Critical appraisal of quantitative studies 1: Is the quality of the study good enough for you to use the findings?

Faith Gibson and Anne-Marie Glenny

KEY POINTS

- Why is critical appraisal necessary?
- The principles of critical appraisal.
- Identifying the appropriate research design for your clinical practice question.
Skills for evidence-based practice

- Assessing the quality of a study that asks a question about the effectiveness of a therapy or intervention.
- Assessing the quality of a study that asks a question about the accuracy of a particular diagnostic test or method of assessment works.
- Assessing the quality of a study that asks a question about finding out the likely pattern and/or outcome of a particular health problem/disease.
- Critical appraisal in practice.

This chapter should be read in conjunction with Chapter 5.

Introduction

On a daily basis nurses and other health-care professionals are faced with a range of important clinical decisions. Practice based on evidence can decrease the uncertainty that patients and health-care professionals experience in a complex and constantly evolving health-care system. Not all published research evidence can be used for making decisions about patient care. Deficiencies in research design can make an intervention look better than it really is (Moher et al 1995). In addition, the location and participants of a particular research study may affect the results in a unique way. It is therefore necessary to assess the quality, importance and applicability of any research evidence that is being consulted to answer a specific clinical question.

The process used to do this is known as critical appraisal and is a core skill for those wanting to use evidence in their practice. Critical appraisal is a discipline for increasing the effectiveness of your reading, by encouraging systematic assessment of reports of research evidence to see which ones can best answer clinical problems and inform ‘best practice’ (O’Rourke 2005). The meaning of research evidence is most fully appreciated when considered within the context of clinical practice, where it can have a direct impact on clinical decisions. Clinical decisions are influenced not only by research evidence but also by clinical expertise and patient preference (Sackett et al 1996). A variety of appraisal approaches can be used to determine the certainty and applicability of knowledge underpinning each of these three aspects of decision making (Stevens 2005). However, in this chapter, we focus on the appraisal of research evidence only, and specifically, quantitative research evidence. So, how do you critically appraise evidence produced through quantitative research designs? What does it demand of you?
Is the study good enough to use the findings?

Critical appraisal can be broken down into three distinct but related parts as illustrated in Figure 4.1.

1. Are the results of the study valid? In other words, is the quality of the study good enough to produce results that can be used to inform clinical decisions?
2. What are the results and what do they mean in my context/for my patient?
3. Will the results help locally? Can I apply them in my clinical setting?

Answering ‘Yes’ or ‘No’ to these questions can prove a challenge to healthcare practitioners. As Oxman et al (1993) observe, research evidence comes in shades of grey, rather than black and white: results may be valid, might show clinically important findings, perhaps will improve the patient’s outcome.

Over the past 20 years or so, researchers and clinicians around the world have been working together to develop standard approaches to addressing these three questions. This work has led to the development of quality criteria for assessing the design of research studies. These criteria have been incorporated into critical appraisal checklists in the form of streamlined guides and toolkits that make the process of assessing studies much easier (see the ‘Appraising’ section of Netting the Evidence (www.shef.ac.uk/scharr/ir/netting/) for examples of quality appraisal guides/toolkits).

Figure 4.1 The three aspects of critical appraisal for evidence-based practice
This and the next chapter (Chapter 5) provide a set of tools that can be used for critical appraisal. The chapters also provide help in developing the skills and knowledge necessary to use these tools. The remainder of this chapter will consider how to assess whether the quality of the study is good enough for the results to be used (i.e. shading in Figure 4.1). Chapter 5 addresses the other two aspects. Both chapters use practical examples to illustrate the process: clinical scenarios are used to generate clinical questions and the relevant published research papers are then appraised. The same examples are used in both chapters. You will obtain more benefit from the examples if you obtain copies of the published research papers.

Is the quality of the study good enough for me to use the results?

One of the tests that all researchers face in the design and conduct of research in real-world settings is that of minimising bias. Bias refers to any influence or action in a study that distorts the findings or slants them away from the true or expected (Burns & Grove 2001). In other words, any factor (for example, the way the study is conducted, analysed or published) that leads to interventions appearing to be effective when in fact they are not, or vice versa. The critical reader of research must always be aware that there may be multiple explanations for the findings reported in a study (Johnston 2005).

When designing a research study, the researchers have to consider a number of questions (Box 4.1). These questions apply in every type of research

<table>
<thead>
<tr>
<th>BOX 4.1 Questions to be considered in the design of a research study (after Sim &amp; Wright 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>What entities or variables to examine</td>
</tr>
<tr>
<td>Under what conditions to examine these entities or variables</td>
</tr>
<tr>
<td>What type(s) of data to collect</td>
</tr>
<tr>
<td>From whom (or what) to collect these data</td>
</tr>
<tr>
<td>At what time points to collect the data</td>
</tr>
<tr>
<td>What method to employ for data collection</td>
</tr>
<tr>
<td>How data will be analysed</td>
</tr>
</tbody>
</table>
design. The decisions made by the researchers in response to each of these questions directly affect the degree to which the results of a study may be affected by bias. The strategies for the minimisation of bias are now well known. Different strategies are required for different research designs and for the different stages of the study (Moore & McQuay 2000). When looked at from the perspective of the research consumer, these bias minimisation strategies become quality criteria that can be used to assess the quality of a study. The strategies adopted by the researchers to minimise bias should be evident in the reporting of the research. Critical appraisal checklists are designed to summarise the bias minimisation strategies and help practitioners to ask the most relevant questions that will lead to a decision about the quality and hence usefulness of a paper. Tables 4.2, 4.3 and 4.4 show worked examples of using quality criteria to assess the worth of three different studies; the quality criteria are explained in detail later in the chapter.

