Quality assessment of primary studies is used at various stages in the review process, from study selection to generation of recommendations for practice and research. This phase provides background information on study quality and describes how to develop quality assessment checklists in a review.

2.5.1 Study quality assessment and bias 2
2.5.2 Bias and quality assessment 2
2.5.3 Quality threshold for study selection 4
2.5.4 Development of quality assessment instruments 6
2.5.5 Quality assessment of effectiveness studies 8
2.5.6 Quality assessment of test accuracy studies 11
2.5.7 Quality assessment of qualitative research 12
2.5.8 Quality assessment of economic evaluations 13
2.5.9 Key points 16
2.5.10 References 17
2.5.1 Study quality assessment

Quality is a construct about which there are differing views.\(^1,2\). Box 5.1 describes some quality constructs. Often both internal and external validity are assessed together during quality assessment. This is because interpretation of the findings of a study depends on design, conduct and analyses (internal validity), as well as on populations, interventions and outcome measures (external validity). These characteristics are related to the way in which the review questions are formulated (see Phase 1.2.2). Assessment of study quality focuses mainly on assessing the internal validity of effectiveness studies. Other quality issues will be covered in test accuracy studies (see Phase 2.5.6), qualitative research (see Phase 2.5.7) and health economic evaluations (see Phase 2.5.8). The information gained from quality assessment is crucial in determining the strength of inferences and in assigning grades to recommendations generated within a review (see Phase 3.8.5.7). Box 5.2 shows how quality assessment can be used at various stages in a review, starting with study selection to data synthesis and interpretation.

2.5.2 Bias and quality assessment

The post-hoc evaluation of bias is usually difficult and often impossible. For this reason, investigators should use methods that try to avoid bias altogether in primary research. However, primary studies often fail to avoid biases,\(^3-5\) therefore reviews should have robust quality assessment procedures. Consideration needs to be given to how the various types of biases shown in Box 5.3 may affect the results of the studies included in the review. Some methods that can be employed in primary studies to

---

Box 5.1
Terminology used in study quality assessment

**Study quality (methodological quality)**

The degree to which a study employs measures to minimise biases, focusing on internal validity.\(^1\) A set of parameters in the design and conduct of a study that reflects the validity of the outcome, related to the external and internal validity and the statistical model used.\(^2\)

**Bias (systematic error)**

A tendency to produce results that depart systematically from the ‘true’ results. Unbiased results are internally valid.

**Internal validity (validity)**

The degree to which the results of a study are likely to approximate to the ‘truth’. It is a prerequisite for external validity.

**External validity (generalisability, applicability)**

The extent to which the effects observed in a study are applicable outside of the study (in routine practice).
A precondition for valid results in effectiveness studies is that comparison groups should be similar at baseline. Unless the groups are balanced for relevant baseline characteristics, differences in outcomes cannot confidently be attributed to the effects of the intervention of interest. The validity of observational studies (e.g. cohort and case-control studies) is threatened, because of their vulnerability to selection bias, and it is generally considered to be lower than that of experimental studies. This is because the experimental method of randomly allocating participants produces comparison groups which are balanced on average for known, unknown and unmeasured confounding variables, whereas it is not possible to adjust for unknown and unmeasured confounding in observational studies. Several empirical studies have compared the effect estimates of studies with observational and experimental designs. While some researchers have found no difference in effect, others have observed larger estimates of effect in observational studies. The general consensus seems to

---

**Box 5.2**

**Reasons for study quality assessment in reviews**

- To determine a minimum quality threshold (study design threshold) for the selection of primary studies
- To explore quality differences as an explanation for heterogeneity in study results
- To weight the study results in proportion to quality in meta-analysis
- To guide the interpretation of findings and to aid in determining the strength of inferences
- To guide recommendations for future research

---

**Box 5.3**

**Types of biases**

**Selection bias (allocation bias)**  
Systematic differences between comparison groups in prognosis or responsiveness to treatment. Randomisation of large numbers of patients with concealment of their allocation to different groups protects against this bias.

