values. Pooling the results in meta-analyses may therefore be inappropriate.
- Statistical analyses and the pooling of results from individual studies of diagnostic tests must take into account that the results consist of pairs that are mutually dependent (eg sensitivity and specificity).
- Diagnostic studies most often measure outcomes as diagnostic accuracy of the test rather than patient outcomes (recovery, improvement etc). However, accurately diagnosing an individual is not enough. If no effective treatment is available, a test is of limited value and may even have negative consequences for the patient.

Terms and measures

Diagnostic accuracy

Diagnostic accuracy refers to the ability of a test to discriminate between individuals with a disease (or more generally a given condition) and without a disease. The test to be evaluated is called the *index test*. Most studies focus on the accuracy of a single test. In practice, however, a test never exists in a vacuum, but is part of a diagnostic process. The role of a test when multiple tests are involved is described in Facts 7.1.

The true presence or absence of the disease or condition is defined by a *reference test* or *reference standard* (also referred to as a “gold standard”). “Reference standard” is the preferred term since most, if not all, medical conditions lack a clear, error-free gold standard. The reference standard should represent the best available method for identifying the disease or condition in question. A reference standard can be a disease, a disease stage, or some other condition, and it might include several steps or follow a patient over time. For conditions where reference tests are lacking or imperfect other methods have been developed (Facts 7.2).

Facts 7.1 Multiple tests.

Often, more than one test is needed to determine the likelihood of disease in a patient seeking care for symptoms. It may be practical to consider the functions of a test in the diagnostic pathway. In principle, multiple tests can be used in two ways:

- Parallel testing (ie all tests are performed at the same time), where a positive test result from *any of the tests* indicates disease.
- Serial (consecutive) testing, where the decision to proceed to the next test depends on the outcome of the previous test(s). Here, *every test* must yield a positive result to establish a diagnosis since the diagnostic process stops if a test yields a negative result (Figure 7.1.1).

Strategy	Sequence of events	Consequences
Parallel testing	Test A *or* Test B *or* Test C is positive	Sensitivity Specificity
Serial testing	Test A *and* Test B *and* Test C are positive	Sensitivity Specificity

Figure 7.1.1 Parallel and serial testing. Parallel testing with multiple tests generally increases sensitivity, while specificity decreases and the percentage of false positive test results increases. Serial testing maximises specificity, while sensitivity decreases, and the percentage of false negative test results increases [5].

In practice, multiple tests are often ordered all at once, especially when rapid results are needed (eg for patients admitted to hospital or emergency cases), and determining the diagnostic pathway and the role of an individual test can be difficult. The diagnostic process can be based on guidelines, but sometimes a consensus is lacking on the optimal order of testing. Hence, one may have to make assumptions about where to place the test in the diagnostic pathway.

Facts 7.2 What to do if the reference standard is imperfect or lacking?

Well-defined criteria for the "true" condition (reference standard or reference test) are said to be valid criteria. An example is a histological examination of a biopsy of breast tissue to establish the presence/absence of breast cancer. In many cases there is no unequivocal or acceptable reference standard. Various solutions have been proposed for when a reference standard is imperfect or lacking [6, 7].

One alternative is to *construct* a reference standard (construct validity):

- **Composite reference standard.** Here, several reference tests (each of which is imperfect) are combined to form a composite measure that is considered to discriminate better between disease/non-disease than the individual reference tests. To determine the presence/absence of disease, *predefined rules* are applied where different definitions can be used depending on the characteristics of the individual reference tests. The simplest model uses two reference tests for all patients. The usual definition states that disease is present if any of the reference tests is positive. An example is a study that examined the accuracy of an index test for antigen analysis (enzyme immunoassay analysis, EIA) to diagnose *Chlamydia trachomatis* infection [8]. Two reference tests were combined: culture and DNA analysis (polymerase chain reaction, PCR). *Chlamydia trachomatis* infection was assumed to be present if either the antigen culture or PCR gave a positive result with EIA. If both reference tests were negative, the diagnosis would be no infection.

- **Panel or consensus diagnosis.** Here the results of different tests are combined with other results or clinical characteristics and prognostic information, which together yield pragmatic validation of the disease. Validation of the reference standard is then based on a large volume of empirical data and is often determined through international consensus procedures involving *expert panels* or a *Delphi process* [7]. An example is the DSM-IV criteria for psychiatric conditions. Another example concerns diastolic heart failure, where the European Society of Cardiology recommends basing the diagnosis on symptoms and clinical findings supported by ECG, radiography, Doppler echocardiography, and biomarkers [9].

- **Statistical models.** Here, clinical information is combined with other test results in statistical models that generate a probability, eg, of heart failure [7].

Another model involves validating the index test through a prospective study design where observations alone, or treatment outcomes, are related to symptoms and tests at baseline. This can be seen as a delayed type of verification. Other endpoints, such as treatment outcomes and relative risk, are often used.

An imperfect reference standard can sometimes be adjusted or corrected for. This requires access to specific information about the deficiency and the magnitude of error. Another possibility is to use an optimistic and a pessimistic estimate of sensitivity and specificity.

Outcome measures

Sensitivity and *specificity* are the classical measures for describing the accuracy of diagnostic tests. They can be used together with the prevalence of the health problem to calculate likelihood ratio, diagnostic odds ratio (DOR), and positive and negative predictive values.

Binary (dichotomous) variables are needed to calculate sensitivity and specificity. When a diagnostic test results in continuous data, eg blood-glucose levels, a threshold (cut-off value) should be defined. Individuals with levels above the threshold are considered to be sick. The cut-off value should reflect the greatest likelihood to discriminate between illness and no illness. This somewhat arbitrary threshold affects the number of true and false positives and negatives that will be found. Lowering the cut-off value for the blood glucose level will yield more false positives and fewer false negatives. Raising the cut-off value will yield the opposite result (Chapter 9).

Diagnostic accuracy is not a fixed, stable feature of a test in any context. Sensitivity and specificity can vary between subgroups of patients, different stages of illness, or different settings (primary care, specialist care, hospital care). They can also vary with different interpretations of the test, and may be dependent on previous tests when multiple tests are used. These factors are important to consider when deciding the inclusions criteria for the review.

The choice of high sensitivity or high specificity

The context in which a diagnostic test is used determines what should be considered as an acceptable accuracy. Although tests rarely yield a diagnosis that is 100% accurate, they can provide enough information to confirm or exclude a diagnosis in a pragmatic way. In other words, a diagnosis may be sufficiently accurate to show that expected benefits of treating the patient outweigh expected negative consequences of not treating. If the sensitivity and specificity of a test are both higher than those of another test, the former is obviously the one to choose. In many cases, however, one has to compromise and choose what is most important: high sensitivity or high specificity. It is then necessary to determine which type of misdiagnosis would have the least serious consequences. Ethical aspects, risks of adverse effects from treatment, and treatment costs have to be considered (Table 7.1).

Table 7.1 Balancing the demands for sensitivity and specificity.

Consequences of incorrect diagnosis	Risk assessment
Low specificity: Healthy people are classified as ill	Risks of treating healthy individuals Example: in foetal diagnostics, or where the treatment involves major risks, the need for specificity is very high.
Low sensitivity: Sick people are classified as healthy	The risk of the individual from not receiving treatment, the risks for the population (eg spread of infection), inherited risks Example: sensitivity should often be high for tests for infectious diseases

Technical accuracy (analytical validity)

It is important to distinguish between analytical validity and diagnostic accuracy. Analytical validity refers to the capacity of a test to provide reliable and accurate information under laboratory conditions and should be the foundation for all tests. An accurate measurement instrument should yield good precision and no systematic error. Precision is determined by repeated measurements of the same sample. The agreement between the measurements is often expressed as the coefficient of variation (standard deviation/mean value). Imperfect precision can also be due to observer variation in interpreting a test. A measurement instrument should also have no systematic error. A systematic error can be due to insufficient calibration of the instrument. If, for example, the plasma concentration of creatinine is measured for estimating kidney function and the analytic method for measuring creatinine is not calibrated against a standard, the risk of systematic error will be considerable. The amount of systematic error is often expressed as bias. In this context, bias refers to the mean or median difference between the estimated (index method) and the “true” value (reference method).

Defining purposes, formulating questions

As with all systematic reviews, well-conceived and clearly formulated questions are essential. This facilitates both the search and the selection of relevant studies.

Different types of questions for diagnostic studies

For pharmaceuticals there is a strictly regulated international standard (hierarchical 4-phase model), where each phase must meet certain conditions before continuing to the next (phase 0 is the initial phase, and phase 4 investigates the long-term effects and adverse effects on patients). Several corresponding models have been proposed for assessing diagnostic tests [10]. One of them clearly uses a hierarchical structure consisting of four types of question that are important to consider.

Stage 1 Question. Do the test results for patients with the target disorder differ from those for healthy people?

Stage 1 studies are often case-control studies where a group of patients known to have the disease are compared with a group definitely known not to have it. The results are often presented as correlations or differences in mean values between sickness and health. A positive outcome in Stage 1 studies opens the way to the next stage.

Stage 2 Question. Are patients with certain test results more likely to have the target disorder than patients with other test results?

Here, the interpretation shifts towards diagnostics. Results are presented in terms of sensitivity and specificity. Again, the design is often a case-control type of study.

Stage 3 Question. Do the test results distinguish between individuals with and without the target disorder among patients in whom it is clinically reasonable to suspect the disorder?

Example 7.1 presents examples of results from Stage 1, Stage 2, and Stage 3 studies. The example shows how Stages 1 and 2 studies, which both have a case-control design, overestimate accuracy, and that Stage 3 studies are necessary to determine the accuracy of a diagnostic test in clinical practice.

Stage 4 Question. Do patients undergoing the test fare better than similar untested patients? This question concerns the actual value of a test for the patient. The value is measured in the health outcomes when the treatment chosen is based on results from a diagnostic test. The benefits are sometimes obvious, eg the correct diagnosis of patients with life-threatening disorders, who thereby receive life saving treatments. More often, however, tests lead to detection of asymptomatic disorders, eg PSA (prostate specific antigen) testing for early detection of prostate cancer. The Stage 4 question in these cases can then be only answered by following patients who are randomised to the diagnostic test/alternative test (or no test).

These four stages or phases are called the architecture of diagnostic research [11]. The available evidence determines which of the question(s) that should be included in the review.

Example 7.1 Illustration of how results vary depending on the stage at which the study is performed [11].

Stage 1 study. In a hospital, the plasma concentration of BNP (B-type natriuretic peptide) precursor was measured in healthy control patients and a convenience (non-systematic) sample of patients with various combinations of elevated blood pressure, ventricular hypertrophy and systolic dysfunction of the left ventricle. Major differences in BNP concentrations were found between the groups.

BNP concentration	Patients known to have disorder	Normal controls
Median (range) concentration (pg/ml) of BNP precursor	493.5 (248.9–909.0)	129.4 (53.6–159.7)

The conclusion was that testing for BNP concentration was a useful diagnostic aid for left ventricular dysfunction.

Stage 2 study. The same method was tested at another hospital on normal control patients and a group of patients with coronary artery disease and varying degrees of left ventricular dysfunction. A cut-off point was chosen that best distinguished patients with severe left ventricular dysfunction from normal controls.

BNP concentration	Patients known to have target disorder	Normal controls
High	39	2
Normal	1	25

Test result (95% confidence interval):

Sensitivity = 98% (87; 100); Specificity = 92% (77; 98)

Positive predictive value = 95% (84; 99); Negative predictive value = 96% (81; 100)

The results are extremely encouraging but are they too optimistic? They are based on patients with established disease compared to healthy individuals.

Stage 3 study. Is the BNP test useful for diagnosing patients with suspected left ventricular dysfunction (LVD)? This was studied in a group of patients referred for suspected heart failure. The patients (n=126) underwent independent, blinded testing using BNP as index test and echocardiography as reference test.

BNP concentration	Patients with LVD on echocardiography	Patients with normal results on echocardiography
High (>17.9 pg/ml)	35	57
Normal (<18 pg/ml)	5	29

Prevalence LVD (pre-test): 40/126 = 32%.
Test result (95% confidence interval):

Sensitivity = 88% (74; 94); Specificity = 34% (25; 44)

Positive predictive value = 38% (29; 48)

Negative predictive value = 85% (70; 94)

The BNP- test thus had lower diagnostic accuracy when used in clinical practice. The authors concluded that routine measurement of BNP would be unlikely to improve the diagnosis of symptomatic left ventricular dysfunction.

Another model stages the diagnostic process in six levels [12, 13]:

1. Technical accuracy (analytical validity)
2. Diagnostic accuracy
3. Effects on clinical decision making
4. Effects of treatment on the patient's well being
5. Effects on the outcome of treatment in patients
6. Cost- benefit, cost-effectiveness.

The model can be useful for distinguishing between different types of studies, but it cannot be viewed as a necessary sequence for assessment since evaluation of tests most likely is a cyclical and repetitive process rather than a linear process [10]. Separate systematic reviews can be conducted for each of these stages.

Levels 3 through 5 in the model correspond to the Stage 4 question, which asks about the clinical value of a diagnostic test, ie "Is the outcome (which can be associated with the test results) better for the patients tested compared to similar patients who are not tested"?

Other important questions are:

- Does the review aim to investigate the analytical validity of a test?
- Which clinical situations (eg primary care, speciality services, hospital care) are relevant?
- Is the aim to investigate screening tests, or does the review address the diagnosis of patients with suspected disease?
- Is the intent to investigate tests used to diagnose established diseases (or conditions) or to investigate the accuracy of a test (or tests) in identifying early signs of disease or in grading disease stages (or all of these)?
- Will the test(s) be used to assess the progression of a disease (eg functional impairment), or to determine the prognosis?
- It may also be practical to attempt to determine the role of the test in the diagnostic process (Facts 7.3).

Facts 7.3 Which function should a new test have?

If a new test is to be studied, which function should it have in the diagnostic pathway? The function could differ from that of the existing test(s). The aim may be to replace, or supplement, an existing test. It might also be a triage test, ie the new test should precede the existing test(s). Figure 7.3.1 illustrates the role and position of the new test in the diagnostic pathway.

Existing situation	Population
Replacement	Initial tests
Add-on	Existing test
Triage	New test

Figure 7.3.1 The role and position of a test in the diagnostic pathway [14].

Existing situation refers to the situation at the outset in the population on which the new test will be tested.

Initial tests refer to the test results or other information (eg patient history and clinical examination) that are available before the test is performed.

Replacement refers to a situation where the accuracy of a new test is compared with the accuracy of an existing test. The comparison can also address the invasiveness of tests or compare costs. An example is mammography screening, where the current test (eg two breast radiologists reading the images) is compared with the new test (eg computer-aided detection plus one breast radiologist reading the images).

Addition to existing test, in the example above, means that patients with a negative test result from the current test are examined using a new test. An example would be positron emission tomography (PET) for detecting metastases, where the current tests are computed tomography and ultrasound.

Triage refers to a situation where the aim is to investigate the accuracy of a test prior to initial tests/existing tests to rule out individuals from further testing. An example is the “Ottawa ankle rules” (simple clinical examination of patients with foot and/or ankle pain), which has a very low percentage of false negative findings and is therefore often used to reduce the number of unnecessary x-ray examinations to confirm or rule out fractures.

In addition to information on where to place the new test in the diagnostic pathway, it is important to know which tests or other information are available before investigating the accuracy of a new test. For instance, if the new test is given only to patients having a certain positive test result, the spectrum of the population changes. A positive test result increases the likelihood of disease, which affects the sensitivity and specificity of the new test (see “Correct patient group” in Appendix 4).

What does a diagnostic test mean for the patient?

Assessing the diagnostic accuracy is an important part of determining the utility of a test, but the clinical value is in improving the health of the patient. The outcome of a diagnostic test can affect treatment, treatment results, and the well-being of the patient (a test result, *per se*, can affect both emotions and behaviour of a patient). But in contrast to interventions, the results of diagnostic tests are intermediary, ie they can influence, but not directly determine, the outcome of treating a patient. The clinical value is therefore important when assessing diagnostic tests. A test with good accuracy may be, but not necessarily is, effective and beneficial for the patient. However, studies that investigate the value of a diagnostic intervention are seldom available, especially for new tests. Neither are adequate and/or independent reference standards available for many diseases. In such instances, the clinical value can be appraised only by following the natural course of a treatment or its outcome.

Ideally SBU should investigate both the diagnostic method and the subsequent intervention to assess the value for the patient. A search could be made for systematic reviews on the intervention methods and refer to their potential benefit in the discussion. Note that even if the review addresses just a limited part of the diagnostic process, the benefits and value of a test must be mentioned and discussed in the report.

Formulating criteria for inclusion and exclusion

In intervention studies, PICO (population, intervention, control, outcome) is used to define criteria for inclusion and exclusion. In diagnostic studies, PICO is used in a corresponding way (population, index test, reference test [corresponds to "control"], outcome). Example 7.2 illustrates inclusion criteria based on PICO that appear in the SBU report on *Methods of Diagnosis and Treatment in Endodontics*.

It pays off to think through the inclusion and exclusion criteria carefully before proceeding to the review phase as they will impact the selection of studies and hence the results.

Example 7.2 Inclusion criteria formulated as PICO [15].

PICO	Inclusion criteria
Population	Patients who can be expected to undergo the examination or test in clinical practice. Permanent teeth
Index tests	Clinical signs or symptoms, other clinical information, clinical tests or biological markers
Reference test (control)	For pulp status in vital tissue: histological examination, or symptoms and clinical/radiographic information in a prospective study design. For determining pulp vitality: as above, or inspection/probing of pulp chamber, or radiographic examination combined with observing continued root development in immature teeth.

PICO	Inclusion criteria
Outcome	Sensitivity, specificity, likelihood ratio, odds ratio, or ROC with or without AUC (area under curve), or data reported so that sensitivity and specificity can be calculated.

Assessing studies that meet the inclusion criteria

Standards for the Reporting of Diagnostic Accuracy Studies (STARD) [16] gives recommendations for designing, implementing, and reporting studies on diagnostic accuracy. STARD consists of 25 items with quality aspects that in many respects are similar to those in intervention studies.

SBU uses a checklist based on the QUADAS tool which primarily is designed to appraise cross-sectional studies [2]. Internationally, this is the most commonly used checklist and corresponds to STARD. The underlying principle for QUADAS is that methodological deficiencies can introduce risk for bias (risk of systematic error due to problems involving study design, conduct, and/or reporting) or limit the applicability of a study. The revised version [17] consists of 11 items (Appendix 4). Each item has three alternative responses: *Yes*, *No*, or *Unclear*. QUADAS also describes how these alternative responses should be interpreted and used. However, depending on the aims of the review some items may be redundant whereas it may be important to add other items. The checklist does not, for instance, pay special attention to studies that compare multiple tests. Additional items are necessary when longitudinal follow-up of the patient is used to verify a diagnosis.

Examples of other items that may need to be added are:

- Was the aim of the study defined in advance?
- Was the cut-off determined in advance?
- Has the technology for the index test remained unchanged since the study was conducted?
- Does the study clearly define what is considered to be a “positive” outcome?
- Did those who performed the test(s) have adequate education/training?
- Was treatment postponed until both the index test and the reference test had been conducted?
- Was observer variation reported, and was it within acceptable limits?
- Were commercial interests involved in financing the study?