Is the study design appropriate to the question?

As shown in Chapter 2, the process of critical appraisal for evidence-based practice starts with the formulation of a question that arises from clinical practice. For critical appraisal purposes, clinical practice questions can, broadly speaking, be categorised into different types. In this and the following chapter, we focus on clinical questions about:

- the effectiveness of a therapy or intervention
- the accuracy of a diagnostic test or method of assessment, and
- the prognosis of a particular disease or health problem (see Chapter 6 for discussion of critical appraisal of questions requiring qualitative research designs).

For each type of question there is a corresponding ‘most appropriate’ research design (overall plan or structure used by the researcher) that can be used to answer the question with a known degree of precision and minimal risk of bias (Blaikie 2000). In this chapter we have focused on the optimal study design for three different types of questions, as presented in Table 4.1. However, it is important to remember that there may be good reasons why researchers choose to use study designs that at first appear to be less appropriate for the research question. Each research project presents unique challenges and a certain degree of flexibility is required by the researcher.
For each research design, specific criteria need to be considered. These will be explored in more detail later in this chapter.

In the case of questions about whether a particular diagnostic test or method of assessment performs well, a study design that compares the accuracy of a new test when used on people with and without the target condition...
Is the study good enough to use the findings?

against a reference standard will be most appropriate (Mant 1999a). Where the question is about the most likely outcome of a particular health problem (i.e. the prognosis), the most appropriate design will be one that measures relevant outcomes in individuals with (and perhaps without) the relevant condition over a sufficient period of time (Mant 1999b). If the clinical question is about whether a particular intervention (e.g. a nurse-led discharge package) produces a certain outcome (e.g. decreased hospital stay), a study that compares length of hospital stay in a group receiving the intervention with length of stay in a group not receiving the intervention is required. There are a number of possible research designs that could be used for such a study but large, multicentred randomised controlled trials (RCTs) are likely to give the best evidence of effectiveness (Gray 1997), provided they are conducted rigorously.

Regardless of the type of study, a rigorous approach to the design, conduct, analysis and reporting stages of the study is important in view of the effect that each of these stages can have on the results. For example, RCTs with methodological shortfalls, such as failure to conceal from the patient and assessor the group to which the patient has been allocated, tend to overestimate treatment effects (Mathews 2000).

Systematic reviews

There has been an explosion in the amount of published research evidence, much of which is not always easily accessible to practitioners. When searching for evidence, it is unlikely that all the studies that address a particular question will be identified. In addition, wading through large numbers of articles can be extremely time consuming and may not be an option for a busy practitioner. Where relevant studies are identified, their conclusions may differ, and this poses a dilemma for the person trying to find a solution to a problem.

These are just some of the reasons for the growing interest in systematic reviews as a form of evidence. A systematic review summarises pertinent research evidence on a defined health question, using explicit and rigorous methods (Cullinan 2005). In contrast to a traditional narrative review, which might simply reflect the findings of a few papers that support an author’s particular point of view (Johnston 2005), a systematic review entails systematic and explicit methods for identifying, assessing and synthesising the available research evidence. This is an important distinction and one which highlights the biases associated with traditional, narrative
reviews. There are numerous advantages to systematic reviews. For the busy practitioner, the most important advantage would seem to be their potential for aiding translation of research evidence into practice (Evans & Pearson 2001): the well-conducted review offers the clinician a rigorous distillation of the pertinent research evidence and recommendations for clinical practice. Systematic reviews can help to overcome many of the practical and methodological limitations of individual studies (Chalmers & Altman 1995), as discussed in Chapter 7.

To date, the majority of systematic reviews have focused on effectiveness of different interventions, where the study design used by the primary authors was the controlled trial. However, systematic reviews of other types of study designs are also being conducted. In this chapter we have not looked at critical appraisal of systematic reviews; however, critical appraisal checklists specific to systematic reviews are available. Chapter 7 provides more information.

The hierarchy of evidence

When looking for evidence about the effectiveness of interventions, properly conducted systematic reviews of RCTs or properly conducted RCTs provide the most powerful form of evidence. The process of randomisation means that the observed differences between the intervention group and the comparison group are more likely to be due to the intervention and not to other factors such as patient, nurse or doctor preference (Mathews 2000). There will, however, always be circumstances where randomisation may be inappropriate or impossible, particularly in health service research (Bowling 1997). An obvious example is studies about harm or prognosis. It would not be ethical to give subjects a substance that was thought to be hazardous to their health. In circumstances where there are clear reasons for not randomising, studies with other designs have a vital role in providing evidence (Martin 2005).