**Performance bias**  
Systematic differences in care provided apart from the intervention being evaluated. Standardisation of the care protocol and blinding of clinicians and participants protects against this bias.

**Measurement bias (detection bias, ascertainment bias)**  
Systematic differences between comparison groups in how outcomes are ascertained. Blinding of study participants and outcome assessors protects against this bias.

**Attrition bias (exclusion bias)**  
Systematic differences between comparison groups in terms of withdrawals or exclusions of participants (e.g. because of side effects of the intervention) from the study sample. Inclusion of such participants in the analysis (in combination with a sensitivity analysis) protects against this bias.
be that in specific reviews, biases due to lack of randomisation can distort effects in either direction\cite{7,15,16} and that it is impossible to predict whether bias has been avoided in any particular non-randomised study.\cite{17}

Performance bias arises as a result of unintended intervention or cointervention not specified in the study protocol. It is associated with a lack of blinding of researchers and study participants.\cite{11,18} In addition, when the outcomes are assessed subjectively, it is important to blind those involved in ascertaining outcomes to prevent detection bias. Attrition bias arises because of inadequacies in accounting for losses of participants due to dropouts, exclusions, etc. leading to missing data on follow-up. When follow-up rates are not explicitly reported, it is impossible to determine if an intention-to-treat analysis has been undertaken or not. A sensitivity analysis may be undertaken by imputing a range of outcomes for the missing data.

2.5.3 Quality threshold for study selection

For study selection, simple assessment based on the appropriateness of study design is often used to guarantee a minimum level of quality (see Phase 2.4.1). The weakest study design that may be included in the review should be clearly stated in the inclusion/exclusion criteria in the protocol (see Phase 1.2.4). This quality threshold for primary studies can be determined by generating a hierarchy of study designs and fixing...
Box 5.5
Hierarchy of study designs for studies of effectiveness

Study design hierarchy

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Experimental studies (e.g. RCT with concealed allocation)</td>
</tr>
<tr>
<td>2.</td>
<td>Quasi-experimental studies (e.g. experimental study without randomisation)</td>
</tr>
<tr>
<td>3.</td>
<td>Controlled observational studies</td>
</tr>
<tr>
<td>3a.</td>
<td>Cohort studies</td>
</tr>
<tr>
<td>3b.</td>
<td>Case control studies</td>
</tr>
<tr>
<td>4.</td>
<td>Observational studies without control groups</td>
</tr>
<tr>
<td>5.</td>
<td>Expert opinion based on pathophysiology, bench research or consensus.</td>
</tr>
</tbody>
</table>

Description of selected study designs

- **Experimental**
  A study in which some conditions, particularly decisions concerning the allocation of participants to different intervention groups, are under the control of the investigator:
  - **Randomised controlled trial**
    Follow-up of participants randomly allocated to intervention or control groups, with a comparison of outcome rates during the time covered. Randomisation (with concealment of allocation sequence) avoids bias because both known and unknown determinants of outcome are on average evenly distributed between intervention and control groups.
  - **Quasi-experimental**
    A study in which the allocation of participants to different intervention groups is controlled by the investigator but the method falls short of genuine randomisation and allocation concealment.

- **Observational**
  A study in which natural variation in interventions or exposure among study participants is investigated to explore the effect of the interventions or exposure on health outcomes:
  - **Cohort study**
    Comparison of outcomes between participants who have received an intervention and a group that has not (i.e. not allocated by investigator) in a follow-up study.
  - **Case-control study**
    Comparison of exposure to interventions between participants with the outcome (cases) and those without the outcome (controls).
  - **Cross-sectional study**
    Examination of the relationship between disease and other variables of interest as they exist in a defined population at one particular time.
  - **Before-and-after study**
    Comparison of findings in study participants before and after an intervention.
  - **Case series**
    Description of a number of cases of an intervention and outcome (without comparison with a control group).

See Box 8.4 for grades of recommendations

See Boxes 5.10 and 5.13 for hierarchies of accuracy studies and economic evaluations

a cut-off level for study selection. The hierarchy of primary study designs depends on the nature of the questions being asked, for example, effectiveness, accuracy, efficiency, etc. (see Phase 1.2.2).

