

BMJ 1997;315:180-183 (19 July)

Education and debate

How to **read** a **paper**: The Medline database

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ **Read** responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by: **[Greenhalgh, T.](#)**
- ▶ Alert me when: [New articles cite this article](#)

▶ Introduction

In 1928, in his introduction to *Sceptical Essays*, Bertrand Russell wrote: "The extent to which beliefs are based on evidence is very much less than believers suppose." Medical beliefs, and the clinical practices that are based on them, are a case in point. Debate continues as to whether scientific evidence alone is sufficient to guide medical decision making, but few doctors would dispute that finding and understanding relevant research based evidence is increasingly necessary in clinical practice. This article is the first in a series that introduces the non-expert to searching the medical literature and assessing the value of medical articles.

- ▲ [Top](#)
- [Introduction](#)
- ▼ [The Medline database](#)
- ▼ [Appendix 1: Evidence based...](#)
- ▼ [Appendix 2: Maximally sensitive...](#)
- ▼ [References](#)

▶ The Medline database

Over 10 million medical articles exist on library shelves. About a third are indexed in the huge Medline database, compiled by the National Library of Medicine of the United States. The Medline database is exactly the same, whichever company is selling it, but the commands differ according to the software.

Vendors of Medline online and on CD ROM include Ovid Technologies (ovid) and Silver Platter Information (WinSPIRS).

Articles can be traced in two ways: by any word listed on the database, including words in the title, abstract, authors' names, and the institution where the research was done; and by a restricted thesaurus of medical titles, known as medical subject heading (MeSH) terms.

▲ Top
▲ Introduction
▪ The Medline database
▼ Appendix 1: Evidence based...
▼ Appendix 2: Maximally sensitive...
▼ References

To illustrate how Medline works, I have worked through some common problems in searching. The scenarios have been drawn up using ovid software.

Problem 1: You are trying to find a known paper

Solution: Search the database by field suffix (title, author, journal, institution, etc) or by textwords.

First, get into the part of the database which covers the approximate year of the **paper's** publication. If you are already in the main Medline menu, select "database" (Alt-B). If you know the approximate title of the **paper** and perhaps the journal where it was published, you can use the title and journal search keys or (this is quicker) the **.ti** and **.jn** field suffixes. The [box](#) shows some other useful field suffixes.

Useful search field suffixes (ovid)		
Syntax	Meaning	Example
.ab	Word in abstract	epilepsy.ab
.au	Author	smith-r.au
.jn	Journal	lancet.jn
.me	Single word, wherever it may appear as a MeSH term	ulcer.me
.ti	Word in title	epilepsy.ti
.tw	Word in title or abstract	epilepsy.tw
.ui	Unique identifier	91574637.ui
.yr	Year of publication	87.yr

Thus, to find a **paper** called something like "Confidentiality and patients' casenotes," which you

remember seeing in the *British Journal of General Practice* a couple of years ago,¹ type the following sequence:

1. confidentiality.ti
2. british journal of general practice.jn
3. 1 and 2

Summary points

Not all medical articles are indexed on Medline, and many that are have been misclassified

Searching by textword can supplement a search by MeSH headings

To increase the sensitivity of a search, use the "explode" command and avoid using subheadings

Scan titles on screen rather than relying on the software to find the most valid or relevant ones

You could do all this in one step:

1. confidentiality.ti and british journal of general practice.jn

This step illustrates the use of the boolean operator "and"; it will give you articles common to both sets. Using "or" will simply add the two sets together.

Note that since 1988 the *British Medical Journal* is abbreviated BMJ in ovid software, and *Journal of the American Medical Association* is JAMA. Other useful field suffixes to try when searching for a known article are author (using the syntax haines-ap.au), institution (for example, manchester.in), or title (for example, evidence-based medicine.ti).

Problem 2: You want to answer a specific question

Solution: Construct a focused (specific) search by combining two or more broad (sensitive) searches.

I was recently asked by the mother of a young girl with anorexia nervosa whose periods had ceased to prescribe oral contraceptives for her so as to stop her bones thinning. This seemed a reasonable request,

though there were ethical problems to consider. But is there any evidence that taking oral contraceptives in these circumstances really prevents long term bone loss? I decided to explore the subject using Medline. To answer this question, you need to search very broadly under "anorexia nervosa," "osteoporosis," and "oral contraceptives." The search described below involves articles from 1992; when replicating it, make sure the database you are searching goes back that far. Type:

1 anorexia nervosa

You have not typed a field suffix (such as .tw), so the ovid system will automatically try to "map" your request to one of its standard medical subject headings (abbreviated MeSH and colloquially known as "mesh terms"). (Note that not all Medline software packages will automatically map your suggestion to MeSH terms. With Silver Platter search software, for example, you need to enter your heading and click the "suggest" button.) For this example, the screen offers you either "eating disorders" or "anorexia nervosa" and asks you to pick the closest one. Choose "anorexia nervosa" (space bar to highlight the text, then press "return").

The screen then asks you whether you want to "restrict to focus." Do you only want articles which are actually about anorexia nervosa, or do you want any article that mentions anorexia nervosa in passing? Let's say we do want to restrict to focus. Next, the screen offers us a choice of subheadings, but we'll ignore these for a moment. Select "Include all subheadings." We could have got this far using a single line command:

2 *anorexia nervosa/

The * shows that the term is a major focus of the article, and the / represents a MeSH term. You should have about 750 articles in this set.

Similarly, to get articles on osteoporosis (which is also a MeSH term), use the following single line command:

3 osteoporosis/

You should get about 2200 articles. Note that in ovid, if you know that the subject you want is an official MeSH term, you can shortcut the mapping process by typing a slash (/) after the word. Note also that we have not used an asterisk here, because osteoporosis may not be the focus of the article we are looking for.

Finally, put in the term "oral contraceptives" (without an asterisk and without a slash) to see what the MeSH term here is. You will be offered "contraceptives, oral," and if you had known this you could have

used the following command:

4 contraceptives, oral/

This set should contain around 1200 articles. You can combine these three sets, either by using their set numbers 1 and 2 and 3 or by typing the single line command:

5 *anorexia nervosa/ and osteoporosis/ and contraceptives, oral/

With this you will have searched over 4000 articles and struck a single bull's eye.² (If you don't find it, check the syntax of your search carefully, then try running the same search through the previous five year database using the Alt-B command.)



Problem 3: You want to get general information quickly about a well defined topic

Solution: Use subheadings and/or the "limit set" options.

Subheadings are the fine tuning of the Medline indexing system; they classify articles on a particular MeSH topic into aetiology, prevention, therapy, and so on. The most useful ones are listed in the [box](#). I try not to use subheadings unless I have unearthed an

unmanageable set of articles, since an estimated 50% of articles in Medline are inadequately or incorrectly classified by subheading. It actually doesn't take long to browse through 50 or so articles on the screen. It is better to do this than to rely on the "limit set" command (see [box](#)) to give you the best of the bunch.

Useful subheadings (ovid)

Syntax	Meaning	Example
/ae	Adverse effects	thalidomide/ae
/co	Complications	measles/co
/ct	Contraindications (of drug)	propranolol/ct
/di	Diagnosis	glioma/di
/dt	Drug therapy	depression/dt
/ed	Education	asthma/ed
/ep	Epidemiology	poliomyelitis/ep
/hi	History	mastectomy/hi
/nu	Nursing	cerebral palsy/nu
/og	Organisation/administration	health service/og
/pc	Prevention and control	influenza/pc
/px	Psychology	diabetes/px
/th	Therapy	hypertension/th
/tu	Therapeutic use (of drug)	aspirin/tu

Useful "limit set" options

	AIM journals	Abstracts
Nursing journals	Local holdings	
Dental journals	English language	
Cancer journals	Male	
Review articles	Human	
Editorials	Publication year	

The articles in this series are excerpts from *How to read a paper: the basics of evidence based medicine*. The book 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.

The option "AIM journals" denotes all journals listed in the Abridged Index Medicus—that is, the "mainstream" medical journals. Alternatively, if you want articles relating to nursing, rather than medical care, you could limit the set to "Nursing journals." This is often a better way of limiting a large set than asking for local holdings. If you are not interested in seeing anything in a foreign language (even though the abstract may be in English), select this option, again bearing in mind that it is a non-systematic (indeed, a very biased) way of excluding articles from your set.³

Note that instead of using the "limit set" function key you can use direct single line commands such as:

9 limit 4 to local holdings

10 limit 5 to human

Problem 4: Your search gives irrelevant articles

Solution: Refine your search as you go along in the light of interim results.

Often, a search uncovers dozens of articles which are irrelevant to your question. The boolean operator "not" can help here. I recently undertook a search to identify articles on surrogate endpoints in clinical pharmacology research. My search revealed hundreds of articles I didn't want—all on surrogate motherhood. The syntax to exclude the unwanted articles is:

1 (surrogate not mother\$).tw

Deciding to use the "not" operator is a good example of how you can (and should) refine your search as you go along—much easier than producing the perfect search off the top of your head. I used the truncation symbol \$ to find all words from a single stem, such as mother, mothers, motherhood, and so on.

Another way of getting rid of irrelevant articles is to narrow your textword search to adjacent words using the "adj" operator. For example, the term "home help" includes two very common words linked in a specific context. Link them as follows:

1 home adj help.tw

Problem 5: The search gives no articles, or too few

Solution: Firstly, don't overuse subheadings or the "limit set" options. Secondly, search under textwords as well as MeSH terms. Thirdly, learn about the "explode" command, and use it routinely.

Many important articles are missed not because we constructed a flawed search strategy but because we relied too heavily on a flawed indexing system. For this reason, you should adopt a "belt and braces" approach and search under textwords as well as by MeSH terms. After all, it is difficult to write an article on the psychology of diabetes without mentioning the words "diabetes," "diabetic," "psychology," or "psychological," so the truncation stems "diabet\$.tw." and "psychol\$.tw." would supplement a search under the MeSH term "diabetes mellitus" and the subheading "/px" (psychology).

Another important strategy for preventing incomplete searches is to use the powerful "explode" command. The MeSH terms are like the branches of a tree with, for example, "asthma" subdividing into "asthma in children," "occupational asthma," and so on. Medline indexers are instructed to index items by using the most specific MeSH terms they can. If you just ask for articles on "asthma" you will miss all the articles indexed under "asthma in children" unless you "explode" the term using the following syntax:

```
1 exp asthma/
```

Problem 6: You don't know where to start searching

Solution: Use the "permuted index" option.

Let's take the term "stress." It comes up often, but searching for particular types of stress would be laborious and searching "stress" as a textword would be too unfocused. We need to know where in the MeSH index the various types of stress lie, and when we see that, we can choose the sort of stress we want to look at. For this, we use the command ptx ("permuted index"):

```
1 ptx stress
```

The screen shows many options, including post-traumatic stress disorders, stress fracture, oxidative stress, stress incontinence, and so on.

The command "ptx" is useful when the term might be found in several subject areas. If your subject is a discrete MeSH term, use the tree command. For example:

```
2 tree epilepsy
```

will show where epilepsy is placed in the MeSH index—as a branch of "brain diseases," which itself branches into generalised epilepsy, partial epilepsy, post-traumatic epilepsy, and so on.

Problem 7: Limiting a set loses important articles but does not exclude those of low methodological quality

Solution: Apply an EBQF (evidence based quality filter).

If your closely focused search still gives you several hundred articles, and if applying subheadings or limit set functions seems to lose valuable (and valid) **papers**, you should insert a quality string designed to limit your set to therapeutic interventions, aetiology, diagnostic procedures, or epidemiology. Alternatively, you could apply search strings to identify the publication type, such as randomised controlled trial, systematic review, or meta-analysis.

These EBQFs (evidence based quality filters), which are listed in [Appendix 1](#), are complex search strategies developed by some of the world's most experienced medical information experts. You can copy them into your personal computer and save them as strategies to be added to your subject searches. Other search strategies that will identify cohort studies, case-control studies, and so on will soon be available from the UK Cochrane Centre, Summertown Pavillion, Middle Way, Oxford OX2 7LG (general@cochrane.co.uk).

Problem 8: Medline hasn't helped

Solution: Explore other medical and paramedical databases.

Entry of articles onto the Medline database is open to human error, both from authors and editors who select key words for indexing, and from the librarians who group articles under subheadings and type in the abstracts. In addition, some sections of indexed journals are not available on Medline (for example, the News section of the *BMJ*). According to one estimate, 40% of material which should be listed on Medline can, in reality, only be accessed by looking through all the journals again, by hand. Furthermore, a number of important medical and paramedical journals are not covered by Medline at all. It is said that Medline lacks comprehensive references in the fields of psychology, medical sociology, and non-clinical pharmacology.