Evidence that some research designs are more powerful than others has given rise to the notion of a hierarchy of evidence (Summerskill 2000). Figure 4.2 illustrates this hierarchy for studies relating to effectiveness of therapies or interventions. The continuum is used to illustrate the increasing risk of bias inherent in each research design. The higher a methodology appears in the hierarchy, the more likely the results of such methods are to represent objective reality and hence the more certainty the practitioner
Is the study good enough to use the findings?

has that the intervention will produce the same health outcomes (Johnston 2005). Such hierarchies provide a useful ‘rule of thumb’ against which to grade studies.

The hierarchy of evidence shown in Figure 4.2 has led some clinicians to expressing concerns that the only questions considered to be important are those about effectiveness of interventions, and that the only valid type of study is the RCT (Evans & Pearson 2001, French 2002). However, it must be emphasised that different research questions have different hierarchies and require evaluation through different study designs. RCTs are the ‘gold standard’ for primary study design upon which to base decisions on the effectiveness of health-care interventions, but they are not necessarily appropriate, or ethical, to answer other questions. For example, it would be inappropriate to conduct a RCT to answer the question ‘How do young people with cancer experience and manage fatigue?’ Given that the researcher would not be intervening in any way, and that the study aims to examine the experience of fatigue, explore influencing factors and management strategies adopted by young people with cancer, a prospective, cohort follow-up design using narrative inquiry would be more appropriate.

**Figure 4.2** The hierarchy of evidence for questions about the effectiveness of an intervention/therapy
Worked example 4.1: Assessing the quality of a study that asks a question about the effectiveness of a therapy or intervention

Scenario

You are part of a nursing team in a children’s hospital delivering care to children with complex needs. The philosophy of your team is one that incorporates the whole family: that of family-centred care. The mother of Henry Pink, one of the children receiving respite care in your unit, shares her worries about an upcoming flight she is taking. The family has been given some money from a local charity to visit Disney World in Florida. The whole family, including Grandma and Grandad, is very excited about this visit. Mum’s worries centre on a journal article she read in a women’s magazine about the risk of deep vein thrombosis and the benefits of wearing elastic stockings. Although all are fit (with no history of thromboembolic problems), Mrs Pink has worries for herself and her parents. You agree to ask one of your colleagues about it, as the colleague recently undertook a long-haul flight to Kenya and investigated this topic prior to flying. Your colleague recalls one particularly helpful paper. You suggest to the child’s mum that you will help her to ‘make sense’ of this paper so that she can reach a decision about any action she might need to take before their holiday.

Your clinical practice question

Does the wearing of elastic compression stockings prevent deep vein thrombosis in passengers undertaking long-haul flights?

Finding the evidence

Your colleague supplies the following paper that looks as though it might be relevant:

Is the study good enough to use the findings?

The paper reports the results of a RCT involving 89 male and 142 female long-haul air passengers, over the age of 50 years, with no history of thromboembolic problems. Passengers were randomised to two groups, with those in the intervention group being issued below-knee graduated elastic compression stockings to be worn for the duration of the flight. The control group did not receive any intervention. The primary outcome was symptomless deep vein thrombosis. Assessments were made of the veins using duplex ultrasonography, and blood samples were taken to measure gene mutations (factor V Leiden and prothrombin), which are known to predispose to venous thromboembolism. In addition, a sensitive D-dimer assay was used to screen for the development of recent thrombosis.

What is the quality of the study?

Table 4.2 shows a worked example of the critical appraisal for the paper referenced above, using the appraisal tool ‘Ten questions to help you make sense of randomised controlled trials’, published by CASP (2002).

Relevance of criteria in Table 4.2

- The importance of randomisation has already been discussed: the aim is to ensure that, as far as possible, the two groups are similar apart from the intervention. This means that any difference in outcome between the two groups is likely to be due to the intervention. A computer-generated number sequence is one example of an appropriate randomisation method.

- The group to which the patient has been allocated must be concealed from the clinician/researcher until the patient has been accepted into the trial. This is often referred to as allocation concealment and is an important factor in reducing bias. If the clinician believes that the patient may benefit from the treatment, and realises that the patient is due to be allocated to the control group, he or she may consciously or subconsciously dissuade the patient from participating in the trial. Ideally, randomisation should be carried out by someone removed from the project using, for example, sequentially numbered sealed, opaque envelopes. Randomisation based on criteria such as date of birth are not recommended as clinicians are able to work out from the date of birth and the allocation sequence which group the patient is to be allocated to.