When assessing the effectiveness of therapy, the question tends to be about how one treatment performs compared to another, when different treatments are available for the same condition. To address this issue, the preferred study design would be one that randomly allocates (concealing the assignment code) the participants with the condition
of interest to alternative therapeutic interventions. This design will serve to remove selection bias described earlier (see Box 5.3). As a result, well-designed experimental studies tend to rank at the top of the study design hierarchy for assessing effectiveness. Next in the hierarchy are quasi-experimental studies where the allocation of participants is controlled, but falls short of genuine randomisation and allocation concealment. It is, however, not feasible to assess every therapeutic intervention on every relevant outcome using an experimental study design, particularly when randomisation is unethical or impractical. This means that when randomised trials are not possible or not available, the next best available evidence should be considered.

Box 5.5 shows a commonly used hierarchy of study designs for reviews of effectiveness. It is based on the degree to which different study designs are inherently susceptible to various biases. Reviewers often focus on randomised studies, but this emphasis may be unwarranted in some circumstances, for example, when literature scoping identifies only a few small randomised studies (see Phase 1.1.2). In this situation it may be prudent to include quasi-experimental and/or observational studies, and use study design as a basis for stratifying the analysis.

### 2.5.4 Development of quality assessment instruments

Once primary studies have been selected based on a minimum quality threshold (see Phases 2.4.3 and 2.5.3), the analysis and interpretation of their results will benefit from detailed assessment of their quality. Even within a particular design there is variability between studies with regard to execution. For example, studies may be biased due to inadequate randomisation, unsuitable comparison interventions, a lack of blind outcome assessment, inadequate follow-up times, inability to define and assess relevant outcomes, unreliable measurement techniques, and inappropriate statistical analyses. The variation in quality of selected primary studies has implications for data synthesis, interpretation of results and generation of inferences in a review.

### Box 5.6

**Quality assessment instruments**

**Individual quality components or items**

Individual aspects of study methodology, e.g. allocation concealment, blinding, follow-up, etc., which have a potential relation to bias in estimation of effect.

**Quality checklists**

Instruments based on a number of quality items, which are not scored numerically.

**Quality scales**

Instruments based on a number of quality items, which are scored numerically to provide a quantitative estimate of overall study quality. All scoring systems tend to be subjective. Scores can be generated by weighting all items equally or by assigning them different weights in relation to their perceived importance.
Classifying studies according to their level of methodological rigour will help to identify those studies which are of better quality. Both poor design of studies and lack of rigour in execution of a study may result in biased estimates of effects. Therefore, quality assessment of selected studies should guide data synthesis (see Phase 7). Where subgroups of studies of different quality exist, they may be analysed separately. Alternatively, an indication of study quality may be used as a variable in a meta-regression (see below). Differences in study quality may provide an explanation for heterogeneity in results (see Phase 2.7.4.4). When heterogeneity exists, reviewers should weight the better quality studies in generating inferences. Meta-analysis may be conducted where the study results are weighted in proportion to quality. An alternative approach is to cumulatively pool studies from high to low quality. These issues are discussed in detail in Phase 7. Finally, the strength of the review's inferences will be guided by the quality of the evidence (see Phase 3.8.5.7).

Quality assessment instruments are usually based on individual aspects or components of study design, conduct and analysis for which there is theoretical evidence of bias (see Box 5.6). These items can be assembled into a checklist, which can be used to systematically evaluate each study. Assigning numerical values to checklist items creates a scale. Checklists and scales offer an overall qualitative and quantitative index of study quality, which cannot be captured by single items alone. In addition, scales can allow quality scores to be generated. There are many generic checklists and scales available. Some emphasise quality of reporting which is not necessarily related to bias, and many tend to ignore external validity. In addition, the available quality instruments tend to focus on specific types of study designs. When reviews include both experimental and observational studies, combined checklists may be considered. At present, however, such reviews generally use separate checklists for separate study designs.