If you wish to broaden your search to other electronic databases, ask your local librarian where you could access the following:

- *AIDSLINE*—Covers AIDS and HIV back to 1980.
- *Allied and Alternative Medicine*—Covers complementary and alternative medicine.
- *American Medical Association Journals*—Provides the full text of JAMA plus 10 specialty journals produced by the American Medical Association; from 1982.

- *ASSIA*—An applied social sciences database covering psychology, sociology, politics, and economics since 1987. All documents have abstracts.
- *Cancer-CD*—A compilation by Silver Platter of cancerlit and Embase cancer related records from 1984. The CD ROM version is updated quarterly.
- *CINAHL*—The nursing and allied health database covering all aspects of nursing, health education, occupational therapy, social services in health care, and other related disciplines from 1983. The CD ROM version is updated monthly.
- *Cochrane Library*—The Cochrane Controlled Trials Register (cctr), Cochrane Database of Systematic Reviews (cdsr), Database of Abstracts of Reviews of Effectiveness (dare), and Cochrane Review Methodology Database (crmd) are updated quarterly; authors of systematic reviews on cdsr undertake to update their own contributions periodically.⁴
- *Current Contents Search*—Indexes journal issues on or before their publication date. It is useful when checking for the very latest output on a subject. Updated weekly; from 1990.
- *Current Research in Britain*—The British national research database of trials in progress.
- *DHData* (formerly DHSS-Data)—The database of the UK's Department of Health indexes articles covering health service and hospital administration; from 1983.
- *Embase*—Focuses on drugs and pharmacology but also includes other biomedical specialties. It is more up to date than Medline and has better European coverage. The CD ROM version is updated monthly.
- *HELMIS*—The Health Management Information Service at the Nuffield Institute of Health, Leeds, UK, indexes articles on health service management.
- *Psychlit*—Produced by the American Psychological Association as the computer searchable version of Psychological Abstracts; covers psychology, psychiatry, and related subjects; journals are included from 1974 and books from 1987 (English language only).
- *Science Citation Index*—Indexes references cited in articles as well as the usual author, title, abstract, and citation of articles themselves. Useful for finding follow up work done on a key article and for tracking down addresses of authors.
- *SHARE*—Based at the King's Fund library in London; published and ongoing research into the

health of, and health services for, black and minority ethnic groups.

- *Toxline*—Information on toxicological effects of chemicals and drugs on living systems; from 1981.
- *UNICORN*—The main database of the King's Fund, London. Covers health, health management, health economics, and social sciences. Particularly strong on primary health care and the health of Londoners.

Acknowledgements

Thanks to Mr Reinhard Wentz, Ms Jane Rowlands, Ms Carol Lefebvre, and Ms Valerie Wildridge for advice on this chapter. I am grateful to Carol Lefebvre of the UK Cochrane Centre for permission to reproduce the EBQFs in [Appendix 1](#).

Appendix 1: Evidence based quality filters for everyday use

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [The Medline database](#)
- [Appendix 1: Evidence based...](#)
- ▼ [Appendix 2: Maximally sensitive...](#)
- ▼ [References](#)

(a) Therapeutic interventions (What works?)

1. exp clinical trials
2. exp research design
3. randomized controlled trial.pt.
4. clinical trial.pt.
5. (single or double or treble or triple).tw.
6. (mask\$ or blind\$).tw.
7. 5 and 6
8. placebos/ or placebo.tw.
9. 1 or 2 or 3 or 4 or 7 or 8

(b) Aetiology (What causes it? What are the risk factors?)

1. exp causality
2. exp cohort studies
3. exp risk
4. 1 or 2 or 3

(c) Diagnostic procedures

1. exp "sensitivity and specificity"
2. exp diagnostic errors
3. exp mass screening
4. 1 or 2 or 3

(d) Epidemiology

1. sn.xs

(This would find all articles indexed under any MeSH term with any of "statistics," "epidemiology," "ethnology," or "mortality" as subheadings.)

Appendix 2: Maximally sensitive search strings (to be used mainly for research)

- [▲ Top](#)
- [▲ Introduction](#)
- [▲ The Medline database](#)
- [▲ Appendix 1: Evidence based...](#)
- [Appendix 2: Maximally sensitive...](#)
- [▼ References](#)

(a) Maximally sensitive qualifying string for randomised controlled trials

1. RANDOMIZED CONTROLLED TRIAL.pt.
2. CONTROLLED CLINICAL TRIAL.pt.
3. RANDOMIZED CONTROLLED TRIALS.sh.
4. RANDOM ALLOCATION.sh.
5. DOUBLE-BLIND METHOD.sh.
6. SINGLE-BLIND METHOD.sh.
7. or/1-6
8. ANIMAL.sh. not HUMAN.sh.
9. 7 not 8

10. CLINICAL TRIAL.pt.
11. exp CLINICAL TRIALS
12. (clin\$ adj25 trial\$.ti,ab.
13. ((single or double or treble or triple) adj25 (blind\$ or mas\$)).ti,ab.
14. PLACEBOS.sh.
15. placebo\$.ti,ab.
16. random\$.ti,ab.
17. RESEARCH DESIGN.sh.
18. or/10-17
19. 18 not 8
20. 19 not 9
21. COMPARATIVE STUDY.sh.
22. exp EVALUATION STUDIES/
23. FOLLOW UP STUDIES.sh.
24. PROSPECTIVE STUDIES.sh.
25. (control\$ or prospectiv\$ or volunteer\$.ti,ab.
26. or/21-25
27. 26 not 8
28. 26 not (9 or 20)
29. 9 or 20 or 28

In these examples, upper case denotes controlled vocabulary and lower case denotes free text terms. Search statements 8, 9, 19, and 27 could be omitted if your search takes too long a time to run.

(b) Maximally sensitive qualifying string for identifying systematic reviews

1. REVIEW, ACADEMIC.pt.
2. REVIEW, TUTORIAL.pt.
3. META-ANALYSIS.pt.
4. META-ANALYSIS.sh.
5. systematic\$ adj25 review\$
6. systematic\$ adj25 overview\$
7. meta-analy\$ or metaanaly\$ or (meta analy\$)
8. or/1-7
9. ANIMAL.sh. not HUMAN.sh.
10. 8 not 9

Search statements 9 and 10 could be omitted if your search seems to be taking a long time to run.

References

1. Caman D, Britten N. Confidentiality and medical records: the patient's perspective. *Br J Gen Pract* 1995;45:485-8.
2. Seeman E, Szmukler GI, Formica C, Tsalamandris C, Mestrovic R. Osteoporosis in anorexia nervosa: the influence of peak bone density, bone loss, oral contraceptive use, and exercise. *J Bone Mineral Res* 1992;7:1467-74.

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [The Medline database](#)
- ▲ [Appendix 1: Evidence based...](#)
- ▲ [Appendix 2: Maximally sensitive...](#)
- [References](#)

3. Moher D, Fortin P, Jadad AR, Juni P, Klassen T, Le Lorier J, et al. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. *Lancet* 1996;347:363-6. [[Medline](#)]
4. Bero L, Rennie D. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995;274:1935-8. [[Medline](#)]

This article has been cited by other articles:

- Wyatt, J, Guly, H (2002). Identifying the research question and planning the project. *Emerg Med J* 19: 318-321
[\[Abstract\]](#) [\[Full text\]](#)
- McQueen, M. J. (2001). Overview of Evidence-based Medicine: Challenges for Evidence-based Laboratory Medicine. *Clin Chem* 47: 1536-1546 [\[Abstract\]](#) [\[Full text\]](#)
- Menz, H. B. (2001). Publication Patterns and Perceptions of the Australian Podiatric Medical Faculty. *J Am Podiatr Med Assoc* 91: 210-218
[\[Abstract\]](#) [\[Full text\]](#)
- Lucassen, P L B J, Assendelft, W J J, van Eijk, J T. M, Gubbels, J W, Douwes, A C, van Geldrop, W J (2001). Systematic review of the occurrence of infantile colic in the community. *Arch. Dis. Child.* 84: 398-403 [\[Abstract\]](#) [\[Full text\]](#)
- Menz, H. B. (2001). The Case for Multiple Database Searching in Podiatric Medicine. *J Am Podiatr Med Assoc* 91: 103-104 [\[Full text\]](#)
- Gehanno, J F, Thirion, B (2000). How to select publications on occupational health: the usefulness of Medline and the impact factor. *Occup Environ Med* 57: 706-709
[\[Abstract\]](#) [\[Full text\]](#)
- Smeenk, F. W J M, van Haastregt, J. C M, de Witte, L. P, Crebolder, H. F J M (1998). Effectiveness of home care programmes for patients with incurable cancer on their quality of life and time spent in hospital: systematic review. *BMJ* 316: 1939-1944 [\[Abstract\]](#) [\[Full text\]](#)
- Egger, M., Smith, G. D., Phillips, A. N (1997). Meta-analysis: Principles and procedures. *BMJ* 315: 1533-1537 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ **Read** responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

Rapid Responses:

Read all [Rapid Responses](#)

Updating your bull's eye?

Samuel Coenen

bmj.com, 2 Jun 1998 [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ 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.

- ▲ [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](#)

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.

- ▲ [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](#)

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 attack²³)

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

- ▲ [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](#)

(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 hypothetico-deductive 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.

- ▲ [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](#)



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?

- ▲ [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](#)

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.

- ▲ [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 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

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

- ▲ [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](#)

The hierarchy of evidence

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 order³⁶:

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.

- ▲ [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](#)

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. Altman DG. The scandal of poor medical research. *BMJ* 1994;308:283-4. [\[Full Text\]](#)
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. [\[Medline\]](#)
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. [\[Medline\]](#)

- [▲ 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](#)

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. [[Medline](#)]
25. Thornton JG, Lilford RJ. Decision analysis for medical managers. *BMJ* 1995;310:791-4. [[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/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.

This article has been cited by other articles:

- Donner-Banzhoff, N., Kunz, R., Rosser, W. (2001). Studies of symptoms in primary care. *Fam. Pract.* 18: 33-38 [[Abstract](#)] [[Full text](#)]
- 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**: **Papers** that go beyond numbers (qualitative research). *BMJ* 315: 740-743 [\[Full text\]](#)**

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



PETER BROWN

[\[View larger version \(170K\)\]](#)

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ 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.

- ▲ [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](#)

▶ 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?

- [▲ 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](#)

▶ 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

- [▲ 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](#)

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.

- ▲ [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](#)

- *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 subject¹⁴

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?

- ▲ [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](#)

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.

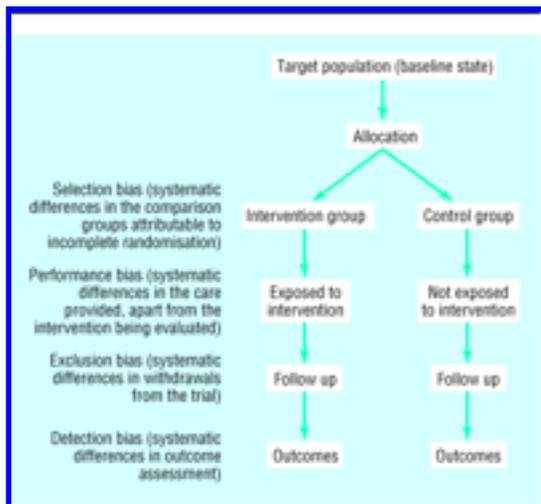


Fig 1 Sources of bias to check for in a randomised controlled trial

View larger version (40K):

[\[in this window\]](#)

[\[in a new window\]](#)

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](#).

- ▲ [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](#)

▶ 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.

- ▲ [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](#)

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.



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.
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

- ▲ [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**

- 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.

