

## **7 Reviewing and grading the evidence**

Studies identified following the literature search need to be reviewed to identify the most appropriate data to help answer the clinical questions and to ensure that the recommendations are based on the best available evidence. This process should be explicit and transparent and should be carried out through a systematic review process. This involves four major steps: selecting relevant studies; assessing their quality; synthesising the results and grading the evidence.

### **7.1 *Selecting studies of relevance***

Before acquiring papers for assessment, the information scientist or the reviewer who carried out the search needs to sift the evidence identified in the search in order to discard irrelevant material. As a preliminary stage, the titles of the retrieved citations should be scanned and those that fall outside the topic of the guideline should be eliminated. A quick check of the remaining abstracts should identify those that are clearly not relevant to the clinical questions and that should be excluded at this stage.

The remaining abstracts should then be scrutinised against the clinical criteria agreed by the GDG. Abstracts that do not meet the inclusion criteria should be eliminated. If there is any doubt about inclusion, this should be resolved by discussion with the GDG. Once the sifting is complete, hard copies of the selected studies can be acquired for assessment. Studies that fail to meet the inclusion criteria should be excluded. Those that meet the criteria can be assessed. Because there is always an element of bias in selecting the evidence, periodic double sifting of a random selection of abstracts should be performed.

The study-selection process should be clearly documented and should detail the inclusion criteria agreed by the GDG that were applied in the selection process.

### **7.2 *Assessing the quality of studies***

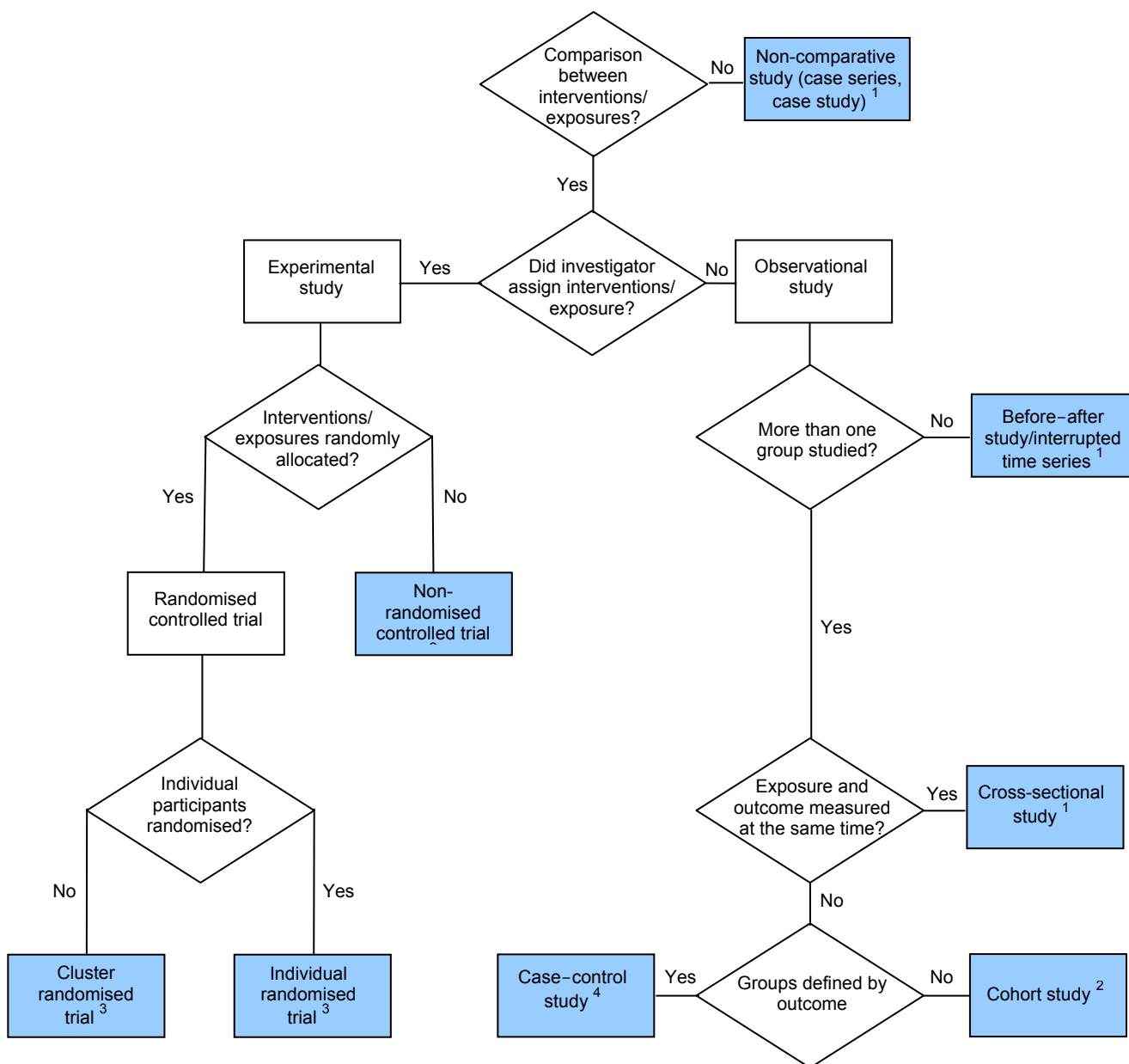
The quality of studies should be evaluated from an assessment of the methods and methodology used. This is a key stage in the guideline development process because the result will affect the level of evidence ascribed to the study. In turn, this will have an impact on the class of recommendations it underpins (see section 7.4).