Quality assessment instruments should have been developed according to the principles used in creating measurement scales (see Box 5.7). However, most instruments have not been developed rigorously. As different checklists and scales emphasise different dimensions of quality, variation in these tools may produce differing assessments for the same studies. In addition, variation in quality scales may also produce differing summary estimates in meta-analysis. This observation can be explained, at least in part, by variations in the purpose, scope and degree of development of the different checklists and scales, and the use of an arbitrary dichotomy (low or high quality) in classifying studies. Scales, in particular, have been criticised for ignoring the direction of bias in their schema. Therefore, in exploring the impact of quality on estimation of effect, it is increasingly considered preferable to use individual components of methodological quality rather than summary scores. The value of this approach is maximised when it includes items related to external validity (see Phase 3.8.5.6).

Quality assessment procedures should be stated in the protocol (see Phase 1.2.5), and items of the proposed checklist or scale should be part of the data extraction form (see Phase 6). Under ideal circumstances, reviewers may be expected to develop specific quality instruments using the guidance shown in Box 5.7 for each new review. Having defined the quality construct, the scope of the review (e.g. effectiveness, accuracy, efficiency etc) should be considered. Following this, reviewers usually take one of the
generic quality tools that best suit their purpose. These tools will often provide good coverage of methodological issues related to study design. Because research in a particular topic area may be susceptible to specific biases, reviewers should be prepared to modify the selected generic quality instrument. They may consider including appropriate additional items or deleting irrelevant items.

The evaluation of quality items is affected by poor reporting of primary studies and tends to be subjective. Therefore, a clear description of how to assess quality should be documented along with the quality checklist. The checklist should be piloted on a practice set of articles to assess the reliability of the quality assessment process (see Phase 2.4.5). If the pilot phase shows inconsistency, a more explicit system of coding the checklist could improve inter-rater agreement (Appendix 2). There seems to be conflicting evidence about the effect of blinding reviewers to names of authors, institutions and journals when scoring quality. As with other aspects of data extraction, it is generally accepted that unmasked independent quality assessment by more than one reviewer should be sufficient in making judgements about study quality.

2.5.5 Quality assessment of effectiveness studies

A hierarchy of study designs is shown in Box 5.5. It is important to remember that RCTs should rank high in the hierarchy only when they are well conducted. Such trials would randomise a sufficient number of participants, conceal allocation code at randomisation, blind participants and investigators to assignment code throughout the investigation, achieve complete follow up for outcomes, and analyse results by intention to treat. If there are no good RCTs, well-conducted quasi-experimental and observational studies should be considered. Such studies assemble groups of participants with the condition of interest, non-randomly allocate them to interventions (quasi-experimental design) or describe the interventions they receive (observational design), complete follow up for outcomes to the end of the study period noting any
additional events they experience, and analyse the results with adjustment for imbalances in relevant measured baseline characteristics. Thus the hierarchy of study designs should be seen as a flexible, not a rigid, rule of thumb. There are numerous quality tools for experimental and observational studies of effectiveness.

2.5.5.1 Experimental studies

Experimental studies have their own hierarchy with high quality RCTs ranked at the top. It is possible that poor quality trials may produce results contrary to good quality trials within a review of RCTs.\cite{18,30,31} There are numerous quality items considered to be important when assessing the validity of randomised studies. One example from a Delphi consensus produced a 9-item criteria list generated after reduction of a total of 206 items in the initial pool (see Box 5.8).\cite{2} Many other checklists and scales have also been published\cite{1,34,43-49} (these references give only a few examples). The same criteria can also be used to assess quasi-experimental studies. Appendix A2.1 shows an example of a system for coding some of the important quality criteria in RCTs.