- Where possible, patients, clinicians and researchers should be ‘blinded’ as to whether a patient is in the treatment or control group. If patients know that they are in the control group, they may feel that they have received
### TABLE 4.2 ASSESSING THE QUALITY OF A STUDY THAT INVESTIGATES WHETHER A PARTICULAR INTERVENTION IS EFFECTIVE: A WORKED EXAMPLE USING THE CASP CHECKLIST ‘TEN QUESTIONS TO HELP YOU MAKE SENSE OF RANDOMISED CONTROLLED TRIALS’ (CASp 2002, WITH PERMISSION, © MILTON KEYNES PRIMARY CARE TRUST, 2002)

<table>
<thead>
<tr>
<th>Type of question</th>
<th>Assessment criteria</th>
<th>Response (Yes / Can’t tell / No)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening question</td>
<td>Did the study ask a clearly focused question?</td>
<td>Yes</td>
<td>This study sought to determine the frequency of deep vein thrombosis (DVT) in the lower limb in middle-aged men and women during long-haul economy class air travel and the efficacy of elastic compression stockings in its prevention.</td>
</tr>
<tr>
<td></td>
<td>Was this a randomised controlled trial (RCT), and was it appropriately so?</td>
<td>Yes</td>
<td>This study compared the frequency and prevention of DVT in two groups randomly allocated to wear or not wear stockings. The study design (RCT) is appropriate for investigating the effectiveness of an intervention.</td>
</tr>
<tr>
<td>Detailed questions</td>
<td>Were participants appropriately allocated to intervention and control groups?</td>
<td>Yes</td>
<td>The author states that passengers were randomly allocated to receive compression stocking (n=115) or no stockings (n=116); however, the method of randomisation is not described. Sealed envelopes were used to conceal from the investigators and passengers which group the passengers were to be allocated to.</td>
</tr>
<tr>
<td>Question</td>
<td>Answer</td>
<td>Reason</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>---------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Were participants, staff and study personnel ‘blind’ to participants’ study group?</td>
<td>Can’t tell</td>
<td>Not possible to blind the passengers or the staff administering the stockings. They all knew who was receiving treatment. It might have been possible to blind the laboratory staff and the ultrasonographer, but there is a high risk that the patients might have revealed to the ultrasonographer which group they were in.</td>
<td></td>
</tr>
<tr>
<td>Were all the participants who entered the trial accounted for at its conclusion?</td>
<td>Yes</td>
<td>Yes, trial profile clearly describes and states intention to treat in the analysis section. Data on all participants originally randomised have been included in the analysis. The authors show the numbers of passengers lost to follow-up for each group. These appear to be equally distributed between groups.</td>
<td></td>
</tr>
<tr>
<td>Were the participants in all groups followed up and data collected in the same way?</td>
<td>Can’t tell</td>
<td>This is unclear; insufficient information is provided about the actual study procedure. It would be difficult to comment on the influence of performance bias.</td>
<td></td>
</tr>
<tr>
<td>Did the study have enough participants to minimise the play of chance?</td>
<td>Can’t tell</td>
<td>This was described as a pilot study and therefore no power calculation was undertaken.</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
TABLE 4.2 (CONTINUED)

<table>
<thead>
<tr>
<th>Type of question</th>
<th>Assessment criteria</th>
<th>Response (Yes / Can’t tell / No)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>How are the results presented and what is the main result?</td>
<td>Present as a proportion of participants experiencing DVT after airline travel. None of the participants wearing stockings had a DVT. Although an increased risk of superficial thrombophlebitis in varicose veins if the stockings are worn is reported, the increase is not statistically significant. 10% (n = 12/116) of participants not wearing stockings developed a symptomless DVT. The authors conclude that wearing an elasticated stocking during long-haul air travel may be associated with a reduction in symptomless DVT.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>How precise are the results?</td>
<td>Confidence intervals are reported for each group, not between groups, but are indicative of clinical significance.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Were all the important outcomes considered so the results can be applied?</td>
<td>No</td>
<td>A number of limitations were addressed in the paper that would need to be considered such as in-flight behaviours, e.g. walking, and drinking water and the effect they may have had on the results. The reported risk of thrombophlebitis in varicose veins if the stockings are worn would also need to be taken into consideration.</td>
<td></td>
</tr>
</tbody>
</table>
substandard care and may, as a result, alter their behaviour. Similarly, clinicians may consciously or subconsciously take compensatory measures for patients who are in the control group (for example, by offering alternative therapies or additional support, etc.). Any difference in the treatment effects between the two groups may be due to this additional attention rather than the intervention. The researcher may have preconceived ideas about the treatment and, where the outcomes of interest are fairly subjective, these preconceptions may influence the way in which the researcher interprets and analyses the data. Clearly, blinding is not possible in all studies but attempts should be made to blind one (single-blind) or all (double-blind/treble-blind) of the above groups of people.

- People drop out of studies for all sorts of reasons: death, relocation to another geographical area, treatment found to be too unpleasant, etc. It is important that the researcher tries to identify whether the reasons relate to the outcomes of interest. The analysis should ideally be done on an ‘intention to treat’ basis: patients are analysed in the groups to which they were randomised regardless of whether they swap from the intervention arm of the trial to the control arm or vice versa. If participants in the treatment group stop taking a drug because they feel worse (and blame the drug), and are then included in the control group, the drug may appear to be more effective than it really is due to exclusion of those patients with poor outcomes from the treatment group.