Several of the items are “black or white”. Others require subjective assessment, eg whether the patients are representative for the intended clinical use (“patient spectrum”). For these items it is important to clearly formulate guidelines for assessment. It is appropriate that the reviewers, working independently, make a pilot checklist based on a selection of at least five studies and then discuss any discrepancies. The level of concordance and important discrepancies should be documented. The ultimate grounds for assessment should be compiled in a manual. Figure 7.1 presents a flow chart for the QUADAS checklist.

Adapt the QUADAS checklist to the questions:

- Are all questions relevant for the review?
- Should the checklist be supplemented with additional questions?

How should the questions be appraised?

- Do some questions have more weight than others?
- Establish the criteria for "yes", "no", "unclear".
- How should questions that involve some measure of subjective appraisal be answered? Write this in a separate document

Testing

- The reviewers, working independently, evaluate 10 to 15 studies.
- The results are compared. Consensus on questions that are appraised differently.

Determine criteria for study quality

- Importance of potential bias is appraised. For instance, does a "no" response to a particular question lower the quality of the study from high to moderate (or from moderate to low)?

Figure 7.1 Flow chart for reviewing diagnostic studies.

Assessing study quality

The methodological quality of a study concerns the internal validity and is assessed as risks of bias. The external validity is often determined by the patient spectrum chosen for a study. Are the results applicable to the population(s) addressed by the review questions? The methodological quality of a study may be good but a closer scrutiny may suggest that its relevance might be questioned. The relevance of a study is appraised separately.

Risks of bias are, to some degree, similar in diagnostic studies and intervention studies. This includes, eg blinding. In intervention studies, those who interpret the outcome of a treatment should be blinded as to which treatment the patient received. In studies of diagnostic accuracy, the interpreters should be blinded as to the outcome of the index test when the outcome of the reference test is determined (and vice versa). But in contrast to intervention studies, where randomisation is common, the population in diagnostic studies often comprises consecutively recruited individuals who receive both the index test and the reference test. Table 7.2 presents the most important sources of bias in diagnostic studies. As shown there, a range of factors involving study design and conduct can lead to bias or variation. However, the size of the actual effect of such bias is uncertain [18].

Table 7.2 Sources of bias in studies on diagnostic accuracy [3,18].

Type of bias	When does it occur?	Under- or over-estimation of diagnostic accuracy
Patients		
Patient spectrum bias	When included patients do not represent the intended spectrum of the severity level of the condition in question	Depends on the difference between the condition in question and the spectrum included in the study

Type of bias	**When does it occur?**	**Under- or over-estimation of diagnostic accuracy**
Selection bias	When the intended patients are not included consecutively or randomly	Often leads to over-estimation
Index test		
Information bias (review bias)	When the results of the index test are interpreted with awareness of the results of the reference test	Often leads to over-estimation. If less information is available compared to that available in clinical practice, it could lead to under-estimation.
Clinical review bias	When access to information on clinical data such as age, sex, and symptoms is available when the index test is interpreted (primarily concerns x-ray images)	Leads to higher sensitivity, but has little influence on specificity
Reference standard		
Classification-error bias	When the reference standard does not correctly classify patients with the condition in question	Depends on whether both tests make the same mistake
Partial-verification bias	When a non-randomised sample of patients do not undergo the reference standard test	Often leads to overestimating sensitivity. Effects on specificity vary
Differential-verification bias	When some of the patients are verified with a second or third reference standard, particularly if this selection depends on the results of the index test	Often leads to overestimation
Incorporation bias	When the index test is incorporated in a (composite) reference standard	Often leads to overestimation
Disease-progression bias	When the patient’s condition changes between administration of the index test and reference standard	Over or under-estimation depending on changes in the patient’s condition
Information (review) bias	When the reference standard is interpreted with knowledge about the results of the index test	Often leads to overestimation
Data analysis		
Excluded data	When the analysis does not include data that cannot be interpreted or information on patient drop-out	Often leads to overestimation

The assessments are based on the degree to which different potential sources of bias are thought to influence study validity. The QUADAS checklist should be seen as an aid in assessing study quality. A consensus between assessors should be aimed for on each question. If the assessors do not agree, another assessor should be asked to decide. After that, the standards for grading the quality of a study as high, moderate, or low are determined.

The aim should be to follow the QUADAS checklist and if items are added they should be as generally applicable as possible. The best way to achieve this is to specify, as far as possible, the criteria for inclusion and exclusion. For instance, it is practical to determine which populations should be included. Should case-control studies be accepted or only studies that use consecutively selected patients with suspected disease? Another example involves specifying the reference standards that should be accepted. Inadequate reporting is common in diagnostic studies. The absence of important information about the question could be a criterion for exclusion. How "accommodating" one should be regarding inclusion and exclusion criteria often depends on the currently available evidence.

Extracting data and constructing tables

Data from studies assessed as having high or moderate quality are tabulated in the same manner as for intervention studies. Table 7.3 presents table headings for diagnostic studies. Relevant parameters for the questions can be included under an appropriate column heading.

Table 7.3 Table headings for studies on diagnostic accuracy.

First author **Year** **Country** **Reference**	**Aim**	**Study design** **Population characteristics** **Setting**	**Index test**	**Reference test**	**Results**	**Study quality** **Relevance** **Comments**

The "aim" presents the authors' purpose(s) of the study. When constructing a table it is valuable, at least initially, to clarify the purpose(s). For instance, this can indicate whether the data extracted from an article corresponds to the original purpose, or only reflect extraneous findings. The design used (eg cross-sectional study, cohort study) is reported under "Study design". Characteristics of the patient population, and how patients were recruited, are also described here. The settings from which patients were recruited (eg primary care, specialist care, hospital care) are described under "Setting". A brief description of index test(s) is presented under "Index test". The same applies to "Reference test". Quantitative outcome measures should be presented under "Results". In addition it may be appropriate to present a brief description based on the authors' conclusions. Study quality and results and reasons for grading down the certainty of the evidence is reported under "Study quality/Relevance/Comments".

Confidence intervals for sensitivity and specificity should be presented. If these are not reported in the original study but can be calculated from presented data, this should be done. An asterisk (*) preceding the calculation indicates that these data were not reported in the original study.

Deficiencies in study design, conduct and/or reporting is a common problem in diagnostic studies [18, 19]. The report should present a graphical summary of study quality based on the QUADAS criteria. This facilitates the subsequent task of assessing the certainty of the evidence with GRADE (Chapter 10). Figure 7.2 presents an example of such a summary which provides a quick overview of where problems concerning study quality might be found.

Representative patient spectrum

Population adequately described

Reference test classifies the target condition correctly

Adequate time interval between index and reference test

Reference test applied to all or to a randomised sample of patients

The same reference test given to all regardless of results of index test

Index test adequately described

Reference test adequately described

Index test interpreted independently of reference test

Reference test interpreted independently of index test

Uninterpretable test results reported

At least two independent examiners of reference test

Reliability of reference test reported

Precision of test results reported

Yes

Unclear

No

Figure 7.2 Reporting of 14 quality criteria (modified after QUADAS criteria) in 18 studies regarding accuracy of dental pulp diagnosis. Percentage distribution of “Yes”, “Unclear”, and “No” for each criterion [15].

The most important problems to evaluate may vary from question to question. Hence, SBU does not recommend standardised scoring. When evaluating the quality of a study it is an advantage to document which factors in a project that carry most weight.

Analysing statistics and compiling data

Analysing data from studies on diagnostic accuracy differs in several ways from analysing studies on therapeutic intervention:

- Diagnostic accuracy is usually quantified with *two* measures: sensitivity and specificity, which cannot be reduced to a single, composite measure (such as diagnostic odds ratio) without losing information. Two other measures are positive and negative prediction values.
- To determine when a test is positive, it is often necessary to select a cut-off value, eg for biochemical tests.
- Studies are often heterogeneous, which can be problematic when compiling data.

Generally, the methodology for statistical synthesis of diagnostic studies is not as well developed as for intervention studies. One should start with a basic, descriptive, compilation of the data. This can be done with paired *forest plots* and a simple *summary ROC (receiver operating characteristic) curve*. Bivariate/hierarchical models for meta-analysis can be used when an adequate number of reasonably homogeneous studies are available.

Analysis of heterogeneity

Heterogeneity in diagnostic studies stems from several sources [20]. It can be random, but usually concerns real heterogeneity between studies. A common reason is that different threshold values (cut-offs) are used to define when a test result is considered to be positive or negative. Other reasons may include differences in the patient spectrum (disease severity or co-morbidities) or partial verification bias (a non-randomised sample of patients did not undergo the reference test). Different technologies for index tests and/or reference tests, differences between observers, different study designs, and implementation can also cause heterogeneity in studies. Heterogeneity is nearly always found in diagnostic studies, and the *reasons* should always be analysed. If heterogeneity is high, statistical synthesis will not be meaningful.

The first step should be to determine the *degree* of heterogeneity. This can be presented graphically with forest plots, where two paired plots are used, one for sensitivity and the other for specificity. Forest plots can also be made for positive and negative prediction values and for likelihood ratios. Example 7.3 presents paired forest plots and a simple compilation of study results.

The forest plots and compilation in Example 7.3 show high heterogeneity for specificity; one study (Seltzer) clearly deviates from the others by having low specificity. Here, the heterogeneity had several causes. Recruitment of patients was reported inadequately in several studies and the reference standard in one study (Seltzer) differed from the others. *Hence, here it is inappropriate to use a meta-analysis (where the values for sensitivity and specificity form a composite measure).* One option may be to use meta-analysis in a subgroup of reasonably homogeneous studies.

Descriptive compilation of heterogeneous studies

Receiver operator characteristic (ROC). In addition to the compilation illustrated in Example 7.3, the results of heterogeneous studies can be presented graphically as a simple ROC curve. This shows how the results are distributed, but not the precision of individual studies. Furthermore, heterogeneity cannot be appraised. Figure 7.3 presents an example.

Paired forest plots and simple ROC curves can be constructed in Cochrane's Review Manager, RevMan (http://ims.cochrane.org/revman). If the initial analysis shows that the studies are too heterogeneous, one should not proceed to more advanced statistical methods.

Example 7.3 Paired "forest plots" of sensitivity and specificity and compilation of sensitivity and specificity in five studies investigating the accuracy of the cold test (a pellet saturated with ethyl chloride is applied to the tooth surface) to determine whether dental pulp is vital or non-vital.

Compilation of studies

Sensitivity				
Study	Sensitivity	(95% CI)	TP/(TP+FN)	TN/(TN+FP)
Evans 1999	0.92	0.82; 0.98	49/53	72/81
Gopikrishna 2007	0.81	0.66; 0.91	34/42	35/38
Kamburoglu 2005	0.94	0.84; 0.99	49/52	40/41
Petersson 1999	0.83	0.64; 0.94	24/29	27/30
Seltzer 1963	0.89	0.65; 0.99	16/18	29/121
Specificity				
Study	Specificity	(95% CI)	TP/(TP+FN)	TN/(TN+FP)
Evans 1999	0.89	0.80; 0.95	49/53	72/81
Gopikrishna 2007	0.92	0.79; 0.98	34/42	35/38
Kamburoglu 2005	0.98	0.87; 0.99	49/52	40/41
Petersson 1999	0.90	0.73; 0.98	24/29	27/30
Seltzer 1963	0.24	0.17; 0.33	16/18	29/121
FN = false negative; FP = false positive; TN = true negative; TP = true positive; CI = confidence interval				

Statistical synthesis of reasonably homogeneous studies

If studies are reasonably homogeneous an average sensitivity, specificity and likelihood ratio can be calculated by pooling data in meta-analyses [21] (ie patient population, index test, and reference test are comparable). Simple methods, such as paired forest plots and ROC, do not take a negative correlation between sensitivity and specificity into consideration. Instead, the results can be pooled by more advanced statistical methods.

Hierarchical models for statistical analysis can be used if there is a sufficient number of reasonably homogenous studies. *Bivariate meta-analysis* takes the underlying correlation between sensitivity and specificity into consideration [22]. *Hierarchical summary ROC analysis* is based on the logarithm of diagnostic odds ratio (DOR), eg taking cut-offs into consideration. Both these methods are considered to be more robust and better suited to meta-analysis of diagnostic data [3, 23], but they require good knowledge of the models and access to statistical software such as STATA, SAS, or SPSS version 19. Chapter 10 of the Cochrane Handbook describes these statistical models in detail [24].

Sensitivity

SROC curve

Symmetric SROC

AUC=0.9437

SE(AUC)=0.0222

Q*=0.8822

SE(Q*)=0.0287

Specificity

AUC = Area under the curve; SE = Standard error; Q* index = the point at which sensitivity and specificity are equal, which is the point on the curve closest to the upper left corner

(NOTE: Use decimal points instead of commas in the numbers presented in the figure.)

Figure 7.3 ROC curve summarising the five studies that examined the accuracy of the cold test to determine whether dental pulp is vital or non-vital. One study (Seltzer) deviates from the others by having a high percentage of false positive results.

Tabulated summary of findings

In addition to the usual tables of included studies, important results should be compiled in a table based on GRADE (Chapter 10). Table 7.4 presents an example from the SBU Alert report on mammography [25].

Patient benefit

While it is possible to compile the results, eg in a GRADE table, the outcome in terms of “patient benefit” is often lacking. This weakness should be mentioned in the discussion (Chapter 10).

Cost-effectiveness

Cost-effectiveness in diagnostic studies is considered in Chapter 11.

Table 7.4 Two methods for mammography screening in diagnosing cancer were compared for accuracy: double reading by two radiologists versus a combination of single reading (1 radiologist) and computer-aided detection (CAD) [25]. The reference standard was breast tissue biopsy or follow-up of the patient. The table is of the same type as that used for GRADE.

Outcome	Sample size (no. of studies)	True positive/proportion recalled: Single reading and CAD (95% CI)	True positive/proportion recalled: Double reading (95% CI)	Absolute difference (95% CI)	Quality of evidence	Comments*
Cancer detection rate	28 204 (1)	0.702% (0.6; 0.8)	0.706% (0.6; 0.8)	0.004% (NS)	+ooo Insufficient	Study quality –1 Indirectness –1 Imprecise data –1
Recall rate	28 204 (1)	3.9% (3.7; 4.1)	3.4% (3.2; 3.6)	0.5% (0.3; 0.8)	+ooo Insufficient	Study quality –1 Indirectness -1 One study –1

CAD = computer-aided detection; CI = confidence interval; NS= No statistically significant difference.
* Study quality = risk of bias, that is sensitivity is probably overestimated due to incomplete follow-up of women with negative test results.
Indirectness = only breast radiologists with long clinical experience took part in the study.
Imprecise data = the difference in sensitivity between double reading and single reading + CAD has wide confidence intervals.

References

1. Wilson JMG, Jungner G. Principles and practice of screening for disease. Geneva: WHO; 1968. http://whqlibdoc.who.int/ php/WHO_PHP_34.pdf

2. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.

3. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM; Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889-97.

4. Westwood ME, Whiting PF, Kleijnen J. How does study quality affect the results of a diagnostic meta-analysis? BMC Med Res Methodol 2005;5:20.

5. Fletcher RH, Fletcher SW. Clinical epidemiology. The essentials. Philadelphia: Lippincott Williams & Wilkins; 2005.

6. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol 2009;62:797-806.

7. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess 2007;11:iii, ix-51.

8. Jang D, Sellors JW, Mahony JB, Pickard L, Chernesky MA. Effects of broadening the gold standard on the performance of a chemiluminometric immunoassay to detect Chlamydia trachomatis antigens in centrifuged first void urine and urethral swab samples from men. Sex Transm Dis 1992;19:315-9.

9. Paulus WJ, Tschöpe C, Sanderson JE, Rusconi C, Flachskampf FA, Rademakers FE, et al. How to diagnose diastolic heart failure: a consensus statement on the diagnosis of heart failure with normal left ventricular ejection fraction by the Heart Failure and Echocardiography Associations of the European Society of Cardiology. Eur Heart J 2007;28:2539-50.

10. Lijmer JG, Leeflang M, Bossuyt PM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13-21.

11. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324: 539-41.

12. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our under-standing of how physicians reason. Science 1959;130:9-21.

13. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88-94.

14. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089-92.

15. SBU. Rotfyllning. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 203. ISBN 978-91-85413-39-3.

16. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.

17. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration, 2009. http://srdta.cochrane.org/

18. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189-202.

19. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061-6.

20. Dinnes J, Deeks J, Kirby J, Roderick P. A methodological review of how heterogeneity has been examined in systematic reviews of diagnostic test accuracy. Health Technol Assess 2005;9:1-113.

21. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001;323:157-62.

22. Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol 2005;58:982-90.

23. Harbord RM, Whiting P, Sterne JA, Egger M, Deeks JJ, Shang A, et al. An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. J Clin Epidemiol 2008;61:1095-103.

24. Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C, editors. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.9.0. The Cochrane Collaboration, 2010. http://srdta.cochrane.org/

25. SBU. Datorassisterad granskning inom mammografiscreening (CAD). Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2011. SBU Alert-rapport nr 2011-05. ISSN 1652-7151. http:// www.sbu.se

Chapter 8

Evaluation and synthesis of studies using qualitative methods of analysis

Available on www.sbu.se/en/method/

Chapter 9

Composite appraisal of results

Assessing health and social service technologies involves determining which alternative interventions are most effective in addressing a given problem or condition. If there are several assessments of the alternative interventions, the results must be compiled and compared in some way. The composite results can then be included in an evidence profile (Chapter 10; GRADE) and serve as part of the foundation for making decisions in the context of evidence-based practice [1]. If statistical analyses are used in the composite analysis, this is referred to as a meta-analysis; if such analyses are not used, the composite analysis is usually referred to as a narrative analysis. Meta-analysis is often used in randomised trials (RCTs) of alternative interventions. It is also used to analyse other types of study, eg in diagnostics and psychometrics. This chapter orientates the reader about:

- what a meta-analysis involves
- problems in meta-analysis and strategies for addressing them:
  - publication bias: “funnel plots” and “trim and fill”
  - heterogeneity: subgroup analysis, “random effects model”, no comparative analysis
- other applications of meta-analysis, eg observational studies, diagnostics, psychometrics.

Some findings carry more weight than others

Meta-analysis normally involves calculating some type of average for multiple study results in order to estimate a single “true” effect. Normally, however, the findings differ in the weight they carry in the analysis. A study’s relative weight normally depends on the number of individuals it includes; the more participants, the greater the findings’ weight in the analysis. In practice it is the spread of the random distribution (the standard error) that decides this; the narrower the spread, the greater the weight (this spread decreases as the number of individuals increases) [2].

A common way of presenting a meta-analysis is as a forest plot. This includes, eg the estimated size of each study’s effect, a composite size of the effect, and confidence intervals for the individual effects and the composite effect. Figure 9.1 presents a forest plot that shows the results of an intervention for homeless people with mental illness and more or less serious problems with substance abuse [3-12]. The intervention consists of an intensive case management (ICM) programme and the control option is usual care (UC). The measure of effect is risk difference[1]. Here, risk difference refers to the additional percentage of the intervention group that has an own domicile at the 12-month follow-up compared to the control group, ie the difference between the two populations. The term “risk” is often also used to describe a positive event such as recovery. The findings from each study are referenced by the first author, the horizontal lines represent the confidence interval, and the rectangle in the middle indicates the size of the effect.

Figure 9.1 Example of a meta-analysis (forest plot): intensive case management (ICM) versus usual care (UC).

----

[1] In medical assessments a common alternative would be to use an odds ratio, or risk ratio, on account of these measures statistical features. We have chosen risk difference because this measure is easiest to understand.