This article has been cited by other articles:

- 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. *Can Med Assoc J* 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](#)]
- Gossop, M., Marsden, J., Daish, P. (1998). Assessing methodological quality of published **papers**. *BMJ* 316: 151a-151 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by: **[Greenhalgh, T.](#)**
- ▶ Alert me when: [New articles cite this article](#)

Related letters in BMJ:

Assessing methodological quality of published **papers**

Michael Gossop, John Marsden, and Peter Daish
BMJ 1998 316: 151. [\[Letter\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



PETER BROWN

[\[View larger version \(176K\)\]](#)

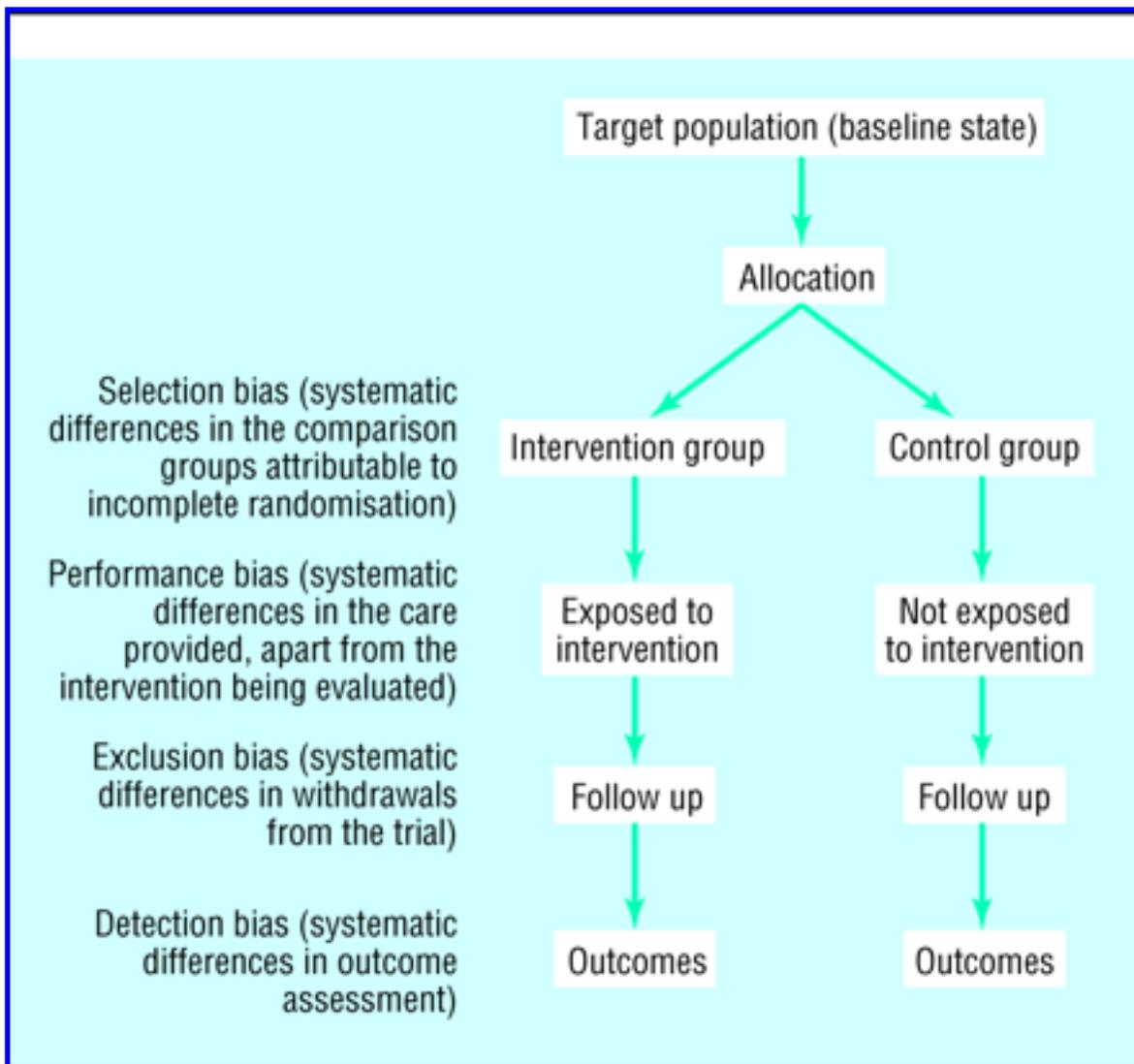


Fig 1 Sources of bias to check for in a randomised controlled trial

[\[View larger version \(200K\)\]](#)

BMJ 1997;315:364-366 (9 August)

Education and debate

How to **read** a **paper**: Statistics for the non- statistician. I: Different types of data need different statistical tests

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ **Read** responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ [See Correction for this article](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ Introduction

As medicine leans increasingly on mathematics no clinician can afford to leave the statistical aspects of a **paper** to the "experts." If you are numerate, try the "Basic Statistics for Clinicians" series in the *Canadian Medical Association Journal*,^{1 2 3 4} or a more mainstream statistical textbook.⁵ If, on the other hand, you find statistics impossibly difficult, this article and the next in this series give a checklist of preliminary questions to help you appraise the statistical validity of a **paper**.

- ▲ [Top](#)
- [Introduction](#)
- ▼ [Have the authors set...](#)
- ▼ [Paired data, tails, and...](#)
- ▼ [References](#)

▶ Have the authors set the scene correctly?

Have they determined whether their groups are comparable, and, if necessary, adjusted for

baseline differences?

Most comparative clinical trials include either a table or a paragraph in the text showing the baseline characteristics of the groups being studied. Such a table should show that the intervention and control groups are similar in terms of age and sex distribution and key prognostic variables (such as the average size of a cancerous lump). Important differences in these characteristics, even if due to chance, can pose a challenge to your interpretation of results. In this situation, adjustments can be made to allow for these differences and hence strengthen the argument.⁶

- ▲ [Top](#)
- ▲ [Introduction](#)
- [Have the authors set...](#)
- ▼ [Paired data, tails, and...](#)
- ▼ [References](#)

Summary points

In assessing the choice of statistical tests in a **paper**, first consider whether groups were analysed for their comparability at baseline

Does the test chosen reflect the type of data analysed (parametric or non-parametric, paired or unpaired)?

Has a two tailed test been performed whenever the effect of an intervention could conceivably be a negative one?

Have the data been analysed according to the original study protocol?

If obscure tests have been used, do the authors justify their choice and provide a reference?

What sort of data have they got, and have they used appropriate statistical tests?

Numbers are often used to label the properties of things. We can assign a number to represent our height, weight, and so on. For properties like these, the measurements can be treated as actual numbers. We can, for example, calculate the average weight and height of a group of people by averaging the measurements. But consider an example in which we use numbers to label the property "city of origin," where 1=London, 2=Manchester, 3=Birmingham, and so on. We could still calculate the average of these numbers for a particular sample of cases, but we would be completely unable to interpret the result. The same would apply if we labelled the property "liking for x" with 1=not at all, 2=a bit, and 3=a lot. Again, we could calculate the "average liking," but the numerical result would be uninterpretable unless we knew that the difference between "not at all" and "a bit" was exactly the same as the difference between "a bit" and "a lot."



PETER BROWN

View larger version (101K):

[\[in this window\]](#)

[\[in a new window\]](#)

All statistical tests are either parametric (that is, they assume that the data were sampled from a particular form of distribution, such as a normal distribution) or non-parametric (they make no such assumption). In general, parametric tests are more powerful than non-parametric ones and so should be used if possible.

Non-parametric tests look at the rank order of the values (which one is the smallest, which one comes next, and so on) and ignore the absolute differences between them. As you might imagine, statistical significance is more difficult to show with non-parametric tests, and this tempts researchers to use statistics such as the r value inappropriately. Not only is the r value (parametric) easier to calculate than its non-parametric equivalent but it is also much more likely to give (apparently) significant results. Unfortunately, it will give a spurious estimate of the significance of the result, unless the data are appropriate to the test being used. More examples of parametric tests and their non-parametric equivalents are given in table [1](#)).

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Table 1 Some commonly used statistical tests

Another consideration is the shape of the distribution from which the data were sampled. When I was at school, my class plotted the amount of pocket money received against the number of children receiving that amount. The results formed a histogram the same shape as figure [2](#)—a "normal" distribution. (The

term "normal" refers to the shape of the graph and is used because many biological phenomena show this pattern of distribution). Some biological variables such as body weight show "skew normal" distribution, as shown in figure 3. (Figure 3) shows a negative skew, whereas body weight would be positively skewed. The average adult male body weight is 70 kg, and people exist who weigh 140 kg, but nobody weighs less than nothing, so the graph cannot possibly be symmetrical.

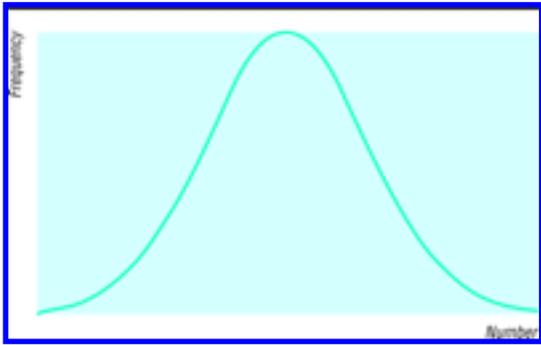


Fig 1 Normal curve

View larger version (8K):

[\[in this window\]](#)

[\[in a new window\]](#)

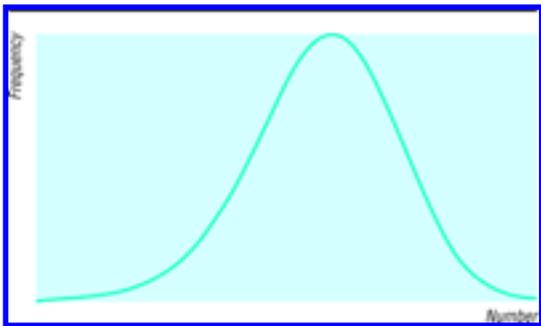


Fig 2 Skewed curve

View larger version (8K):

[\[in this window\]](#)

[\[in a new window\]](#)

Non-normal (skewed) data can sometimes be transformed to give a graph of normal shape by performing some mathematical transformation (such as using the variable's logarithm, square root, or reciprocal). Some data, however, cannot be transformed into a smooth pattern. For a very **readable** discussion of the normal distribution see chapter 7 of Martin Bland's *Introduction to Medical Statistics*.⁵

Deciding whether data are normally distributed is not an academic exercise, since it will determine what type of statistical tests to use. For example, linear regression will give misleading results unless the points on the scatter graph form a particular distribution about the regression line—that is, the residuals (the perpendicular distance from each point to the line) should themselves be normally distributed.

Transforming data to achieve a normal distribution (if this is indeed achievable) is not cheating: it simply ensures that data values are given appropriate emphasis in assessing the overall effect. Using tests based on the normal distribution to analyse non-normally distributed data, however, is definitely cheating.

If the authors have used obscure statistical tests, why have they done so and have they referenced them?

The number of possible statistical tests sometimes seems infinite. In fact, most statisticians could survive with a formulary of about a dozen. The rest should generally be reserved for special indications. If the **paper** you are **reading** seems to describe a standard set of data which have been collected in a standard way, but the test used has an unpronounceable name and is not listed in a basic statistics textbook, you should smell a rat. The authors should, in such circumstances, state why they have used this test, and give a reference (with page numbers) for a definitive description of it.

Are the data analysed according to the original protocol?

If you play coin toss with someone, no matter how far you fall behind, there will come a time when you are one ahead. Most people would agree that to stop the game then would not be a fair way to play. So it is with research. If you make it inevitable that you will (eventually) get an apparently positive result you will also make it inevitable that you will be misleading yourself about the justice of your case.⁷

(Terminating an intervention trial prematurely for ethical reasons when subjects in one arm are faring particularly badly is a different matter and is discussed elsewhere.⁷)

Raking over your data for "interesting results" (retrospective subgroup analysis) can lead to false conclusions.⁸ In an early study on the use of aspirin in preventing stroke, the results showed a significant effect in both sexes combined, and a retrospective subgroup analysis seemed to show that the effect was confined to men.⁹ This conclusion led to aspirin being withheld from women for many years, until the results of other studies¹⁰ showed that this subgroup effect was spurious.

This and other examples are included in Oxman and Guyatt's, "A consumer's guide to subgroup analysis," which reproduces a useful checklist for deciding whether apparent subgroup differences are real.¹¹

▶ Paired data, tails, and outliers

Were paired tests performed on paired data?

Students often find it difficult to decide whether to use a paired or unpaired statistical test to analyse their data. There is no great mystery about this. If you measure something twice on each subject—for example, blood pressure measured when the subject is lying and when standing—you will probably be interested not just in the average difference of lying versus standing blood pressure in the entire sample, but in how much each individual's blood pressure changes with position. In this situation, you have what is called "paired" data, because each measurement beforehand is paired with a measurement afterwards.

In this example, it is using the same person on both occasions which makes the pairings, but there are other possibilities (for example, any two measurements of bed occupancy made of the same hospital ward). In these situations, it is likely that the two sets of values will be significantly correlated (for example, my blood pressure next week is likely to be closer to my own blood pressure last week than to the blood pressure of a randomly selected adult last week). In other words, we would expect two randomly selected paired values to be closer to each other than two randomly selected unpaired values. Unless we allow for this, by carrying out the appropriate paired sample tests, we can end up with a biased estimate of the significance of our results.