- Look to see if the investigators estimated how many people needed to be studied in order to answer their research question. Studying more people than is necessary wastes resources, whilst studying too few people might lead to results that reflect chance alone, rather than the real situation. To illustrate, in a trial of 10 participants, where five are randomised to receive drug A and five to receive drug B, and where the outcomes are the same for both groups, this could mean one of two things: either there is no difference between the drugs or one of the drugs is more effective but, because of the small numbers of participants, this difference between the drugs is not shown (Kirkwood 1988). A ‘power calculation’ will provide an estimation of the required sample size. The power of a study is the probability of obtaining a significant result if the difference between outcomes in two groups is ≥ the smallest worthwhile clinical difference specified.

- Demographic and health status details for the two groups are of interest. Significant differences between the two groups, for example differences in age, co-morbid conditions, gender or disease severity, could potentially affect the results of the study. The groups should ideally be similar, on average, for any variables that are likely to influence outcome. Similarity between groups is not always achieved by randomisation, even where the methods of randomisation are adequate.
Skills for evidence-based practice

- It is helpful if the intervention is described in sufficient detail to allow clinicians to reproduce it in their own setting. In addition, the primary outcome in which the investigators would expect to see a clinically important difference should be given, along with details of how the outcome was to be measured. Measurement instruments which have been validated outside the study and found to measure what they purport to measure, and which are sensitive, appropriate and acceptable, inspire more confidence than measurement instruments that have not been validated.

The quality assessment bottom line

This study met most of the quality criteria. The main exception is the lack of blinding; however, this is a common feature of studies of non-pharmacological interventions as it can be difficult or impossible to conceal from the patient and the clinician what is being done. It is often possible to blind the individual who is measuring or assessing the outcome – in this study, the laboratory staff and the ultrasonographer. There was, however, a risk that passengers might have revealed to the ultrasonographer which group they were in. If this was the case, one could argue that the ultrasonographer may (consciously or subconsciously) have been more vigilant when examining non-stocking passengers, and this could potentially impact on the study results.

Overall, we can conclude that the difference between the outcomes in the two groups was unlikely to be subject to a high level of bias (for the criteria examined), which suggests that we can trust the results of this study. The applicability of the study and what the results mean for our patients are considered in Chapter 5.

Worked example 4.2: Assessing the quality of a study that asks a question about the performance or accuracy of a particular diagnostic test or method of assessment

Scenario

Thomas Davies, a 45-year-old man, has been referred to the practice nurses by his general practitioner who is ‘fed up’ with seeing him when
Is the study good enough to use the findings?

there is ‘nothing wrong with him’. You notice from his records that Mr Davies has been registered with the practice for 5 years. He hardly ever visited the practice until the last 3 months and since then he seems to have visited nearly every week. During your interview Mr Davies tells you that he lost his job as a supervisor in a warehouse 6 months ago and has been unable to find work since. When you question him further, he tells you that he has never had any mental health problems but he says that during the last month he has felt really down. He also feels that he has little interest in anything, so much so that some days he just sits in his armchair all day unless his wife ‘nags him, to get out from under her feet’. From the information Mr Davies provides, you think he might have clinical depression. You wonder how accurately a patient’s self-reported loss of interest and ‘feeling down’ diagnoses clinical depression.

Your clinical practice question

In patients with clinical depression, how accurately does patient response to questions about feeling down or depressed and loss of interest diagnose clinical depression?

Finding the evidence

You phone the local hospital library and give them your question to run a systematic search for you. They phone you a few days later and tell you that they have not found any summarised evidence, but the following article may be useful:


The paper reports the results of a diagnostic accuracy study of 1025 people attending a primary care setting in New Zealand. The study assesses the accuracy of one type of assessment instrument (asking two questions about feeling down and about loss of interest or pleasure in doing things during the past month, and a question as to whether help was needed) compared with a different and longer assessment instrument (the mood module of the Composite International Diagnostic Interview [CIDI]) in diagnosing clinical depression.
What is the quality of the study?

Table 4.3 shows a worked example of the critical appraisal for the paper referenced above, using the appraisal tool ‘Twelve questions to help you make sense of a diagnostic test study’, published by CASP (1998).

Relevance of criteria in Table 4.3

- The ‘new’ test should be compared against the method that is currently regarded as ‘the best’ (i.e. the reference standard) and both tests should be applied in all participants.
- An appropriate spectrum of patients (i.e. patients with mild, moderate and severe forms of the condition) should ideally be included in the study, with details of the proportions of each of these groups given. A test may be able to identify people who are severely ill, but not those with a mild form of the condition.
- Ideally, a consecutive set of participants who fulfil the inclusion criteria should be tested. This ensures that individuals are not inappropriately ‘selected out’ of the study, thereby affecting the results and conclusions of the study.
- It is recommended that the clinician or investigator is ‘blinded’ to the results of the test that is carried out first. If the clinician suspects from the initial test that the patient does not have the disease in question, he or she may decide to avoid subjecting the patient to the second test. Blinding also avoids the conscious and unconscious bias of causing the reference standard to be overinterpreted when the diagnostic test is positive and underinterpreted when it is negative (Sackett et al 2000).
- Reliability (reproducibility) of a test needs to be considered: the results of tests carried out by different individuals or by the same individual at different times should remain unchanged, provided the true underlying variable being measured remains the same. Disagreement between two examiners is called interobserver variability, and disagreement within one examiner over time is called intraobserver variability. In tests that are not reproducible, it is difficult to know whether a true measurement is being obtained. This creates problems in research and in clinical practice.