---

**Box 5.8**

**Quality criteria for assessment of experimental studies**\cite{2}

1. **Was the assignment to the treatment groups really random?**
   - Adequate approaches to sequence generation
     - Computer-generated random numbers
     - Random numbers tables
   - Inadequate approaches to sequence generation
     - Use of alternation, case record numbers, birth dates or week days

2. **Was the treatment allocation concealed?**
   - Adequate approaches to concealment of randomisation
     - Centralised or pharmacy-controlled randomisation
     - Serially-numbered identical containers
     - On-site computer based system with a randomisation sequence that is not readable until allocation
     - Other approaches with robust methods to prevent foreknowledge of the allocation sequence to clinicians and patients
   - Inadequate approaches to concealment of randomisation
     - Use of alternation, case record numbers, birth dates or week days
     - Open random numbers lists
     - Serially numbered envelopes (even sealed opaque envelopes can be subject to manipulation)

3. **Were the groups similar at baseline in terms of prognostic factors?**
4. **Were the eligibility criteria specified?**
5. **Were outcome assessors blinded to the treatment allocation?**
6. **Was the care provider blinded?**
7. **Was the patient blinded?**
8. **Were the point estimates and measure of variability presented for the primary outcome measure?**
9. **Did the analyses include an intention to treat analysis**
2.5.5.2 Observational studies

Observational studies can also be grouped into a hierarchy. Cohort studies, involving concurrent evaluation of groups receiving the different interventions being compared, are generally regarded as more valid than studies that make comparisons with historical controls. This reflects a tendency for there to be many more differences between two groups separated in time than there are between two groups at the same point in time. Cohort studies that are planned in advance and followed prospectively should be less biased than studies undertaken retrospectively. Data collection in prospective studies is planned and therefore more likely to be uniformly reliable and complete. In addition, in contrast to retrospective studies, the selection of the participants in prospective studies is unlikely to be influenced by their outcomes.

Case-control studies are more susceptible to bias than cohort studies. If case-control studies are included in reviews they may be graded according to the suitability of the choice of the control group. Matching is used to make groups comparable for one or more potential confounding factors, for example by ensuring approximately equal numbers of confounding variables in each group, or by individual matching (where one or more controls are identified that share some of the same characteristics as an individual case). It is also important to consider that the treatment effects may be underestimated due to over-matching on factors that are related to allocation to the intervention. Other issues related to validity include the assessment of exposure to interventions being performed blind to the outcome, and the selection of cases and controls being blind to measures of exposure.

A third group of studies have a before-and-after design, where the same participants are evaluated before and after an intervention with no additional comparator or 'control'. The comparison is made within the single group of participants. Here it is often very hard to conclude whether any differences seen can be attributed to the intervention. However, when RCTs are not feasible, large differences in before-and-after studies may provide some indication of effect, e.g. when evaluating the impact of change in health policy. An alternative design is time series, where measurements are made sequentially over time, before and after the intervention. Such studies may provide more reliable information than simple before-and-after studies. In time series, analysis centres on identifying underlying trends and periodicity. The statistically important feature is the correlation between observations within a series, which must be appropriately modelled if correct conclusions are to be drawn. In using a time series design the researcher should be able to quantify a stable baseline level or trend before the intervention. This trend is projected into the future and the effect of the intervention is manifested as the difference between the actual and projected outcomes.

Checklists are available for epidemiological studies which assess potential links between exposures to risk factors and harm. These checklists can be modified to assess the strength of evidence from observational studies assessing effectiveness, as the same methodological issues are relevant (see Box 5.9).

On some occasions there may be no studies that compare interventions, but there may be separate reports of patients who have received different interventions being followed up (e.g. a case series of patients receiving different prostheses in hip replacement). The validity of the effect estimates in these studies depends on factors such as the recruitment
of inception cohorts and the length of follow-up. Checklists which assess articles on prognosis can be used to evaluate these articles.\textsuperscript{54, 55} In a review it may be possible to draw some tentative conclusions by comparing the cohorts from different studies, but it is rarely clear whether the groups in the different studies are comparable, and great caution is needed with such comparisons. In general, inferences derived from such studies should be considered to generate hypotheses for testing in future research.

