

[How to read a paper: Assessing the methodological quality of published papers](#)
[Trisha Greenhalgh, senior lecturer^a](#)

Introduction

Before changing your practice in the light of a published research paper, you should decide whether the methods used were valid. This article considers five essential questions that should form the basis of your decision.

Question 1: Is the study original?

Only a tiny proportion of medical research breaks entirely new ground, and an equally tiny proportion repeats exactly the steps of previous workers. The vast majority of research studies will tell us, at best, that a particular hypothesis is slightly more or less likely to be correct than it was before we added our piece to the wider jigsaw. Hence, it may be perfectly valid to do a study which is, on the face of it, "unoriginal." Indeed, the whole science of meta-analysis depends on the literature containing more than one study that has addressed a question in much the same way.

The practical question to ask, then, about a new piece of research is not "Has anyone ever done a similar study?" but "Does this new research add to the literature in any way?" For example:

Is this study bigger, continued for longer, or otherwise more substantial than the previous one(s)?

Is the methodology of this study any more rigorous (in particular, does it address any specific methodological criticisms of previous studies)?

Will the numerical results of this study add significantly to a meta-analysis of previous studies?

Is the population that was studied different in any way (has the study looked at different ages, sex, or ethnic groups than previous studies)?

Is the clinical issue addressed of sufficient importance, and is there sufficient doubt in the minds of the public or key decision makers, to make new evidence "politically" desirable even when it is not strictly scientifically necessary?

Question 2: Whom is the study about?

Before assuming that the results of a paper are applicable to your own practice, ask yourself the following questions:

How were the subjects recruited? If you wanted to do a questionnaire survey of the views of users of the hospital casualty department, you could recruit respondents by advertising in the local newspaper. However, this method would be a good example of recruitment bias since the sample you obtain would be skewed in favour of users who were highly motivated and liked to read newspapers. You would, of course, be better to issue a questionnaire to every user (or to a 1 in 10 sample of users) who turned up on a particular day.

Who was included in the study? Many trials in Britain and North America routinely exclude patients with coexisting illness, those who do not speak English, those taking certain other medication, and those who are illiterate. This approach may be scientifically "clean," but since clinical trial results will be used to guide practice in relation to wider patient groups it is not necessarily logical.¹ The results of

pharmacokinetic studies of new drugs in 23 year old healthy male volunteers will clearly not be applicable to the average elderly woman.

Who was excluded from the study? For example, a randomised controlled trial may be restricted to patients with moderate or severe forms of a disease such as heart failure—a policy which could lead to false conclusions about the treatment of mild heart failure. This has important practical implications when clinical trials performed on hospital outpatients are used to dictate "best practice" in primary care, where the spectrum of disease is generally milder.

Were the subjects studied in "real life" circumstances? For example, were they admitted to hospital purely for observation? Did they receive lengthy and detailed explanations of the potential benefits of the intervention? Were they given the telephone number of a key research worker? Did the company that funded the research provide new equipment which would not be available to the ordinary clinician? These factors would not necessarily invalidate the study itself, but they may cast doubt on the applicability of its findings to your own practice.

Question 3: Was the design of the study sensible?

Although the terminology of research trial design can be forbidding, much of what is grandly termed "critical appraisal" is plain common sense. I usually start with two fundamental questions:

What specific intervention or other manoeuvre was being considered, and what was it being compared with? It is tempting to take published statements at face value, but remember that authors frequently misrepresent (usually subconsciously rather than deliberately) what they actually did, and they overestimate its originality and potential importance. The examples in the [box](#) use hypothetical statements, but they are all based on similar mistakes seen in print.

What outcome was measured, and how? If you had an incurable disease for which a pharmaceutical company claimed to have produced a new wonder drug, you would measure the efficacy of the drug in terms of whether it made you live longer (and, perhaps, whether life was worth living given your condition and any side effects of the medication). You would not be too interested in the levels of some obscure enzyme in your blood which the manufacturer assured you were a reliable indicator of your chances of survival. The use of such surrogate endpoints is discussed in a later article in this series.²

Examples of problematic descriptions in the methods section of a paper;		
What the authors said	What they should have said (or should have done)	An example of:
"We measured how often GPs ask patients whether they smoke."	"We looked in patients' medical records and counted how many had had their smoking status recorded."	Assumption that medical records are 100% accurate.
"We measured how doctors treat low back pain."	"We measured what doctors say they do when faced with a patient with low back pain."	Assumption that what doctors say they do reflects what they actually do.