The quality assessment bottom line

This study appears to meet the quality assessment criteria, therefore we can trust the results. The applicability of the study and what the results mean for our patients are considered in Chapter 5.
**TABLE 4.3 ASSESSING THE QUALITY OF A STUDY THAT INVESTIGATES THE PERFORMANCE OR ACCURACY OF A PARTICULAR TEST/METHOD OF ASSESSMENT: A WORKED EXAMPLE USING THE CASP CHECKLIST ‘TWELVE QUESTIONS TO HELP YOU MAKE SENSE OF A DIAGNOSTIC TEST STUDY’ (CAS 1998, WITH PERMISSION, © MILTON KEYNES PRIMARY CARE TRUST, 1998)**

<table>
<thead>
<tr>
<th>Type of question</th>
<th>Assessment criteria</th>
<th>Response (Yes / Can’t tell / No)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening question</td>
<td>Was there a clear question for this study to address?</td>
<td>Yes</td>
<td>This study sought to determine the accuracy of an assessment instrument (two written screening questions for depression with the addition of a question asking about whether help is needed) in diagnosing depression.</td>
</tr>
<tr>
<td></td>
<td>Was there a comparison with an appropriate reference– standard?</td>
<td>Yes</td>
<td>Responses to two screening questions and a question inquiring about help required were compared to a validated instrument, the CIDI.</td>
</tr>
<tr>
<td>Detailed questions</td>
<td>Did all patients get the diagnostic test and the reference standard?</td>
<td>Yes</td>
<td>It appears as if all participants entering the study were invited to complete both the CIDI and the other assessment tool.</td>
</tr>
<tr>
<td></td>
<td>Could the results of the test of interest have been influenced by the results of the reference standard?</td>
<td>No</td>
<td>The tests were performed independently. The research assistant did not look at the results of the screening questions until after the CIDI was completed and therefore would not have known if the participant had been diagnosed as depressed prior to administering the reference standard.</td>
</tr>
</tbody>
</table>

*(Continued)*
<table>
<thead>
<tr>
<th>Type of question</th>
<th>Assessment criteria</th>
<th>Response (Yes / Can't tell / No)</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is the disease status of the tested population clearly described?</td>
<td>No</td>
<td>Very limited information is provided about the participants, other than they were attending a general practitioner clinic and were receiving no psychotropic drugs.</td>
<td></td>
</tr>
<tr>
<td>Were the methods for performing the test described in sufficient detail?</td>
<td>No</td>
<td>Very little information is provided about the actual protocol followed.</td>
<td></td>
</tr>
<tr>
<td>What are the results?</td>
<td></td>
<td>When compared with the CIDI, the two screening questions with the help question had a sensitivity of 96% and a specificity of 89%. Overall, the two screening questions with the help question had a good sensitivity and an excellent specificity for screening depression.</td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>Response</td>
<td>Explanation</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>How sure are we about these results?</td>
<td>Can't tell</td>
<td>Confidence intervals are included, although they are described for each group rather than between groups.</td>
<td></td>
</tr>
<tr>
<td>Can the results be applied to your patient or population of interest?</td>
<td>Can't tell</td>
<td>Difficult to say as so little information is provided about the population.</td>
<td></td>
</tr>
<tr>
<td>Can the test be applied to your patient or population of interest?</td>
<td>No</td>
<td>No intra- or interobserver reliability established for the test, therefore variability of the assessment remains unknown.</td>
<td></td>
</tr>
<tr>
<td>Were all outcomes important to the individual population or considered?</td>
<td>Can't tell</td>
<td>Although the questions are reported to have good sensitivity and specificity for major depression, limited description of the population does little to establish how far the tests were evaluated in an appropriate spectrum of patients.</td>
<td></td>
</tr>
<tr>
<td>What would be the impact of using this test on your patients/population?</td>
<td>Brevity of the questionnaire may be useful to consider in clinical practice; further testing is required.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Worked example 4.3: Assessing the quality of a study that asks a question about the likely pattern and/or outcome of a particular health problem/disease

Scenario

Florence Barrett, a 33-year-old mother of two, has stopped by the doctor’s surgery on her way to work to collect a repeat prescription for oral contraception. You look in her case notes and notice that she has not had a pill review check this year and you invite her to return for a review the following week. You also ask her whether she is still smoking and she tells you that she smokes between 25 and 30 cigarettes per day. You note from her records that she has been using oral contraception for just over 1 year since the birth of her second child. You recall reading something about there being a higher risk of myocardial infarction (MI) in pill users who smoke and decide that before her visit you will investigate the following clinical question.

Your clinical question

Are women who smoke and take oral contraception at higher risk of MI than women who smoke but use other forms of contraception?

