BMJ 1997;315:243-246 (26 July)

Education and debate

How to read a paper : getting your bearings (deciding what the paper is about)

Trisha Greenhalgh, *senior lecturer* [a]

[a] Unit for Evidence-Based Practice and Policy, Department of Primary Care and Population Sciences, University College London Medical School/Royal Free Hospital School of Medicine, Whittington Hospital, London N19 5NF
p.greenhalgh@ucl.ac.uk

The science of "trashing" papers

It usually comes as a surprise to students to learn that some (perhaps most) published articles belong in the bin, and should certainly not be used to inform practice.[1] The first box shows some common reasons why papers are rejected by peer reviewed journals.

Why were papers rejected for publication?

- The study did not address an important scientific issue
- The study was not original (someone else had already done the same or a similar study)
- The study did not actually test the authors' hypothesis
- A different type of study should have been done
- Practical difficulties (in recruiting subjects, for example) led the authors to compromise on the original study protocol
- The sample size was too small
- The study was uncontrolled or inadequately controlled
- The statistical analysis was incorrect or inappropriate
- The authors drew unjustified conclusions from their data
- There is a significant conflict of interest (one of the authors, or a sponsor, might benefit

financially from the publication of the paper and insufficient safeguards were seen to be in place to guard against bias)

- The paper is so badly written that it is incomprehensible

Most papers now appearing in medical journals are presented more or less in standard IMRAD format: Introduction (why the authors decided to do this research), Methods (how they did it, and how they analysed their results), Results (what they found), and Discussion (what the results mean). If you are deciding whether a paper is worth reading, you should do so on the design of the methods section and not on the interest of the hypothesis, the nature or potential impact of the results, or the speculation in the discussion.

Critical appraisal

The assessment of methodological quality (critical appraisal) has been covered in detail in many textbooks on evidence based medicine,[2 3 4 5 6] and in Sackett and colleagues' Users' Guides to the Medical Literature in *JAMA*.[7 8 9 10 11 12 13 14 15 16 17 18 19 20 21] If you are an experienced journal reader, the structured checklists produced by these authors will be largely self explanatory. If you are not, try these preliminary questions.

Question 1: Why was the study done, and what clinical question were the authors addressing?

The introductory sentence of a research paper should state, in a nutshell, what the background to the research is. For example, "Grommet insertion is a common procedure in children, and it has been suggested that not all operations are clinically necessary." This statement should be followed by a brief review of the published literature.

Unless it has already been covered in the introduction, the hypothesis which the authors have decided to test should be clearly stated in the methods section of the paper. If the hypothesis is presented in the negative, such as "the addition of metformin to maximal dose sulphonylurea therapy will not improve the control of type 2 diabetes," it is known as a null hypothesis.

Summary points

Many papers published in medical journals have potentially serious methodological flaws

When deciding whether a paper is valid and relevant to your practice, first establish what specific clinical question it addressed

Questions to do with drug treatment or other medical interventions should be addressed by double blind, randomised controlled trials

Questions about prognosis require longitudinal cohort studies, and those about causation require either cohort or case-control studies

Case reports, though methodologically weak, can be produced rapidly and have a place in alerting practitioners to adverse drug reactions

The authors of a study rarely actually believe their null hypothesis when they embark on their research. Being human, they have usually set out to show a difference between the two arms of their study. But the way scientists do this is to say, "Let's assume there's no difference; now let's try to disprove that theory." If you adhere to the teachings of Karl Popper, this hypotheticodeductive approach (setting up falsifiable hypotheses which you then proceed to test) is the very essence of the scientific method.[22]

Question 2: What type of study was done?

First, decide whether the paper describes a primary study, which reports research first hand, or a secondary (or integrative) one, which attempts to summarise and draw conclusions from primary studies. Primary studies, the stuff of most published research in medical journals, usually fall into one of three categories:

- Experiments, in which a manoeuvre is performed on an animal or a volunteer in artificial and controlled surroundings;

- Clinical trials, in which an intervention, such as a drug treatment, is offered to a group of patients who are then followed up to see what happens to them; or
- Surveys, in which something is measured in a group of patients, health professionals, or some other sample of individuals.

The second box shows some common jargon terms used in describing study design.