The study by Lehmann et.al gave the following results: the risk difference is 14 per cent, ie at the 12-month follow-up, 14 per cent more in the intervention group than in the control group had a stable domicile of their own. However, the confidence interval, from –3 to 32 per cent, overlaps the zero line, which means that the difference is within the statistical margin of error. In other words, the results are not statistically significant. The diamond symbol (rhombus) furthest down the list denotes the composite effect, a risk difference of 7 per cent, and the confidence interval, from 3 to 12 per cent.

The "Weight" column presents each study's weight in the analysis. The study by Lipton et.al has the smallest weight (not quite 1.8 per cent) and the study by Conrad et.al has the largest weight (41.2 per cent). Note that the heavier the results, the shorter the confidence interval. This is because the larger the standard error, the wider the confidence interval.

Figure 9.1 illustrates the advantage of meta-analyses. Firstly, meta-analyses result in a *single* composite effect from the 10 studies (the rhombus furthest down in Figure 9.1). Having a single composite effect with one confidence interval instead of ten different effects with ten confidence intervals facilitates the interpretation of an assessment's results. Secondly, the precision of the estimated effect is normally greater than that of the individual results. Including a larger number of individuals reduces the risk of missing a "true" effect.[2]

Even so, due to certain problems, the average effect from a meta-analysis is not always a reliable estimate of the "true" effect. First, the studies in the meta-analyses may not be a representative sample because of publication bias. Usually this means that the size of the effect is somewhat overestimated. Second, the results may be based on studies that are not sufficiently similar as regards, eg the composition of the population, local context, intervention content, control conditions, effects measurements, and study design. This problem is usually referred to as clinical heterogeneity [13] and can give rise to over- or underestimation of the "true" effect. The following section describes how meta-analysis can be used to address such problems, ie publication bias and heterogeneity.

----

[2] The risk of type II errors (beta errors) decreases in meta-analyses since the power of the statistical test increases.

Publication bias and funnel plots

In Figure 9.2 the results from Figure 9.1 are transformed into a *funnel plot*. Effect size is plotted against the horizontal axis, distribution (standard error) against the vertical axis. Note that the values on the vertical axis are inverted, so that the larger a result, the smaller its distribution. The results from Lipton et.al (on the far right) have an effect of 44 per cent and the widest distribution of the included studies. The dotted triangle is an aid to the interpretation of the results; the vertical midline represents the position of the composite effect of 7 per cent.

The model is partly based on two assumptions: (a) that publication of results from larger studies (with a narrow distribution) is easier to achieve than results from smaller studies, and (b) that results with a large effect favouring the assessed intervention are easier to get published than results that are either not significant or do not favour the intervention [2,14]. Publication difficulties could lead to small, non-positive results never being published, to publication taking longer, or to publication occurring in journals that are not indexed in reference databases (and consequently are difficult to find). Awareness of such difficulties can also lead to selective reporting in a particular study, that is, that the authors report only the statistically significant results that favour the intervention and refrain from reporting other results. Reporting bias differs from publication bias; it refers to tendentious reporting within a study, ie a tendency to report only results that favour the intervention. If reporting bias is more common in small studies than in large, it would be similar to publication bias. Financial interests can also underlie this type of selective reporting if those who assess the intervention might benefit from the assessment finding that the intervention is effective.

If the assumptions described above are accurate, there should be relatively few study results in the lower left corner of the triangle (ie small studies that do not favour the intervention, or are not statistically significant). If there is no publication bias, the results should be distributed symmetrically around the estimated composite effect. Figure 9.2 reveals signs of publication bias. This means that the risk of 7 per cent might be an overestimation of the "true" effect.

To determine the extent to which the effects are overestimated, the most extreme results favouring the intervention can be *trimmed* before recalculating the effect size. To avoid overestimating the width of the confidence interval, new hypothetical results can be included as *fill.* This way of dealing with publication bias has been developed into a statistical method, known as "trim and fill", that involves an iterative process [2]

Figure 9.2 Funnel plot with signs of publication bias.
Risk difference's standard error
Risk difference

Removing the findings from Lipton et.al, for example, does not alter the estimated effect of 7 per cent but lowers the upper threshold of the confidence interval from 0.12 to 0.11. When the study by Bond et. al is also excluded, the effect decreases to 6 per cent, and the confidence interval changes from 0.02 to 0.10: the results are still statistically significant. The methodological exercise described above can provide some indication of the consistency of the results and the possible extent of publication bias in this example.

Heterogeneity can be clarified and studied

This section addresses how meta-analyses can help manage the problem of heterogeneity [2, 13]. Although all but one of the findings in Figure 9.1 have a positive effect, the results are not consistent. For example, the size of the effect varies greatly, from 44 per cent (Lipton et.al) to -4 per cent (Clarke et. al). This lack of consistency can be quantified by using different measures of heterogeneity, such as I^2 and Q. Q is a weighted measure based on the extent to which each finding deviates from the composite effect. Using a chi-square-test, heterogeneity is shown to be statistically significant in the example since $p=0.07 <0.10$ (to be on the cautious side, as a rule of thumb 0.10 is used as the cut-off). I^2 captures the percentage of the total variance explained by the real differences between the studies' effect sizes. As a rule of thumb, I^2 is usually designated as follows: low heterogeneity = 0.25, moderate heterogeneity = 0.50, and high heterogeneity =0.75 [2].

Assume that the different results are based on studies that are similar to each other in terms of intervention, control conditions, assessment design and outcomes. Further assume that the populations differ between the studies but the positive effects are nevertheless highly consistent. Under such circumstances, the composite results suggest that the intervention's estimated effectiveness is relatively stable regardless of subgroups in the population (all else being equal). In Figure 9.1, however, the results are not consistent, as revealed by statistical heterogeneity.

The lack of consistency could have clinical and methodological explanations. One possibility is that different patient groups react differently to the intensive case management (ICM) intervention. ICM has been developed primarily for people with mental illness, eg schizophrenia (MI= mental illness), and may function differently for patients whose main problem involves heavy drug abuse (SA, ie substance abuse) or both mental illness and heavy substance abuse (MISA). A strategy for managing heterogeneity would therefore involve analysing the importance of different subgroups.

In Figure 9.3 the results have been divided into two subgroups but a total synthesis has not been performed. This group division points to no heterogeneity in the SA/MISA group and greater heterogeneity in the MI group. This might indicate that ICM functions differently in the two groups of patients, worse in the SA/MISA group and better in the MI group compared to UC. As the percentage of the total variance explained by the two subgroups is more than moderately large (60.7 per cent), the division into subgroups may be appropriate. Since the heterogeneity in the MI group increases and differences between the subgroups are not statistically significant (p= 0.11), it would perhaps be appropriate to proceed with additional subgroups in the MI group, or to report the results separately for the individual studies. However, other alternatives could possibly explain heterogeneity.

The reason for the heterogeneity could lie in a methodological problem. This problem may occur when the control conditions consist of usual care and the assessed intervention consists of a combination of several more or less active components. The problem arises when components of a new, perhaps more effective, intervention have started to disseminate and become integrated in usual care (a type of contamination). If this is the case, the effects of ICM compared to UC should decrease over time as UC becomes increasingly similar to ICM.

Figure 9.3 Subgroups: mental illness and substance abuse. Intensive case management (ICM) versus usual care (UC).

Figure 9.4 shows that dividing the results into two halves by the median (between 1997 and 1998) for the time period covered reveals that new subgroups have been formed. With the new division, the heterogeneity in both subgroups disappears, the difference between the subgroups becomes statistically significant (p=0.003), and the percentage of the total variance explained by the two subgroups is 88.4 per cent.

If the assumption regarding contamination is correct, the reduced effect over time should be attributed primarily to an improvement in UC's results, while the ICM group's results are largely unchanged. Summing each individual in the respective groups from the two time intervals gives the following results:

- Of the participants who received UC, during the 1988-97 period 56 per cent (170/306=0.56) had a stable domicile at the 12-month follow-up, while the figure for the 1998-2006 period was 66 per cent (311/474=0.66). This represents an improvement of 10 percentage points.
- Of the participants receiving ICM, 70 per cent had a stable domicile at the 12-month follow-up in both time intervals (165/236=0.70 and 360/516=0.70).

Figure 9.4 Subgroups from studies in 1988-97 and 1998-2006. Intensive case management (ICM) versus usual care (UC).

These two results suggest that the control groups may have been contaminated over time. In this case, when a variation in effect size might be explained by a continuous variable, it should also be possible to use meta-regression as an analytical tool instead of two time periods [2].

Figures 9.3 and 9.4 illustrate what subgroups analysis can do as a strategy in managing the problem of heterogeneity. The causes of heterogeneity are probably manifold and the example above shows that both a heterogeneous patient population and methodological problems may be underlying factors. There could also be other causes.

Heterogeneity can be included in the meta-analysis model

Up to now we have used a fixed effects model (FEM), see Figures 9.1–9.4 [2]; "Fixed" is indicated under the column heading "Risk Difference". This model builds on the assumption that all results refer to a random selection from one and the same population for which there is a single "true" effect. Another common way of addressing heterogeneity is to use a model built on other assumptions. This alternative model is called the random effects model (REM) [2]. An assumption behind this model is that every study result is based on a random selection from several populations of results, with a "true" effect for each study. In practice, this means that small, outlying studies will carry more weight than in a fixed effects model. It can be noted that the less heterogeneous results, the smaller the differences in results between the models.

From Figure 9.5 it can be seen how the REM changes the results compared to Figure 9.1. First, the effects increase from 7 per cent to 10 per cent and the confidence interval has both shifted and become wider: 0.04 to 0.16 instead of 0.03 to 0.12. Furthermore, the strongest result in Figure 9.1 (from the study by Conrad et. al) decreases from 41.2 per cent to 19.8 per cent and the weakest result (from the study by Lipton et.al) increases from 1.9 per cent to 3.4 per cent.

Dividing the results into subgroups (Figures 9.3 and 9.4) and including the heterogeneity in the meta-analysis model (Figure 9.5) are different ways of addressing heterogeneity. The impact of these different strategies becomes clear when the results are interpreted. The results in Figure 9.5, an effect of 10 per cent in risk difference, build on the assumption that there are 10 different populations, each with its own true effect for each study. The 10 different results are assumed to comprise random samples of studies from the respective populations. The composite effect of 10 per cent is therefore an estimate not of *one* true effect but of the mean value of a distribution of estimated "true" effects. The division into subgroups (Figures 9.3 and 9.4), instead of using REM, assumes that there are two populations, one for each group, and two "true" effects. These populations are judged to be so different that it is not meaningful to include their results in the same composite analysis.

Figure 9.5 Random effects model. Intensive case management (ICM) versus usual care (UC).

High clinical heterogeneity and no composite statistics

Each individual result is based on studies that can differ from one another regarding patient populations (eg composition, risk factors), interventions (eg content, including adjunct treatments, implementation), control conditions (eg content, including adjunct treatments, implementation), outcomes (eg definitions, measurements methods, follow-up time), and study design (eg allocation methods, handling interruption of treatment). If the differences are too great, one can simply choose not to combine the results in a single estimate of effect size. Using a forest plot to summarise the results may be helpful when interpreting the findings (Figure 9.6).

In Figure 9.6, all the results have been obtained with the same statistical outcome measure (risk difference), including confidence intervals. Instead of presenting separate figures, or describing the effects in the text alone, this provides an overview of the material. Not calculating a composite effect indicates that doing so would be inappropriate. If the material is too complex and heterogeneous, composite statistics would falsely imply a level of precision. Hence, composite analyses of the results, such as those presented in Figure 9.6, must be narrative, not statistical. This means that the complete picture in Figure 9.6 has to interpreted and summarised in words.

Figure 9.6 Forest plot, not composite.

Analytical tools or components in the evidence profile

Meta-analyses can be used in various ways. In the example above, meta-analysis serves as an analytical tool that can lead to a better understanding of the data we are working with. When it comes to writing the final report and compiling the evidence in GRADE tables to construct an evidence profile (Chapter 10), the meta-analyses that are to be included must be chosen with care.

This applies to the study results, the choice of model (fixed effects model, random effects model), possible subgroups, and whether or not a composite analysis is appropriate. Also, the choice of statistical effect measures (odds ratio, risk ratio, risk difference, hazard ratio, etc.) needs to be justified. These measures have different statistical characteristics and are not always equally appropriate. In the example above, the risk difference was chosen for pedagogic reasons in that risk difference is intuitively easy to understand.

These choices can play an important role when deciding between alternative interventions in health care. If Figure 9.1 or Figure 9.5 is chosen, the results favour ICM over UC (all else being equal) regardless of the subgroup concerned (mental illness or heavy abuse with or without mental illness). Taking possible publication bias into consideration (Figure 9.2) does not alter the picture, though the expected effects are somewhat lower. Choosing Figure 9.3 would mean that ICM is preferable if the main problem is mental illness. This is less obvious if heavy substance abuse is involved (all else being equal). If instead the meta-analysis in Figure 9.4 is chosen as part of the evidence profile, it is doubtful whether ICM would be preferable to UC, regardless of whether the main problem is mental illness or substance abuse. It seems that UC has improved so much during the past decade that it no longer differs from ICM.

Let us assume that Figure 9.6 is ultimately used as part of an evidence profile. In this case, the results are found to come from studies that differ too much to make a comparative analysis meaningful. The key issue here is which results are most relevant for decision making in Swedish practice (compare generalisability in GRADE). The ten outcomes in Figure 9.6 might not all be equally relevant if patient population, intervention, control alternative, outcome and design are considered in detail. Perhaps the study by Morse et.al best captures the alternatives regarding practice, or perhaps some other study results are more relevant. The results that are ultimately chosen obviously determine which decision will be supported, so this choice must be justified systematically and transparently.

It should be emphasised that the research questions we try to answer with meta-analyses can differ widely. This depends, in part, on how the questions are specified and what the current research fields look like. The research field in the example above has many studies with few participants, complex and often insufficiently described interventions and control conditions, and effect measures that are not always reliable. The research questions are also relatively broad. Things may look completely different with a more narrowly defined question in a methodologically stronger research field.

Suppose we want to know how two alternative platelet inhibitors (ticagrelor and clopidogrel) affect total mortality in people with acute coronary syndrome. At present there are only two randomised trials which address this issue [15, 16]. The first, PLATO, included just over 18,000 participants from more than 740 different centres worldwide, while the second, DISPERSE2, had 990 participants from

132 centres. Figure 9.7 presents the results in terms of relative risks (per cent mortality in the ticagrelor group divided by per cent mortality in the clopidogrel group)[3].

The results of the larger study are statistically significant and favour ticagrelor, while the smaller study shows a non-significant and minimum excess risk (3 persons) for the ticagrelor group. Although there is no statistical heterogeneity, the two studies send different messages and present certain clinical and methodological differences. First, the total percentage of deaths in PLATO is 4.9 per cent versus 1.7 per cent in DSIPERSE2, which suggests that the patients in DISPERSE2 might have been somewhat healthier. The target group in DISPERSE2 was patients having acute coronary syndrome without ST segment elevation, while the target group in PLATO included patients with ST segment elevation. Second, PLATO had a 12-month follow-up period, DISPERSE2 only 3 months. The DISPERSE2 study also reported substantially fewer events.

Figure 9.7 Ticagrelor versus clopidogrel.

Overall, this could mean that the studies differ too much to warrant a composite evidence profile. If a follow-up period of 12 months or longer is found to be necessary for accurate results, the DISPERSE2 study can be eliminated from the evidence profile.

---

[3] The ticagrelor study presents the outcome as a hazard ratio, which is a better alternative than risk ratio since it considers the time to the event. We used risk ratio because this information is not available in the smaller study,.

From a purely statistical standpoint, the DISPERSE2 study is irrelevant because including it does not markedly alter the estimated effects and the confidence interval. The study carries a weight of only 1 per cent.

It may seem pointless to conduct a meta-analysis using only two studies that might be too different for composite analysis. Nevertheless, meta-analysis does play a role as an analytical tool. It helps to clarify the differences between the two study results. This could increase the awareness of clinical and methodological differences that went unnoticed previously. Finally, the relative statistical weight of the study results becomes clearer. Ultimately, this information can help to determine what to include in the evidence profile. If the two studies are found to be sufficiently similar, composite data should be included in the evidence profile to simplify presentation and interpretation (Chapter 10, GRADE).

Meta-analyses from observational studies

Meta-analyses can also be conducted with results from *observational studies*, though this is less common and requires more work than for randomised trials. The basic principle is the same: Interventions are compared to control conditions. However, several problems involving principles and practice make this more difficult and more labour-intensive that when using randomised trials. Observational studies vary widely in methodological design. Variations can result from, eg whether there is a matched comparison (control) group at baseline (measurements prior to intervention) instead of doing matching retrospectively by some kind of multivariate methodology, the number of comparison groups, and the number of times measurements are made. Campbell and Stanley [17] present 14 variations and Shadish et. al [18] describe around 20 designs, divided into four categories: a) observational studies with no comparison group and no measurements at baseline, b) observational studies with a comparison group and measurements at baseline, c) discontinuous time series, and d) regression discontinuity design.

All design alternatives for studies should include a model that helps manage problems concerning the risk of selection bias. Selection bias may arise when intervention and control groups are not sufficiently similar, eg regarding risks and protective factors. For meta-analyses based on observational studies to be feasible, data must be available in a format that compares the intervention group with the control group after adjusting for possible differences. The control group can be matched against the intervention group at baseline, eg by using known risk and protective factors to make the groups as similar as possible. In other cases, statistical models can help in trying to achieve similarity retrospectively.

If a study aims to assess an intervention compared to a control alternative, it may be possible to use the study results in a meta-analysis. If the primary aim has been to test a causal model instead of making such an assessment, it can be more difficult to use the results in a meta-analysis, particularly if the statistical information is insufficient (eg number of individuals, mean values, measures of distribution).

Using meta-analysis with observational studies as an analytical tool is not necessarily associated with any major problem of principle; for example, it could involve handling heterogeneity. Using composite statistics as part of an evidence profile may, however, be more risky when it involves outcomes from observational studies rather than randomised trials. The adjustments made might address different background factors in the individual studies, so they might not be sufficiently similar to analyse together. A forest plot without combination of data may, however, be made. Because of systematic deficiencies in randomised studies, observational studies may be an acceptable option in some cases, eg studies of long-term adverse effects [19].

Example 9.1 illustrates the complexity of observational studies that are not sufficiently similar at baseline.

Example 9.1 Observational study with differences in baseline data
An observational study aimed to investigate whether multidisciplinary care (MDC) affected mortality in elderly patients with chronic renal disease [20]. In a logistic regression, the dependent variable was assignment to MDC and several risk factors were independent variables. With the help of this model the authors could calculate a propensity score that a given patient would receive MDC. After a propensity score had been calculated for each patient, the patients were matched pairwise. The survival curves of those who received MDC and those who did not were then compared. The findings showed that those who received MDC clearly had a lower momentary risk of dying compared to the control group, with a hazard ratio of 0.50 (95% CI = 0.35 - 0.71).

Acquiring an overview of the similarities and differences of the included observational studies may require tabulating more information than would normally be included in tables for randomised studies, eg identifying the variables contained in the model one uses to manage selection bias and the model itself. Table 9.1 exemplifies this with studies of programmes using multidisciplinary teams, compared to usual care, for sick elderly people.