Finding the evidence

You begin a search on PubMed (MEDLINE via the internet) using the ‘clinical queries’ function and, selecting the prognosis filter, enter the terms ‘myocardial infarction AND oral contraceptives’. Twenty citations are identified, one of which is:

Is the study good enough to use the findings?

This is a review article but does not meet the criteria of a systematic review. Neither does it provide raw data. One of the references cited is:


The paper reports the results of a cohort study involving 17,032 women using contraception who were followed up for 20–26 years. One of the outcome measures used is the rate of MI. This is compared in women who had used oral contraception and those that had not. Subgroup analysis compares the rate of MI in women who smoke and use oral contraception and women who smoke and do not use oral contraception.

What is the quality of the study?

Table 4.4 shows a worked example of the critical appraisal for the paper referenced above that uses the quality criteria from the JAMA guidelines on critical appraisal (Laupacis et al 1994). A brief explanation of why the criteria are important is provided below, but readers are referred to the JAMA guidelines for a more detailed discussion.

Relevance of criteria in Table 4.4

- It is important that the participants in the study truly have the disorder of interest, and that they are entering the study at a common point in the course of their disease. In a study aiming to identify the risk of renal disease in people with diabetes, for example, if some of the participants already have undiagnosed mild kidney damage at the start of the study, this could influence the results in a negative way.

- Length of follow-up should be adequate for all possible outcomes (especially negative outcomes) to become manifest. If participants, for example people who have smoked 20 or more cigarettes a day for 1 year, are followed up for 4 years to establish risk of lung cancer, the conclusions of the study are likely to be different than if followed up for 15 years.

- People are inevitably lost to follow-up and the reasons for this should be explored. If participants are lost to follow-up through death rather than because they feel better, and this information is known to the researcher, he or she can take this into consideration when presenting the results.
### TABLE 4.4 ASSESSING THE QUALITY OF A STUDY ABOUT THE LIKELY PATTERN OR OUTCOME OF A PARTICULAR HEALTH PROBLEM OR DISEASE

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Quality appraisal question</th>
<th>Worked example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design</td>
<td>Was a group of participants followed up prospectively over a period of time?</td>
<td>Yes. A cohort of 17,032 women who used contraception were recruited when they were between the ages of 25 and 39 years and followed up for between 20 and 26 years.</td>
</tr>
<tr>
<td>Sample</td>
<td>Was a defined, representative sample of patients assembled at a common (usually early) point in the course of their life and/or disease?</td>
<td>Yes. 17,032 British, Caucasian, married women, who used contraception and were aged between 25 and 35, were recruited from 17 large family planning clinics in England and Scotland.</td>
</tr>
<tr>
<td>Measures</td>
<td>Were appropriate endpoints specified in advance of the study?</td>
<td>Yes. The outcome measures used were occurrence of angina, myocardial infarction or stroke that was associated with hospital admission, referral to hospital or death.</td>
</tr>
<tr>
<td>Researcher</td>
<td>What is the researcher/investigator’s relation to those being investigated?</td>
<td>The study investigators were academic staff at the University of Oxford. The study was not funded by any manufacturers of contraception.</td>
</tr>
<tr>
<td>Data collection</td>
<td>Follow-up of adequate duration.</td>
<td>Were the patients followed up for long enough?</td>
</tr>
<tr>
<td>All cases accounted for at end of study.</td>
<td>Were endpoints given for all participants?</td>
<td>No. End measurements were not obtained in women who dropped out of the study. Reasons for their withdrawal are not always given.</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Effects of potential confounders accounted for.</td>
<td>Does the data analysis take into consideration the effect of any potentially confounding variables?</td>
</tr>
</tbody>
</table>

NB: These questions are specifically for the critical appraisal of cohort studies.
Outcomes measurement can be a source of bias, especially where the outcome is a subjective one, for example quality of life. The outcome should therefore be clearly defined in advance. In addition, where outcome measurement requires a degree of judgement, the person taking the measurement or doing the assessment should be blind to the patient's condition.

When observing health outcomes over time, it is important to take account of the factors or variables that can affect health. In longitudinal studies, time itself acts as a confounding variable. As people get older they develop more illness regardless of any other factors. The effect of such confounding variables can be taken into account in the process of data analysis.

The quality assessment bottom line

The only area in which the quality of the study is questionable is in respect of the number of women who were lost to follow-up. The loss of patients is often unavoidable in large-scale longitudinal studies such as this one. From our point of view, we are concerned about the effect this loss of patients may have on the validity of the conclusions made in the study. In all, 1740 women were lost to follow-up, which is approximately 10% of the original sample. Is this too many? One way to decide is to use the ‘5 and 20’ rule. Fewer than 5% loss probably leads to little bias, greater than 20% loss seriously threatens validity (Sackett et al 2000).

A second approach is to ask a series of ‘what if?’ questions, known as sensitivity analysis. We can use different combinations of ‘what if’ scenarios to examine the possible effects of adding the missing cases back into the data analysis. The key thing here is to consider whether the cases lost to follow-up would have a different pattern of outcomes from those who remained. For our particular clinical question, we would ask ourselves the question of whether the heavy smoking, oral contraception users who dropped out were any more likely to have an adverse outcome than those who remained. The results of the study and its applicability to Mrs Barrett are considered in more detail in Chapter 5.