Was a two tailed test performed whenever the effect of an intervention could conceivably be a negative one?

The term "tail" refers to the extremes of the distribution—the areas at the outer edges of the bell in figure 2. Let's say that the graph represents the diastolic blood pressures of a group of people of which a random sample are about to be put on a low sodium diet. If a low sodium diet has a significant lowering effect on blood pressure, subsequent blood pressure measurements on these subjects would be more likely to lie within the left tail of the graph. Hence we would analyse the data with statistical tests designed to show whether unusually low **readings** in this patient sample were likely to have arisen by chance.

But on what grounds may we assume that a low sodium diet could only conceivably put blood pressure down, but could never do the reverse, put it up? Even if there are valid physiological reasons in this particular example, it is certainly not good science always to assume that you know the direction of the effect which your intervention will have. A new drug intended to relieve nausea might actually exacerbate it, or an educational leaflet intended to reduce anxiety might increase it. Hence, your statistical analysis should, in general, test the hypothesis that either high or low values in your dataset have arisen by chance. In the language of the statisticians, this means you need a two tailed test, unless you have very convincing evidence that the difference can only be in one direction.

Were "outliers" analysed with both common sense and appropriate statistical adjustments?

Unexpected results may reflect idiosyncrasies in the subject (for example, unusual metabolism), errors in measurement (faulty equipment), errors in interpretation (mis**reading** a meter **reading**), or errors in

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [Have the authors set...](#)
- [Paired data, tails, and...](#)
- ▼ [References](#)

calculation (misplaced decimal points). Only the first of these is a "real" result which deserves to be included in the analysis. A result which is many orders of magnitude away from the others is less likely to be genuine, but it may be so. A few years ago, while doing a research project, I measured several different hormones in about 30 subjects. One subject's growth hormone levels came back about 100 times higher than everyone else's. I assumed this was a transcription error, so I moved the decimal point two places to the left. Some weeks later, I met the technician who had analysed the specimens and he asked, "Whatever happened to that chap with acromegaly?"

Statistically correcting for outliers (for example, to modify their effect on the overall result) requires sophisticated analysis and is covered elsewhere.⁶

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.

Acknowledgements

I am grateful to Mr John Dobby for educating me on statistics and for repeatedly checking and amending this article. Responsibility for any errors is mine alone.

References

1. Guyatt G, Jaenschke R, Heddle, N, Cook D, Shannon H, Walter S. Basic statistics for clinicians. 1. Hypothesis testing. *Can Med Assoc J* 1995;152:27-32.
2. Guyatt G, Jaenschke R, Heddle, N, Cook D, Shannon H, Walter S. Basic statistics for clinicians. 2. Interpreting study results: confidence intervals. *Can Med Assoc J* 1995;152:169-73.
3. Jaenschke R, Guyatt G, Shannon H, Walter S, Cook D, Heddle, N. Basic statistics for clinicians: 3. Assessing the effects of treatment: measures of association. *Can Med Assoc J* 1995;152:351-7.
4. Guyatt G, Walter S, Shannon H, Cook D, Jaenschke R, Heddle, N. Basic statistics for clinicians.

- [▲ Top](#)
- [▲ Introduction](#)
- [▲ Have the authors set...](#)
- [▲ Paired data, tails, and...](#)
- **References**

4. Correlation and regression. *Can Med Assoc J* 1995;152:497-504.
5. Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 1987.
6. Altman D. *Practical statistics for medical research*. London: Chapman and Hall, 1995.
7. Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Statistics in Medicine* 1987;7:1231-42.
8. Stewart LA, Parmar MKB. Bias in the analysis and reporting of randomized controlled trials. *Int J Health Technology Assessment* 1996;12:264-75.
9. Canadian Cooperative Stroke Group. A randomised trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978;299:53-9.
10. Antiplatelet Trialists Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. *BMJ* 1988;296:320-1.
11. Oxman, AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med* 1992;116:79-84.

This article has been cited by other articles:

- Redmond, A. C., Keenan, A.-M. (2002). Understanding Statistics: Putting P-Values into Perspective. *J Am Podiatr Med Assoc* 92: 297-305 [\[Abstract\]](#) [\[Full text\]](#)
- 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 [\[Abstract\]](#) [\[Full text\]](#)
- Perneger, T. V (1998). What's wrong with Bonferroni adjustments. *BMJ* 316: 1236-1238 [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Read](#) responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ [See Correction for this article](#)
- ▶ Search Medline for articles by: [Greenhalgh, T.](#)
- ▶ Alert me when: [New articles cite this article](#)

Rapid Responses:

Read all [Rapid Responses](#)

How to read a paper: Statistics for the non-statistician. Comment to Greenhalgh

Emili Garcia-Berthou

bmj.com, 30 Sep 1998 [\[Full text\]](#)

Comment on the correction

Kathleen M. Koehler

bmj.com, 23 May 2002 [\[Full text\]](#)



PETER BROWN

[\[View larger version \(190K\)\]](#)

Table 1 Some commonly used statistical tests

Parametric test	Example of equivalent non-parametric test	Purpose of test	Example
Two sample (unpaired) <i>t</i> test	Mann-Whitney U test	Compares two independent samples drawn from the same population	To compare girls' heights with boys' heights
One sample (paired) <i>t</i> test	Wilcoxon matched pairs test	Compares two sets of observations on a single sample	To compare weight of infants before and after a feed
One way analysis of variance (<i>F</i> test) using total sum of squares	Kruskall-Wallis analysis of variance by ranks	Effectively, a generalisation of the paired <i>t</i> or Wilcoxon matched pairs test where three or more sets of observations are made on a single sample	To determine whether plasma glucose level is higher one hour, two hours, or three hours after a meal
Two way analysis of variance	Two way analysis of variance by ranks	As above, but tests the influence (and interaction) of two different covariates	In the above example, to determine if the results differ in male and female subjects
χ^2 test	Fisher's exact test	Tests the null hypothesis that the distribution of a discontinuous variable is the same in two (or more) independent samples	To assess whether acceptance into medical school is more likely if the applicant was born in Britain
Product moment correlation coefficient (Pearson's <i>r</i>)	Spearman's rank correlation coefficient (r_s)	Assesses the strength of the straight line association between two continuous variables.	To assess whether and to what extent plasma HbA1 concentration is related to plasma triglyceride concentration in diabetic patients
Regression by least squares method	Non-parametric regression (various tests)	Describes the numerical relation between two quantitative variables, allowing one value to be predicted from the other	To see how peak expiratory flow rate varies with height

Multiple regression by least squares method	Non-parametric regression (various tests)	Describes the numerical relation between a dependent variable and several predictor variables (covariates)	To determine whether and to what extent a person's age, body fat, and sodium intake determine their blood pressure
---	---	--	--

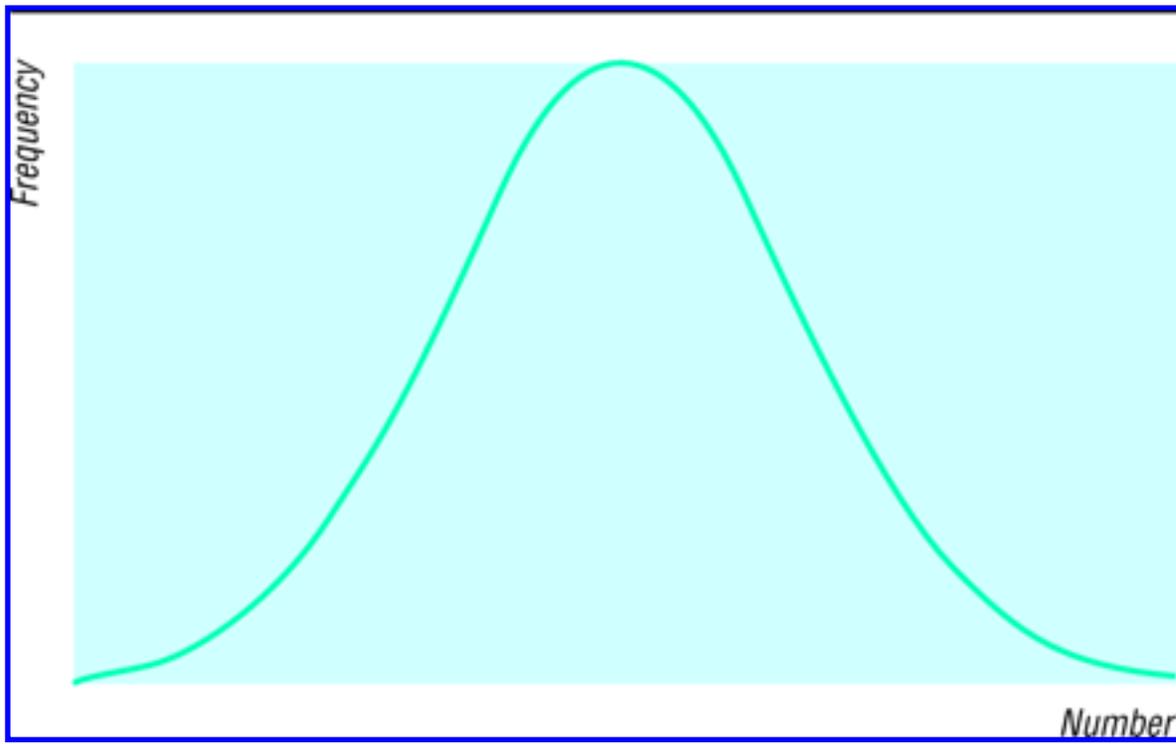


Fig 1 Normal curve

[\[View larger version \(54K\)\]](#)

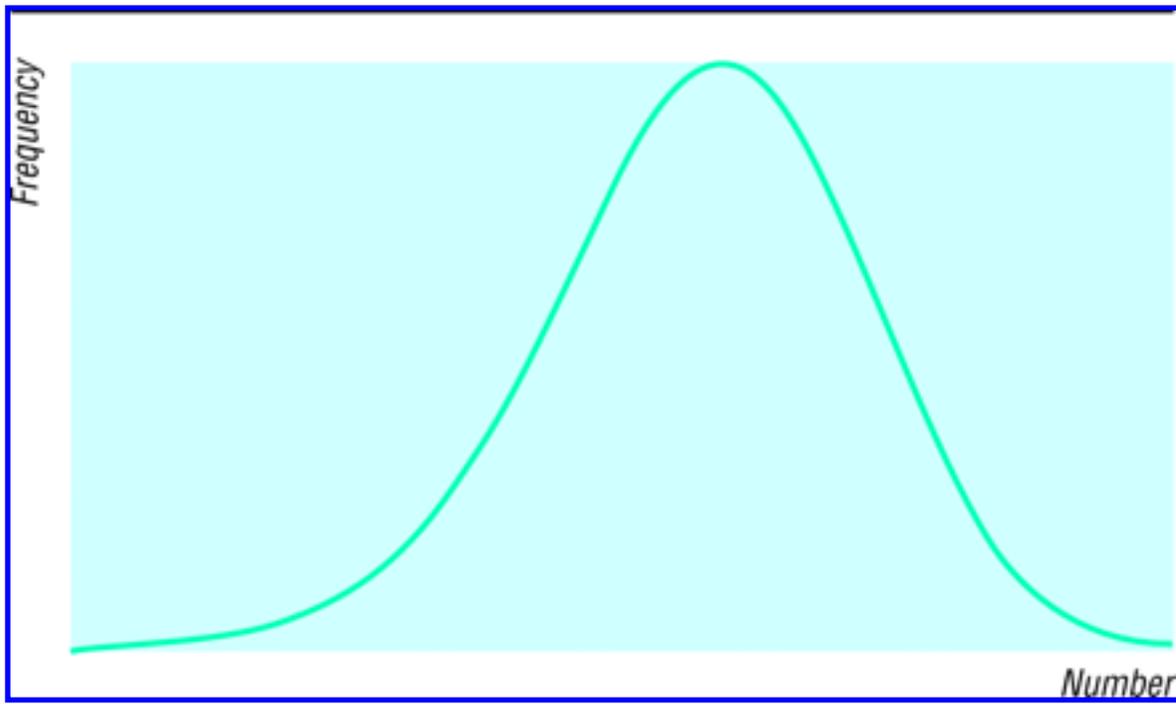


Fig 2 Skewed curve

[\[View larger version \(54K\)\]](#)

BMJ 1997;315:422-425 (16 August)

Education and debate

How to **read** a **paper**: Statistics for the non-statistician. II: "Significant" relations and their pitfalls

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by: **[Greenhalgh, T.](#)**
- ▶ Alert me when: [New articles cite this article](#)

Introduction

This article continues the checklist of questions that will help you to appraise the statistical validity of a **paper**. The first of this pair of articles was published last week.¹

- ▲ [Top](#)
- [Introduction](#)
- ▼ [Correlation, regression, and...](#)
- ▼ [Probability and confidence](#)
- ▼ [The bottom line](#)
- ▼ [Summary](#)
- ▼ [References](#)

Correlation, regression, and causation

Has correlation been distinguished from regression, and has the correlation coefficient (r value) been calculated and interpreted correctly?