"We compared a nicotine-replacement patch with placebo."	"Subjects in the intervention group were asked to apply a patch containing 15 mg nicotine twice daily; those in the control group received identical-looking patches."	Failure to state dose of drug or nature of placebo.
"We asked 100 teenagers to participate in our survey of sexual attitudes."	"We approached 147 white American teenagers aged 12-18 (85 males) at a summer camp; 100 of them (31 males) agreed to participate."	Failure to give sufficient information about subjects. (Note in this example the figures indicate a recruitment bias towards females.)
"We randomised patients to either 'individual care plan' or 'usual care'."	"The intervention group were offered an individual care plan consisting of ...; control patients were offered"	Failure to give sufficient information about intervention. (Enough information should be given to allow the study to be repeated by other workers.)
"To assess the value of an educational leaflet, we gave the intervention group a leaflet and a telephone helpline number. Controls received neither."	If the study is purely to assess the value of the leaflet, both groups should have been given the helpline number.	Failure to treat groups equally apart from the specific intervention.
"We measured the use of vitamin C in the prevention of the common cold."	A systematic literature search would have found numerous previous studies on this subject ¹⁴	Unoriginal study.

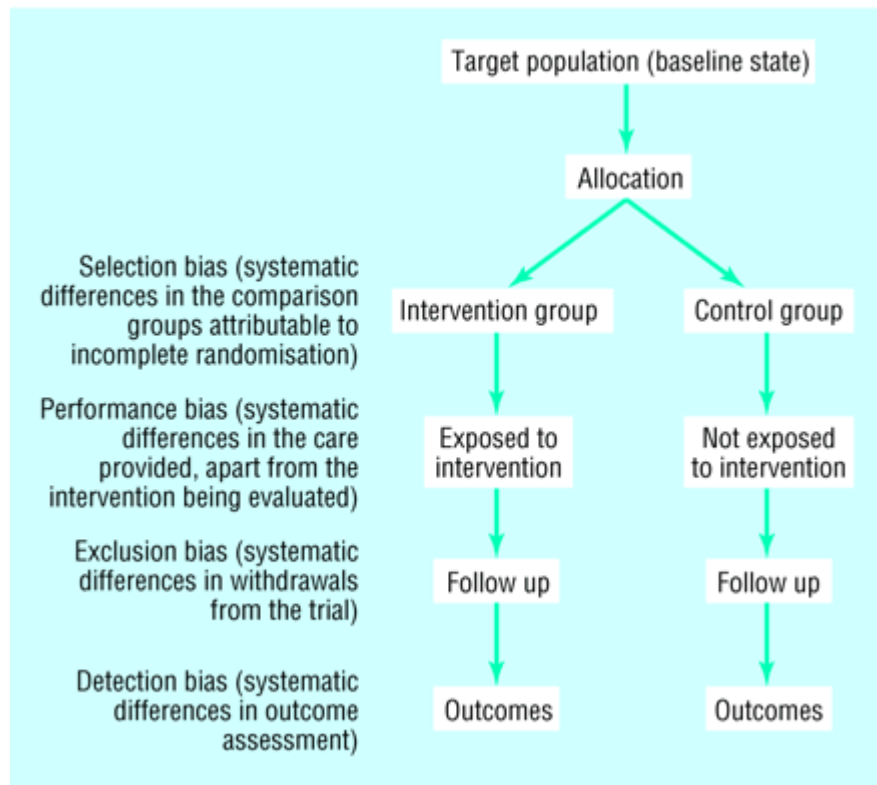
The measurement of symptomatic effects (such as pain), functional effects (mobility), psychological effects (anxiety), or social effects (inconvenience) of an intervention is fraught with even more problems. You should always look for evidence in the paper that the outcome measure has been objectively validated—that is, that someone has confirmed that the scale of anxiety, pain, and so on used in this study measures what it purports to measure, and that changes in this outcome measure adequately reflect changes in the status of the patient. Remember that what is important in the eyes of the doctor may not be valued so highly by the patient, and vice versa.³

Question 4: Was systematic bias avoided or minimised?