Table 9.1 Model, variables, and matching procedure.

If the matching methods are found to be sufficiently similar, the meta-analyses can be conducted in the same way as for randomised studies (assuming they are sufficiently similar in other essential respects). If the differences are too great, forest plots can be used but without the composite effects (Figure 9.6).

Meta-analyses from studies on diagnostics and psychometrics

Diagnostics and psychometrics comprise another area of application for meta-analyses [2]. This is a complex, developing area. Diagnostic and psychometric studies can be randomised or observational [25]. The meta-analyses may involve various aspects of current clinical processes, eg test accuracy or patient benefits. Test accuracy can be compared using a validated reference test of actual patient conditions. Furthermore, a coherent diagnostic strategy that includes multiple tests and treatments can be compared with alternative diagnostic strategies.

For example, randomised trials can be used if patient benefit is a main focus and alternative diagnostic strategies are compared. In this case, current methods for meta-analysis of results from studies of intervention effects are used. On account of the complex intervention and comparison alternatives, the composite weights need to be interpreted cautiously. If the test's accuracy as regards sensitivity and specificity is central to the review, other analytical methods would be needed than those for composite weighting of intervention effects [26, 27]. Furthermore, other methods from diagnostic studies are consolidated, eg diagnostic odds ratios, summary ROC curves, and AUC (area under the curve). Chapter 7, on diagnostic studies, describes these in detail.

Regarding the accuracy of *psychometric tests*, meta-analyses have addressed the correlation between instruments, eg a questionnaire answered by patients and one answered by clinicians. At least two methods have been used to weigh these correlations, one developed by Hedges and Olkin [2], the other by Hunter and Schmidt [28]. The first uses the distribution, the second the number of individuals.

Finally it should be mentioned that methods of meta-analysis are being developed to deal with a lack of direct comparisons of relevance for practice. For instance, one might want to know the outcome of comparing two interventions when each of them is only compared with a third alternative. An example is two platelet inhibitors, ticagrelor [15] and prasugrel [29], for individuals with acute coronary syndrome. Both of these medicines are compared with clopidogrel, but not head to head. Since clopidogrel is a common denominator, it might be possible to estimate the effects of a hypothetical head-to-head comparison of ticagrelor with prasugrel [30]. Other examples illustrate attempts to estimate similar hypothetical effects, where entire networks of related results are combined in network meta-analyses [31, 32]. SBU's position in these types of meta-analyses is that they can be useful as analytical tools, but only in exceptional cases should they be included in an evidence profile (in part, because the necessary statistical assumptions are seldom met).

About software

There are a number of software programmes that can be used for meta-analysis. The simplest programme, available free of charge on the Internet, is Review Manager (RevMan), developed through the Cochrane Collaboration (www.cochrane.org). This programme complies with internationally established conventions but cannot yet handle hazard ratios and more complex types of meta-analyses, eg meta-regression, network meta-analyses, diagnostic meta-analyses, meta-analysis of correlations. Comprehensive Meta-Analysis (CMA) includes more functions than RevMan (eg meta-regression), but is subject to a fee (www.meta-analysis.com). Meta Disk, which was developed especially for meta-analyses in diagnostics, is still freely available via the Internet (www.hrc.es/investigacion/metadisk_en.htm). The programme with the greatest potential, but is subject to a charge and requires the most experience, is the general statistics program STATA (www.stata.com). Excel (part of the Office package) can also serve many functions and should be kept in mind.

A rapidly developing field

Meta-analyses and related methods are developing rapidly. Older methods are being modified and new ones are emerging. Meta-analyses have made most progress regarding intervention effects, but have not come so far with regard to eg diagnostics. In this context, it is important to follow developments via the international HTA network, for instance the Cochrane Collaboration, the GRADE working group, and others involved in developing conventions, systems, and transparency. The PRISMA statement (a refinement of the QUORUM statement) is of particular importance in working with meta-analyses. Three changes in the past decade can be emphasised: a) the focus has shifted from individual studies to outcomes (which can include results from multiple studies) that appraise the risk of bias; b) the importance of context and external validity has been emphasised more than previously; and c) old forms of evidence hierarchies are starting to be problematised (which implies that findings from observational studies may possibly be found to have a low risk of bias).

References

1. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidencebased medicine and patient choice. ACP J Club 2002;136:A11-4.
2. Borenstein M, Hedges LV, Higgins JPT, et al. Introduction to meta-analysis. Chichester: John Wiley & Sons Ltd; 2009.
3. Bond GR, Witheridge TF, Dincin J, Wasmer D. Assertive community treatment for frequent users of psychiatric hospitalsin a large city: A controlled study. Am J Community Psychol 1990;18:865-91.
4. Clarke GN, Herinckx HA, Kinney RF, Paulson RI, Cutler DL, Lewis K, Oxman E. Psychiatric ospitalizations,
5. arrests, emergency room visits, and homelessness of clients with serious and persistent mental illness: findings from a randomized trial of two ACT programs vs. usual care. Ment Health Serv Res 2000;2:155-64.
6. Conrad KJ, Hultman CI, Pope AR, Lyons JS, Baxter WC, Daghestani AN, et al. Case managed residential care for homeless addicted veterans: Results of a true experiment. Medical Care 1998;36:40-53.
7. Cox GB, Walker RD, Freng SA, Short BA, Meijer L, Gilchrist L. Outcome of a controlled trial of the effectiveness of intensive case management for chronic public inebriates. J Stud Alcohol 1998;59:523-32.
8. Lehman AF, Dixon LB, Kernan E, DeForge BR, Postrado LT. A randomized trial of assertive community treatment for homeless persons with severe mental illness. Arch Gen Psychiatry 1997;54:1038-43.
9. Rosenheck R, Kasprow W, Frisman L, Liu-Mares W. Cost-effectiveness of supported housing for homeless persons with mental illness. Arch Gen Psychiatry 2003;60:940-51.
10. Lipton FR, Nutt S, Sabatini A. Housing the homeless mentally ill: a longitudinal study of a treatment approach. Hosp Community Psychiatry 1988;39:40-5.

11. Morse GA, Calsyn RJ, Allen G, Tempelhoff B, Smith R. Experimental comparison of the effects of three treatment programs for homeless mentally ill people. Hosp Community Psychiatry 1992;43:1005-10.
12. Morse GA, Calsyn RJ, Dean Klinkenberg W, Helminiak TW, Wolff N, Drake RE, et al. Treating homeless clients with severe mental illness and substance use disorders: costs and outcomes. Community Ment
13. Health J 2006;42:377-404.
14. 12. Sosin MR, Bruni M, Reidy M. Paths and impacts in the progressive independence model: a homelessness and substance abuse intervention in Chicago. J Addict Dis 1995;14:1-20.
15. 13. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions Version 5.1.0. The Cochrane Collaboration. Available from www.cochrane-handbook. org; 2008.
16. 14. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database of Systematic Reviews 2009, Issue 1. Art. No.: MR000006. DOI: 10.1002/14651858. MR000006.pub3.
17. 15. Wallentin L, Becker RC, Budaj A, Cannon CP, Emanuelsson H, Held C, et al. Ticagrelor versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2009;361:1045-57.
18. 16. Cannon CP, Husted S, Harrington RA, Scirica BM, Emanuelsson H, Peters G, et al. Safety, tolerability, and initial efficacy of AZD6140, the first reversible oral adenosine diphosphate receptor antagonist,
19. compared with clopidogrel, in patients with non-ST-segment elevation acute coronary syndrome: primary results of the DISPERSE-2 trial. J Am Coll Cardiol 2007;50:1844-51.
20. 17. Campbell DS, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally & Company; 1963.
21. 18. Shadish WC, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Boston/New York: Houghton Mifflin Company; 2002.
22. 19. Golder S, Loke YK, Bland M. Meta-analy-ses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med 2011;8:e1001026.
23. 20. Hemmelgarn BR, Manns BJ, Zhang J, Tonelli M, Klarenbach S, Walsh M, Culleton BF. Association between multidisciplinary care and survival for elderly patients with chronic kidney disease. J Am Soc Nephrol 2007;18:993-9.
24. 21. Wong RY, Chittock DR, McLean N, Wilbur K. Discharge outcomes of older medical in-patients a pecialized acute care for elders unit compared with non-specialized units. Canadian Journal of Geriatrics
25. 2006;9:96-101.
26. 22. Meissner P, Andolsek K, Mears P, Fletcher B. Maximizing the functional status of geriatric patients in an acute community hospital setting. Gerontologist 1989;29:524-8.
27. 23. Stewart M, Suchak N, Scheve A, Popat-Thakkar V, David E, Laquinte J, Gloth FM 3rd. The impact of a geriatrics evaluation and management unit compared to standard care in a community teaching hospital.
28. Md Med J 1999;48:62-7.
29. 24. Zelada MA, Salinas R, Baztán JJ. Reduction of functional deterioration during hospitalization in an acute geriatric unit. Arch Gerontol Geriatr 2009;48:35-9. in an acute geriatric unit. Arch Gerontol Geriatr 2009;48:35-9.
30. 25. Brozek JL, Akl EA, Jaeschke R, Lang DM, Bossuyt P, Glasziou P, et al. Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies. Allergy 2009;64:1109-16.
31. 26. Gatsonis C, Paliwal P. Meta-analysis of diagnostic and screening test accuracy evaluations: methodologic primer. AJR Am J Roentgenol 2006;187:271-81.

32. 27. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A. Meta-DiSc: a software for meta-analysis of test accuracy data.BMC Med Res Methodol 2006;6:31.
33. 28. Field AP. Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. Psychol Methods 2001;6:161-80.
34. 29. Wiviott SD, Braunwald E, McCabe CH, Montalescot G, Ruzyllo W, Gottlieb S, et al. Prasugrel versus clopidogrel in patients with acute coronary syndromes. N Engl J Med 2007;357:2001-15.
35. 30. Biondi-Zoccai G, Lotrionte M, Agostoni P, Abbate A, Romagnoli E, Sangiorgi G, et al. Adjusted indirect comparison metaanalysis of prasugrel versus ticagrelor for patients with acute coronary syndromes. Int J Cardiol 2011;150:325-31.
36. 31. Woods BS, Hawkins N, Scott DA. Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: a tutorial. BMC Med Res Methodol
37. 2010;10:54.
38. 32. Wandel S, Jüni P, Tendal B, Nüesch E,Villiger PM, Welton NJ, et al. Effects of glucosamine, hondroitin, or placebo in patients with osteoarthritis of hip or knee: network meta-analysis. BMJ 2010;341:c4675.

Chapter 10

Grading evidence

The final step in the assessment is to appraise the whole body of scientific evidence. SBU uses a system partly based on international GRADE system [1,2] to classify the strength of the scientific evidence.

GRADE is a system that is continually being improved by the GRADE Working Group, which includes representation from SBU. GRADE was developed as a response to the wide variety of systems that were being used to grade evidence and the strength of recommendations. This multiplicity that preceded the development of GRADE led to some confusion and many found that reports occasionally missed important steps in the process or failed to clarify them.

In principle, GRADE is built on experience from other grading systems but focuses more clearly on the risk-benefit perspective. International bodies such as WHO, NICE, Cochrane Collaboration and BMJ have already adopted GRADE. In Sweden, the system is used by, for example, SBU and the National Board of Health and Welfare.

The quality of scientific evidence is rated in GRADE on a four-point scale, ie high, moderate, low and very low quality of evidence. The SBU-system uses a different terminology compared to the GRADE working group that replaces "high quality" with "strong evidence", "moderate quality" with "moderately strong evidence", "low quality" with "limited evidence", and "very low quality" with "insufficient evidence". This we do because it is closer our previous nomenclature and we feel it is slightly more descriptive. Thus, SBU classifies the strength of evidence as *strong* (++++), *moderately strong* (+++0), *limited* (++00) and *insufficient* (+000). Limited evidence may be sufficient to apply a method in clinical practice if other criteria are fulfilled, eg reasonable cost-effectiveness. Insufficient evidence means that more research is necessary before the method can be applied on a large scale. Strong scientific evidence is so solid that new research is unlikely to result in different conclusions. Limited scientific evidence implies a greater likelihood that new studies will modify the conclusions.

This chapter describes the procedure for using the SBU-system partly based on GRADE. Note that the system currently applies to intervention/treatment studies. However, it can be used in a similar way for studies on causal associations.

The GRADE system also includes a recommendation component that SBU does not use.

A table that summarises the findings gives a good overview of the quality of evidence

An appropriate initial step that facilitates future work is to summarise the results in a table; an example is given in Table 10.1. The table should present the pooled-effects of all important outcome measures separately (eg mortality, function and quality of life). In some cases, outcomes that are more in the nature of surrogate measures (eg HbA_{1c} and blood pressure) are also noted. Besides positive effects, outcomes should include negative consequences such as adverse effects and complications. The various outcome measures should be tabulated in descending order of importance. As Table 10.1 indicates, the columns “Scientific evidence” and “Comments” are filled in at a later stage after the overall appraisal.

Table 10.1 Summary of findings. Effect of antibiotic prophylaxis compared to placebo in maxillofacial surgery.
Outcome
Wound infection in surgery of jaw fractures
Study design No. patients (no. studies)
RCT 461 (3)
Mean risk in standard group (min-max)
39 % (20 %-62 %)
Relative risk (95 % CI)
RR 0.25 (0.15-0.41)
Absolute effect per 1000 patients
259 fewer
Scientific evidence
Comments

In this example of wound infection in surgical treatment of jaw fractures (Table 10.1) three randomised controlled trials (RCTs), with a total of 461 patients, were assessed. A meta-analysis that would yield a pooled ~~composite~~ effect in numerical terms proved feasible [3]. In many cases, however, precise data on mean values and risk differences cannot be generated. If data are too heterogeneous, the results cannot be combined in a pooled meta-analysis with an absolute or relative risk difference. Instead, the effects can be presented as minimum-maximum variation.

If a sufficient number of well-executed RCTs is available, observational studies are normally not included in the appraisal of positive effects. However, when randomised trials are either not available or deficient, observational studies can provide important supplementary information and contribute to the overall grading of evidence in both positive and negative directions. The table of results then presents RCTs and observational studies separately. When appraising risks, it is often important to include observational studies since in most cases RCTs are not designed to answer questions about long-term risks.

Preliminary strength of evidence

As mentioned above, each outcome in the table of results is graded by the strength of the evidence.

The GRADE process as well as the SBU-process starts with a preliminary grading of the evidence. This is based entirely on the design of the studies that make up the scientific evidence. The preliminary strength of evidence can then be adjusted upwards or downwards depending on a number of quality factors (described below). If the evidence is based mainly on randomised trials, where there is least risk of systematic errors, the preliminary grade will be *strong* in the SBU-system.

For assessments of diagnostic accuracy, the GRADE working group considered several years ago-that even in observational studies the preliminary grade for the evidence should start from high quality-[4]. This is debatable but SBU currently accepts it. It is, however, essential to analyse whether improved diagnostics will ultimately benefit the patient.

Facts 10.1 Preliminary strength of evidence based on study design, and reasons for down- or upgrading the evidence.
Preliminary strength of evidence for intervention studies according to the SBU-system:
Strength of evidence
Strong
Moderate
Limited
Insufficient
Study design
Randomised controlled studies
Observational studies
Case studies, etc.

The strength of evidence may then be down- or upgraded based on the following according to the SBU-system:
Downgrade if

- Deficient study quality (max -2)
- Deficient consistency between studies (max -2)
- Deficient generalisability/relevance (max -2)
- Deficient precision (max -2)
- High risk of publication bias (max -2)

The GRADE working group uses the terms “risk of bias”, “inconsistency”, “indirectness”, “imprecision”, and “publication bias” respectively
Upgrade if

- Large effects, no probable confounders (max +2)
- Clear dose-response relationship (max +1)
- The absence of confounders in the analysis should lead to better treatment results in the control group, ie high probability that the effects are underestimated (max +1)

The GRADE working group uses the following expressions: “large effect”, “dose response evidence of a gradient”, “all plausible residual confounding would reduce a demonstrated effect or would suggest a spurious effect if no effect was observed”

Eight factors influence the final strength of evidence

In the final assessment of the strength of evidence, the expert panel appraises the scientific evidence's validity. The preliminary assessment of the strength of evidence is downgraded if the evidence is uncertain with regard to:

- study quality
- consistency/coherence
- generalisability/relevance
- data precision
- risk of publication bias.

The GRADE working group uses the terms "risk of bias", "inconsistency", "indirectness", "imprecision", and "publication bias" respectively

The strength of evidence can be downgraded one or two steps for each factor depending on the magnitude of the deficiencies. If the deficiencies are minor, this can be noted without downgrading. If minor deficiencies apply to several factors, this could lead to downgrading the strength of evidence one step. However, bear in mind that observational studies are already downgraded on account of their study design and should therefore generally not be further downgraded for a lack of control of confounders. Serious deficiencies in controlling for confounders can, however, be a reason for downgrading observational studies by a total of one step. Deficiencies involving consistency, generalisability, precision, and risk of publication bias can be reasons for downgrading observational studies.

In some instances there may be reasons for *raising* the strength of evidence one or two steps. This may apply when the scientific evidence comes from large, well-executed observational studies with good control of confounders. The three factors that can increase the strength of evidence are:

- large effects
- dose-response relationship
- high probability that the effect in the study is underestimated.

The GRADE working group uses the following expressions: "large effect", "dose response evidence of a gradient", "all plausible residual confounding would reduce a demonstrated effect or would suggest a spurious effect if no effect was observed

Study quality according to the SBU-system

During the assessment phase, the quality of each study is appraised individually. This step involves a comprehensive evaluation of the material. Traditional factors such as randomisation, blinding and attrition are considered first. How well has the study been executed in terms of the overall randomisation? Are some studies well executed while the randomisation process is not clear in other studies? Other factors on the check lists may also be important to consider, as are topic-specific, methodological problems that the experts have identified.

For cohort studies and other observational studies, a key question concerns the comparability of the trial and control groups. This means that the appraisal of overall study quality is highly dependent on the extent to which the studies have controlled for confounders (Chapter 6).

Example 10.1 Risk of overestimating the effects.
Observational studies indicated that oestrogen replacement therapy reduced the risk of cardiovascular disease. Since oestrogen replacement therapy is known to be more common in women from higher socioeconomic groups, the observer effects are likely to be overestimated and confidence in the association would then be downgraded.

Diagnostic studies can be downgraded if recruitment is not consecutive, if evaluators are not blinded, and if the study has some type of verification bias.

In the SBU-system-classification, the strength of evidence can be adjusted downwards one step if the evidence has serious quality deficiencies, and two steps if the limitations are very serious. Note that GRADE*, per se*, does not require the evidence to be of at least moderate quality. In SBU's process, which excludes low-quality studies, the evidence is seldom downgraded two steps on the basis of methodological deficiencies.

Consistency/coherence according to the SBU-system

Here we appraise the extent to which the studies produce the same results. Do they point in the same direction? Is the size of the effect comparable in the different studies? Meta-analyses can be helpful in appraising the degree of consistency.

Consistency will depend on the studies' similarity as regards population, exactly how the intervention was performed, and which control group was used. Another important factor is whether one and the same research group conducted a sizeable proportion of the studies.

Generally, the total material's reliability increases if the studies were conducted by different research groups using different populations and if their results consistently point in the same direction. If the studies reveal excessive or subnormal risks, the strength of the evidence can be downgraded one step. The same applies if the size of effects varies substantially between studies, leading to greater uncertainty.