For many non-statisticians, the terms "correlation" and "regression" are synonymous, and refer vaguely to a mental image of a scatter graph with dots sprinkled messily along a diagonal line sprouting from the intercept of the axes. You would be right in assuming that if two things are not correlated, it will be meaningless to attempt a regression. But regression and correlation are both precise statistical terms which serve quite different functions.¹

- ▲ [Top](#)
- ▲ [Introduction](#)
- [Correlation, regression, and...](#)
- ▼ [Probability and confidence](#)
- ▼ [The bottom line](#)
- ▼ [Summary](#)
- ▼ [References](#)

The r value (Pearson's product-moment correlation coefficient) is among the most overused statistical instrument. Strictly speaking, the r value is not valid unless the following criteria are fulfilled:

Summary points

An association between two variables is likely to be causal if it is strong, consistent, specific, plausible, follows a logical time sequence, and shows a dose-response gradient

A P value of <0.05 means that this result would have arisen by chance on less than one occasion in 20

The confidence interval around a result in a clinical trial indicates the limits within which the "real" difference between the treatments is likely to lie, and hence the strength of the inference that can be drawn from the result

A statistically significant result may not be clinically significant. The results of intervention trials should be expressed in terms of the likely benefit an individual could expect (for example, the absolute risk reduction)

- The data (or, more accurately, the population from which the data are drawn) should be normally distributed. If they are not, non-itemmetric tests of correlation should be used instead.¹
- The two datasets should be independent (one should not automatically vary with the other). If they are not, a paired t test or other paired test should be used.
- Only a single pair of measurements should be made on each subject. If repeated measurements are made, analysis of variance should be used instead.²
- Every r value should be accompanied by a P value, which expresses how likely an association of this strength would be to have arisen by chance, or a confidence interval, which expresses the range within which the "true" r value is likely to lie.

Remember, too, that even if the r value is appropriate for a set of data, it does not tell you whether the relation, however strong, is causal (see below).



PETER BROWN

View larger version (149K):

[\[in this window\]](#)

[\[in a new window\]](#)

The term "regression" refers to a mathematical equation that allows one variable (the target variable) to be predicted from another (the independent variable). Regression, then, implies a direction of influence, although—as the next section will argue—it does not prove causality. In the case of multiple regression, a far more complex mathematical equation (which, thankfully, usually remains the secret of the computer that calculated it) allows the target variable to be predicted from two or more independent variables (often known as covariables).

The simplest regression equation, which you may remember from your school days, is $y=a+bx$, where y is the dependent variable (plotted on the vertical axis), x is the independent variable (plotted on the horizontal axis), and a is the y intercept. Not many biological variables can be predicted with such a simple equation. The weight of a group of people, for example, varies with their height, but not in a linear way. I am twice as tall as my son and three times his weight, but although I am four times as tall as my newborn nephew I am much more than six times his weight. Weight, in fact, probably varies more closely with the square of someone's height than with height itself (so a quadratic rather than a linear regression would probably be more appropriate).

Of course, even when the height-weight data fed into a computer are sufficient for it to calculate the regression equation that best predicts a person's weight from their height, your predictions would still be pretty poor since weight and height are not all that closely correlated. There are other things that influence weight in addition to height, and we could, to illustrate the principle of multiple regression, enter data on age, sex, daily calorie intake, and physical activity into the computer and ask it how much each of these covariables contributes to the overall equation (or model).

The elementary principles described here, particularly the criteria for the r value given above, should help you to spot whether correlation and regression are being used correctly in the **paper** you are **reading**. A more detailed discussion on the subject can be found elsewhere.^{[2](#) [3](#)}

Have assumptions been made about the nature and direction of causality?

Remember the ecological fallacy: just because a town has a large number of unemployed people and a very

high crime rate, it does not necessarily follow that the unemployed are committing the crimes. In other words, the presence of an association between A and B tells you nothing at all about either the presence or the direction of causality. To show that A has caused B (rather than B causing A, or A and B both being caused by C), you need more than a correlation coefficient. The [box](#) gives some criteria, originally developed by Sir Austin Bradford Hill, which should be met before assuming causality.⁴

Tests for causation⁴

- Is there evidence from true experiments in humans?
- Is the association strong?
- Is the association consistent from study to study?
- Is the temporal relation appropriate (did the postulated cause precede the postulated effect)?
- Is there a dose-response gradient (does more of the postulated effect follow more of the postulated cause)?
- Does the association make epidemiological sense?
- Does the association make biological sense?
- Is the association specific?
- Is the association analogous to a previously proved causal association?

Probability and confidence

Have "P values" been calculated and interpreted appropriately?

One of the first values a student of statistics learns to calculate is the P value—that is, the probability that any particular outcome would have arisen by chance. Standard scientific practice, which is entirely arbitrary, usually deems a P value of less than 1 in 20 (expressed as $P < 0.05$, and equivalent to a betting odds of 20 to 1) as "statistically significant" and a P value of less than 1 in 100 ($P < 0.01$) as "statistically highly significant."

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [Correlation, regression, and...](#)
 - [Probability and confidence](#)
- ▼ [The bottom line](#)
- ▼ [Summary](#)
- ▼ [References](#)

By definition, then, one chance association in 20 (this must be around one major published result per journal issue) will seem to be significant when it is not, and one in 100 will seem highly significant when it is really what my children call a "fluke." Hence, if you must analyse multiple outcomes from your data set, you need to make a correction to try to allow for this (usually achieved by the Bonferroni method^{5 6}).

A result in the statistically significant range ($P < 0.05$ or $P < 0.01$, depending on what is chosen as the cut off) suggests that the authors should reject the null hypothesis (the hypothesis that there is no real difference between two groups). But a P value in the non-significant range tells you that either there is no difference between the groups or that there were too few subjects to demonstrate such a difference if it existed—but it does not tell you which.

The P value has a further limitation. Guyatt and colleagues, in the first article of their "Basic Statistics for Clinicians" series on hypothesis testing using P values, conclude: "Why use a single cut off point [for statistical significance] when the choice of such point is arbitrary? Why make the question of whether a treatment is effective a dichotomy (a yes-no decision) when it would be more appropriate to view it as a continuum?"⁷ For a better assessment of the strength of evidence, we need confidence intervals.

Have confidence intervals been calculated, and do the authors' conclusions reflect them?

A confidence interval, which a good statistician can calculate on the result of just about any statistical test (the t test, the r value, the absolute risk reduction, the number needed to treat, and the sensitivity, specificity, and other key features of a diagnostic test), allows you to estimate for both "positive" trials (those that show a statistically significant difference between two arms of the trial) and "negative" ones (those that seem to show no difference), whether the strength of the evidence is strong or weak, and whether the study is definitive (obviates the need for further similar studies). The calculation and interpretation of confidence intervals have been covered elsewhere.⁸

If you repeated the same clinical trial hundreds of times, you would not get exactly the same result each time. But, on average, you would establish a particular level of difference (or lack of difference) between the two arms of the trial. In 90% of the trials the difference between two arms would lie within certain broad limits, and in 95% of the trials it would lie between certain, even broader, limits.

Now, if (as is usually the case) you conducted only one trial, how do you know how close the result is to the "real" difference between the groups? The answer is you don't. But by calculating, say, the 95% confidence interval around your result, you will be able to say that there is a 95% chance that the "real" difference lies between these two limits. The sentence to look for in a **paper** should **read** something like: "In a trial of the treatment of heart failure, 33% of the patients randomised to ACE inhibitors died, whereas 38% of those randomised to hydralazine and nitrates died. The point estimate of the difference between the groups [the best single estimate of the benefit in lives saved from the use of an ACE inhibitor] is 5%. The 95% confidence interval around this difference is -1.2% to 12%."

More likely, the results would be expressed in the following shorthand: "The ACE inhibitor group had a 5% (95% CI -1.2% to 12%) higher survival."

In this particular example, the 95% confidence interval overlaps zero difference and, if we were expressing the result as a dichotomy (that is, is the hypothesis "proved" or "disproved"?) we would classify it as a negative trial. Yet as Guyatt and colleagues argue, there probably is a real difference, and it probably lies closer to 5% than either -1.2% or 12%. A more useful conclusion from these results is that "all else being equal, an ACE inhibitor is the appropriate choice for patients with heart failure, but the strength of that inference is weak."⁹

Note that the larger the trial (or the larger the pooled results of several trials), the narrower the confidence interval—and, therefore, the more likely the result is to be definitive.

In interpreting "negative" trials, one important thing you need to know is whether a much larger trial would be likely to show a significant benefit. To determine this, look at the upper 95% confidence limit of the result. There is only one chance in 40 (that is, a 2½% chance, since the other 2½% of extreme results will lie below the lower 95% confidence limit) that the real result will be this much or more. Now ask yourself, "Would this level of difference be clinically important?" If not, you can classify the trial as not only negative but also definitive. If, on the other hand, the upper 95% confidence limit represented a clinically important level of difference between the groups, the trial may be negative but it is also non-definitive.

The use of confidence intervals is still relatively uncommon in medical **papers**. In one survey of 100 articles from three of North America's top journals (the *New England Journal of Medicine*, *Annals of Internal Medicine*, and the *Canadian Medical Association Journal*), only 43 reported any confidence intervals, whereas 66 gave a P value.⁷ An even smaller proportion of articles interpret their confidence intervals correctly. You should check carefully in the discussion section to see whether the authors have correctly concluded not only whether and to what extent their trial supported their hypothesis, but also whether any further studies need to be done.

▶ The bottom line

Have the authors expressed the effects of an intervention in terms of the likely benefit or harm which an individual patient can expect?

It is all very well to say that a particular intervention produces a "statistically significant difference" in outcome, but if I were being asked to take a new medicine I would want to know how much better my chances would be (in terms of any particular outcome) than they would be if I didn't take it. Four simple calculations (if you can add, subtract, multiply, and divide you will be able to follow this section) will enable you to answer this question objectively and in a way that means something to the non-statistician. These calculations are the relative risk reduction, the absolute risk reduction, the number needed to treat, and the odds ratio.

To illustrate these concepts, and to persuade you that you need to know about them, consider a survey which Tom Fahey and his colleagues conducted recently.¹⁰ They wrote to 182 board members of district health

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [Correlation, regression, and...](#)
- ▲ [Probability and confidence](#)
- The bottom line
- ▼ [Summary](#)
- ▼ [References](#)

authorities in England (all of whom would be in some way responsible for making important health service decisions), asking them which of four different rehabilitation programmes for heart attack victims they would prefer to fund:

Programme A reduced the rate of deaths by 20%;

Programme B produced an absolute reduction in deaths of 3%;

Programme C increased patients' survival rate from 84% to 87%;

Programme D meant that 31 people needed to enter the programme to avoid one death.

Let us continue with the example shown in table [1](#)), which Fahey and colleagues reproduced from a study by Salim Yusuf and colleagues.¹¹ I have expressed the figures as a two by two table giving details of which treatment the patients received in their randomised trial and whether they were dead or alive 10 years later.

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Table 1 Bottom line effects: treatment and outcome¹⁰

Simple mathematics tells you that patients receiving medical treatment have a chance of $404/1324=0.305$ or 30.5% of being dead at 10 years. Let us call this risk x . Patients randomised to coronary artery bypass grafting have a chance of $350/1325=0.264$ or 26.4% of being dead at 10 years. Let us call this risk y .

The relative risk of death—that is, the risk in surgically treated patients compared with medically treated controls—is y/x or $0.264/0.305=0.87$ (87%).

The relative risk reduction—that is, the amount by which the risk of death is reduced by the surgery—is $100\%-87\%$ ($1-y/x$)=13%.

The absolute risk reduction (or risk difference)—that is, the absolute amount by which surgical treatment reduces the risk of death at 10 years—is $30.5\%-26.4%=4.1\%$ (0.041).