2.5.6 Quality assessment of test accuracy studies

In the screening and diagnosis domain, research often focuses on the evaluation of test accuracy, i.e. the degree to which a test correctly identifies the presence or absence of disease (see Phase 1.2.2.2). The quality issues in this type of literature are quite different to those in effectiveness studies.\textsuperscript{56} The preferred study design for assessing test accuracy is one which prospectively recruits all eligible participants, uses the test and

<table>
<thead>
<tr>
<th>Box 5.9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Some quality criteria for assessment of observational studies</strong></td>
</tr>
</tbody>
</table>

### Cohort studies
- Is there sufficient description of the groups and the distribution of prognostic factors?
- Are the groups assembled at a similar point in their disease progression?
- Is the intervention/treatment reliably ascertained?
- Were the groups comparable on all important confounding factors?
- Was there adequate adjustment for the effects of these confounding variables?
- Was a dose-response relationship between intervention and outcome demonstrated?
- Was outcome assessment blind to exposure status?
- Was follow-up long enough for the outcomes to occur?
- What proportion of the cohort was followed-up?
- Were drop-out rates and reasons for drop-out similar across intervention and unexposed groups?

### Case-control studies
- Is the case definition explicit?
- Has the disease state of the cases been reliably assessed and validated?
- Were the controls randomly selected from the source of population of the cases?
- How comparable are the cases and controls with respect to potential confounding factors?
- Were interventions and other exposures assessed in the same way for cases and controls?
- How was the response rate defined?
- Were the non-response rates and reasons for non-response the same in both groups?
- Is it possible that over-matching has occurred in that cases and controls were matched on factors related to exposure?
- Was an appropriate statistical analysis used (matched or unmatched)?

### Case series
- Is the study based on a representative sample selected from a relevant population?
- Are the criteria for inclusion explicit?
- Did all individuals enter the survey at a similar point in their disease progression?
- Was follow-up long enough for important events to occur?
- Were outcomes assessed using objective criteria or was blinding used?
- If comparisons of sub-series are being made, was there sufficient description of the series and the distribution of prognostic factors?
the ‘reference standard’ investigation to confirm or refute the presence of disease, and determines the accuracy with which the test identifies disease. There are several tools to assess the quality of test accuracy studies,\textsuperscript{57-59} and Box 5.10 shows a modified version of a recently published hierarchy of evidence for such studies.\textsuperscript{60} Several elements of this evidence hierarchy have been shown to have a direct relationship to bias. For example, bias in estimation of test accuracy has been associated with inadequacies in description of population and diagnostic test, inappropriate use of reference standard, lack of blinding and case-control study design.\textsuperscript{61}

In addition to evaluation of test accuracy, primary research may also assess effectiveness of testing (and consequent therapeutic) strategies, considering the benefits and harms of correct as well as incorrect treatment (and non-treatment). Several study designs can be used to assess the effectiveness of tests, including RCTs, observational studies with modelling of the contribution of tests and treatments to outcomes, and clinical decision analysis. However, it is not feasible to assess the effectiveness of all tests using randomised trials. In addition, observational studies that use modelling to determine the impact of tests and treatments on important health outcomes are not always available. Observational cohort studies that explore the effectiveness of testing by comparing a tested population with the rest of the population are particularly susceptible to lead time bias. This happens because a test simply detects a disease earlier. It does not actually improve survival. It may appear in an observational study that the test has increased survival but this is simply because the persons tested have lived with a diagnosis for longer. In general, the information gathered on the effectiveness of testing strategies can be assessed for quality according to the hierarchy of evidence shown in Box 5.5 and the principles described in Phase 2.5.4.

### 2.5.7 Quality assessment of qualitative research

The growing interest in the contribution of qualitative research findings to the health and social care evidence base\textsuperscript{62} has led to the inclusion of findings from qualitative research in systematic reviews of effectiveness alongside quantitatively measured

---

**Box 5.10**

**Hierarchy of evidence for test accuracy studies**

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>A blind comparison with reference standard among an appropriate broadly defined sample of consecutive patients</td>
</tr>
<tr>
<td>2.</td>
<td>Any one of the following ⇒ Narrow population spectrum</td>
</tr>
<tr>
<td>3.</td>
<td>Any two of the following ⇒ Differential use of reference standard</td>
</tr>
<tr>
<td>4.</td>
<td>Any three or more of the following ⇒ Reference standard not blind Case control study design</td>
</tr>
<tr>
<td>5.</td>
<td>Expert opinion with no explicit critical appraisal, based on physiology, bench research or first principles.</td>
</tr>
</tbody>
</table>