In some cases, the differences can be explained by the studies' dissimilarities, eg that they investigated different populations. In such cases, it may be more appropriate to break down the material and formulate conclusions for each of the individual populations.

Generalisability/relevance according to the SBU-system

Generalisability refers to the extent to which the scientific evidence is relevant and can be applied to Swedish conditions. Important questions to consider include: how well does the population reflect that observed in everyday Swedish practice? Is the intervention performed in the same manner as in Sweden? Is the control group relevant? Is the health care environment reasonably similar?

An example of deficient generalisability would be a control group that receives a treatment which is not available in Sweden, making it impossible to compare the intervention's effectiveness with that of standard Swedish practice.

If generalisability and relevance are deficient, the strength of the evidence can be adjusted downwards one or two steps. As with study quality, studies that are less relevant to Swedish conditions have probably already been eliminated on the basis of the checklist for relevance (Chapter 5 & Appendix 1). A special case arises when just one study is available to measure the effects with an important outcome measure. SBU considers that generalisability usually requires at least two studies. This means that in such a case the evidence grade would be adjusted downwards one step. An exception might be, eg evidence involving a large and very well-executed, randomised multicentre study.

Precision of data according to the SBU-system

This criterion estimates the uncertainty of the overall effects. Few observations and wide confidence intervals in the different studies would lead to lower precision. Precision depends on the number of events, the number of people in the groups and the relative risk reduction.

A way of determining whether precision is uncertain is to perform a power analysis based on the total number of observations in the included studies. If the number of observations in these studies falls below the number required to yield statistically significant results, there may be reason to downgrade because of inadequate precision. If the studies are very small, one needs to be very observant of whether the differences in baseline data are large even though the outcome is statistically significant. If the baseline data differ substantially between the groups, there may be reason to downgrade on the basis of data precision.

Risk of publication bias according to the SBU-system

Publication bias indicates that published studies do not include all the scientific evidence. The risk of publication bias increases, eg if the evidence comes entirely from small studies that have been conducted by the same research group and have substantial methodological deficiencies. Early assessments of new methods often belong to this category. Similar situations arise when the evidence is comprised of corporate-sponsored studies, eg pharmaceutical studies. Also, if the first author in all studies of a new method is its inventor, there is reason to check for publication bias.

The effects of selective publication are well illustrated in a study by the Swedish Medical Products Agency [6]. Systematic reviews [7] and many other studies [8,9] uniformly suggest that studies sponsored by industry or other parties with an interest in the findings overestimate the effects of their products. For instance, cost-effectiveness ratios below USD 20,000 per quality-adjusted life year gained were indicated more than twice as often in analyses by pharmaceutical companies compared with non-industry sponsored studies [8,9].

Publication bias is often difficult to confirm, but some methods can facilitate evaluation. One approach involves using central registers (eg www.controlled-trials.com and www.clinicaltrials.gov) of on-going clinical trials. The evaluation should include a check of which studies are included in the registers.

The registers date from the turn of the millennium and can provide valuable information on studies initiated during the past decade. Studies that, according to the register, were completed several years ago could be a possible source of publication bias. Moreover, the existence of large ongoing studies might increase the risk that their results may affect the strength of evidence in the future. So registers could also provide some guidance in appraising uncertainty.

Another way to identify publication bias is to search the databases for conference abstracts. The presence of abstracts of studies that have not been published within a reasonable number of years is a strong indicator of publication bias. This, however, is a labour-intensive process.

Funnel plots (Appendix 9) may be helpful in appraising risk when many studies have been found. As a rule of thumb, at least five studies are necessary for the analysis to be meaningful.

Effect size according to the SBU-system

Effect size is a criterion that can be used if there are at least two large, well-executed, observational studies that are adequately controlled for confounders. High effects increase the probability of a causal association.

According to GRADE, the overall strength of evidence can be raised one step if the pooled effect, defined as relative risk (RR), exceeds 2 (RR>2 or RR<0.5). The strength of evidence can be raised two steps if RR>5.0 (or RR<0.2). If study quality has previously been appraised as insufficient, one can refrain from raising the strength of evidence on the grounds of large effect size.

Note that if the outcome measure is an odds ratio (OR), the effects can be overestimated if the outcomes are common (>10 %). However, the thresholds may need to be adjusted. For a description of the differences between RR and OR and for a formula to convert OR to RR and vice versa, see the section on statistics (Appendix 9).

Example 10.2 Large effects can raise the strength of evidence according to the SBU-system.

A meta-analysis of observational studies showed that bicycle helmets reduced the risk of head trauma, reporting an odds ratio of 0.31 (0.26-0.37) [5]. This is a large effect, leading to the strength of evidence being adjusted one step upwards. Unless there are reasons for a downward adjustment because of deficiencies in the evidence, the scientific evidence would be graded moderately strong (+++0).

Dose-response relationship according to the SBU-system.

This criterion is also limited to large, well-executed, observational studies. Dose-response can apply to both effects and risks. A dose-response relationship increases the credibility of an intervention having an effect. Dose-response can apply to pharmaceuticals and other interventions. Effects can be measured within a study and between studies. In general, a dose-response relationship is more credible when shown within a study than in comparisons between studies.

A dose-response relationship can raise the strength of evidence one step.

Example 10.3 Dose-response relationship can raise the strength of evidence.

The SBU report "Dietary Treatment of Diabetes" [10] addressed the risk of myocardial infarction in people with diabetes and different levels of alcohol consumption. Three large observational studies with a total of 10,312 patients all reported a dose-response relationship showing that the group with higher alcohol consumption had a substantially lower relative risk than the group with little or no alcohol consumption. Hence, the expert panel raised the strength of evidence to moderately strong because of the dose-response relationship.

High probability that effects are underestimated according to the SBU-system

In isolated cases, the strength of evidence can be adjusted upwards if there is a high probability that the studies have underestimated the effects. This could happen if the study was unable to adjust for certain confounders, suggesting that the effects have been underestimated.

Example 10.4 Underestimated effect can raise the quality of evidence.

A systematic review covering 38 million patients showed a higher mortality rate at private for-profit hospitals than at private non-profit hospitals [11]. The GRADE Working Group reported that it is probable that the patients at non-profit hospitals were sicker, while the for-profit hospitals had more resources and more patients who were well-insured. Thus, evidence that the mortality rate is actually higher at for-profit hospitals was reinforced.

Overall grading of the strength of evidence according to the SBU-system

In justifying their appraisal, the project group should provide written comments on how each of the above-mentioned factors affects the evidence. A table of evidence can be prepared to facilitate this work (Table 10.2). The table should show the preliminary assessment of the strength of the evidence and the adjustments in the next step.

Using wound infection as the outcome measure in the example of antibiotic prophylaxis in maxillofacial surgery, the first step involves preparing a table of evidence (Table 10.2) as a basis for the summary of findings (Table 10.3).

In this case, the reasons for downgrading included deficient study quality and the overall deficiencies resulting from reporting few outcomes and using different antibiotics.

Table 10.2 Table of evidence. Effect of antibiotic prophylaxis compared to placebo in maxillofacial surgery.

Studies/Patients	3/461	
Design		RCT ++++
Study quality	-1*	
Consistency	0	
Generalisability		0**
Imprecise data		-1***
Publication bias		0
Effect size	0	
Dose-response		0
Confounders	0	

*no blinding, one study did not report attrition
** different antibiotics
*** few outcomes, together with **-1.

Table 10.3 presents the summary findings for antibiotic prophylaxis .Here, relative risk was high and could have motivated an upgrade, but the expert group decided that this was not justified on account of the deficiencies in study quality.

When the scientific evidence for all outcomes has been appraised, the summary of findings provides an overall profile of the scientific evidence on a given question.

Table 10.3 Summary of findings. Effects of antibiotic prophylaxis compared to placebo in maxillofacial surgery.

Outcome/effect measure
Wound infection in surgery of jaw fractures
Study design No. patients (no. of studies)
RCT 461 (3)
Mean risk in standard group (95 % CI)
39 % (20%-62%)
Relative risk (95 % CI)
RR 0.25 (0.15-0.41)
Absolute effect per 1000 patients
259 fewer

Scientific evidence
++00 Limited
Comments
Weakness in study quality (-1)
Few outcomes (-1)

In general, SBU considers is that a single study of limited size is not sufficient for any evidence. In these cases there is great uncertainty about precision and no other study is available to confirm the results. Any exception to this rule, such as a large multicentre study or other strong reasons, has to be motivated.

Interpreting the strength of evidence according to the SBU-system

Even if the assessment does not result in a recommendation, it does provide guidance for health services. If the scientific evidence is insufficient, this indicates a need for more research before health services can use the method routinely. A method supported by limited evidence can be used in health services provided it meets other requirements for an acceptable balance between risks and benefits, cost-effectiveness, and ethical acceptability. A method supported by moderately strong or strong evidence can probably be applied provided it is supported by the economic, ethical and social analyses in the assessment.

Facts 10.2 Study quality, strength of evidence and conclusions (according to the SBU-system)

Study quality indicates the scientific quality of an individual study and its ability to answer a particular question in a reliable way.

Strength of evidence is based on an appraisal of the strength of the collective scientific evidence for answering a particular question in a reliable way. SBU uses the GRADE system, which has been developed internationally. Every effect measure starts with an overall appraisal based on study design. The strength of evidence may be influenced by the presence of factors that strengthen or weaken the evidence, such as study quality, relevance, consistency, generalisability, effect size, data precision, risk of publication bias, and other aspects, eg dose-response relationship.

The strength of evidence is graded on four levels:

Strong scientific evidence (++++)
Based on studies of high or moderate quality with no factors that would weaken an overall appraisal.

Moderately strong scientific evidence (+++0)
Based on studies of high or moderate quality with isolated factors that weaken an overall appraisal.

Limited scientific evidence (++00)

Based on studies of high or moderate quality with factors that weaken an overall appraisal.

Insufficient scientific evidence (+000)
The scientific evidence is insufficient when scientific evidence is absent, available studies have low quality or studies of similar quality report conflicting results.

The stronger the evidence, the less likely it is that the results will be influenced by new research findings in the foreseeable future.

Conclusions

SBU's conclusions present an overall appraisal of benefits, risks and cost-effectiveness.

References

1. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.

2. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines 1. Introduction – GRADE evidence profiles and summary of finding tables. J Clin Epidemiol 2011;64:383-94.

3. SBU. Antibiotikaprofylax vid kirurgiska ingrepp. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 200. ISBN 978-91-85413-36-2.

4. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106-10.

5. Thompson DC, Rivara F, Thompson R. Helmets for preventing head and facial injuries in bicyclists. Cochrane Database of Systematic Reviews 1999, Issue 4. Art. No.: CD001855. DOI: 10.1002/14651858. CD001855.

6. Melander H, Ahlqvist-Rastad J, Meijer G, Beermann B. Evidence b(i)ased medicine-selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003;326:1171-3.

7. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 2003;326:1167-70.

8. Jörgensen AW, Maric KL, Tendal B, Faurschou A, Götzche PC. Industry-supported meta-analysis compared with meta-analysis with non-profit or no support: differences in methodological quality and conclusions. BMC Med Res Methodol 2008;8:60.

9. Bell CM, Urbach DR, Ray JG, Bayoumi A, Rosen AB, Greenberg D, Neumann PJ. Bias in published cost effectiveness studies: systematic review. BMJ 2006; 332:699-703.

10. SBU. Mat vid diabetes. En systematisk litteraturöversikt. Stockholm: Statens beredning för medicinsk utvärdering (SBU); 2010. SBU-rapport nr 201. ISBN 978-91-85413-37-9.

11. Devereaux PJ, Choi PT, Lacchetti C, Weaver B, Schünemann HJ, Haines T, et al. A systematic review and meta-analysis of studies comparing mortality rates of private for-profit and private not-for-profit hospitals. CMAJ 2002; 166:1399-406.

Chapter 11

Health economics

Introduction

The SBU evidence base should result in more health for the money/resources invested. Therefore, SBU's mission to assess medical methods (health technology assessment) includes assessments from an economic perspective. Economy plays an important role because society's resources are limited. Together with a growing demand for health services and social care [1, 2], this leads to a gap between what the health and social welfare system can offer and what its citizens demand. Consequently, priorities have to be established for the treatments on which the available resources are to be spent. Health economic evaluations that compare the costs and effects of two or more alternatives in a structured way are an aid for decision-makers to determine whether a method generates sufficient health in proportion to its costs.

Analyses of economic aspects constitute an important part of SBU projects. This chapter first describes the task of compiling and assessing the quality of the health economic literature and how SBU works with in-house economic evaluations. It then presents some health economic concepts and methods that constitute the basis of health economic evaluations.

SBU's work with health economic evaluations

Health economic aspects of SBU's projects are usually considered in one or more of the following forms:

- Cost-of-illness and burden-of-illness
- Health economic evaluations
    - Systematic reviews of available literature on cost-effectiveness
    - In-house cost-effectiveness analyses
- Budget impact analyses

Cost-of-illness and burden-of-illness

Illness and ill health can be described and measured from different perspectives; based on the individual's criteria (the individual's/patient's own experience (self-reported morbidity, illness)) or on medical criteria (diagnosed morbidity, disease).

One approach to providing an overall (comprehensive) description of illness and of changes in the illness panorama involves calculating the societal costs of illness and injuries. This results in what are usually known as cost-of-illness studies (COI) [3, 4]. Another approach involves calculating the loss of healthy years by using measures that combine quantity and quality of life, most frequently DALYs but also QALYs. DALYs (disability-adjusted life years), which are recommended by the WHO (World Health Organization), combine premature death and the severity of various health problems (1= death, 0= full health) using standardised methods [5]. DALYs have been used, for example, in Global Burden of Disease 2010 [6] to describe the losses of health for different populations in terms of specific diseases, injuries and risk factors. QALYs (quality-adjusted life years) measure remaining years of life and quality of life (1=full health, 0= death) and are mostly used as a measure of health in health economic evaluations (see more information below), but also to compare health between populations.

The societal costs or the burden of illness due to different diseases provide information about the magnitude of the problem in a society but say little about the cost-effectiveness of different methods. Consequently such studies cannot be used to decide how resources should be distributed in health and social care [7, 8].

Health economic evaluation

Systemic reviews and health economic evaluation

The first step in SBU’s work to describe the cost-effectiveness of alternative methods is to perform a systemic review of the published literature. The literature search is based on the terms that are used to search the medical literature plus economic search terms. The relevance of studies is judged on the basis of the project’s PICOs (see Chapter 3) and whether the studies include economic analyses. Thereafter, the quality of the studies and their applicability to Swedish circumstances in health and social care services are assessed.

The quality of health economic evaluations depends on the quality of the study data and the principles used to calculate costs and effects. The economic evaluation cannot be better than available data allow. There are a number of check lists for assessing the quality of economic analyses [9-11]. The most commonly used is by Drummond and collaborators [10]. There are other similar checklists [9, 11] and specific checklists have been developed for assessing the quality of health economic models [12]. Based on these checklists and experience from previous work, SBU has developed two checklists for quality assessment, one for empirical studies and the other for modelling studies (Appendixes 7 and 8). They have the same foundation but are adjusted to better reflect specific issues regarding the types of study design. The checklists are supplemented with three questions on the risk of conflicts of interest, identical to the ones in the checklists for medical studies. Regarding the assessment of the quality of the data items used in models (for example, epidemiological data, costs and quality of life), we refer to Cooper and collaborators [13], who report a quality hierarchy of data that is possible to use in models.

Applicability to Swedish circumstances is assessed on how well data items of the health economic analysis correspond with Swedish data. Differences in costs, mortality, quality of life, and epidemiological data, such as the prevalence of the health problem, all influence the results of the health economic analysis [14, 15]. In general, it would be preferable to obtain all data from Swedish data sources of good quality [13]. The majority of health economic analyses are, however, performed in other countries. It is therefore an important part of SBU’s assessment to consider the extent to which an analysis based on Swedish data would give similar results. The setting in which the evaluated method was implemented can also differ from ordinary Swedish circumstances. For example, certain methods may be used in primary care in some countries and in hospital care in other countries. As the health service setting influences the results of the health economic analysis, it is necessary to assess whether the organisational circumstances in the study are similar to Swedish circumstances.

When using the checklists it is important to remember that only a few health economic analyses fully meet the criteria. This does not mean that studies which do not fulfill all criteria are of no use. One should just be aware of the shortcomings when interpreting the results. The quality of studies is rated in an overall judgment as high, moderate or low.

In-house analyses

Sometimes a systematic review cannot yield answers about economic issues. This is particularly the case when there are just a few health economic studies in a certain field and when the results of available empirical studies from other countries are not applicable to Swedish circumstances. If trustworthy data on costs and effects can be obtained, SBU can make in-house and adapted analyses of cost-effectiveness. Such analyses are based on the established clinical evidence obtained from the medical review. The analyses may then be supplemented with calculations of the costs of alternatives.

Depending on the complexity of the issue and access to data, these analyses may be more or less comprehensive. For example, if there is evidence that two treatments have equal effects, it may suffice to compare their costs to establish which is the more cost-effective. In other cases a more comprehensive analysis of both costs and effects may be necessary. Sometimes it is appropriate to use available data to perform modelling analyses. These analyses, produced within the project, are often based on available clinical studies but adapted to Swedish circumstances (ie costs, epidemiology). These analyses should comply with the required standard for publication in scientific journals. The project's medical experts should be consulted for advice on the relevance and validity of the data used in the calculations. Irrespective of the form of the modelling analysis, in-house modelling exercises should be subjected to a detailed sensitivity analysis as well as to internal and external peer reviews.

What is a health economic evaluation?

In health economic evaluations, two or more alternative methods are compared in terms of costs as well as effects with the aim of establishing which of them is most cost-effective [16]. Thus cost-effectiveness is a relative concept. In some cases the most relevant alternative is "no treatment". It is common to distinguish between five different types of health economic evaluation. All include costs but differ as to how effects are described and valued, see Table 11.1.

Table 11.1 Types of method for health economic evaluations.

Type of evaluation	Outcome (effect) measure	How the results are presented
Cost- Minimisation Analysis (CMA)	No outcome measure since the effects are presumed to be identical	Presents only costs
Cost-Consequences Analysis (CCA)	Several outcome measures	Presents costs and effects
Cost-Effectiveness Analysis (CEA)	Physical entities, eg life-years, number of individuals with positive results, average reduction of risk markers	Cost per effect, ie per gained life-year (LYS), per adolescent graduating from compulsory school, per unit improvement in a depression scale
Cost-Utility Analysis (CUA)	A measure that combines life-years with health or quality of life, eg quality-adjusted life-years (QALYs)	Cost per gained quality-adjusted life-year (QALY, DALY)
Cost-Benefit Analysis (CBA)	The effects, eg life-years or reduced pain, valued as benefits in monetary terms, eg Swedish krona (willingness-to-pay)	Net social benefit

The first four types of evaluation are actually variations on the same methodology and are sometimes referred to as cost-effectiveness analyses. In contrast to the cost-benefit analysis, they do not value the effect in monetary terms.

In the *cost-minimisation analysis* the effects are assumed to be equivalent and the alternatives are therefore only compared with regard to their costs.

In a *cost-consequences analysis* the costs and several different effects are presented, for example the number of home visits, the ability to walk and the quality of life of the patient's family, but without aggregating them. The method allows the decision-makers to choose the most relevant aspects, based on the specific situation, and draw their own conclusions.