Terms used to describe design features of clinical research studies

*Parallel group comparison*Each group receives a different treatment, with both groups being entered at the same time; results are analysed by comparing groups

*Paired (or matched) comparison*Subjects receiving different treatments are matched to balance potential confounding variables such as age and sex; results are analysed in terms of differences between subject pairs

*Within subject comparison*Subjects are assessed before and after an intervention and results analysed in terms of changes within the subjects

*Single blind*Subjects did not know which treatment they were receiving

*Double blind*Neither did the investigators

*Crossover*Each subject received both the intervention and control treatments (in random order), often separated by a washout period with no treatment

*Placebo controlled*Control subjects receive a placebo (inactive pill) which should look and taste the same as the active pill. Placebo (sham) operations may also be used in trials of surgery

*Factorial design*A study which permits investigation of the effects (both separately and combined) of more than one independent variable on a given outcome (for example, a 2x2 factorial design tested the effects of placebo, aspirin alone, streptokinase alone, or aspirin plus streptokinase in acute heart attack[23])

Secondary research is made up of:

- Overviews, which may be divided into:

[Non-systematic] reviews, which summarise primary studies;

Systematic reviews, which do this according to a rigorous and predefined methodology; and

Meta-analyses, which integrate the numerical data from more than one study.

- Guidelines, which draw conclusions from primary studies about **how** clinicians should be behaving.
- Decision analyses, which use the results of primary studies to generate probability trees to be used by health professionals and patients in making choices about clinical management.[24 25 26]
- Economic analyses, which use the results of primary studies to say whether a particular course of action is a good use of resources.

Question 3: Was this design appropriate to the research?

This question is best addressed by considering what broad field of research is covered by the study. Most research studies are concerned with one or more of the broad fields shown in the box below.

Broad fields of research

- *Therapy:* testing the efficacy of drug treatments, surgical procedures, alternative methods of service delivery, or other interventions. Preferred study design is randomised controlled trial

- *Diagnosis:* demonstrating whether a new diagnostic test is valid (can we trust it?) and reliable (would we get the same results every time?). Preferred study design is cross sectional survey in which both the new test and the gold standard are performed
- *Screening:* demonstrating the value of tests which can be applied to large populations and which pick up disease at a presymptomatic stage. Preferred study design is cross sectional survey
- *Prognosis:* determining what is likely to happen to someone whose disease is picked up at an early stage. Preferred study design is longitudinal cohort study
- *Causation:* determining whether a putative harmful agent, such as environmental pollution, is related to the development of illness. Preferred study design is cohort or case-control study, depending on how rare the disease is, but case reports may also provide crucial information

Randomised controlled trials

In a randomised controlled trial, participants are randomly allocated by a process equivalent to the flip of a coin to either one intervention (such as a drug) or another (such as placebo treatment or a different drug). Both groups are followed up for a specified period and analysed in terms of outcomes defined at the outset (death, heart attack, serum cholesterol level, etc). Because, on average, the groups are identical apart from the intervention, any differences in outcome are, in theory, attributable to the intervention.

Some trials comparing an intervention group with a control group are not randomised trials. Random allocation may be impossible, impractical, or unethical—for example, in a trial to compare the outcomes of childbirth at home and in hospital. More commonly, inexperienced investigators compare one group (such as patients on ward A) with another (such as patients on ward B). With such designs, it is far less likely that the two groups can reasonably be compared with one another on a statistical level.

A randomised controlled trial should answer questions such as the following:

- Is this drug better than placebo or a different drug for a particular disease?
- Is a leaflet better than verbal advice in helping patients make informed choices about the treatment options for a particular condition?

