Supported by sanofi-aventis

What is critical appraisal?

Amanda Burls MBBS BA MSc FFPH Director of the Critical Appraisal Skills Programme, Director of Postgraduate Programmes in Evidence-Based Health Care, University of Oxford

- **Critical appraisal** is the process of carefully and systematically examining research to judge its trustworthiness, and its value and relevance in a particular context.
- The **Critical Appraisal Skills Programme** aims to help people develop the necessary skills to make sense of scientific evidence, and has produced appraisal checklists covering **validity, results** and **relevance.**
- Different research questions require different study designs. The best design for studies evaluating the effectiveness of an intervention or treatment is a **randomised controlled trial.**
- Studies are also subject to **bias** and it is important that researchers take steps to minimise this bias; for example, use of a **control group, randomisation** and **blinding.**
- **Odds ratios, risk ratios** and **number needed to treat** are methods of analysing results in order to determine if an intervention is effective.
- **Systematic reviews,** which collect, appraise and combine evidence, should be used when available.

For further titles in the series, visit:
www.whatisseries.co.uk

Date of preparation: February 2009

NPR09/1113

What is critical appraisal?

Critical appraisal is the process of carefully and systematically examining research to judge its trustworthiness, and its value and relevance in a particular context. It is an essential skill for evidence-based medicine because it allows clinicians to find and use research evidence reliably and efficiently (see *What is evidence-based medicine?*[1] for further discussion).

All of us would like to enjoy the best possible health we can. To achieve this we need reliable information about what might harm or help us when we make healthcare decisions. Research involves gathering data, then collating and analysing it to produce meaningful information. However, not all research is good quality and many studies are biased and their results untrue. This can lead us to draw false conclusions.

So, how can we tell whether a piece of research has been done properly and that the information it reports is reliable and trustworthy? How can we decide what to believe when research on the same topic comes to contradictory conclusions? This is where critical appraisal helps.

If healthcare professionals and patients are going to make the best decisions they need to be able to:

- Decide whether studies have been undertaken in a way that makes their findings reliable
- Make sense of the results
- Know what these results mean in the context of the decision they are making.

What makes studies reliable?

'Clinical tests have shown...'

Everyday we meet statements that try to influence our decisions and choices by claiming that research has demonstrated that something is useful or effective. Before we believe such claims we need to be sure that the study was not undertaken in a way such that it was likely to produce the result observed regardless of the truth.

Imagine for a moment that you are the maker of the beauty product 'EverYoung' and you want to advertise it by citing research suggesting that it makes people look younger; for example, 'nine out of every ten woman we asked agreed that "EverYoung" makes their skin firmer and younger looking.'

You want to avoid making a claim that is not based on a study because this could backfire should it come to light. Which of the following two designs would you choose if you wanted to maximise the probability of getting the result you want?

A. Ask women in shops who are buying 'EverYoung' whether they agree that it makes their skin firmer and younger looking?

B. Ask a random sample of women to try 'EverYoung' and then comment on whether they agree it made their skin firmer and younger looking?

Study A will tend to select women who are already likely to believe that the product works (otherwise they would not be parting with good money to buy it). This design thus increases the chance of a woman being surveyed agreeing with your statement. Such a study could find that nine out of ten women agreed with the statement even when study B shows that nine out of ten women who try the product do not believe it helps. Conducting a study in a way that tends to lead to a particular conclusion, regardless of the truth, is known as **bias.** Bias can be defined as 'the systematic deviation of the results of a study from the truth because of the way it has been conducted, analysed or reported'. Key sources of bias are shown in Table 1,[2] while further discussion can be found on the CONSORT Statement website.[3]

When critically appraising research, it is important to first look for biases in the study; that is, whether the findings of the study might be due to the way the study was designed and carried out, rather than reflecting the truth. It is also important to remember that no study is perfect and free from bias; it is therefore necessary to

Table 1. Key sources of bias in clinical trials[2]

Selection bias	Biased allocation to comparison groups
Performance bias	Unequal provision of care apart from treatment under evaluation
Detection bias	Biased assessment of outcome
Attrition bias	Biased occurrence and handling of deviations from protocol and loss to follow up

systematically check that the researchers have done all they can to minimise bias, and that any biases that might remain are not likely to be so large as to be able to account for the results observed. A study which is sufficiently free from bias is said to have **internal validity.**

Different types of question require different study designs

There are many sorts of questions that research can address.

- **Aetiology:** what caused this illness?
- **Diagnosis:** what does this test result mean in this patient?
- **Prognosis:** what is likely to happen to this patient?
- **Harm:** is having been exposed to this substance likely to do harm, and, if so, what?
- **Effectiveness:** is this treatment likely to help patients with this illness?
- **Qualitative:** what are the outcomes that are most important to patients with this condition?