A *cost-effectiveness analysis* uses a one-dimensional measure of effect, such as the number of cured patients or those with a positive outcome, the number of days without pain, the reduction in risk markers, or the number of life-years saved (LYS).

A *cost-utility analysis* relates the costs to a measure that combines survival and quality of life, for example the number of quality-adjusted life years, QALYs. Sometimes cost-utility analyses are called cost-effectiveness analyses with QALY as the effect measure.

In a cost-benefit analysis the effects are valued in monetary terms and the analysis accordingly indicates the method's profitability; the evaluated method results in a net social benefit if the benefits are greater than the costs. This type of analysis has been considered difficult or impossible to apply in health care on account of the practical and ethical difficulties of valuing the effects in monetary terms [17, 18]. Methodological developments in recent years, such as new methods to estimate willingness-to-pay (WTP), have to some extent made cost-benefit analyses more useful but ethical and methodological problems still remain. The value of the effect in monetary terms is often estimated by asking patients about their willingness to pay for the effect. Since willingness-to-pay for health is related to income there have been concerns that treatments directed at higher income populations would be valued higher than treatments for populations with lower incomes. As a parallel to this it can be mentioned that other Swedish national agencies in the transport and environmental fields employ cost-benefit analyses, including the monetary value of life (the value of a statistical life), to estimate the societal net benefit of investment alternatives.

Different types of evaluation answer to different issues. The issue should determine the choice of method but access to relevant data also plays a part. If the evaluation is to form the basis for choosing between treatment options (ie two alternative medicines) with the same therapeutic effect and no differences in side effects, the natural choice is a cost-minimisation analysis. If the issue is choosing between alternative methods that primarily affect mortality, a cost-effectiveness analysis with life-years as effect measure may be sufficient. If, on the contrary, the issue is alternative treatments of a condition that is not life-threatening, it is necessary to consider effects on quality of life. Here the appropriate method would be a cost-utility analysis.

What is cost-effectiveness?

Data are needed on both costs and effects in order to decide which of two methods is the more cost-effective. If a new method entails lower costs and higher effects than the old method, the new method is "dominant" and choosing it is a simple matter from a health economic view. Often, however, methods that are more effective are also more costly. The nine possible alternatives when methods are compared in terms of costs and effects are shown in Table 11.2.

For alternatives 2, 3 and 6, the old method is cost-effective and should remain in use. The opposite applies to alternatives 4, 7 and 8, that is, the new method is more cost-effective and should therefore be implemented instead. For alternative 5 the methods do not differ in either costs or effects and there are no reasons for switching to the new method. Additional analyses may be necessary for alternative 1 and definitely for alternative 9, see also Figure 11.1.

Table 11.2. A decision matrix for cost-effectiveness.

New method compared with old	**Lower effect**	**Same effect**	**Higher effect**
Lower cost	1.Situation unclear, consider doing incremental analysis	4.Implement the new method	7.Implement the new method
Same cost	2. Keep the old method	5.The methods are equivalent	8.Implement the new method
Higher cost	3. Keep the old method	6. Keep the old method	9. Situation unclear, consider doing incremental analysis

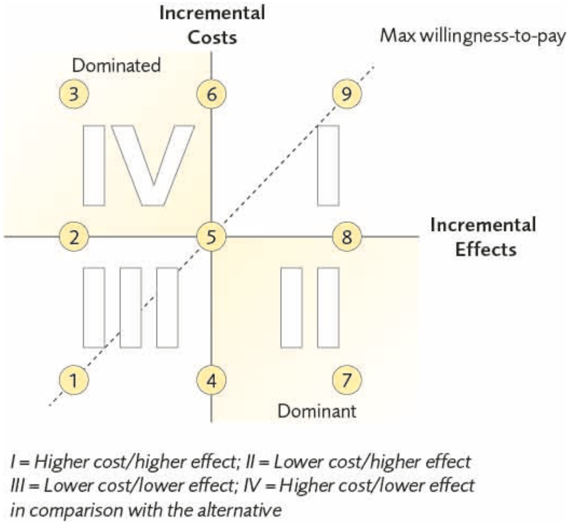
The result of a health economic analysis is often presented as an incremental cost-effectiveness ratio (ICER), which is the ratio between the difference in the alternatives' costs and the difference in their effects:

ICER= Cost A - Cost B/ Effect A - Effect B

Cost A is the costs arising from method A, Cost B the costs for method B; Effect A is the effects that follow from method A, Effect B the effects from method B. This means that the ICER calculates the costs of achieving one more unit of effect (eg a gained life-year or an adolescent graduating) by switching from one method to another. The new method is cost-effective if society's willingness to pay for an effect is higher than the extra costs needed to achieve the effect with that particular method.

The result can also be depicted in the cost-effectiveness plain, where the differences in incremental costs and incremental effects are plotted in a figure with four quadrants (Figure 11.1). Since quadrants IV and II supply obvious answers (the old method dominates in IV and the new method dominates in II) the focus, in general, is on quadrants I and III. The plots lie in these when the new method entails higher effects but also higher costs, or lower costs but also lower effects compared with the old method.

Figure 11.1 The cost-effectiveness plain



I = Higher cost/higher effect; II = Lower cost/higher effect III = Lower cost/lower effect; IV = Higher cost/lower effect in comparison with the alternative

Given that one knows how much society is willing to pay for one unit of effect, a limit can be inferred to determine whether ICERs are considered cost-effective. This limit is depicted as a line that passes through quadrants I and III; all methods whose incremental costs and incremental effects lie on the right of this line will be considered cost-effective.

Choice of perspective for the analysis

The general recommendation is that the analysis should be conducted from a societal perspective, so that it reflects the costs and effects of all societal sectors and thus does not result in suboptimisation between sectors. Using a societal perspective implies that costs and effects are considered irrespective of the sector in which they originate. However, it might be of interest to describe how costs and effects are distributed between different parties, such as patient, county council, community, state and other societal sectors.

Costs

An economic analysis includes both costs and cost savings expressed in monetary terms. Costs arise when resources are utilized in a certain treatment. If the treatment has a positive effect in terms of reduced illness, this can imply future cost savings. One example is an antismoking program that carries initial costs but eventually results in cost savings, as the health care sector does not have to spend resources on treating smoking-related diseases.

From a societal perspective, all relevant costs associated with a method should be identified, quantified and valued. In health economics the theoretically correct concept of cost is the *opportunity cost,* that is, the value of what can be achieved with the best alternative use of the resources. In practice, one often has to resort to using market prices or costs derived from the health services' accounts.

Costs related to health and social care services can be divided into different categories: direct health and social care costs, direct other costs and indirect costs [19]. Direct costs consist of the consumption of resources that arises as a direct consequence of treatment and care, while indirect costs are resources that are lost indirectly due to illness or treatment, for example reduced ability to work. Examples of direct costs are personnel time, material, buildings, aids and medical devices, and costs for the patient/family. Which costs to include depends on the type of method that is evaluated. In certain cases, costs for sectors of society other than those delivering the services can be most important for the analysis. The most important indirect cost is the reduced production due to disease.

Basic data for calculating costs can be found in Swedish registers or statistical databases. For example the Swedish Board of Health and Welfare maintains health data registers and databases containing information about the number of episodes of care, numbers of surgical operations, days of inpatient care, mean days of inpatient care as well as consumption of drugs in different age groups distributed by diagnoses, type of surgical operation or DRGs (Diagnosis-related groups). The Swedish Association of Local Authorities and Regions (SALAR) keeps two databases on costs: KPP, which contains data on cost per patient at certain hospitals, and KPB, which contains costs per user for social care of elderly and persons with disabilities in certain municipalities. Regional healthcare prices and remunerations are also published by some regional committees. A further source is the national quality registers, many of which contain quite comprehensive data on treatments and patients.

Estimating the value of production

Costs due to reduced production arise mainly when a person cannot participate in remunerated work due to illness or treatment, but also when a person's productivity is lower than previously as a consequence of illness or injury. The latter is usually called sickness presence or "presenteeism". Some older studies, particularly cost of illness (COI) studies, also include reduced production due to premature death, usually before the age of 65.

Another situation is that a treatment contributes to persons previously on sick leave being able to start working again, which results in increased production. To clarify this aspect SBU has chosen to use the expression "influence on production", even though it is often called production loss in the literature.

There are two methods for estimating the value of production: the human capital method and the friction cost method [16]. The human capital method assumes that production can be valued as a market price, that is, the wage rate plus payroll taxes. A drawback with the human capital method is that it can result in overestimates. This is a well-known controversy in the health economic literature and several Dutch health economists recommend the friction cost method instead [20, 21]. Friction is the period (including costs) that passes before a previously unemployed person replaces another worker in full. The method can be regarded as a pragmatic approach to the fact that influence on production is often concentrated to short periods of sick leave or to the start of sick leave before another person can take over the work.

The analysis usually does not include the influence on production from individuals who are not of working age. This has been criticised since retired persons often contribute to informal production, which should also be valued and included in the health economic analysis [22].

To include influence on production only for those capable of working can also conflict with the Swedish national priority-setting principle of human dignity. The principle states that priorities in Swedish health and welfare services shall be set "independently of personal characteristics and functions in the society" [22, 23], where age is considered to be a personal characteristic. If influence on production is included in the analysis only for individuals under age 65, measures directed at the elderly could be assigned lower priority. It has therefore been recommended that the results of health economic analyses are presented both with and without costs from influence on production [16, 22], an approach that SBU follows.

If wages are used to calculate the value of production, they should be standard or mean wages, since inter-group differences in wage rates can arise on other grounds than the actual value of production. One example is the difference in wage rates between men and women. If production by men and women, respectively, is valued differently based on official data, a treatment aimed at men will result in a lower ICER than the same treatment for women. This in turn will lead to higher priority for treatments aimed at men. To avoid this, production can be valued at the average national wage rate, instead of the average salary for men and women, respectively.

Quality-adjusted life-years (QALYs)

A common recommendation is that health economic analysis should use quality-adjusted life-years (QALY) as a measure of effect [24 26]. QALYs value both the duration (survival) and the quality of life, that is, both life-years and health status, including possible side effects. Quality-of-life is valued on a scale from 0 to 1, where 0=death and 1= full health. For example, five years with a quality-of-life of 0.7 will result in 3.5 QALYs [5*0.7). The major advantage of QALYs is that they can be used to compare completely different interventions. This assumes, however, that high quality and generally valid quality-of-life weights (QALY-weights) are available.

QALY-weights can be estimated with direct or indirect methods. The direct methods are used to estimate the value of different health states. The value is estimated on a scale where 0 denotes death and 1 denotes full health. The indirect methods use questionnaires, often called quality-of-life instruments, which are connected to valuations, often called tariffs, that have been elicited with a direct method.

The most common direct methods are standard gamble (SG) [27], time trade-off (TTO) [28] and the visual analogue scale (VAS) [29]. All three can be used either to ask patients to value their own quality-of-life or to ask the general population to value hypothetical conditions. SG and TTO are based on scenarios that the individual can choose between, while VAS is based on a linear scale where the individual marks how he or she values a health state, from best possible to worst possible health.

The most common indirect instruments are EQ-5D [30], SF-6D [31] and HUI-3 [32]. The questionnaires cover different aspects of health and the tariffs for converting the answers to the questionnaire into QALY-weights have been developed in different ways. The population groups that perform the valuations also differ, with three possible groups: the general population, patients who value their health status, and experts. In general, patients report a higher quality-of-life for health states than the general population.

Most tariffs in current use are based on valuations from the general population. For example, the most commonly used tariff for EQ-5D (commonly used in Sweden) is based on valuations made by 3395 British people who valued the different combinations of alternatives in the EQ-5D instrument with the TTO method [34]. A Swedish tariff based on patient valuations was published recently [35]. It values most health states higher than the British tariffs. The new Swedish tariff is generated from patients' valuations of their own health state, whereas the British tariff used the general population's valuations of hypothetical conditions. The choice of tariff can accordingly influence the results of health economic evaluations.

The methods used to estimate the quality-of-life weights reportedly yield different values for a number of diseases and conditions [37-43], which in turn can affect costs per QALY. This might be due to differences in the health states included in the questionnaires, in the eliciting method used, in the statistical techniques used to estimate the tariffs, and in the population group from which the tariffs were elicited. Since tariffs from different countries also differ [44], the QALY weights should be obtained via direct measurements in the respective country.

The value of a QALY – How do we know that a method is cost-effective?

A method is considered cost-effective in comparison to another method if its incremental cost-effectiveness ratio is lower than the society's willingness to pay for a QALY. The limit of the society's willingness to pay is often called the threshold value. Ideally, the willingness-to-pay for a QALY should represent the opportunity cost, that is, the value of what would have to be sacrificed if resources were redistributed from methods currently included in the health and welfare budgets to the new method [24, 25]. If the introduction of the new method implies excluding another method that produces 1 QALY for 500 000 krona, to motivate the substitution the new method must produce 1 QALY for less than 500 000 krona. Otherwise we would receive less health from the new method than we obtained with the current distribution of resources. To use this line of reasoning we would need to know the costs per QALY for all methods that are funded from the society's tax revenues. As that is not possible in practice, the society's willingness-to-pay for a QALY is usually assumed to amount to a certain threshold value [45-47].

In England and Wales, NICE apply a threshold of £20 000 -£30 000 (about 225 000-335 000 Swedish krona) to determine whether a method is cost-effective [24]. In Sweden there is no such determined threshold. In its guidelines for clinical practice, however, the Swedish National Board of Health and Welfare has defined limits on what the Board considers low and high costs, respectively, for a QALY. A low cost per QALY is defined as below 100 000 kronor, a high cost per QALY as more than 500 000 kronor and a very high cost per QALY as above 1 million kronor [48, 49].

Model analyses

Economic evaluations are dependent on the extent to which a method's costs and effects are known. The evaluations often use data collected in connection with a specific trial, commonly called trial-based economic evaluations. SBU calls them empirical health economic studies since they are based only on specifically collected primary data. It happens quite often, however, that the data are not sufficient for answering the health economic question, ie the introduction of new methods for diagnosis or treatment. In that case, the primary data can be supplemented by other data in a so-called model analysis.

A health economic model seeks to clarify a decision problem using the best available information, not to replace empirical studies. In model analysis, diverse types of data that have been collected previously, often called secondary data, are used together with primary data from trials. Models in health economic evaluations are primarily used in situations where costs and effects are affected over a longer period than is possible for a trial follow-up. Moreover, modelling is often of help in the following situations [50]:

- When relevant clinical evaluations are lacking or do not include data on costs and QALYs.
- To extrapolate from intermediate outcomes to final outcomes, ie from blood pressure to cardiac infarction.
- When it is not possible for ethical reasons to set up controlled clinical trials.
- When the costs of performing sufficiently large empirical studies are unreasonably high in relation to the potential value of the additional information.
- When costs collected in clinical trials do not reflect clinical practice or are not applicable to Swedish health care circumstances.

The most common techniques used in model analysis in health economics are *decision-tree*s (Figure 11.2) or *Markov- models* (Figure 11.3) [50]. These two methods are similar in many respects but a decision-tree depicts a sequence of events during a fixed time period. The technique is suitable for evaluating diseases of a more temporary character with relevant events restricted to a relatively short period of time.

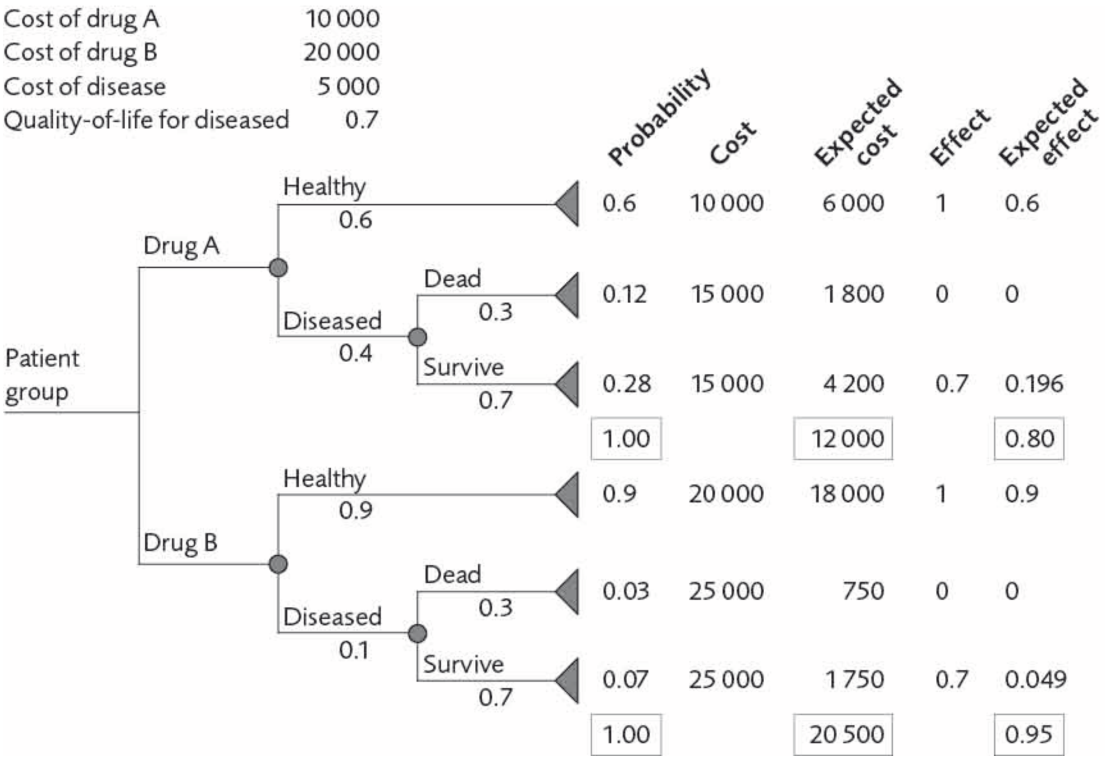


Figure 11.2 Decision-tree. Example: comparison of two alternative drugs, A and B.

In Figure 11.2, two alternative drugs (A and B) are compared using a decision-tree. The model starts with a choice between the alternatives, shown as two arms. From these treatment arms, possible treatment outcomes are depicted as branches. The probability of the outcome is given under each branch. All pathways end in so called terminal nodes (triangles). The parameter values are given in the figure's upper left corner. The probability of following each pathway, given the choice of treatment strategy, is presented in the first column to the right of the tree. The remaining columns present the cost of the strategy, the expected cost of the pathway, the effect from the strategy and the expected effect of the pathway. Framed values in the third and fifth columns show the total expected cost and total expected effect of the two alternatives, A and B. The incremental cost-effectiveness ratio (ICER), that is, the additional cost per effect if one chooses drug B instead of drug A is (20 500 – 12 000)/(0.95-0.80)= 56 667 kronor.