Before assessing the quality of studies selected from the search it is important to determine the study design so that the appropriate criteria can be applied in their assessment (see section 7.2.1). Because it is sometimes difficult to establish the exact design used in studies NICE has developed an algorithm to help systematic reviewers classify study design for answering questions of effectiveness (See Figure 7.1). Algorithms for questions of diagnosis and prognosis are in preparation.

**Figure 7.1. Algorithm for classifying study design for questions of effectiveness**

**Key**

- 1 = currently no checklist
- 2 = cohort study checklist, see Appendix D
- 3 = RCT checklist, see Appendix C
- 4 = case-control study checklist, see Appendix E



**7.2.1 Published studies**

The published studies selected from the search should be assessed for their methodological rigour against a number of criteria. Because these criteria will differ according to the study type, a range of checklists have been designed to provide a consistent approach to the assessment and its reporting. NICE

recommends the checklists developed originally by the MERGE (Method for Evaluating Research and Guideline Evidence) Group in Australia and modified by the Scottish Intercollegiate Guidelines Network (SIGN) (see section 7.5). These checklists may be used to assess the selected studies. Health-economics studies should be assessed with the Drummond checklist (see Appendix G). All these checklists are presented in Appendices B to H, together with explanatory notes on their use. The overall assessment of each study is graded using a code '++', '+' or '-', based on the extent to which the potential biases have been minimised. This is used as a basis for classifying the recommendations (see Chapter 11).

To minimise any potential bias in the assessment, independent assessment by two reviewers on a random selection of papers is desirable. Any differences arising from this should be discussed fully at the GDG meeting.

### **7.2.2 Unpublished data and studies in progress**

Unpublished data may be obtained in the course of the review, particularly from stakeholders. NCCs are not routinely expected to search the grey literature. Any unpublished data should be subjected to an assessment of quality in the same way as published studies. Authors should be contacted and requested to provide the necessary information so that the reviewers can complete the relevant quality checklist, or to provide details on individual patient data.

### **7.2.3 Published guidelines**

Relevant published guidelines may be identified in the search. These are either NICE guidelines or other guidelines.

#### **7.2.3.1 NICE guidelines**

These should be fully referenced and the evidence underpinning the recommendations should be left unchanged, provided it is not out of date. The wording of the recommendation may be changed to reflect the topic of the guidelines, but the guidance should not go beyond the evidence base, and the grading of the recommendation should not be changed.

If there is new published evidence that would significantly alter the existing recommendations, the NCC should follow the methodology for early update that is described in 15.3. The recommendation should be graded accordingly to reflect the evidence base.

#### **7.2.3.2 Other guidelines**

Other relevant published guidelines identified in the search should be assessed for quality using the AGREE instrument<sup>1</sup> to ensure they have

---

<sup>1</sup> AGREE Collaboration, Development and validation of an international appraisal instrument for assessing the quality of clinical practice guidelines: the AGREE project. *Quality and Safety in Health Care* 2003; 12(1): 18-23.

sufficient documentation to be considered. There is no cut-off point for accepting or rejecting a guideline and each group will need to set its own parameters. These should be documented in the methods section of the full guideline along with a summary of the assessment. The results should be presented as an appendix in the full guideline.

Reviews of evidence from other guidelines that cover clinical questions formulated by the GDG can be considered as evidence provided:

- the review of evidence is assessed using the appropriate checklist from the technical manual and is judged to be of high quality
- they are accompanied by the evidence statement and evidence table(s)
- the evidence is updated according to the methodology for early update that is described in section 15.3.