*See Box 8.4 for grades of recommendations*  
*See Boxes 5.5 and 5.13 for hierarchies of effectiveness studies and economic evaluations*
outcomes (see Phase 1.2.2.3). Whilst the nature and application of procedures to minimise bias in quantitative research may be problematic, it is desirable and theoretically possible to have a structured approach to quality assessment (see Phase 2.5.4). Qualitative researchers also readily accept the need for clear and transparent approaches for judging the quality or credibility of research. For example, it has recently been noted that ‘the distinguishing mark of all ‘good’ research is the awareness and acknowledgement of error and, that what flows from this is the necessity of establishing procedures which will minimise the effect such errors may have on what counts as knowledge.63 However, it is much less clear whether a single hierarchy of evidence can be developed or, indeed, if it is feasible at all.

In relation to quality assessment some commentators take an extreme view, arguing for example, that quality ‘cannot be determined by following prescribed formulas64 or that it is ‘fruitless to try to set standards for qualitative research per se.65 Others, accepting the need for structured procedures, argue for more rigorous use and reporting of analytical approaches which improve reliability and validity in qualitative research.66 Another suggested approach is to audit the research process from beginning to end, a laborious and impractical process in most situations.67 Others have suggested there are general questions that can be asked to judge validity and reliability in qualitative research, but that these are not readily codified.62

Despite these disagreements, structured approaches to judging validity and reliability in qualitative research are being developed. In a recent publication, eight frameworks containing a large number of appraisal criteria were identified.63 None of these criteria appeared in all frameworks although there was a considerable degree of overlap between the main criteria described. Such overlaps are common in the general quality criteria used in both qualitative and quantitative research appraisal.68 Box 5.11 provides three quality appraisal frameworks. Popay and colleagues suggest that before proceeding to a detailed assessment of methodological soundness, an important primary question relates to the appropriateness of the methods used.69 The other frameworks do not include this step, the application of which arguably requires some experience of qualitative research. Problems are likely to arise not simply in relation to which criteria should be included in quality appraisal, but in how they should be applied. It has been noted that most frameworks fail to specify how judgements should be made or whether or not a standard has been reached.70 Given these uncertainties, it is not possible to generate guidance at this stage on hierarchies of qualitative evidence.

2.5.8 Quality assessment of economic evaluations
Systematic reviews are increasingly being required to evaluate efficiency in addition to clinical effectiveness.73 The types of economic evaluations to be included in a review of efficiency depend on the nature of questions being asked in the review (see Phase 1.2.2.4). To qualify for inclusion in a review of full economic evaluations, a study would need to explicitly analyse both the costs and the clinical outcomes of the intervention being considered in comparison to the costs and outcomes of at least one alternative.74
Once economic evaluations have been selected for inclusion in the review, their quality needs to be assessed. Critical appraisal of an economic evaluation using checklists allows a thorough exploration of the methods of analysis used, so that informed judgements can be made about the validity of the study results. Quality assessment may also provide an insight into heterogeneity of results between economic evaluations.
A quality checklist, like the one shown in Box 5.12, will ensure that all relevant methodological points are appraised for each economic evaluation. In this checklist, items 1-3 can be used to develop criteria for study selection. When these criteria are applied strictly, evaluations assessing only the costs and not the effectiveness of an intervention (cost studies) would usually be excluded from a review. Item 4 is usually the issue being tackled in the effectiveness part of the systematic review. Items 5-9 allow reviewers to assess the validity of the conclusions about efficiency. Items 10 and 11 allow the issues of inclusivity and generalisability for the target setting to be assessed. These items can be used to develop a hierarchy of economic evidence (see Box 5.13). Should the reviewer wish to include a more detailed analysis of each study, more comprehensive checklists are available, which can be adapted to produce a quality scoring system.78