It should be remembered, however, that randomised trials have several disadvantages (see box).[27] Remember, too, that the results of a trial may have limited applicability as a result of exclusion criteria (rules about who may not be entered into the study), inclusion bias (selection of subjects from a group unrepresentative of everyone with the condition), refusal of certain patient groups to give consent to be included in the trial,[28] analysis of only predefined "objective" endpoints which may exclude important qualitative aspects of the intervention, and publication bias (the selective publication of positive results).[29]

Randomised controlled trial design

Advantages

- Allows rigorous evaluation of a single variable (effect of drug treatment versus placebo, for example) in a precisely defined patient group (postmenopausal women aged 50-60 years)
- Prospective design (data are collected on events that happen after you decide to do the study)

- Uses hypotheticodeductive reasoning (seeks to falsify, rather than confirm, its own hypothesis)
- Potentially eradicates bias by comparing two otherwise identical groups (but see below)
- Allows for meta-analysis (combining the numerical results of several similar trials at a later date)

Disadvantages

- Expensive and time consuming; hence, in practice:
- Many randomised controlled trials are either never done, are performed on too few patients, or are undertaken for too short a period
- Most are funded by large research bodies (university or government sponsored) or drug companies, who ultimately dictate the research agenda
- Surrogate endpoints are often used in preference to clinical outcome measures may introduce "hidden bias," especially through:
- Imperfect randomisation (see above)
- Failure to randomise all eligible patients (clinician only offers participation in the trial to patients he or she considers will respond well to the intervention)
- Failure to blind assessors to randomisation status of patients

There is now a recommended format for reporting randomised controlled trials in medical journals.[30] You should try to follow it if you are writing one up yourself.

Cohort studies

In a cohort study, two (or more) groups of people are selected on the basis of differences in their exposure to a particular agent (such as a vaccine, a drug, or an environmental toxin), and followed up to see how many in each group develop a particular disease or other outcome. The follow up period in cohort studies is generally measured in years (and sometimes in decades), since that is how long many diseases, especially cancer, take to develop. Note that randomised controlled trials are usually begun on patients (people who already have a disease), whereas most cohort studies are begun on subjects who may or may not develop disease.

PETER BROWN

View larger version (145K):
[in this window]
[in a new window]

A special type of cohort study may also be used to determine the prognosis of a disease (what is likely to happen to someone who has it). A group of patients who have all been diagnosed as having an early stage of the disease or a positive result on a screening test is assembled (the inception cohort) and followed up on repeated occasions to see the incidence (new cases per year) and time course of different outcomes.

The world's most famous cohort study, which won its two original authors a knighthood, was undertaken by Sir Austin Bradford Hill, Sir Richard Doll, and, latterly, Richard Peto. They followed up 40 000 British doctors divided into four cohorts (non-smokers, and light, moderate, and heavy smokers) using both all cause mortality (any death) and cause specific mortality (death from a particular disease) as outcome measures. Publication of their 10 year interim results in 1964, which showed a substantial excess in both lung cancer mortality and all cause mortality in smokers, with a "dose-response" relation (the more you smoke, the worse your chances of getting lung cancer), went a long way to showing that the link between smoking and ill health was causal rather than coincidental.[31] The 20 year and 40 year results of this momentous study (which achieved an impressive 94% follow up of those recruited in 1951 and not known to have died) illustrate both the perils of smoking and the strength of evidence that can be obtained from a properly conducted cohort study.[32 33]

A cohort study should be used to address clinical questions such as:

- Does high blood pressure get better over time?
- What happens to infants who have been born very prematurely, in terms of subsequent physical development and educational achievement?

Case-control studies

In a case-control study, patients with a particular disease or condition are identified and "matched" with controls (patients with some other disease, the general population, neighbours, or relatives). Data are then collected (for example, by searching back through these people's medical records or by asking them to recall their own history) on past exposure to a possible causal agent for the disease. Like cohort studies, case-control studies are generally concerned with the aetiology of a disease (what causes it) rather than its treatment. They lie lower down the hierarchy of evidence (see below), but this design is usually the only option for studying rare conditions. An important source of difficulty (and potential bias) in a case-control study is the precise definition of who counts as a "case," since one misallocated subject may substantially influence the results. In addition, such a design cannot show causality—the association of A with B in a case-control study does not prove that A has caused B.

A case-control study should be used to address clinical questions such as:

- Does the prone sleeping position increase the risk of cot death (the sudden infant death syndrome)?
- Does whooping cough vaccine cause brain damage?
- Do overhead power cables cause leukaemia?