The GDG should create its own evidence summaries or statements. Evidence tables from other guidelines should be referenced with a direct link to the source website address or a full reference to the published document. The GDG should formulate its own recommendations taking into consideration the whole body of evidence. The recommendations should be classified using the system described in section 11.3 to reflect the evidence base.

Recommendations from other guidelines should not be quoted verbatim. The exceptions are recommendations from NHS policy (for example, National Service Frameworks).

## **7.3 Summarising the evidence**

### **7.3.1 Data extraction and evidence tables**

Data should be extracted to a standard template, for inclusion in an evidence table. Evidence tables help identify the similarities and differences between studies, including key characteristics of the study population and interventions or outcome measures; this provides a basis for comparison. They also help determine if it is possible to calculate a mean estimate of effect. In some circumstances and if the necessary data are available, it may be appropriate to carry out a meta-analysis. A full description of data synthesis, including meta-analysis and extraction methods, is available from the report produced by the Centre for Review and Dissemination (*Report Number 4*, 2nd edition – see ‘Further reading’). Sensitivity analysis could be used to investigate the impact of missing data.

The information to be extracted may vary depending on the clinical question, the level of detail and the analysis needed. Appendix I provides a template for information that should be included in evidence tables related to intervention studies. Information from studies of the accuracy of diagnostic tests should be reported using the template provided in Appendix J.

### **7.3.2 Conducting a meta-analysis**

Synthesis of outcome data through meta-analysis (usually of RCTs only) is appropriate provided there are sufficient relevant and valid data with measures of outcome that are comparable. Where such data are not available, the analysis may have to be restricted to a qualitative overview of individual studies. Forest plots are a useful tool to illustrate the individual study population results. The characteristics and limitations of the data (that is, population, intervention, setting, sample size and validity of the evidence) need to be fully reported.

Before any statistical pooling is carried out, an assessment of the degree of, and the reasons for, heterogeneity in the study results should be undertaken – that is, variability in the effects between studies that may suggest that individual studies reflect different study circumstances. Statistical heterogeneity of study results can be addressed using a random (as opposed to fixed) effects model. Known clinical heterogeneity (for example patient characteristics, or intervention dose or frequency) can be managed by judicious use of methods such as subgroup analyses and meta-regression. For methodological heterogeneity (for example, where different trials are of different quality), the results of sensitivity analyses (varying the studies in the meta-analysis) should be reported.

Forest plots should include lines for studies that are believed to contain eligible data even if the data are missing from the analysis in the published study. An estimate of the proportion of eligible data that are missing (because some studies will not include all relevant outcomes) will be needed for each analysis.

### **7.3.3 Levels of evidence**

#### **7.3.3.1 Intervention studies**

Studies that meet the minimum quality criteria should be ascribed a level of evidence to help the guideline developers and the eventual users of the guideline understand the type of evidence on which the recommendations have been based.

There are many different methods of assigning levels to the evidence and there has been considerable debate about what system is best. A number of initiatives are currently under way to find an international consensus on the subject. NICE has previously published guidelines using different systems and is now examining a number of systems in collaboration with the NCCs and academic groups throughout the world to identify the most appropriate system for future use.

Until a decision is reached on the most appropriate system for the NICE guidelines, the Institute advises the NCCs to use the system for evidence shown in Table 7.1.

**Table 7.1 Levels of evidence for intervention studies.** Reproduced with permission from the Scottish Intercollegiate Guidelines Network; for further information, see 'Further reading'.

Level of evidence	Type of evidence
1 <sup>++</sup>	High-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
1 <sup>+</sup>	Well-conducted meta-analyses, systematic reviews of RCTs, or RCTs with a low risk of bias
1 <sup>-</sup>	Meta-analyses, systematic reviews of RCTs, or RCTs with a high risk of bias*
2 <sup>++</sup>	High-quality systematic reviews of case-control or cohort studies High-quality case-control or cohort studies with a very low risk of confounding, bias or chance and a high probability that the relationship is causal
2 <sup>+</sup>	Well-conducted case-control or cohort studies with a low risk of confounding, bias or chance and a moderate probability that the relationship is causal
2 <sup>-</sup>	Case-control or cohort studies with a high risk of confounding bias, or chance and a significant risk that the relationship is not causal*
3	Non-analytic studies (for example, case reports, case series)
4	Expert opinion, formal consensus
*Studies with a level of evidence '-' should not be used as a basis for making a recommendation (see section 7.4)	

It is the responsibility of the GDG to endorse the final levels given to the evidence, although it may delegate this process to the systematic reviewers.