The number needed to treat—how many patients need coronary artery bypass grafting in order to prevent, on average, one death after 10 years—is the reciprocal of the absolute risk reduction: $1/ARR=1/0.041=24$.

Yet another way of expressing the effect of treatment is the odds ratio. Look back at the two by two table and

you will see that the "odds" of dying compared with the odds of surviving for patients in the medical treatment group is $404/921=0.44$, and for patients in the surgical group is $350/974=0.36$. The ratio of these odds will be $0.36/0.44=0.82$.

The general formulas for calculating these "bottom line" effects of an intervention, taken from Sackett and colleagues' latest book,¹² are shown in the [box](#).

The outcome event can be desirable (cure, for example) or undesirable (an adverse drug reaction). In the latter case, it is semantically preferable to refer to numbers needed to harm and the relative or absolute increase in risk.

Calculating the "bottom line" effects on an intervention

Group	Outcome event		Total
	Yes	No	
Control group	a	b	a+b
Experimental group	c	d	c+d

Control event rate (CER)=risk of outcome event in control group= $a/(a+b)$

Experimental event rate (EER)=risk of outcome event in experimental group= $c/(c+d)$

Relative risk reduction (RRR)=(CER—EER)/CER

Absolute risk reduction (ARR)=CER—EER

Number needed to treat (NNT)= $1/ARR=1/(CER—EER)$

Odds ratio =

$$\frac{(\text{odds of outcome event } v \text{ odds of no event) in intervention group}}{(\text{odds of outcome event } v \text{ odds of no event) in control group}}$$

Summary

It is possible to be seriously misled by taking the statistical competence (and/or the intellectual honesty) of

authors for granted. Some common errors committed (deliberately or inadvertently) by the authors of **papers** are given in the final [box](#).

▲	Top
▲	Introduction
▲	Correlation, regression, and...
▲	Probability and confidence
▲	The bottom line
▪	Summary
▼	References

Ten ways to cheat on statistical tests when writing up results

- Throw all your data into a computer and report as significant any relation where $P < 0.05$
- If baseline differences between the groups favour the intervention group, remember not to adjust for them
- Do not test your data to see if they are normally distributed. If you do, you might get stuck with non-itemmetric tests, which aren't as much fun
- Ignore all withdrawals (drop outs) and non-responders, so the analysis only concerns subjects who fully complied with treatment
- Always assume that you can plot one set of data against another and calculate an "*r* value" (Pearson correlation coefficient), and assume that a "significant" *r* value proves causation
- If outliers (points which lie a long way from the others on your graph) are messing up your calculations, just rub them out. But if outliers are helping your case, even if they seem to be spurious results, leave them in
- If the confidence intervals of your result overlap zero difference between the groups, leave them out of your report. Better still, mention them briefly in the text but don't draw them in on the graph—and ignore them when drawing your conclusions
- If the difference between two groups becomes significant four and a half months into a six month trial, stop the trial and start writing up. Alternatively, if at six months the results are "nearly significant," extend the trial for another three weeks
- If your results prove uninteresting, ask the computer to go back and see if any particular subgroups behaved differently. You might find that your intervention worked after all in Chinese women aged 52-61
- If analysing your data the way you plan to does not give the result you wanted, run the figures through a selection of other tests

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.

Acknowledgements

I am grateful to Mr John Dobby for educating me on statistics and for repeatedly checking and amending this article. Responsibility for any errors is mine alone.

References

1. **Greenhalgh** T. Statistics for the non-statistician. I. Different types of data need different statistical tests. *BMJ* 1997;315:000-0.
2. Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 1987.
3. Guyatt G, Walter S, Shannon H, Cook D, Jaenschke R, Heddle, N. Basic statistics for clinicians: 4. Correlation and regression. *Can Med Assoc J* 1995;152:497-504.
4. Haines A. Multi-practice research: a cohort study. In: Jones R, Kinmonth AL, eds. *Critical reading for primary care*. Oxford: Oxford University Press, 1995:124. (Originally published as: Bradford Hill A. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295-300.)
5. Altman D. *Practical statistics for medical research*. London: Chapman and Hall, 1995:210-2.
6. Pocock SJ, Geller XPL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987;43:487-98.
7. Guyatt G, Jaenschke R, Heddle, N, Cook D, Shannon H, Walter S. Basic statistics for clinicians. 1. Hypothesis testing. *Can Med Assoc J* 1995;152:27-32.
8. Gardner MJ, Altman DG, eds. *Statistics with confidence: confidence intervals and statistical guidelines*. London: BMJ Books, 1989.
9. Guyatt G, Jaenschke R, Heddle, N, Cook D, Shannon H, Walter S. Basic statistics for clinicians. 2. Interpreting study results: confidence intervals. *Can Med Assoc J* 1995;152:169-73.
10. Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding the results of clinical trials and systematic reviews. *BMJ* 1995; 311:1056-60.
11. Yusuf S, Zucker D, Peduzzi P, Liher LD, Takaro T, Kennedy WJ, *et al*. Effect of coronary artery bypass surgery on survival: overview of ten year results form randomized trials by the coronary artery surgery

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [Correlation, regression, and...](#)
- ▲ [Probability and confidence](#)
- ▲ [The bottom line](#)
- ▲ [Summary](#)
- [References](#)

trials collaboration. *Lancet* 1994;344:563-70.

12. Sackett DL, Richardson WS, Rosenberg WMC, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. London: Churchill-Livingstone, 1996.

This article has been cited by other articles:

- Redmond, A. C., Keenan, A.-M. (2002). Understanding Statistics: Putting P-Values into Perspective. *J Am Podiatr Med Assoc* 92: 297-305 [[Abstract](#)] [[Full text](#)]
- Leung, W-C (2001). Balancing statistical and clinical significance in evaluating treatment effects. *Postgrad Med J* 77: 201-204 [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond](#) to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



PETER BROWN

[\[View larger version \(268K\)\]](#)

Table 1 Bottom line effects: treatment and outcome¹⁰

Treatment	Outcome at 10 years	
	Dead	Alive
Medical treatment (n=1325)	404	921
Coronary artery bypass grafting (n=1324)	350	974

BMJ 1997;315:480-483 (23 August)

Education and debate

How to read a paper: Papers that report drug trials

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ "Evidence" and marketing

If you prescribe drugs, the pharmaceutical industry is interested in you and is investing a staggering sum of money trying to influence you. The most effective way of changing the prescribing habits of a clinician is through personal representatives (known in Britain as "drug reps" and in North America as "detailers"), who travel round with a briefcase full of "evidence" in support of their wares.¹

- ▲ [Top](#)
- ["Evidence" and marketing](#)
- ▼ [Making decisions about treatment](#)
- ▼ [Surrogate end points](#)
- ▼ [How to get evidence...](#)
- ▼ [References](#)

Pharmaceutical "reps" do not tell nearly as many lies as they used to (drug marketing has become an altogether more sophisticated science), but they have been known to cultivate a shocking ignorance of basic epidemiology and clinical trial design when it suits them.² It often helps their case, for example, to present the results of uncontrolled trials and express them in terms of before and after differences in a particular outcome measure.³ The recent correspondence in the *Lancet* and *BMJ* on placebo effects should remind you why uncontrolled before and after studies are the stuff of teenage magazines, not hard science.^{4 5 6 7 8 9 10 11 12}

Making decisions about treatment

Sackett and colleagues have argued that before giving a drug to a patient the doctor should:

- ▲ [Top](#)
- ▲ ["Evidence" and marketing](#)
 - [Making decisions about treatment](#)
- ▼ [Surrogate end points](#)
- ▼ [How to get evidence...](#)
- ▼ [References](#)

Summary points

Pharmaceutical "reps" are now much more informative than they used to be, but they may show ignorance of basic epidemiology and clinical trial design

The value of a drug should be expressed in terms of safety, tolerability, efficacy, and price

The efficacy of a drug should ideally be measured in terms of clinical end points that are relevant to patients; if surrogate end points are used they should be valid

Promotional literature of low scientific validity (such as uncontrolled before and after trials) should not be allowed to influence practice

- identify, for this patient, the ultimate objective of treatment (cure, prevention of recurrence, limitation of functional disability, prevention of later complications, reassurance, palliation, relief of symptoms, etc);
- select the most appropriate treatment, using all available evidence (this includes considering whether the patient needs to take any drug at all); and
- specify the treatment target (to know when to stop treatment, change its intensity, or switch to some other treatment).¹³

For example, in treating high blood pressure, the doctor might decide that:

- the ultimate objective of treatment is to prevent (further) target organ damage to brain, eye, heart, kidney, etc (and thereby prevent death);
- the choice of specific treatment is between the various classes of antihypertensive drug selected on

the basis of randomised, placebo controlled and comitemtive trials—as well as non-drug treatments such as salt restriction; and

- the treatment target might be a phase V diastolic blood pressure (right arm, sitting) of less than 90 mm Hg, or as close to that as tolerable in the face of drug side effects.

If these three steps are not followed (as is often the case—for example in terminal care), therapeutic chaos can result.

▶ Surrogate end points

A surrogate end point may be defined as a variable which is relatively easily measured and which predicts a rare or distant outcome of either a toxic stimulus (such as a pollutant) or a therapeutic intervention (a drug, surgical procedure, piece of advice, etc) but which is not itself a direct measure of either harm or clinical benefit. The growing interest in surrogate end points in medical research, and particularly by the pharmaceutical industry, reflects two important features of their use:

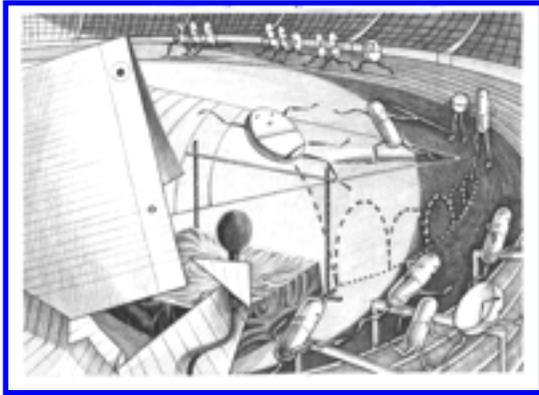
- they can considerably reduce the sample size, duration, and, therefore, cost, of clinical trials; and
- they can allow treatments to be assessed in situations where the use of primary outcomes would be excessively invasive or unethical.

In the evaluation of pharmaceutical products, commonly used surrogate end points include:

- pharmacokinetic measurements (for example, concentration-time curves of a drug or its active metabolite in the bloodstream);
- in vitro (laboratory) measures such as the mean inhibitory concentration of an antimicrobial against a bacterial culture on agar;
- macroscopic appearance of tissues (for example, gastric erosion seen at endoscopy);
- change in levels of (alleged) serum markers of disease (for example, prostate specific antigen¹⁴);

- ▲ [Top](#)
- ▲ ["Evidence" and marketing](#)
- ▲ [Making decisions about treatment](#)
 - [Surrogate end points](#)
- ▼ [How to get evidence...](#)
- ▼ [References](#)

- radiological appearance (for example, shadowing on a chest x ray film).



PETER BROWN

View larger version (138K):

[\[in this window\]](#)

[\[in a new window\]](#)

But surrogate end points have some drawbacks. Firstly, a change in the surrogate end point does not itself answer the essential preliminary questions: "what is the objective of treatment in this patient?" and "what, according to valid and reliable research studies, is the best available treatment for this condition?" Secondly, the surrogate end point may not closely reflect the treatment target—in other words, it may not be valid or reliable. Thirdly, overreliance on a single surrogate end point as a measure of therapeutic success usually reflects a narrow clinical perspective. Finally, surrogate end points are often developed in animal models of disease, since changes in a specific variable can be measured under controlled conditions in a well defined population. However, extrapolation of these findings to human disease is likely to be invalid.^{15 16 17}

The features of an ideal surrogate end point are shown in the [box](#). If the "rep" who is trying to persuade you of the value of the drug cannot justify the end points used, you should challenge him or her to produce additional evidence.