Markov models are constructed around mutually exclusive health states, see Figure 11.3. Each health state is associated with a certain cost and a certain QALY-weight. The models always contain an initial health state, such as a mild stage of disease, and a final, terminating, health state, often death. The arrows in the figure represent the transition probabilities, that is, the risks of moving between the different health states. In modern Markov models these risks can vary over time, for example increase with the patient's age. Markov models are more useful for analysing decision problems that extend over a long period of time, for example treatment of chronic diseases, and are therefore the most common type of health economic model

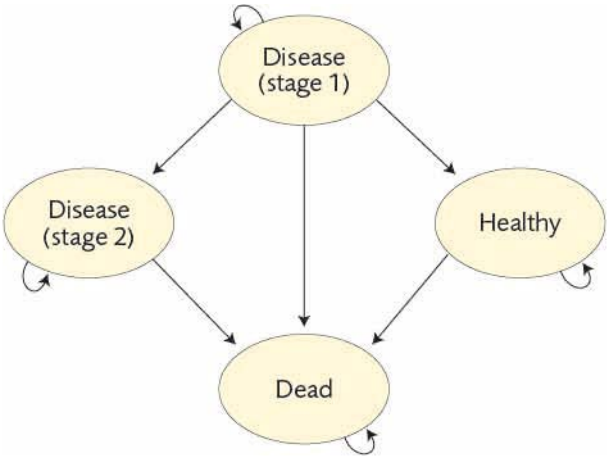


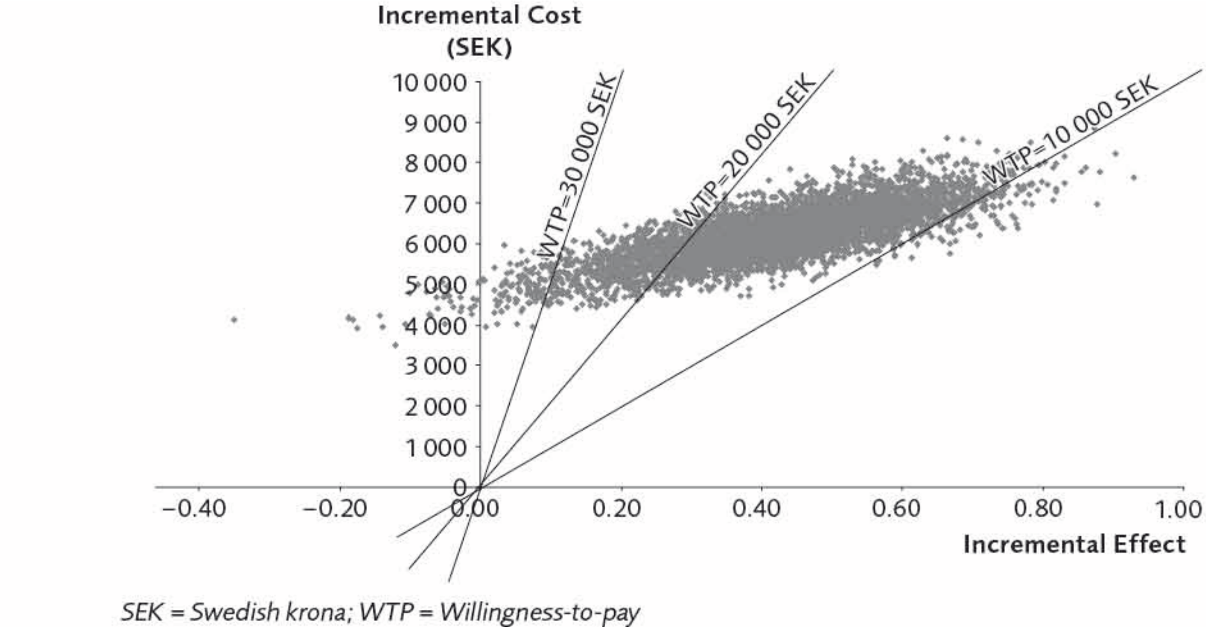
Figure 11.3 Markov-model.

Besides decision-trees and Markov models, discrete event simulation (DES) models [51] have recently become more common. Instead of different health states, as in the Markov model, these models are based on events that might occur at specific points in time. Examples of events are: a patient becomes sick, a primary care visit or the start of a treatment. Several events can occur at the same time and each of them carries consequences in terms of costs, quality-of-life and/or changes in the risks of future events. DES-models have, however, been criticised as they require comprehensive data that might be difficult to obtain. Another criticism of these models is that performing so-called probabilistic sensitivity analyses is difficult and time-consuming (see below) [52], although not all modelers share that view [51].

Sensitivity analyses

Sensitivity analyses [19], to describe the uncertainty in the results, are a very important part of health economic analyses. In a sensitivity analysis, one or more variables are varied in order to investigate how this affects the results. When there is a considerable change in the results – for example, the cost-effectiveness ratio becomes higher than the societal willingness-to-pay threshold – the results are said to be sensitive to that variable. It is, for example, prudent to increase the costs of implementing the new method if one suspects that they might become higher than was assumed in the base-case analysis. There might also be uncertainty about some assumptions used in a model, for example the proportion of patients that need to undergo further examinations or how frequently patients need a control visit. In these situations, a variety of scenarios can be tested to study to what extent they alter the overall results. Another approach is to determine the parameter values at which the results change from being below the threshold of cost-effectiveness to exceed it or vice versa.

Probabilistic sensitivity analyses, or PSA, are common in modeling studies [50]. They imply that the analysis considers the uncertainty around the model variables. Each variable is assigned a statistical distribution (ie a normal, beta or gamma distribution), based on the variability around the specific variable (ie based on the standard deviation of the data). The model is then run numerous times (often 1 000 to 10 000 times), with different combinations of possible variable values, in order to calculate the expected cost per effect. For each model run, a value from each distribution is randomly selected and the result is calculated. Figure 11.4 illustrates the result of a model that was run 5 000 times. The lines in the figure indicate different levels of willingness-to-pay for one unit of effect. Apart from the mean model result, a PSA can estimate the probability that the evaluated method is cost-effective. This is based on the proportion of model runs that end up to the right of the line that represent the willingness-to-pay threshold. For example, the figure shows that about 90 percent of the estimates are found to the right of the line for a willingness-to-pay of 30 000 kronor per effect unit. This means that if we are prepared to pay 30 000 kronor to gain one more effect unit, the probability of the method being cost-effective is about 90 percent.



SEK = Swedish krona; WTP = Willingness-to-pay

Figure 11.4 Cost-effectiveness plane with probabilistic sensitivity analysis.

Another analysis of uncertainty that could also be presented in a cost-effectiveness plane is based on data from empirical trial-based studies and involves bootstrapping costs and effects [9]. In trial-based empirical studies, patient-level data on costs and effects can be used to calculate a large number of cost-effectiveness ratios. This is achieved by drawing patient effect and cost data randomly from the two study groups and calculating the difference. This is done many times, often 1 000 times, and each time the patient data are replaced so that they can be selected again. Thus, bootstrapping results in a large number of cost-effectiveness ratios and a type of confidence (credibility) interval can be obtained by calculating the ICER values between which 95 percent of the draws lie. A similar method can be used to describe the empirical (non-parametric) distribution of costs and effects separately, both from trials and from model simulations.

Budget impact analysis

Cost-effectiveness analyses can be supplemented by a budget impact analysis in order to provide further information for those responsible for a new method's financing implications. The analysis differs from cost-effectiveness analyses in that it aims to describe how particular budgets are affected by a method and the expected consequences for involved entities. It does not assess whether there is a reasonable balance between the method's costs and effects and cannot be used for optimising the distribution of societal resources. The ISPOR Task Force has published guidelines for budget impact analyses [53].

Health economics and evidence

Health economic evaluations are based on theories from the science of economics. This means that the evaluations are grounded on theories about human behavior as well as norms and values. The idea behind health economic evaluation is that it should support decision-making. For that reason the analyses and statistical tests differ from those used to assess the clinical outcomes of interventions in health and social care.

As the results from health economic models are based on numerous data sources and assumptions, they should not be interpreted as evidence, but as an estimate of a method's impact on costs and health. From SBU's point of view, on the other hand, it is important that the clinical outcomes on which the model is based are statistically significant. Health economics outcome measures in randomised controlled studies (ie number of days of care) can be evidence-graded just like the clinical outcome measures. It is important, however, that the evidence grading is performed on single outcome measures, not on the ICER, since that is formed from several weighted outcome measures [9].

To conclude, we want to point out that health economic analyses are crucial for allocating scarce resources in the public sector, to ensure they are used in the best possible way. If resources are spent on measures which are not cost-effective, measures that give more effect per krona will not be implemented. In such cases our resources are not being used in an optimal way.

References

1. Stevens A, Milne R, Burls A. Health technology assessment: history and demand. J Public Health Med 2003;25:98-101.
2. Newdick C. Who should we treat? rights, rationing, and resources in the NHS. New York, Oxford University Press; 2005.
3. Rice DP. Estimating the cost of illness. Am J Public Health Nations Health 1967;57:424-40.
4. Hodgson TA, Meiners MR. Cost-of-illness methodology: a guide to current practices and procedures. ilbank Mem Fund Q Health Soc 1982;60:429-62.
5. Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M, Mokdad A, et al. Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010. Lancet 2012;380:2129-43.
6. Murray CJ, Vos T, Lozano R, Naghavi M,Flaxman AD, Michaud C, et al. Disability-adjusted life years (DALYs) for 291diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012;380:2197-223.
7. Drummond M. Cost-of-illness studies: a major headache? Pharmacoeconomics 1992;2:1-4.
8. Byford S, Torgerson DJ, Raftery J. Economic note: cost of illness studies. BMJ 2000;320:1335.
9. Brunetti M, Ruiz F, Lord J, Pregno S, Oxman A. Chapter 10: Grading economic evidence. . In: Schemilt I, Mugford M,Vale L, Marsch K, Donaldson C, editors. Evidence-based decisions and economics: health care, social welfare, education and criminal justice. Oxford: Wiley-Blackwell; 2010.
10. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. BMJ 1996;313:275-83.
11. Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on Health Economic Criteria. Int J Technol Assess Health Care2005;21:240-5.
12. Philips Z, Ginnelly L, Sculpher M,Claxton K, Golder S, Riemsma R, et al. Review of guidelines for good practice in decision-analytic modelling in health technology assessment. Health Technol Assess 2004;8:iii-iv, ix-xi, 1-158.
13. Cooper N, Coyle D, Abrams K, Mugford M, Sutton A. Use of evidence in decision models: an appraisal of health technology assessments in the UK since 1997. J Health Serv Res Policy 2005;10:245-50.
14. Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health 2009;12:409-18.
15. Mulligan J-A, Fox-Rushby J. Transferring cost-effectiveness data across space and time. IN: Fox-Rushby J & Cairns J (eds).Economic evaluation. Open University Press, 2005.
16. Drummond M, Sculpher M, Torrance G, O'Brien B, Stoddart G. Methods for the economic evaluation of health care programmes. Oxford, Oxford University Press; 2005.
17. Bayoumi AM. The measurement of contingent valuation for health economics. Pharmacoeconomics 2004;22:691-700.
18. Liljas B, Blumenschein K. On hypothetical bias and calibration in cost-benefit studies. Health Policy 2000;52:53-70.
19. Gold M, Siegel J, Russell L, MC W. Costeffectiveness in Health and Medicine. New York, Oxford University Press; 1996.

20. Johannesson M, Karlsson G. The friction cost method: a comment. J Health Econ 1997;16:249-55; discussion 257-9.
21. Koopmanschap MA, Rutten FF, van Ineveld BM, van Roijen L. The friction cost method for measuring indirect costs of disease. J Health Econ 1995;14:171-89.
22. Sahlén K-G, Löfgren C, Lindholm L. Är det lönsamt med prevention efter 65? Ålderns betydelse i hälsoekonomiska utvärderingar. Stockholm: Statens folkhälsoinstitut; 2006.
23. Sculpher M. The role and estimation of productivity costs in economic evaluation. In: Drummond M, McGuire A, editors. Economic evaluation in health care. Merging theory with practice. Oxford: Oxford University Press; 2001.
24. National Institute for Health and Clinical excellence. Guide to the methods of technology appraisal. London: National Institute for Health and Clinical excellence (NICE); 2008.
25. International Society for Pharmacoeconomics and Outcomes Research. Pharmacoeconomic guidelines around the world. In. ISPOR, Lawrenceville (NJ); 2008.
26. Tand- och läkemedelsförmånsverket (TLV). Läkemedelsförmånsnämndens allmänna råd om ekonomiska utvärderingar. LFNAR 2003:2: Tand- och läkemedelsförmånsverket (TLV); 2003.
27. von Neumann J, Morgenstern O. Theory of games and economic behaviour. Princeton, NJ, Princeton University Press; 1944.
28. Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res 1972;7:118-33.
29. Patrick DL, Bush JW, Chen MM. Methods for measuring levels of well-being for a health status index. Health Serv Res 1973;8:228-45.
30. The EuroQol Group. EuroQol--a new facility for the measurement of healthrelated quality of life. The EuroQol Group. Health Policy 1990;16:199-208.
31. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271-92.
32. Feeny D, Furlong W, Torrance GW, Goldsmith CH, Zhu Z, DePauw S, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. Med Care 2002;40:113-28.
33. Bernfort L. QALY som effektmått inom vården. Möjligheter och begränsningar, CMT Rapport 2012:2 2012.
34. Dolan P, Gudex C, Kind P, Williams A.A social tariff for EuroQol: results from a UK general public survey. York: Centre Centre for health Economics, University of York; 1995.
35. Burstrom K, Sun S, Gerdtham UG, Henriksson M, Johannesson M, Levin LA, et al. Swedish experience-based value sets for EQ-5D health states. Qual Life Res 2014;23:431-42.
36. Norrlid H, Dahm P, Ragnarson Tennvall G. Evaluation of the cost-effectiveness of buprenorphine in treatment of chronic pain using competing EQ-5D weights. Scandinavian Journal of Pain (Available online 14 August 2014).
37. McDonough CM, Tosteson AN. Measuring preferences for cost-utility analysis: how choice of method may influence decision-making. Pharmacoeconomics 2007;25:93-106.
38. Seymour J, McNamee P, Scott A, Tinelli M. Shedding new light onto the ceiling and floor? A quantile regression approach to compare EQ-5D and SF-6D responses. Health Econ 2010;19:683-96.
39. Kopec JA, Willison KD. A comparative review of four preference-weighted measures of health-related quality of life. J Clin Epidemiol 2003;56:317-25.

40. Heintz E, Wirehn AB, Peebo BB, Rosenqvist U, Levin LA. QALY weights for diabetic retinopathy--a comparison of health state valuations with HUI-3, EQ-5D, EQ-VAS, and TTO. Value Health 2012;15:475-84.
41. Bleichrodt H, Johannesson M. Standard gamble, time trade-off and rating scale: experimental results on the ranking properties of QALYs. J Health Econ 1997;16:155-75.
42. Puhan MA, Schunemann HJ, Wong E, Griffith L, Guyatt GH. The standard gamble showed better construct validity than the time trade-off. J Clin Epidemiol 2007;60:1029-33.
43. Stiggelbout AM, Kiebert GM, Kievit J, Leer JW, Stoter G, de Haes JC. Utility assessment in cancer patients: adjustment of time tradeoff scores for the utility of life years and comparison with standard gamble scores. Med Decis Making 1994;14:82-90.
44. Karlsson JA, Nilsson JA, Neovius M, Kristensen LE, Gulfe A, Saxne T, et al. National EQ-5D tariffs and qualityadjusted life-year estimation: comparison of UK, US and Danish utilities in south Swedish rheumatoid arthritis patients. Ann Rheum Dis 2011;70:2163-6.
45. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. Pharmacoeconomics 2008;26:733-44.
46. Devlin N, Parkin D. Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. Health Econ 2004;13:437-52.
47. Rawlins MD, Culyer AJ. National Institute for Clinical Excellence and its value judgments. BMJ 2004;329:224-7.
48. Carlsson P, Anell A, Eliasson M. Hälsoekonomi får allt större roll för sjukvårdens prioriteringar Läkartidningen 2006;103:3617-3623.
49. Socialstyrelsen. Bilaga 4, Metod, Nationella riktlinjer för diabetesvården 2010 – Stöd för styrning och ledning Stockholm: Socialstyrelsen; 2010.
50. Briggs A, Claxton K, Sculpher M. Decision Modelling for Health Economic Evaluation. New York, Oxford University Press; 2006.
51. Caro JJ, Moller J, Getsios D. Discrete event simulation: the preferred technique for health economic evaluations? Value Health 2010;13:1056-60.
52. Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, et al. Probabilistic sensitivity analysis for NICE technology assessment: not an optional extra. Health Econ 2005;14:339-47.
53. Mauskopf JA, Sullivan SD, Annemans L, Caro J, Mullins CD, Nuijten M, et al. Principles of good practice for budget impact analysis: report of the ISPOR Task Force on good research practices- budget impact analysis. Value Health 2007;10:336-47.

Chapter 12

Ethical and social aspects

Background

Ethical and social aspects should be taken into consideration at every stage of a health technology assessment – right from the initial selection of assessment topic, and then continuously throughout the entire assessment process. However, this chapter focuses on how to systematically assess ethical and social aspects related to the application of a given intervention interventions [1]. Research ethics and environmental aspects are also addressed at the end of this chapter.

What is meant by ethical and social aspects?

Ethics (from the Greek word 'ethos', meaning 'character' or 'custom') is synonymous with moral philosophy, or practically speaking, the area of philosophical science that attempts to answer questions such as "What is good?", "What is right?" and "How should one behave?" Ethical aspects in a healthcare context relate primarily to what benefits or harms an intervention may cause the individual patient. Questions relating to respect for the patient's autonomy and integrity are also brought into play, as well as equality when it comes to who should be offered various types of care interventions. Above all else, the focus is on the patient perspective, but other perspectives – such as those of other patients or patient groups, healthcare workers, healthcare as a whole or society – may also be considered.

Social aspects may relate to the causes of disease, to the consequences of disease, and to the use of various interventions. For example, a new intervention may impact an individual in terms of their opportunities to lead a normal life with respect to accommodations, family-life or relationships, while also being associated with organizational changes or costs for the health care provider. This may also involve patients' opportunities to choose their own lifestyle. When assessing care interventions, social aspects may also relate to which resources are required, and how they are divided up among the population [2]. Ethical and social aspects often overlap, and are therefore usually dealt with together in SBU's assessments.

Working with ethical and social aspects in projects

In SBU's work to assess different interventions from an ethical and social perspective, a model consisting of three main stages is used – see Figure 12.1. The number of stages implemented in an SBU project depends on factors such as the character of the ethical and social aspects identified during the introductory stage.

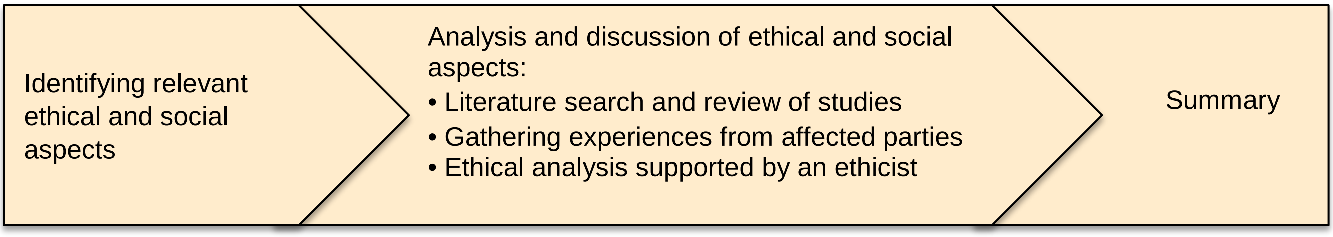

Identifying relevant ethical and social aspects
Analysis and discussion of ethical and social aspects:
• Literature search and review of studies
• Gathering experiences from affected parties
• Ethical analysis supported by an ethicist
Summary

Figure 12.1 Overview of SBU's working model for assessing interventions within health care from an ethical and social perspective.