Cross sectional surveys

We have probably all been asked to take part in a survey, even if only one asking us which brand of toothpaste we prefer. Surveys conducted by epidemiologists are run along the same lines: a representative sample of subjects (or patients) is interviewed, examined, or otherwise studied to gain answers to a specific clinical question. In cross sectional surveys, data are collected at a single time but may refer retrospectively to experiences in the past—such as the study of casenotes to see how often patients' blood pressure has been recorded in the past five years.

A cross sectional survey should be used to address clinical questions such as:

- What is the "normal" height of a 3 year old child?

- What do psychiatric nurses believe about the value of electroconvulsive therapy in severe depression?
- Is it true that half of all cases of diabetes are undiagnosed?

A memorable example of **a** case report

A doctor notices that two newborn babies in his hospital have absent limbs (phocomelia). Both mothers had taken **a** new drug (thalidomide) in early pregnancy. The doctor wishes **to** alert his colleagues worldwide **to** the possibility of drug related damage as quickly as possible.[35]

Case reports

- Top
- The science of "trashing"...
- Critical appraisal
- Randomised controlled trials
- Cohort studies
- Case-control studies
- Cross sectional surveys
- Case reports
- The hierarchy of evidence
- References

A case report describes the medical history of **a** single patient in the form of **a** story: "Mrs B is **a** 54 year old secretary who developed chest pain in June 1995...." Case reports are often run together **to** form **a** case series, in which the medical histories of more than one patient with **a** particular condition are described **to** illustrate an aspect of the condition, the treatment, or, most commonly these days, adverse reaction **to** treatment. Although this type of research is traditionally considered **to** be "quick and dirty" evidence, **a** great deal of information can be conveyed in **a** case report that would be lost in **a** clinical trial or survey .[34]

The hierarchy of evidence

- Top
- The science of "trashing"...
- Critical appraisal
- Randomised controlled trials
- Cohort studies
- Case-control studies
- Cross sectional surveys
- Case reports
- The hierarchy of evidence
- References

Standard notation for the relative weight carried by the different types of primary study when making decisions about clinical interventions (the "hierarchy of evidence") puts them in the following order[36]:

1. Systematic reviews and meta-analyses
2. Randomised controlled trials with definitive results (confidence intervals that do not overlap the threshold clinically significant effect)
3. Randomised controlled trials with non-definitive results (a point estimate that suggests a clinically significant effect but with confidence intervals overlapping the threshold for this effect)
4. Cohort studies
5. Case-control studies
6. Cross sectional surveys
7. Case reports.

The articles in this series are excerpts from ***How to read a paper****: the basics of evidence based medicine*. The book includes chapters on searching the literature and implementing evidence based findings. It can be ordered from the BMJ Bookshop: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

References

- Top
- The science of "trashing"...