Features of the ideal surrogate end point

- The surrogate end point should be reliable, reproducible, clinically available, easily quantifiable, affordable, and show a "dose-response" effect (the higher the level of the surrogate end point, the greater the probability of disease)
- It should be a true predictor of disease (or risk of disease) and not merely express exposure to a covariable. The relation between the surrogate end point and the disease should have a biologically plausible explanation
- It should be sensitive—a "positive" result in the surrogate end point should pick up all or most patients at increased risk of adverse outcome
- It should be specific—a "negative" result should exclude all or most of those without increased risk of adverse outcome
- There should be a precise cut off between normal and abnormal values
- It should have an acceptable positive predictive value—a "positive" result should always or usually mean that the patient thus identified is at increased risk of adverse outcome
- It should have an acceptable negative predictive value—a "negative" result should always or usually mean that the patient thus identified is not at increased risk of adverse outcome
- It should be amenable to quality control monitoring
- Changes in the surrogate end point should rapidly and accurately reflect the response to treatment. In particular, levels should normalise in states of remission or cure

One important example of the invalid use of a surrogate end point is the CD4 cell count in monitoring progression to AIDS in HIV positive subjects. The CONCORDE trial was a randomised controlled trial comparing early and late start of treatment with zidovudine in patients who were HIV positive but clinically asymptomatic.¹⁸ Previous studies had shown that starting treatment early led to a slower decline in the CD4 cell count (a variable which had been shown to fall with the progression of AIDS), and it was assumed that a higher CD4 cell count would reflect improved chances of survival.

However, the CONCORDE trial showed that, although CD4 cell counts fell more slowly in the treatment group, the three year survival rates were identical in the two groups. This experience confirmed a warning that was issued earlier by authors suspicious of the validity of this end point.¹⁹ Subsequent

research in this field has attempted to identify a surrogate end point that correlates with real therapeutic benefit—that is, delayed progression of asymptomatic HIV infection to clinical AIDS, and longer survival time after the onset of AIDS.^{20 21} Using multiple regression analysis, investigators in the USA found that a combination of markers (percentage of CD4:C29 cells, degree of fatigue, age, and haemoglobin concentration) was the best predictor of progression.²⁰

Other examples of surrogate end points which have seriously misled researchers include ventricular premature beats as a predictor of death from serious cardiac arrhythmias,^{22 23} blood concentrations of antibiotics as a predictor of clinical cure of infection,²⁴ and plaques seen on magnetic resonance imaging in monitoring the progression of multiple sclerosis.²⁵

Before surrogate end points can be used in the marketing of pharmaceuticals, those in the industry must justify the utility of these measures by showing a plausible and consistent link between the end point and the development or progression of disease. It would be wrong to suggest that the pharmaceutical industry develops surrogate end points with the deliberate intention to mislead the licensing authorities and health professionals. However, the industry does, theoretically, have a vested interest in overstating its case on the significance of these end points. Given that much of the data relating to the validation of surrogate end points are not currently presented in published clinical **papers**, and that the development of such markers is often a lengthy and expensive process, one author has suggested setting up a data archive that would pool data across studies.²⁶

▶ How to get evidence out of a drug rep

Any doctor who has ever given an audience to a "rep" who is selling a non-steroidal anti-inflammatory drug will recognise the argument that "this NSAID reduces the incidence of gastric erosion in comparison to its competitors." The question to ask the rep is not "what is the incidence of endoscopic signs of gastric erosion in volunteers who take this drug?" but "what is the incidence in clinical practice of potentially life threatening gastric bleeding in patients who take this drug?" Other questions, collated from recommendations in *Drug and Therapeutics Bulletin*²⁷ and other sources,^{1 3} are listed below.

- ▲ [Top](#)
- ▲ ["Evidence" and marketing](#)
- ▲ [Making decisions about treatment](#)
- ▲ [Surrogate end points](#)
 - [How to get evidence...](#)
- ▼ [References](#)

- See representatives only by appointment. Choose to see only those whose product interests you, and confine the interview to that product
- Take charge of the interview. Do not hear out a rehearsed sales routine but ask directly for the

information below

- Request independent published evidence from reputable, peer reviewed journals
- Do not look at promotional brochures, which may contain unpublished material, misleading graphs, and selective quotations
- Ignore anecdotal "evidence," such as the fact that a medical celebrity is prescribing the product
- Using the STEP acronym, ask for evidence in four specific areas:

Safety—the likelihood of long term or serious side effects caused by the drug (remember that rare but serious adverse reactions to new drugs may be poorly documented)

Tolerability—best measured by comparing the pooled withdrawal rates between the drug and its most significant competitor

Efficacy—the most relevant dimension is how the product compares with your current favourite

Price—should take into account indirect as well as direct costs

- Evaluate the evidence stringently, paying particular attention to the power (sample size) and methodological quality of clinical trials, and the use of surrogate end points. Do not accept theoretical arguments in the drug's favour ("longer half life," for example) without direct evidence that this translates into clinical benefit
- Do not accept the newness of a product as an argument for changing to it. Indeed, there are good scientific arguments for doing the opposite²⁸
- Decline to try the product via starter packs or by participating in small scale, uncontrolled "research" studies
- Record in writing the content of the interview and return to these notes if the "rep" requests another audience

Checklist for evaluating information provided by a drug company

- Does this material cover a subject which interests me and is clinically important in my practice?
- Has this material been published in independent peer reviewed journals? Has any significant evidence been omitted from this presentation or withheld from publication?
- Does the material include high-level evidence such as systematic reviews, meta-analyses, or double-blind randomised controlled trials against the drug's closest competitor given at optimal dosage?
- Have the trials or reviews addressed a clearly focused, important and answerable clinical question which reflects a problem of relevance to patients? Do they provide evidence on safety, tolerability, efficacy and price?
- Has each trial or meta-analysis defined the condition to be treated, the patients to be included, the interventions to be compared and the outcomes to be examined?
- Does the material provide direct evidence that the drug will help my patients live a longer, healthier, more productive, and symptom-free life?
- If a surrogate outcome measure has been used, what is the evidence that it is reliable, reproducible, sensitive, specific, a true predictor of disease, and rapidly reflects the response to therapy?
- Do trial results indicate whether (and how) the effectiveness of the treatments differed and whether there was a difference in the type or frequency of adverse reactions? Are the results expressed in terms of numbers needed to treat, and are they clinically as well as statistically significant?
- If large amounts of material have been provided by the representative, which three **papers** provide the strongest evidence for the company's claims?

In conclusion, it is often more difficult than you are being led to believe to weigh the potential benefits of a drug against its risks to the patient and cost to the taxpayer.²⁹ The difference between the science of critical appraisal and the pharmaceutical industry's well rehearsed tactics of marketing and persuasion should be borne in mind when you are considering "evidence" presented by those with a commercial conflict of interest.

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 Publishing Group: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 for UK members, £14.95 for non-members.

Acknowledgements

I am grateful to Dr Andrew Herxheimer for advice on this article.

References

1. Shaughnessy AF, Slawson DC. Pharmaceutical representatives. *BMJ* 1996;312:1494-5. [\[Full Text\]](#)
2. Bardelay D. Visits from medical representatives: fine principles, poor practice. *Prescrire International* 1995;4:120-2.
3. Bero LA, Rennie D. Influences on the quality of published drug studies. *Int J Health Technology Assessment* 1996;12:209-37.
4. Kleijnen J, de Craen AJ, van Everdingen J, Krol L. Placebo effect in double-blind clinical trials: a review of interactions with medications. *Lancet* 1994;344:1347-9. [\[Medline\]](#)
5. Joyce CR. Placebo and complementary medicine. *Lancet* 1994;344:1279-81. [\[Medline\]](#)
6. Laporte JR, Figueras A. Placebo effects in psychiatry. *Lancet* 1994;344:1206-9. [\[Medline\]](#)
7. Johnson AG. Surgery as a placebo. *Lancet* 1994;344:1140-2. [\[Medline\]](#)
8. Thomas KB. The placebo in general practice. *Lancet* 1994;344:1066-7. [\[Medline\]](#)
9. Chaput de Saintonge DM, Herxheimer A. Harnessing placebo effects in health care. *Lancet* 1994;344:995-8.
10. Gotzsche PC. Is there logic in the placebo? *Lancet* 1994;344:925-6. [\[Medline\]](#)
11. Rothman KJ. Placebo mania. *BMJ* 1996;313:3-4. [\[Full Text\]](#)
12. McQuay H, Moore A, Double DB, Georgiou A, Korkia P. Placebo mania. *BMJ* 1996;313:1008-9. [\[Full Text\]](#)
13. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology—a basic science for clinical medicine*. London, Little, Brown, 1991:187-248.

- ▲ [Top](#)
- ▲ ["Evidence" and marketing](#)
- ▲ [Making decisions about treatment](#)
- ▲ [Surrogate end points](#)
- ▲ [How to get evidence...](#)
- **References**

14. Bostwick DG, Burke HB, Wheeler TM, Chung LW, Bookstein R, Pretlow TG, et al. The most promising surrogate endpoint biomarkers for screening candidate chemopreventive compounds for prostatic adenocarcinoma in short-term Phase II clinical trials. *J Cell Biochem* 1994;56(suppl 19):283-9.
15. Gøtzsche P, Liberati A, Torri V, Rosetti L. Beware of surrogate outcome measures. *Int J Health Technology Assessment* 1996;12:238-46.
16. Lipkin M. Summary of recommendations for colonic biomarker studies of candidate chemopreventive compounds in Phase II clinical trials. *J Cell Biochem* 1994;56(suppl 19):94-8.
17. Kimbrough RD. Determining acceptable risks: experimental and epidemiological issues. *Clin Chem* 1994;40:1448-53.
18. CONCORDE Co-ordinating Committee. CONCORDE MRC/ANRS randomised double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet* 1994;343:871-81. [[Medline](#)]
19. Jacobson MA, Bacchetti P, Kolokathis A et al. Surrogate markers for survival in patients with AIDS and AIDS related complex treated with zidovudine. *BMJ* 1991;302:73-8. [[Medline](#)]
20. Blatt SP, McCarthy WF, Bucko-Krasnicka B, Melcher GP, Boswell RN, Dolan J, et al. Multivariate models for predicting progression to AIDS and survival in HIV-infected patients. *J Infect Dis* 1995;171:837-44.
21. Tsoukas CM, Bernard NF. Markers predicting progression of HIV-related disease. *Clin Microbiol Rev* 1994;7:14-28.
22. Epstein AE, Hallstrom AO, Rogers WJ, Liebson PR, Seals AA, Anderson JL, et al. Mortality following ventricular arrhythmia suppression by encainide, flecainide and moricizine after myocardial infarction. *JAMA* 1993; 270, 2451-55.
23. Lipicky RJ, Packer M. Role of surrogate endpoints in the evaluation of drugs for heart failure. *J Am Coll Cardiol* 1993;22(suppl A):179-84.
24. Hyatt JM, McKinnon PS, Zimmer GS, Schentag JJ. The importance of pharmacokinetic/pharmacodynamic surrogate markers to outcome. Focus on antibacterial agents. *Clin Pharmacokinetics* 1995;28:143-60.
25. Interferon beta-1b—hope or hype? *Drug Ther Bull* 1996;34:9-11. [[Medline](#)]
26. Aickin M. If there is gold in the labelling index hills, are we digging in the right place? *J Cell Biochem* 1994;56(suppl 19):91-3.
27. Getting good value from drug reps. *Drug Ther Bull* 1983;21:13-5. [[Medline](#)]
28. Ferner RE. Newly licensed drugs. *BMJ* 1996;313:1157-8. [[Full Text](#)]
29. Risk:benefit analysis of drugs in practice. *Drug Ther Bull* 1995;33:33-5. [[Medline](#)]

This article has been cited by other articles:

- McCormack, J., **Greenhalgh**, T. (2001). Seeing what you want to see in randomized controlled trials: versions and perversions of UK Prospective Diabetes Study data. *eWJM* 174: 123-127

[\[Full text\]](#)

- McCormack, J., **Greenhalgh**, T. (2000). Seeing what you want to see in randomised controlled trials: versions and perversions of UKPDS data. *BMJ* 320: 1720-1723 [\[Full text\]](#)
- Behrens, R. H, Erny, S., Maradit, H., Houston, S., Keystone, J. S, Kain, K. C, Croft, A., Garner, P. (1998). Mefloquine to prevent malaria. *BMJ* 316: 1980a-1980 [\[Full text\]](#)
- Goran, M. I., Gower, B. A., Nagy, T. R., Johnson, R. K. (1998). Developmental Changes in Energy Expenditure and Physical Activity in Children: Evidence for a Decline in Physical Activity in Girls Before Puberty. *Pediatrics* 101: 887-891 [\[Abstract\]](#) [\[Full text\]](#)
- Croft, A., Garner, P. (1997). Mefloquine to prevent malaria: a systematic review of trials. *BMJ* 315: 1412-1416 [\[Abstract\]](#) [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

Related letters in BMJ:

Mefloquine to prevent malaria

Ron H Behrens, Samuel Erny, Hilal Maradit, Stan Houston, Jay S Keystone, Kevin C Kain, Ashley Croft, and Paul Garner
BMJ 1998 316: 1980. [\[Letter\]](#)