Systematic bias is defined as anything that erroneously influences the conclusions about groups and distorts comparisons.⁴ Whether the design of a study is a randomised controlled trial, a non-randomised comparative trial, a cohort study, or a case-control study, the aim should be for the groups being compared to be as similar as possible except for the particular difference being examined. They should, as far as possible, receive the same explanations, have the same contacts with health professionals, and be assessed the same number of times by using the same outcome measures. Different study designs call for different steps to reduce systematic bias:

Randomised controlled trials

In a randomised controlled trial, systematic bias is (in theory) avoided by selecting a sample of participants from a particular population and allocating them randomly to the different groups. Figure 1 summarises sources of bias to check for.



Non-randomised controlled clinical trials

I recently chaired a seminar in which a multidisciplinary group of students from the medical, nursing, pharmacy, and allied professions were presenting the results of several in house research studies. All but one of the studies presented were of comparative, but non-randomised, design—that is, one group of patients (say, hospital outpatients with asthma) had received one intervention (say, an educational leaflet) while another group (say, patients attending GP surgeries with asthma) had received another intervention (say, group educational sessions). I was surprised how many of the presenters believed that their study was, or was equivalent to, a randomised controlled trial. In other words, these commendably enthusiastic and committed young researchers were blind to the most obvious bias of all: they were comparing two groups which had inherent, self selected differences even before the intervention was applied (as well as having all the additional potential sources of bias of randomised controlled trials). As a general rule, if the paper you are looking at is a non-randomised controlled clinical trial, you must use your common sense to decide if the baseline differences between the intervention and control groups are likely to have been so great as to invalidate any differences ascribed to the effects of the intervention. This is, in fact, almost always the case.^{5 6}

Cohort studies

The selection of a comparable control group is one of the most difficult decisions facing the authors of an observational (cohort or case-control) study. Few, if any, cohort studies, for example, succeed in identifying two groups of subjects who are equal in age, sex mix, socioeconomic status, presence of coexisting illness, and so on, with the single difference being their exposure to the agent being studied. In practice, much of the "controlling" in cohort studies occurs at the analysis stage, where complex statistical adjustment is made for baseline differences in key variables. Unless this is done adequately, statistical tests of probability and confidence intervals will be dangerously misleading.⁷

This problem is illustrated by the various cohort studies on the risks and benefits of alcohol, which have consistently found a "J shaped" relation between alcohol intake and mortality. The best outcome (in terms of premature death) lies with the cohort who are moderate drinkers.⁸ The question of whether "teetotallers" (a group that includes people who have been ordered to give up alcohol on health grounds, health faddists, religious fundamentalists, and liars, as well as those who are in all other respects comparable with the group of moderate drinkers) have a genuinely increased risk of heart disease, or whether the J shape can be explained by confounding factors, has occupied epidemiologists for years.⁸

Case-control studies

In case-control studies (in which the experiences of individuals with and without a particular disease are analysed retrospectively to identify putative causative events), the process that is most open to bias is not the assessment of outcome, but the diagnosis of "caseness" and the decision as to when the individual became a case.

A good example of this occurred a few years ago when a legal action was brought against the manufacturers of the whooping cough (pertussis) vaccine, which was alleged to have caused neurological damage in a number of infants.⁹ In the court hearing, the judge ruled that misclassification of three brain damaged infants as "cases" rather than controls led to the overestimation of the harm attributable to whooping cough vaccine by a factor of three.⁹

Question 5: Was assessment "blind"?