1. Altman DG. The scandal of poor medical research. *BMJ* 1994;308:283-4. [Free Full Text]

2. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine.* London, Little, Brown, 1991.
3. Sackett DL, Richardson WS, Rosenberg WMC, Haynes RB. *Evidence-based medicine: how to practice and teach EBM.* London: Churchill-Livingstone, 1996.
4. Crombie IM. *The pocket guide to critical appraisal.* London: BMJ Publishing Group, 1996.
5. Fletcher RH, Fletcher SW, Wagner EH. *Clinical epidemiology: the essentials.* 3rd ed. Baltimore: Williams and Williams, 1996.
6. Rose G, Barker DJP. *Epidemiology for the uninitiated.* 3rd ed. London: BMJ Books, 1993.
7. Oxman AD, Sackett DS, Guyatt GH. Users' guides to the medical literature. I. How to get started. *JAMA* 1993;270:2093-5.
8. 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;270:2598-601.
9. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA* 1994;271:59-63.
10. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? *JAMA* 1994;271:389-91. [Medline]
11. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What were the results and will they help me in caring for my patients? *JAMA* 1994;271:703-7.
12. Levine M, Walter S, Lee H, Haines T, Holbrook A, Moyer V. Users' guides to the medical literature. IV. How to use an article about harm. *JAMA* 1994;271:1615-9. [Medline]
13. Laupacis A, Wells G, Richardson WS, Tugwell P. Users' guides to the medical literature. V. How to use an article about prognosis. *JAMA* 1994;271:234-7. [Abstract]
14. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. *JAMA* 1994;272:1367-71.
15. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. A. Are the results of the study valid? *JAMA* 1995;273:1292-5. [Medline]
16. Richardson WS, Detsky AS. Users' guides to the medical literature. VII. How to use a clinical decision analysis. B. What are the results and will they help me in caring for my patients? *JAMA* 1995;273:1610-3.
17. Hayward RSA, Wilson MC, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. A. Are the recommendations valid? *JAMA* 1995;274:570-4.
18. Wilson MC, Hayward RS, Tunis SR, Bass EB, Guyatt G. Users' guides to the medical literature. VIII. How to use clinical practice guidelines. B. Will the recommendations help me in caring for my patients? *JAMA* 1995;274:1630-2. [Medline]
19. Naylor CD, Guyatt GH. Users' guides to the medical literature. XI. How to use an article about a clinical utilization review. *JAMA* 1996;275:1435-9. [Medline]
20. Drummond MF, Richardson WS, O'Brien BJ, Levine M, Heyland D. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? *JAMA* 1997;277:1552-7. [Medline]
21. O'Brien BJ, Heyland D, Richardson WS, Levine M, Drummond MF. Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. B. What are the results and will they help me in caring for my patients? *JAMA* 1997;277:1802-6. [Medline]
22. Popper K. *Conjectures and refutations: the growth of scientific knowledge.* New York: Routledge and Kegan Paul, 1963.
23. Randomised trial of intravenous streptokinase, aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. (ISIS-2 Collaborative Group). *Lancet* 1988;ii:349-60.
24. Thornton JG, Lilford RJ, Johnson N. Decision analysis in medicine. *BMJ* 1992;304:1099-103.
25. Thornton JG, Lilford RJ. Decision analysis for medical managers. *BMJ* 1995;310:791-4. [Free Full Text]
26. Dowie J. "Evidence-based", "cost-effective", and "preference-driven" medicine. *J Health Serv Res*

Policy 1996;1:104-13.
27. Bero LA, Rennie D. Influences on the quality of published drug studies. *Int J Health Technology Assessment* 1996;12:209-37.
28. MacIntyre IMC. Tribulations for clinical trials. Poor recruitment is hampering research. *BMJ* 1991;302:1099-100.
29. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet* 1991;337:867-72.
30. Altman D. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ* 1996; 313:570-1.
31. Doll R, Hill AB. Mortality in relation to smoking: ten years' observations on British doctors. *BMJ* 1964;i:1399-414, 1460-7.
32. Doll R, Peto R. Mortality in relation to smoking: ten years' observations on British doctors. *BMJ* 1976;ii:1525-36.
33. Doll R, Peto R, Wheatley K, Gray R, Sutherland I. Mortality in relation to smoking: 40 years' observations on male British doctors. *BMJ* 1994;309:901-11. [Abstract/Free Full Text]
34. MacNaughton J. Anecdotes and empiricism. *Br J Gen Pract* 1995; 45:571-2.
35. McBride WG. Thalidomide and congenital abnormalities. *Lancet* 1961;ii:1358.
36. Guyatt GH, Sackett DL, Sinclair JC, Hayward R, Cook DJ, Cook RJ. Users' guides to the medical literature. IX. A method for grading health care recommendations. *JAMA* 1995;274:1800-4.

Related Article

Sexual abstinence only programmes to prevent HIV infection in high income countries: systematic review
Kristen Underhill, Paul Montgomery, and Don Operario
BMJ 2007 335: 248. [Abstract] [Full Text] [PDF]

This article has been cited by other articles:

- Hammond, R., Harry, T. C (2008). Efficacy of antiretroviral therapy in Africa: effect on immunological and virological outcome measures - a meta-analysis. *Int J STD AIDS* 19: 291-296 [Abstract] [Full text]
- Underhill, K., Montgomery, P., Operario, D. (2007). Sexual abstinence only programmes to prevent HIV infection in high income countries: systematic review. *BMJ* 335: 248-248 [Abstract] [Full text]
- Holosko, M. J. (2006). A Suggested Authors' Checklist for Submitting Manuscripts to Research on Social Work Practice. *Research on Social Work Practice* 16: 449-454 [Abstract]
- Glenny, A-M., Harrison, J. E. (2003). **How to** ... interpret the orthodontic literature. *J. Orthod.* 30: 159-164 [Full text]
- Galli, M., Luciani, D., Bertolini, G., Barbui, T. (2003). Lupus anticoagulants are stronger risk factors for thrombosis than anticardiolipin antibodies in the antiphospholipid syndrome: a systematic review of the literature. *Blood* 101: 1827-1832 [Abstract] [Full text]
- Gonzalez, L. S III (2003). Referees make journal clubs fun. *BMJ* 326: 106-106 [Full text]
- Donner-Banzhoff, N., Kunz, R., Rosser, W. (2001). Studies of symptoms in primary care. *Fam Pract* 18: 33-38 [Abstract] [Full text]
- Lotze, T. E., Paolicchi, J. (2000). Vertebral Artery Dissection and Migraine Headaches in Children. *J Child Neurol* 15: 694-696 [Abstract]
- Bruinsma, F., Venn, A., Lancaster, P., Speirs, A., Healy, D. (2000). Incidence of cancer in children born after in-vitro fertilization. *Hum Reprod* 15: 604-607 [Abstract] [Full text]

- Greenhalgh, T., Taylor, R. (1997). **How to read a paper**: **Paper**s that go beyond numbers (qualitative research). *BMJ* 315: 740-743 [Full text]

BMJ 1997;315:305-308 (2 August)

Education and debate

How to read a paper: Assessing the methodological quality of published papers

Trisha Greenhalgh, *senior lecturer* [a]

[a] Unit for Evidence-Based Practice and Policy, Department of Primary Care and Population Sciences, University College London Medical School/Royal Free Hospital School of Medicine, Whittington Hospital, London N19 5NF

Correspondence to: p.greenhalgh@ucl.ac.uk

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: Was 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?

Top
Introduction
Question 1: Was the...
Question 2: Whom is...
Question 3: Was the...
Question 4: Was systematic...
Question 5: Was assessment...
Question 6: Were preliminary...
References

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?

Top
Introduction
Question 1: Was the...
Question 2: Whom is...
Question 3: Was the...
Question 4: Was systematic...
Question 5: Was assessment...
Question 6: Were preliminary...
References

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 form 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.

PETER BROWN

View larger version (135K):
[in this window]
[in a new window]

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

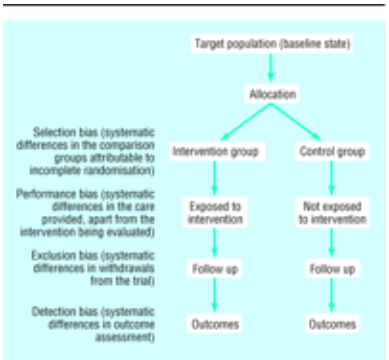
outcome measure has been objectively validated—that is, that someone has confirmed that the scale of anxiety, pain, and so on used in this study measures what it purports to measure, and that changes in this outcome measure adequately reflect changes in the status of the patient. Remember that what is 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 2 summarises sources of bias to check for.



View larger version (40K):
[in this window]
[in a new window]

Fig 1 Sources of bias to check for in a randomised controlled trial

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 years[12]—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 ß 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.

View larger version (83K):
[in this window]
[in a new window]

Are these results credible?
BMJ/PREUSS/SOUTHAMPTON UNIVERSITY TRUST

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?

The articles in this series are excerpts from ***How to read a paper**: the basics of evidence based medicine*. The book includes chapters on searching the literature and implementing evidence based findings. It can be ordered from the BMJ Bookshop: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

References

1. Bero LA, Rennie D. Influences on the quality of published drug studies. *Int J Health Technology Assessment* 1996;12:209-37.
2. Greenhalgh T. Papers that report drug trials. In: ***How to read a paper**: the basics of evidence based medicine.* London: BMJ Publishing Group, 1997:87-96.

Top
Introduction
Question 1: Was the...
Question 2: Whom is...
Question 3: Was the...