Identifying relevant ethical and social aspects

For every project, potentially relevant ethical and social aspects should be considered already when formulating the research questions. As an initial step in this work, it is therefore important – preferably, at one of the first project group meetings – that ethical and social issues that may be associated with the interventions to be evaluated are identified and described. SBU has produced a manual including guiding questions to support this (Appendix 9). This guidance focuses primarily on ethical issues, but since ethical and social issues often overlap certain social aspects are also included.

Within SBU's projects, project groups should begin identifying relevant ethical and social aspects with a general group discussion. This initial step is important to avoid focusing the group's attention solely on the empirical questions to be assessed, thereby missing issues that might otherwise be identified spontaneously.

The project group may then use SBU's guidance to systematically assess whether additional issues were missed in the initial discussion. The questions in the guidance may be of varying importance to different projects, therefore only those issues that are deemed relevant to the intervention in question need to be assessed. The process may need to be repeated as the scientific basis for the benefits and risks is clarified throughout the project.

The structure of the guidance

The guidance is based on the work of the Norwegian philosopher Hofmann [3,4], that has been supplemented with questions from work published by the European Network for Health Technology Assessment (EUnetHTA) [5] and the International Network of Agencies for Health Technology Assessment (INAHTA) [6], and has then been adapted according to the Swedish context and SBU's working methods.

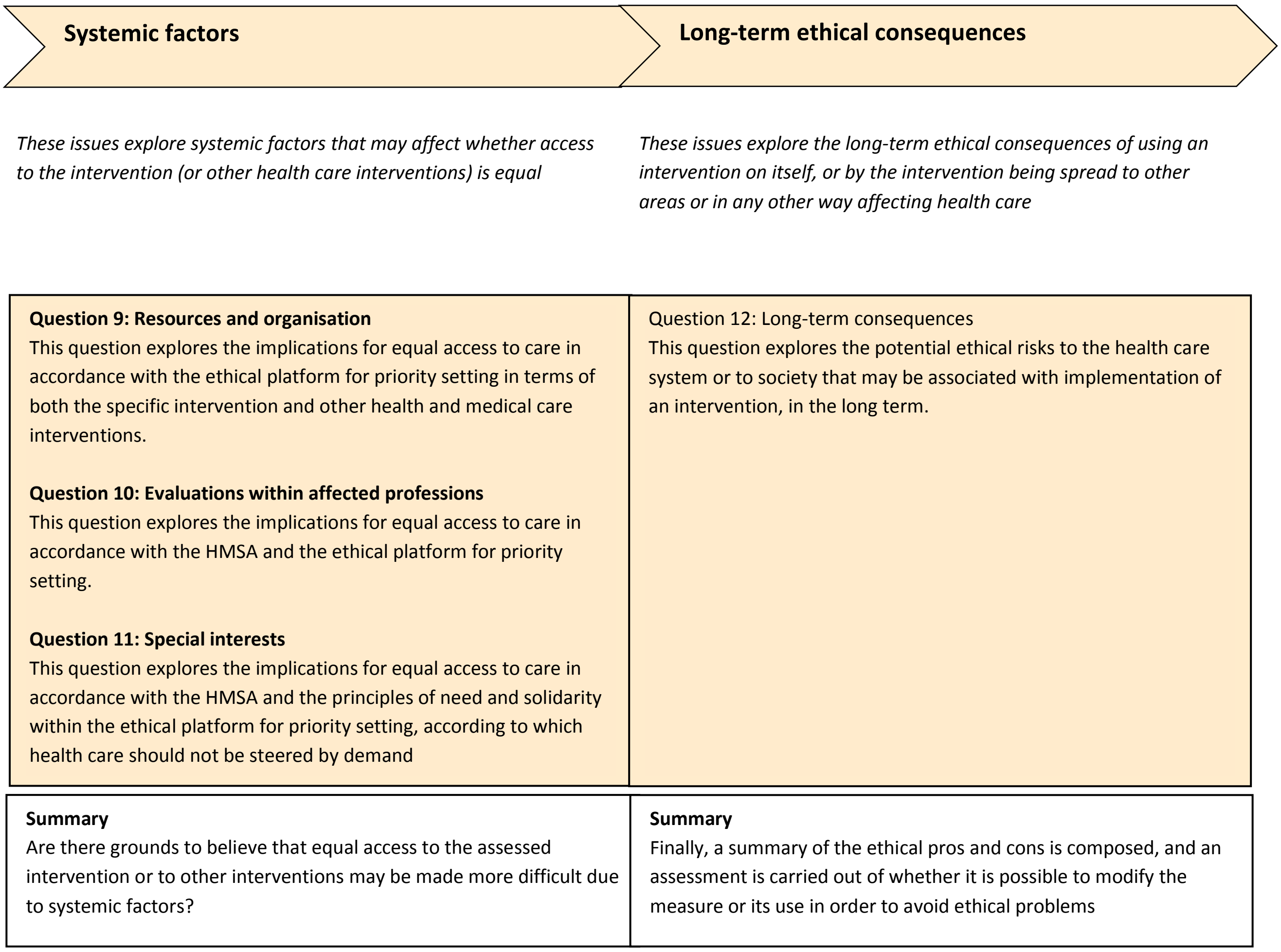
The guidance contains questions related to values and norms in Swedish health and medical care legislation (the Swedish Health and Medical Services Act, the Swedish Patient Data Act, the Swedish Patient Safety Act, etc.). The list has been divided up into four different fields of questions with a concluding summary (Figure 12.2). The first field relates to how the intervention contributes to a good standard of health, which is the ethically motivated objective of health and medical care. The second explores values that set boundaries for how the objective relating to good standard of health can be achieved, including equality, justice, autonomy, integrity and cost-effectiveness. The third field deals with various systemic factors – such as resource and organisational aspects, professional values and special interests – which may affect patients' opportunities to access the intervention, or how implementation might affect the availability of other interventions. The fourth field deals with the long-term ethical consequences of the measure. The guidance concludes with questions to help the assessors summarise the ethical arguments for and against the intervention, and to examine whether changes can be made to the intervention to address any identified ethical problems.

The impact of the intervention on health

According to the Swedish Health and Medical Services Act (HMSA), one of the main objectives of health care is to assure good health for the population. In order to be able to assess whether an intervention should actually be used, its effect on individuals' health in a broad sense needs to be determined. This includes evaluating whether the measure has negative consequences that despite a positive effect, might mean that its use is not justified. If there is insufficient scientific evidence to draw any conclusions about an interventions effect, specific ethical issues arise that should be taken into consideration. For example, there may be ethical and/or methodological problems with carrying out further research. In which case, the question arises as to whether there may be ethical reasons to use an intervention despite having inconclusive scientific support. When there is evidence based on studies of dubious quality in terms of research ethics, it may be worth pointing this out – not least because, if the evidence is weak, it may prove difficult to confirm the results of earlier studies. If these questions have already been addressed in other stages of the assessment, these results provide a starting point for subsequent ethical reasoning.

This section of the guidance also includes the impact of the intervention on the health of third parties, which indirectly appears to have support in the HMSA in view of the fact that the objective is good health for the population. Here, third parties refer primarily to the patient's close relatives, although a public health perspective is also included.

Impact on health	**Compatibility with ethical values**
These issues deal with an intervention's effect on health outcomes, and the urgency of the intervention in question in relation to those objectives. The questions provide a starting point for subsequent ethical reasoning.	*These issues deal with the extent to which the use of an intervention is compatible with ethical values and norms.*
Question 1: Effect on health from the patient's perspective This question is linked to the Swedish Health and Medical Services Act (HMSA), the Swedish Patient Safety Act and the principles of need and solidarity provided by the ethical platform for priority setting **Question 2: Evidence gaps** This question is linked to the HMSA's requirement that measures should be based on scientific evidence and professional practice. **Question 3: The severity of the condition** This question is linked to the HSL and the principles of need and solidarity provided by the ethical platform for priority setting. **Question 4: The impact on the health of third parties** This question is linked to the HMSA	**Question 5: Equality and justice** This question is linked to the HMSA and the principle of human values provided by the ethical platform for priority setting **Question 6: Autonomy** This question is linked to the HMSA **Question 7: Integrity** This question is linked to the HMSA and the Swedish Patient Data Act **Question 8: Cost-effectiveness** This question is linked to the HMSA and the principle of cost-effectiveness provided by the ethical platform for priority setting
Summary What are the risks and benefits of the intervention?	**Summary** Is use of this intervention compatible with applicable ethical values?



Systemic factors

These issues explore systemic factors that may affect whether access to the intervention (or other health care interventions) is equal

Question 9: Resources and organisation
This question explores the implications for equal access to care in accordance with the ethical platform for priority setting in terms of both the specific intervention and other health and medical care interventions.

Question 10: Evaluations within affected professions
This question explores the implications for equal access to care in accordance with the HMSA and the ethical platform for priority setting.

Question 11: Special interests
This question explores the implications for equal access to care in accordance with the HMSA and the principles of need and solidarity within the ethical platform for priority setting, according to which health care should not be steered by demand

Summary
Are there grounds to believe that equal access to the assessed intervention or to other interventions may be made more difficult due to systemic factors?

Long-term ethical consequences

These issues explore the long-term ethical consequences of using an intervention on itself, or by the intervention being spread to other areas or in any other way affecting health care

Question 12: Long-term consequences
This question explores the potential ethical risks to the health care system or to society that may be associated with implementation of an intervention, in the long term.

Summary
Finally, a summary of the ethical pros and cons is composed, and an assessment is carried out of whether it is possible to modify the measure or its use in order to avoid ethical problems

Figure 12.2 Overview of the structure of the guidance, and the included questions' links to Swedish legislation and the ethical platform for priority setting. The questions included in the guidance can be found in attachment 9.

In order to ensure that a good standard of health is an accessible goal for everyone, the HMSA emphasizes that care should be given according to need. This refers to the principle of need and solidarity set out in the ethical platform for priority setting. Based on the fact that the objective is to provide equal opportunities and equal outcomes when it comes to the health of the population, those with a greater need should be given priority over those whose need is less urgent. This means that someone with a less serious condition may be given a lower priority than someone with a more serious condition. The size of the need (also referred to as the severity of the condition) should, however, be balanced against the effect of the intervention. It is clearly stated in the HMSA that a person cannot be considered to be in need of an intervention that offers no benefit. However, the ethical platform does not state whether this also involves a demand for evidence of a benefit.

The severity of the condition is also significant for determining the extent to which associated ethical problems, affecting both the patient and relatives, can be accepted. If this relates to a more effective intervention directed towards a serious and thus more urgent condition, there is reason to accept greater ethical problems or to make other considerations than in the case of a milder condition.

Here, ‘patients’ refers to the individual or group who gains access to the intervention, even if not all of these are patients in a real sense. For example, in connection with pregnancy we are of the opinion that there are two patients – the mother and the foetus – who both should be taken into consideration.

Compatibility of the intervention with ethical values

After having presented the risk/benefit profile of an intervention, it should be assessed whether it is compatible with prevailing ethical values in health and medical care. If not, the intervention may be deemed unsuitable for use. It should also be considered whether the intervention can be modified or only offered under certain circumstances in order to be more compatible with these ethical values. In some cases, however, it may be that a degree of incompatibility can be accepted, depending, for example,

on the severity of the condition to be treated, how effective the intervention is, and whether there are alternative interventions available. The ethical values discussed under this heading are based on Swedish health and medical care legislation. When assessing the compatibility of the intervention with these values, a degree of guidance can be provided by Swedish legislation and other declarations, as well as guidelines and steering documents that the Swedish government has adopted or that are established within the Swedish health and medical care context.

Under this heading, we also deal with the impact of an intervention on these ethical principles with respect to third parties, particularly close relatives. This does not have explicit support in Swedish health and medical care legislation, but it appears to be an important part of the ethical practice of care, and may therefore need to be taken into consideration. However, in the event of conflict between how the intervention affects these ethical values for the patient and how it affects third parties, the patient’s interests is usually given priority.

Systemic factors with ethical implications

After having assessed whether there are ethical arguments for or against using the intervention, there should also be an assessment of whether there are systemic factors that could affect the use of the intervention – or other interventions – and thus patients’ equal access to care. According to the HMSA, care should be offered to the entire population on equal terms. Where there are no studies to assess these aspects, a more experience-based assessment is required.

Long-term ethical consequences

Even if a particular intervention is not associated with any of the above mentioned ethical issues, its implementation may still lead to developments that have problematic ethical consequences in the long term.

Analysis and discussion of ethical and social aspects

Once the project group – with help from the guidance – has identified the relevant ethical and social issues, they are discussed in greater depth. If it is considered important in the project, additional information can be obtained through a literature search and by gathering experiences from affected groups. A more in-depth ethical analysis can also be carried out with the help of an ethicist.

Literature search and review of studies

To identify previous analyses in connection with ethical and social aspects, or to provide answers to empirical questions which may have arisen in the introductory analysis, it may be useful to carry out a systematic literature search. For more information on searching for literature that is relevant to ethical assessment, see Droste et al. [7]. A search for relevant literature is carried out in cooperation with an information specialist. Apart from medical databases, a search of IBSS and PsycInfo may also be relevant. Identified studies are reviewed and assessed with regard to study quality and relevance.

It is important to bear in mind that the social environment varies with cultural, economic and social conditions, which may affect the transferability of the results in the identified studies to Swedish circumstances. For more details on how to review studies, see Chapters 6-8.

Gathering experiences from affected parties

It may also be desirable to gather information about affected parties' experiences. These experiences may act as a source of knowledge about how the intervention may affect the parties.

In some cases it may also be appropriate that the groups that will be affected by the report and the ethical implications of the results (such as patients, relatives and various groups of health care personnel) are invited to give their views on the ethical analysis of the evaluation model before the report is published [8].

Ethical analysis with support from ethical expertise

It should also be considered whether a deeper ethical analysis needs to be carried out with the help of a professional ethicist. Involving ethical expertise to carry out a more in-depth analysis will be particularly important when ethical issues of greater importance or of a fundamentally interesting nature are identified. As SBU has established cooperation with the Swedish National Council on Medical Ethics (SMER), it is also possible in certain cases to refer the ethical analysis to this council [5]. This is particularly true for questions that would benefit from a discussion at the national level, or from an overall societal perspective. Once the ethical questions have been identified by the project group, possibly using documentation of established practice and a literature search, an ethicist can lead the group in further discussion about how best to illustrate these issues.

Research ethics

During the process of reviewing the scientific medical and health economics studies, the project group should also determine whether the underlying research has been carried out in an ethically acceptable manner, for example in accordance with the internationally recognised Declaration of

Helsinki [9]. A checklist for assessing research ethics aspects has been proposed by Weingarten et al [10]. It recommends that the following information should be provided: whether the study

participants have received adequate information and given their consent, whether an ethical committee reviewed and approved the study, and finally how the research was financed. Any link to commercial or other special interests should also be noted. In principle, the results from such studies may be used, provided that they are of sufficient quality and the results are deemed to be relevant and valuable. However, the project group should formulate and discuss the research ethics problems that arise. The discussion can also be summarised in the chapter discussing future research (Example 12.1).

Example 12.1 Research ethics problem.

In studies of artificial sweeteners to avoid caries, the study participants – children aged 10 to 14 – were instructed to chew chewing gum containing artificial sweeteners several times a day. In this way, the children got used to daily consumption of sweet chewing gum. The studies were carried out in developing countries where products containing artificial sweeteners are expensive. Hence, there was a risk that the children would continue to chew chewing gum after the end of the study, but using chewing gum containing sucrose instead.

Environmental aspects

What is meant by environmental aspects?

Environmental aspects are activities linked to environmental impact. All authorities should carry out an environmental assessment to quantify the direct environmental impact of its operations. The results of this assessment will provide decision-making data for the authority's continued environmental work, and form the basis for the authority's environmental policy, environmental targets and action plans. Each year a report is sent to the Swedish Ministry of Health and Social Affairs and the Swedish Environmental Protection Agency containing details of concrete environmental goals and action plans addressing aspects that should be improved. The extent to which previous years' environmental goals were met are also presented.

For many years, SBU has submitted annual reports on the effects of activities with a direct environmental impact in its operations, such things as the authority's business travel (expressed in tonnes of carbon dioxide emissions), paper consumption, and electricity use.

Environmental aspects in SBU's reports

However, SBU's biggest environmental impact is probably indirect – i.e. through its results and conclusions. The conclusions of the reports can influence which interventions are used in health care, and may therefore impact the environment. For example, pharmaceuticals may contain biologically active components that can have varying degrees of environmental impact. Medical devices can have other types of environmental impact. The impact on the environment should therefore also be described where applicable in SBU's assessments.

Examples of questions:

- How does the intervention affect the environment (e.g. emissions, transportation, manufacturing process, etc.)?
- Is there environmental legislation that is relevant to the method?
- Is there environmental commitment in society that comes into conflict with the intervention?
- Can the implementation of the intervention have an adverse influence on public confidence in health care service from an environmental perspective?

The following databases may be searched for relevant literature:

- PubMed
- Academic Search Elite
- Scopus

More information is available via the following links:

- The environmental impact of pharmaceuticals:
  - www.fass.se (environmental information is available for those pharmaceuticals with a green ‘M’)
  - www.janusinfo.se (Stockholm County Council’s full classification list of pharmaceuticals)
  - www.mistrapharma.se (research into pharmaceutical environmental effects)
- The environmental impact of chemicals:
  - http://www.kemi.se/en(the Swedish Chemicals Agency’s prioritisation guide, PRIO)
  - www.sll.se (Stockholm County Council’s work to phase out hazardous chemicals)

References

1. Braunack-Mayer AJ. Ethics and health technology assessment: Handmaiden and/or critic? Int J Technol Assess Health Care 2006;22:307-21.
2. Leboux P, Williams-Jones B. Mapping the integration of social and ethical issues in health technology assessment. Int J Technol Assess Health Care 2007;23:9-16.
3. Hofmann B. Towards a procedure for integrating moral issues in health technology assessment. Int J of Technol Assess Health Care 2005;21:312-8.
4. Hofmann B. Etikk i vurdering av helsetiltak. Utvikling av en metode for å synliggjøre etiske utfordringer i vurdering av helsetiltak. ("Ethics in the evaluation of health measures. Developing a method to show ethical challenges in the evaluation of health measures.") Report no. 26–2008. Oslo: The Norwegian Knowledge Centre for the Health Services; 2008.
5. Saarni SI, Braunack-Mayer A, Hofmann B, et al. Different methods for ethical analysis in health technology assessment: An empirical study. Int J of Technol Assess Health Care 2011;27:305-12.
6. Burls A, Caron L, de Langavant GC, et al. Tackling ethical issues in health technology assessment: A proposed framework. Int J of Technol Assess Health Care 2011;27:230-7.
7. Droste S, Dintsios CM, Gerber A. Information on ethical issues in health technology assessment. How and where to find them. Int J Technol Assess Health Care 2010;26:441-9.
8. Autti-Rämö I, Mäkelä M. Ethical evaluation in health technology assessment reports: An eclectic approach. Int J Technol Assess Health Care 2007;23:1-8.
9. The Declaration of Helsinki. The World Medical Association. 2008 [quoted 3 September 2014]; Available from: http://www.wma.net/en/30publications/10policies/b3/17c.pdf
10. Weingarten MA, Paul M, Leibovici L. Assessing ethics of trials in systematic reviews. BMJ 2004;328:1013-4.