In quality assessment of economic evaluations, the reviewer should be cognisant of the type of evaluation being assessed: trial-based (using a single source of evidence) or model-based (using multiple sources in the literature). If trial-based evaluations are being assessed then the internal validity of the study design should be the point of focus for clinical effectiveness. Alternatively, if a model-based evaluation is being assessed, the quality of the literature review, the methods of deriving input parameters for the model, the representativeness of the model in terms of clinical practice and sensitivity analyses should be the points of focus. In these evaluations, the decision analytic models commonly employ decision trees or Markov models, which aim at reflecting more realistic clinical decision-making processes by taking into account all possible strategies and outcomes.80 For example, at different times during follow-up, patients may change from one health state to another. Markov models also take into account the natural history of the condition under review and probabilities associated with health state transitions.81

---

**Box 5.12**

**Checklist for assessing economic evaluations**

1. Is there a well defined question?
2. Is there comprehensive description of alternatives?
3. Are all important and relevant costs and outcomes for each alternative identified?
4. Has clinical effectiveness been established?
5. Are costs and outcomes measured accurately?
6. Are costs and outcomes valued credibly?
7. Are costs and outcomes adjusted for differential timing?
8. Is there an incremental analysis of costs and consequences?
9. Were sensitivity analyses conducted to investigate uncertainty in estimates of cost or consequences?
10. How far do study results include all issues of concern to users?
11. Are the results generalisable to the setting of interest in the review?

*Based on Drummond's checklist*77
2.5.9 Key points about study quality assessment

- Quality assessment of research involves appraisal of a study's internal validity, i.e. the degree to which its design, conduct and analysis have minimised biases or errors. For practical reasons, study quality assessment in reviews often covers both internal and external validity.
- Initially, quality assessment can be used to determine a minimum quality threshold (study design threshold) for the selection of primary studies to be included in a review.
- Detailed quality assessment is then used to scrutinise the quality of included studies in order to explore quality differences as an explanation for heterogeneity in study results. This aids interpretation of the results and allows the generation of inferences to inform practice and research.
- Checklists and scales are available, or can be developed, to assess the quality of included studies. Their components should be selected with due consideration for the scope and purpose of quality assessment. Components should capture both generic methodological issues and issues specific to the subject area being reviewed. The prevalent view is that for exploring the impact of quality, individual quality components are preferable to quality scores.

---

**Box 5.13**

**Hierarchy of economic evidence**

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Evaluation of important alternative interventions comparing all clinically relevant outcomes against appropriate cost measurement, and including a clinically sensible sensitivity analysis</td>
</tr>
<tr>
<td>2.</td>
<td>Evaluation of important alternative interventions comparing a limited number of outcomes against appropriate cost measurement, but including a clinically sensible sensitivity analysis</td>
</tr>
<tr>
<td>3.</td>
<td>Evaluation of important alternative interventions comparing all clinically relevant outcomes against inappropriate cost measurement, but including a clinically sensible sensitivity analysis</td>
</tr>
<tr>
<td>4.</td>
<td>Evaluation without a clinically sensible sensitivity analysis</td>
</tr>
<tr>
<td>5.</td>
<td>Expert opinion with no explicit critical appraisal, based on economic theory</td>
</tr>
</tbody>
</table>

*See Box 8.4 for grades of recommendations*

*See Boxes 5.5 and 5.10 for hierarchies of effectiveness and accuracy studies*
2.5.10 References


47. Guyatt GH, Sackett DL, Cook DJ. Users’ guides to the medical literature. II: How to use an article about therapy or prevention, A: are the results of the study valid? *JAMA* 1993;270:2598-2601.

48. Guyatt GH, Sackett DL, Cook DJ. Users’ guides to the medical literature. II: How to use an article about therapy or prevention, B: what are the results and will they help me in caring for my patients? *JAMA* 1994;271:59-63.


69. Popay J, Rogers A, Williams G. Rationale and standards in the systematic review of qualitative literature in health services research. *Qualitative Health Research* 1998;8:341-351.


75. How to read clinical journals, VII: To understand an economic evaluation (part A). *Can Med Assoc J* 1984;130:1428-1433.

76. How to read clinical journals, VII: To understand an economic evaluation (part B). *Can Med Assoc J* 1984;130:1542-1549.