Critical appraisal in practice

The critical appraisal tools given above are designed to help develop skills and knowledge of critical appraisal. Other tools produced by different organisations are available (see the ‘Netting the Evidence’ website for more information: www.shef.ac.uk/scharr/it/netting/).
Although initially it may feel otherwise, critical appraisal gets easier with practice and people quickly become adept at recognising whether they will be able to use a paper or not. Answers to the questions outlined in Tables 4.2–4.4 are often included in the methodology sections. If not, there is a high probability they are not there at all and so it may not be worth bothering to look at the rest of the paper.

It is important to remember that research is a real-world process. This means that researchers are often forced to compromise on certain aspects of the research and to modify the study design for lots of legitimate reasons. There is no such thing as a perfect research study. Similarly, it is also important to recognise that research is done on samples of a population who will never be identical to patients attending a clinic, the patients in a ward, etc. It is unrealistic to search for perfect research studies where the study population exactly matches a patient group. People who are new to critical appraisal may feel that no research is good enough and quickly judge a paper as not being relevant to a particular population. The trick is to identify studies that can be applied to a specific clinical context where the design is good enough for the results to be trusted.

The tools in this chapter contain the basic questions necessary for assessing the quality of research studies in the practice environment. More sophisticated critical appraisal tools and techniques are available for people who are becoming more experienced or are doing critical appraisal on questions not covered in this chapter (for example, those published as a series in the Journal of the American Medical Association, the so-called JAMA user guides). Details of where to find these can also be found at the ‘Netting the Evidence’ website.

If the study is good enough to use and can be applied in the clinical setting of interest, the next step is to work out what to do about it. If a change in practice is required, how should that change be brought about? Simply telling colleagues that the evidence says they should be doing B rather than A has been demonstrated to be a rather ineffective change method (Mulhall 1999). Methods for increasing the chance of successfully changing practice are discussed in Chapters 9–11.

Presenting the results of critical appraisal to colleagues in a systematic way, which makes explicit the process used to come to these conclusions, is an important part of preparing for change (Melnyk 2005). The critical appraisal tools used above provide a method for generating a summary of a study. Another way of doing this is to use CATs or critically appraised...
Skills for evidence-based practice

topics, which are one-page summaries of research papers. These were developed by the Centre for Evidence-Based Medicine in Oxford and examples of completed CATs can be accessed via their website (cebm.net/cats.asp). They also produce CATmaker software, which is a computer program that can be used to make CATs. Raw data from the study can be entered into the program which will calculate some useful numbers for summarising the results in terms of effects on individual patients, e.g. the number needed to treat (NNT). This is discussed in more detail in Chapter 5.

Summary

This chapter has established why critical appraisal is necessary and important for evidence-based practice. The principles of research design and method underpinning the use of quality criteria to assess studies have been outlined. Critical appraisal for evidence-based practice is a three-part process comprising assessment of the quality of the study, assessment of the applicability of the study and interpretation of the study results for the individual patient. This chapter has focused on the assessment of the quality of a study and provided practical examples of how this can be done for three different types of clinical question that require the use of different research designs.

Scenario 4.1 leads to a clinical question on the effectiveness of elasticated stockings to influence frequency and prevention of DVT after flying long-haul economy class. The most appropriate research design for this question (in the absence of a systematic review) is a randomised controlled trial. Table 4.2 gives an example of the assessment of the quality of a trial investigating this topic.

Scenario 4.2 leads to a clinical question about whether patients’ self-reported feelings and a request for help predict clinical depression. The most appropriate study design for this question is a study investigating diagnostic accuracy. Table 4.3 gives an example of the assessment of the quality of a published diagnostic study on this topic.

Scenario 4.3 generates a clinical question about whether users of oral contraception are at greater risk of MI. The most appropriate study design for this question is a cohort study. Table 4.4 gives an example of the assessment of quality of a published cohort study that addresses this question.
Is the study good enough to use the findings?

Acknowledgements

We give special thanks to Mark Newman and Tony Roberts whose contributions to the previous edition have provided the foundation for this chapter.

References

* Papers used for critical appraisal examples.


Burns N, Grove S K 2001 The practice of nursing research: conduct, critique and utilization, 4th edn. W B Saunders, Philadelphia


Skills for evidence-based practice


Oxman A D, Sackett D L, Guyatt G H and the Evidence-Based Medicine Working Group 1993 How to get started. Based on the Users’ Guides to
Evidence-Based Medicine and reproduced with permission from Journal of the American Medical Association 270(17):2093–2095. Available online at: www.cche.net/usersguides/start.asp 29 Jan 2006


Further reading

Internet resources

There are numerous websites that contain materials for critical appraisal. However, these sites often change their website address and new sites are opening up all the time. We have therefore provided you with the address for the most comprehensive and up-to-date information about evidence-based practice on the web, ‘Netting the Evidence’. This site was established and is maintained by Andrew Booth at Sheffield University. www.shef.ac.uk/~scharr/ir/netting/

Books


Skills for evidence-based practice