Different questions require different study designs. To find out what living with a condition is like, a **qualitative study** that explores the subjective meanings and experiences is required. In contrast, a qualitative study relying only on the subjective beliefs of individuals could be misleading when trying to establish whether an intervention or treatment works. The best design for effectiveness studies is the **randomised controlled trial (RCT),** discussed below. A hierarchy of evidence exists, by which different methods of collecting evidence are graded as to their relative levels of validity.[4] When testing a particular treatment, subjective anecdotal reports of benefit can be misleading and qualitative studies are therefore not appropriate. An extreme example was the fashion for drinking Radithor® a century ago. The death of one keen proponent, Eben Byer, led to the 1932 *Wall Street Journal* headline, 'The Radium Water Worked Fine until His Jaw Came Off.'[5]

A **cross-sectional survey** is a useful design to determine how frequent a particular condition is. However, when determining an accurate prognosis for someone diagnosed with, say, cancer, a cross-sectional survey (that observes people who have the disease and describes their condition) can give a biased result. This is because by selecting people who are alive, a cross-sectional survey systematically selects a group with a better prognosis than average because it ignores those who died. The design needed for a prognosis question is an **inception cohort** – a study that follows up a recently diagnosed patient and records what happens to them.

It is important to recognise that different questions require different study designs for critical appraisal; first, because you need to choose a paper with the right type of study design for the question that you are seeking to answer and, second, because different study designs are prone to different biases. Thus, when critically appraising a piece of research it is important to first ask: did the researchers use the right sort of study design for their question? It is then necessary to check that the researchers tried to minimise the biases (that is, threats to internal validity) associated with any particular study design; these differ between studies.

The Critical Appraisal Skills Programme (CASP) aims to help people develop the skills they need to make sense of scientific evidence. CASP has produced simple critical appraisal checklists for the key study designs. These are not meant to replace considered thought and judgement when reading a paper but are for use as a guide and *aide memoire*.

All CASP checklists cover three main areas: validity, results and clinical relevance. The validity questions vary according to the type of study being appraised, and provide a method to check that the biases to which that particular study design is prone have been minimised. (The first two questions of each checklist are screening questions. If it is not possible to answer 'yes' to these questions, the paper is unlikely to be helpful and, rather than read on, you should try and find a better paper.)[6]

Effectiveness studies – the randomised controlled trial

Validity

'The art of medicine consists in amusing the patient while nature cures the disease.' – Voltaire

The fact that many illnesses tend to get better on their own is one of the challenges researchers face when trying to establish whether a treatment – be it a drug, device or surgical procedure – is truly effective. If an intervention is tested by giving it to a patient (such an experiment is known as a **trial),** and it is shown that the patient improves, it is often unclear whether this is because the intervention worked or because the patient would have got better anyway. This is a well-known problem when testing treatments and researchers avoid this bias by comparing how well patients given the intervention perform with how well patients not given the intervention perform (a **control group).** Trials in which there is a comparison group not given the intervention being tested are known as **controlled** trials.

It is important that the intervention and control groups are similar in all respects apart from receiving the treatment being tested. Otherwise we cannot be sure that any difference in outcome at the end is not due to pre-existing differences. If one group has a significantly different average age or social class make-up, this might be an explanation of why that group did better or worse. Most of the validity questions on the CASP RCT checklist are concerned with whether the researchers have avoided those things we know can lead to differences between the groups.

The best method to create two groups that are similar in all important respects is by deciding entirely by chance into which group a patient will be assigned. This is known as **randomisation.** In true randomisation all patients have the same chance as each other of being placed into any of the groups.

If researchers are able predict which group the next patient enrolled into the trial will be in, it can influence their decision whether to enter the patient into the trial or not. This can subvert the randomisation and produce two unequal groups. Thus, it is important that allocation is concealed from researchers.

Sometimes even randomisation can produce unequal groups, so another CASP question asks whether baseline characteristics of the group were comparable.

Even when the groups are similar at the start, researchers need to ensure that they do not begin to differ for reasons other than the intervention. To prevent patients' expectations influencing the results they should be **blinded,** where possible, as to which treatment they are receiving; for example, by using a placebo. Blinding of staff also helps stop the groups being treated differently and blinding of researchers stops the groups having their outcomes assessed differently.

It is also important to monitor the dropout rate, or treatment withdrawals, from the trial, as well as the number of patients lost to follow-up, to ensure that the composition of groups does not become different. In addition, patients should be analysed in the group to which they were allocated even if they did not receive the treatment they were assigned to (intention-to-treat analysis). Further discussion can be found on the CONSORT Statement website.[3]

Table 2. How to calculate odds ratios, risk r

	Number of patients	Number of events (eg cured)
Intervention	1,000	150
Control	1,000	100

These potential biases are the subject of the validity questions of the RCT checklist. In the other checklists, the validity questions cover the biases to which each individual study design is prone. The checklists are available online.[6]

Results

It is only worth thinking about what the findings of a study mean if the study design and methods are valid.