### 7.3.3.2 Diagnostic studies

The system described above covers studies of treatment effectiveness. However, it is less appropriate for studies reporting diagnostic tests of accuracy. In the absence of a validated ranking system for this type of test, NICE has developed a hierarchy for evidence of accuracy of diagnostic tests that takes into account the various factors likely to affect the validity of these studies (Table 7.2). Because this hierarchy has not been systematically tested, NICE recommends that the NCCs use the system when appropriate, on a pilot basis, and report their experience to the Institute.

**Table 7.2 Levels of evidence for studies of the accuracy of diagnostic tests.** Adapted from *The Oxford Centre for Evidence-based Medicine Levels of Evidence* (2001) and the *Centre for Reviews and Dissemination Report Number 4* (2001).

Levels of evidence	Type of evidence
Ia	Systematic review (with homogeneity) <sup>*</sup> of level-1 studies <sup>†</sup>
Ib	Level-1 studies <sup>†</sup>
II	Level-2 studies <sup>‡</sup> Systematic reviews of level-2 studies
III	Level-3 studies <sup>§</sup> Systematic reviews of level-3 studies
IV	Consensus, expert committee reports or opinions and/or clinical experience without explicit critical appraisal; or based on physiology, bench research or 'first principles'
<p><sup>*</sup> Homogeneity means there are no or minor variations in the directions and degrees of results between individual studies that are included in the systematic review.</p> <p><sup>†</sup> Level-1 studies are studies:</p> <ul style="list-style-type: none"> <li>• that use a blind comparison of the test with a validated reference standard (gold standard)</li> <li>• in a sample of patients that reflects the population to whom the test would apply.</li> </ul> <p><sup>‡</sup> Level-2 studies are studies that have <b>only one</b> of the following:</p> <ul style="list-style-type: none"> <li>• narrow population (the sample does not reflect the population to whom the test would apply)</li> <li>• use a poor reference standard (defined as that where the 'test' is included in the 'reference', or where the 'testing' affects the 'reference')</li> <li>• the comparison between the test and reference standard is not blind</li> <li>• case-control studies.</li> </ul> <p><sup>§</sup> Level-3 studies are studies that have <b>at least two or three</b> of the features listed above<sup>§</sup>.</p>	

#### 7.4 Using the quality checklists to grade the evidence

In the quality assessment, each paper receives a quality rating coded as '++', '+', or '-'. Usually, studies rated as '-' should not be used as a basis for making a recommendation. If good-quality studies are available to help answer the clinical question, and their outcomes are consistent, the '-'-rated studies should be rejected. If there is a body of reasonable, but fairly weak, evidence showing a consistent effect and there are '-' studies that show the same effect, the '-'-rated studies should be included in the evidence table to demonstrate the extent of consistent evidence. If the '-' studies suggest a different outcome they should be left in the evidence table for further discussion with the GDG; they should not be used to support the recommendation as their inclusion as supporting evidence would weaken and downgrade the recommendation.

#### 7.5 Further reading

NHS Centre for Reviews and Dissemination (2001) *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out*

or commissioning reviews. *CRD Report Number 4*. 2nd edition. NHS Centre for Reviews and Dissemination, University of York. Available from: [www.york.ac.uk/inst/crd/report4.htm](http://www.york.ac.uk/inst/crd/report4.htm)

Drummond MF, O'Brien B, Stoddart GL et al. (1997) Critical assessment of economic evaluation. In: *Methods for the Economic Evaluation of Health Care Programmes*. 2nd edition. Oxford: Oxford Medical Publications.

Edwards P, Clarke M, DiGuseppi C et al. (2002) Identification of randomized trials in systematic reviews: accuracy and reliability of screening records. *Statistics in Medicine* 21:1635–40.

Eccles M, Mason J (2001) How to develop cost-conscious guidelines. *Health Technology Assessment* 5.

Khan KS, Kunz R, Kleijnen J, Antes G (2003) *Systematic Reviews to Support Evidence-based Medicine. How to Review and Apply Findings of Healthcare Research*. London: Royal Society of Medicine Press.

Scottish Intercollegiate Guidelines Network (2002) *SIGN 50. A Guideline Developer's Handbook*. Edinburgh: Scottish Intercollegiate Guidelines Network.