Even the most rigorous attempt to achieve a comparable control group will be wasted effort if the people who assess outcome (for example, those who judge whether someone is still clinically in heart failure, or who say whether an x ray is "improved" from last time) know which group the patient they are assessing was allocated to. If, for example, I knew that a patient had been randomised to an active drug to lower blood pressure rather than to a placebo, I might be more likely to recheck a reading which was surprisingly high. This is an example of performance bias, which, along with other pitfalls for the unblinded assessor, is listed in figure 2.

Question 6: Were preliminary statistical questions dealt with?

Three important numbers can often be found in the methods section of a paper: the size of the sample; the duration of follow up; and the completeness of follow up.

Sample size

In the words of statistician Douglas Altman, a trial should be big enough to have a high chance of detecting, as statistically significant, a worthwhile effect if it exists, and thus to

be reasonably sure that no benefit exists if it is not found in the trial.¹⁰ To calculate sample size, the clinician must decide two things.

The first is what level of difference between the two groups would constitute a clinically significant effect. Note that this may not be the same as a statistically significant effect. You could administer a new drug which lowered blood pressure by around 10 mm Hg, and the effect would be a significant lowering of the chances of developing stroke (odds of less than 1 in 20 that the reduced incidence occurred by chance).¹¹ However, in some patients, this may correspond to a clinical reduction in risk of only 1 in 850 patient years¹²—a difference which many patients would classify as not worth the effort of taking the tablets. Secondly, the clinician must decide the mean and the standard deviation of the principal outcome variable.

Using a statistical nomogram,¹⁰ the authors can then, before the trial begins, work out how large a sample they will need in order to have a moderate, high, or very high chance of detecting a true difference between the groups—the power of the study. It is common for studies to stipulate a power of between 80% and 90%. Underpowered studies are ubiquitous, usually because the authors found it harder than they anticipated to recruit their subjects. Such studies typically lead to a type II or β error—the erroneous conclusion that an intervention has no effect. (In contrast, the rarer type I or α error is the conclusion that a difference is significant when in fact it is due to sampling error.)

Duration of follow up

Even if the sample size was adequate, a study must continue long enough for the effect of the intervention to be reflected in the outcome variable. A study looking at the effect of a new painkiller on the degree of postoperative pain may only need a follow up period of 48 hours. On the other hand, in a study of the effect of nutritional supplementation in the preschool years on final adult height, follow up should be measured in decades.

Completeness of follow up

Subjects who withdraw from ("drop out of") research studies are less likely to have taken their tablets as directed, more likely to have missed their interim checkups, and more likely to have experienced side effects when taking medication, than those who do not withdraw.¹³ The reasons why patients withdraw from clinical trials include the following:

Incorrect entry of patient into trial (that is, researcher discovers during the trial that the patient should not have been randomised in the first place because he or she did not fulfil the entry criteria);

Suspected adverse reaction to the trial drug. Note that the "adverse reaction" rate in the intervention group should always be compared with that in patients given placebo. Inert tablets bring people out in a rash surprisingly frequently;

Loss of patient motivation;

Withdrawal by clinician for clinical reasons (such as concurrent illness or pregnancy);

Loss to follow up (patient moves away, etc);

Death.

Simply ignoring everyone who has withdrawn from a clinical trial will bias the results, usually in favour of the intervention. It is, therefore, standard practice to analyse the results of comparative studies on an intention to treat basis.¹⁴ This means that all data on patients originally allocated to the intervention arm of the study—including those who withdrew before the trial finished, those who did not take their tablets, and even those who subsequently received the control intervention for whatever reason—should be analysed along with data on the patients who followed the protocol throughout.

Conversely, withdrawals from the placebo arm of the study should be analysed with those who faithfully took their placebo.

In a few situations, intention to treat analysis is not used. The most common is the efficacy analysis, which is to explain the effects of the intervention itself, and is therefore of the treatment actually received. But even if the subjects in an efficacy analysis are part of a randomised controlled trial, for the purposes of the analysis they effectively constitute a cohort study.

Summary points

The first essential question to ask about the methods section of a published paper is: was the study original?

The second is: whom is the study about?

Thirdly, was the design of the study sensible?

Fourthly, was systematic bias avoided or minimised?

Finally, was the study large enough, and continued for long enough, to make the results credible?