3. Dunning M, Needham G. *But will it work, doctor? Report of conference held in Northampton, 22-23 May 1996.* London: King's Fund, 1997.
4. Rose G, Barker DJP. *Epidemiology for the uninitiated.* 3rd ed. London: BMJ Publishing Group, 1994.
5. Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. *N Engl J Med* 1983;309:1358-61.
6. Colditz GA, Miller JA, Mosteller JF. How study design affects outcome in comparisons of therapy. I. Medical. *Statistics in Medicine* 1989;8:441-54.
7. Brennan P, Croft P. Interpreting the results of observational research: chance is not such a fine thing. *BMJ* 1994;309:727-30.
8. Maclure M. Demonstration of deductive meta-analysis: alcohol intake and risk of myocardial infarction. *Epidemiol Rev* 1993;15:328-51.
9. Bowie C. Lessons from the pertussis vaccine trial. *Lancet* 1990;335:397-9. [Medline]
10. Altman D. *Practical statistics for medical research.* London: Chapman and Hall, 1991:456.
11. Medical Research Council Working Party. MRC trial of mild hypertension: principal results. *BMJ* 1985;291:97-104.
12. MacMahon S, Rogers A. The effects of antihypertensive treatment on vascular disease: re-appraisal of the evidence in 1993. *J Vascular Med Biol* 1993;4:265-71.
13. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology—a basic science for clinical medicine.* London: Little, Brown, 1991:19-49.
14. Stewart LA, Parmar MKB. Bias in the analysis and reporting of randomized controlled trials. *Int J Health Technology Assessment* 1996;12:264-75.
15. Knipschild P. Some examples of systematic reviews. In: Chalmers I, Altman DG, eds. *Systematic reviews.* London: BMJ Publishing Group, 1995:9-16.

Related Article

Assessing methodological quality of published papers
Michael Gossop, John Marsden, and Peter Daish
BMJ 1998 316: 151. [Extract] [Full Text]

This article has been cited by other articles:

- Fitch, C., Malik, A., Lelliott, P., Bhugra, D., Andiappan, M. (2008). Assessing psychiatric competencies: what does the literature tell us about methods of workplace-based assessment?. *Adv. Psychiatr. Treat.* 14: 122-130 [Abstract] [Full text]
- Zaccai, J H (2004). How to assess epidemiological studies. *Postgrad. Med. J.* 80: 140-147 [Abstract] [Full text]
- Athanasiou, T., Al-Ruzzeh, S., Kumar, P., Crossman, M.-C., Amrani, M., Pepper, J. R., Del Stanbridge, R., Casula, R., Glenville, B. (2004). Off-pump myocardial revascularization is associated with less incidence of stroke in elderly patients. *Ann. Thorac. Surg.* 77: 745-753 [Abstract] [Full text]
- Guevara, J. P, Wolf, F. M, Grum, C. M, Clark, N. M (2003). Effects of educational interventions for self management of asthma in children and adolescents: systematic review and meta-analysis. *BMJ* 326: 1308-1309 [Abstract] [Full text]
- Brown, J. P., Josse, R. G. (2003). Lignes directrices de pratique clinique 2002 pour le diagnostic et le traitement de l'osteoporose au Canada. *CMAJ* 168: SF1-38 [Abstract] [Full text]
- Brown, J. P., Josse, R. G. (2002). 2002 clinical practice guidelines for the diagnosis and management of osteoporosis in Canada. *CMAJ* 167: s1-34 [Abstract] [Full text]
- Redmond, A. C., Keenan, A.-M., Landorf, K. (2002). 'Horses for Courses': The Differences Between Quantitative and Qualitative Approaches to Research. *J. Am. Podiatr. Med. Assoc.* 92: 159-169

[Abstract] [Full text]

- Devereaux, P.J., Manns, B. J., Ghali, W. A., Quan, H., Guyatt, G. H. (2001). Reviewing the reviewers: the quality of reporting in three secondary journals. *CMAJ* 164: 1573-1576 [Abstract] [Full text]
- Giacomini, M. K (2001). The rocky road: qualitative research as evidence. *Evid. Based Med.* 6: 4-6 [Full text]
- McInnes, E., Duf, L., McClarey, M. (1999). Challenges in updating a systematic review. *Journal of Research in Nursing* 4: 66-71 [Abstract]
- Gossop, M., Marsden, J., Daish, P. (1998). Assessing methodological quality of published papers. *BMJ* 316: 151a-151 [Full text]