Results are presented in many different ways. In RCTs, cohort studies and case-control studies, two groups are compared and the results are often expressed as a relative risk (for example, dividing the outcome in the intervention group by the outcome in the control group). If the outcome is measured as the odds of an event occurring (for example, being cured) in a group (those with the event / those without the event), then the relative risk is known as the **odds ratio (OR).** If it is the frequency with which an event occurs (those with the event / the total number in that group), then the relative risk is known as the **risk ratio (RR).** When there is no difference between the groups, the OR and the RR are 1. A relative risk (OR or RR) of more than 1 means that the outcome occurred more in the intervention group than the control group (if it is a desired outcome, such as stopping smoking, then the intervention worked; if the outcome is not desired, for example death, then the control group performed better). Similarly, if the OR or RR is less than 1, then the outcome occurred less frequently in the intervention group.[7]

Results are usually more helpful when they are presented as **risk differences.** In this case you subtract the proportion of events in the control group from that in the intervention group. The risk difference can also be presented as the number needed to treat (NNT). This is the number of people to whom the treatment would have to be given – rather than the control – to produce one *extra* outcome of interest.[8]

Table 2 gives an example of how to calculate these metrics.

There will always be some uncertainty about the true result because trials are only a sample of possible results. The confidence interval (CI) gives the range of where the truth might lie, given the findings of a study, for a given degree of certainty (usually 95% certainty). P-values report the probability of seeing a result such as the one obtained if there were no real effect. P-values can range from 0 (absolutely impossible) to 1 (absolutely certain). A p-value of less than 0.05 means that a result such as the one seen would occur by chance on less than 1 in 20 occasions. In this circumstance a result is described as **statistically significant.** This does not mean that it is necessarily important.

Clinical relevance

It is important to consider whether the study is applicable to the decision being made for a particular patient or population. Any important differences between the participants in the trial and the patient or population in question that might change the effectiveness of an intervention must be identified.

It is also important to think about whether the researchers considered all the important outcomes. It is no use establishing that patients had less pain but neglecting to observe that they could be dying more often simply because this outcome was not measured.

Many interventions and processes that are used in everyday clinical practice have potential benefits and adverse consequences

atios and number needed to treat

	Odds of cure	Odds ratio	Risk of cure (frequency)	Risk ratio	Risk difference	Number needed to treat
	150/850	$\frac{150/850}{100/900} = 1.59$	150/1,000	$\frac{150/1,000}{100/1,000} = 1.5$	(150/1,000) – (100/1,000) = 0.05 (= 5%)	1/0.05 (= 100/5) = 20
	100/900		100/1,000			

and it is important that these are weighed against each other judiciously. For example, if one patient has a major bleed for every five patients prevented from having a stroke when patients are given anticoagulants, then this intervention may be beneficial. However, if five patients have a major bleed for every stroke prevented, then the intervention may not be worthwhile. In both cases the treatment prevents strokes, but in the latter the likelihood of harm outweighs the benefit.

Costs are usually not reported in a trial but if a treatment is very expensive and only gives a small health gain, it may not be a good use of resources. Usually an economic evaluation is necessary to provide information on cost-effectiveness, but sometimes a 'back-of-the-envelope' calculation can be performed. If the cost of treating one patient and the NNT can be established, these values can be multiplied to give a rough idea of the likely order of cost for producing one unit of benefit.

Systematic reviews

Decisions are most beneficial when informed by a consideration of all the available evidence. Given the limited time available to decision-makers, systematic reviews – which collect, appraise and combine evidence – should be used when available. If possible, good quality, up-to-date systematic reviews should be used as opposed to an individual study.[4] The CASP checklist for appraising a systematic review is available online.[6]

Conclusions

When reading any research – be it a systematic review, RCT, economic evaluation or other study design – it is important to remember that there are three broad things to consider: validity, results, relevance. It is always necessary to consider the following questions.

- Has the research been conducted in such a way as to minimise bias?
- If so, what does the study show?
- What do the results mean for the particular patient or context in which a decision is being made?

References

1. Belsey J. *What is evidence-based medicine?* London: Hayward Medical Communications, 2009.
2. Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ* 2001; **323:** 42–46.
3. *The CONSORT Statement.* www.consort-statement.org/?o=1011 (last accessed 2 February 2009)
4. Hemingway P, Brereton N. *What is a systematic review?* London: Hayward Medical Communications, 2009.
5. *A Little Poison Can Be Good For You: The received wisdom about toxins and radiation may be all wet.* http://money.cnn.com/magazines/fortune/fortune_archive/2003/06/09/343948/index.htm (last accessed 22 January 2009)
6. *Critical Appraisal Skills Programme checklists.* www.phru.nhs.uk/Pages/PHD/resources.htm (last accessed 22 January 2009)
7. Moore A, McQuay H. Clinical trials. In: *Bandolier's little book of making sense of the medical evidence.* Oxford: Oxford University Press, 2006.
8. Moore A. *What is an NNT?* London: Hayward Medical Communications, 2009.

What is critical appraisal?

First edition published 2003
Authors: Alison Hill and Claire Spittlehouse.

Published by Hayward Medical Communications, a division of Hayward Group Ltd.

Supported by *sanofi-aventis*