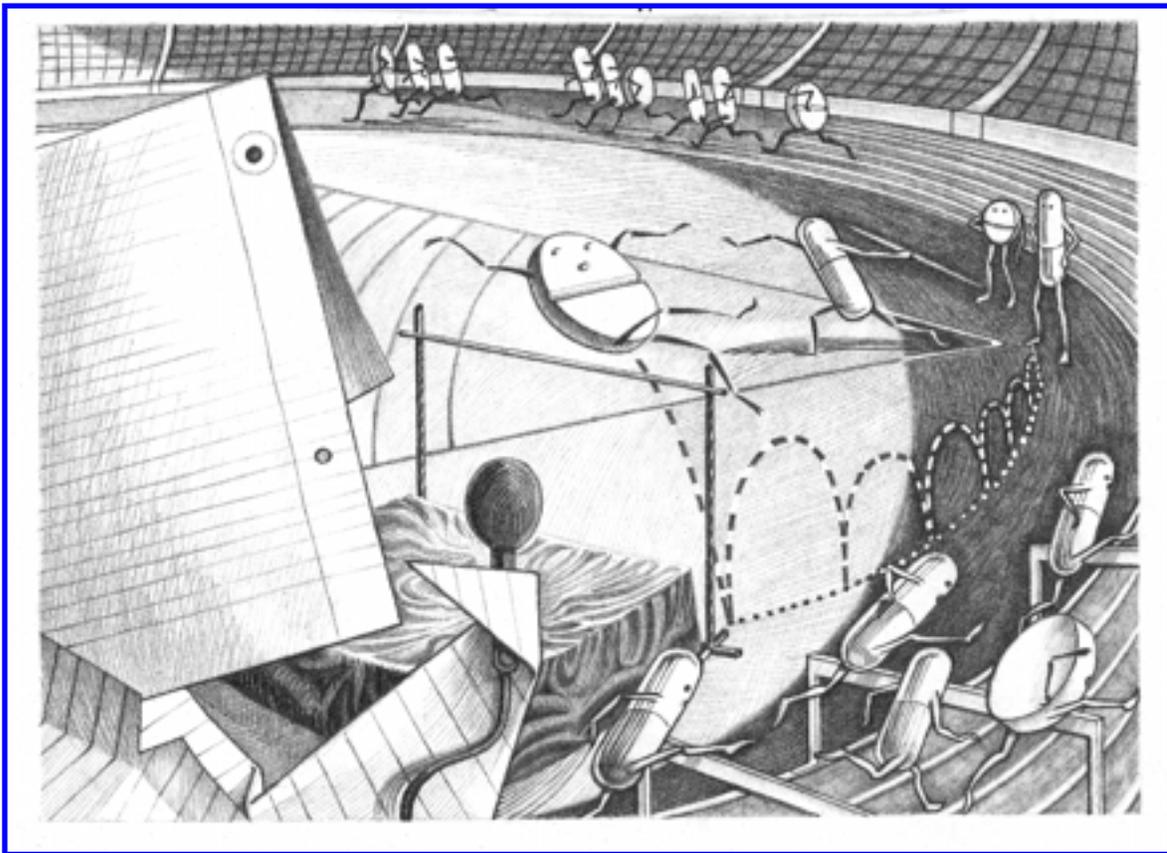
[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



PETER BROWN

[\[View larger version \(176K\)\]](#)

BMJ 1997;315:540-543 (30 August)

Education and debate

How to **read** a **paper**: **Papers** that report diagnostic or screening tests

Trisha **Greenhalgh**, *senior lecturer*^a

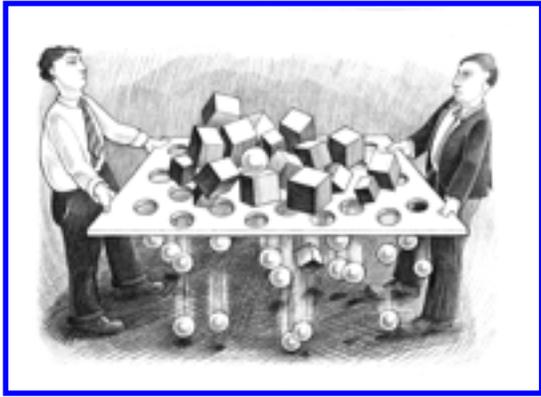
^a Unit for Evidence-Based Practice and Policy Department of Primary Care and Population Sciences University College London Medical School/Royal Free Hospital School of Medicine Whittington Hospital London N19 5NF p.greenhalgh@ucl.ac.uk

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ **Read** responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ Ten men in the dock

If you are new to the concept of validating diagnostic tests, the following example may help you. Ten men are awaiting trial for murder. Only three of them actually committed a murder; the seven others are innocent of any crime. A jury hears each case and finds six of the men guilty of murder. Two of the convicted are true murderers. Four men are wrongly imprisoned. One murderer walks free.

- ▲ [Top](#)
- [Ten men in the...](#)
- ▼ [Validating tests against a...](#)
- ▼ [Does the **paper** validate...](#)
- ▼ [A note on likelihood...](#)
- ▼ [References](#)



PETER BROWN

View larger version (109K):

[\[in this window\]](#)

[\[in a new window\]](#)

This information can be expressed in what is known as a two by two table (table [1](#)). Note that the "truth" (whether or not the men really committed a murder) is expressed along the horizontal title row, whereas the jury's verdict (which may or may not reflect the truth) is expressed down the vertical row.

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Table 1 Two by two table showing outcome of trial for 10 men accused of murder

These figures, if they are typical, reflect several features of this particular jury:

- the jury correctly identifies two in every three true murderers;
- it correctly acquits three out of every seven innocent people;
- if this jury has found a person guilty, there is still only a one in three chance that they are actually a murderer;
- if this jury found a person innocent, he or she has a three in four chance of actually being innocent; and
- in five cases out of every 10 the jury gets it right.

These five features constitute, respectively, the sensitivity, specificity, positive predictive value, negative predictive value, and accuracy of this jury's performance. The rest of this article considers these five features applied to diagnostic (or screening) tests when compared with a "true" diagnosis or gold standard. A sixth feature—the likelihood ratio—is introduced at the end of the article.

▶ Validating tests against a gold standard

Our window cleaner told me that he had been feeling thirsty recently and had asked his general practitioner to be tested for diabetes, which runs in his family. The nurse in his surgery had asked him to produce a urine specimen and dipped a stick in it. The stick stayed green, which meant, apparently, that there was no sugar in his urine. This, the nurse had said, meant that he did not have diabetes.

- ▲ [Top](#)
- ▲ [Ten men in the...](#)
 - [Validating tests against a...](#)
- ▼ [Does the paper validate...](#)
- ▼ [A note on likelihood...](#)
- ▼ [References](#)

Summary points

New tests should be validated by comparison against an established gold standard in an appropriate spectrum of subjects

Diagnostic tests are seldom 100% accurate (false positives and false negatives will occur)

A test is valid if it detects most people with the target disorder (high sensitivity) and excludes most people without the disorder (high specificity), and if a positive test usually indicates that the disorder is present (high positive predictive value)

The best measure of the usefulness of a test is probably the likelihood ratio—how much more likely a positive test is to be found in someone with, as opposed to without, the disorder

I had trouble explaining that the result did not necessarily mean this, any more than a guilty verdict necessarily makes someone a murderer. The definition of diabetes, according to the World Health Organisation, is a blood glucose level above 8 mmol/l in the fasting state, or above 11 mmol/l two hours after a 100 g oral glucose load, on one occasion if the patient has symptoms and on two occasions if he or she does not.¹ These stringent criteria can be termed the gold standard for diagnosing diabetes (although

purists have challenged this notion²).

The dipstick test, however, has some distinct practical advantages over the fullblown glucose tolerance test. To assess objectively just how useful the dipstick test for diabetes is, we would need to select a sample of people (say 100) and do two tests on each of them: the urine test (screening test) and a standard glucose tolerance test (gold standard). We could then see, for each person, whether the result of the screening test matched the gold standard (see table [2](#)). Such an exercise is known as a validation study.

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Table 2 Two by two table notation for expressing the results of validation study for diagnostic or screening test

The validity of urine testing for glucose in diagnosing diabetes has been looked at by Andersson and colleagues,³ whose data I have adapted for use (expressed as a proportion of 1000 subjects tested) in table [3](#).

View this table:

[\[in this window\]](#)

[\[in a new window\]](#)

Table 3 Two by two table showing results of validation study of urine glucose testing for diabetes against gold standard³

From the calculations of important features of the urine dipstick test for diabetes ([box](#)), you can see why I did not share the window cleaner's assurance that he did not have diabetes. A positive urine glucose test is only 22% sensitive, which means that the test misses nearly four fifths of people who have diabetes. In the presence of classical symptoms and a family history, the window cleaner's baseline chances (pretest likelihood) of having the condition are pretty high and is reduced to only about four fifths of this (the negative likelihood ratio, 0.78; see below) after a single negative urine test. This man clearly needs to undergo a more definitive test.

Features of diagnostic test that can be calculated by comparison with gold standard in validation study

Feature of the test	Alternative name	Question addressed	Formula (see table 2)
Sensitivity	True positive rate (positive in disease)	How good is this test at picking up people who have the condition?	$a / (a+c)$
Specificity	True negative rate (negative in health)	How good is this test at correctly excluding people without the condition?	$d / (b+d)$
Positive predictive value	Post-test probability of a positive test	If a person tests positive, what is the probability that he or she has the condition?	$a / (a+b)$
Negative predictive value	Post-test probability of a negative test	If a person tests negative, what is the probability that he or she does not have the condition?	$d / (c+d)$
Accuracy	—	What proportion of all tests have given the correct result? (true positives and true negatives as a proportion of all results)	$(a+d) / (a+b+c+d)$
Likelihood ratio of a positive test	—	How much more likely is a positive test to be found in a person with the condition than in a person without it?	$\text{sensitivity} / (1-\text{specificity})$

Likelihood ratio of a —
negative test

How much more (1-sensitivity)/specificity?
likely is a negative
test to be found in a
person without the
condition than in a
person with it

▶ Does the **paper** validate the test?

The 10 questions below can be asked about a **paper** that claims to validate a diagnostic or screening test. In preparing these tips, I have drawn on several sources.[4](#) [5](#) [6](#) [7](#) [8](#)

Question 1: Is this test potentially relevant to my practice?

Sackett and colleagues call this the utility of the test.[6](#) Even if this test were 100% valid, accurate, and reliable, would it help me? Would it identify a treatable disorder? If so, would I use it in preference to the test I use now? Could I (or my patients or the taxpayer) afford it? Would my patients consent to it? Would it change the probabilities for competing diagnoses sufficiently for me to alter my treatment plan?

Question 2: Has the test been compared with a true gold standard?

You need to ask, firstly, whether the test has been compared with anything at all. Assuming that a "gold standard" test has been used, you should verify that it merits the description, perhaps by using the questions listed in question 1. For many conditions, there is no gold standard diagnostic test. Unsurprisingly, these tend to be the conditions for which new tests are most actively sought. Hence, the authors of such **papers** may need to develop and justify a combination of criteria against which the new test is to be assessed. One specific point to check is that the test being validated in the **paper** is not being used to define the gold standard.

Question 3: Did this validation study include an appropriate spectrum of subjects?

Although few investigators would be naive enough to select only, say, healthy male medical students for their validation study, only 27% of published studies explicitly define the spectrum of subjects tested in terms of age, sex, symptoms or disease severity, and specific eligibility criteria.[7](#) Importantly, the test should be verified on a population which includes mild and severe disease, treated and untreated subjects, and those with different but commonly confused conditions.[6](#)

- ▲ [Top](#)
- ▲ [Ten men in the...](#)
- ▲ [Validating tests against a...](#)
- [Does the **paper** validate...](#)
- ▼ [A note on likelihood...](#)
- ▼ [References](#)

Calculating the important features of screening test

Feature	Formula	Data (see table 3)	Value
Sensitivity	$a / (a+c)$	6/27	22.2%
Specificity	$d / (b+d)$	966/973	99.3%
Positive predictive value	$a / (a+b)$	6/13	46.2%
Negative predictive value	$d / (c+d)$	966/973	97.8%
Accuracy	$(a+d) / (a+b+c+d)$	972/1000	97.2%
Likelihood ratio:			
Positive test	Sensitivity/ (1-specificity)	22.2/0.7	32
Negative test	(1-sensitivity)/specificity	77.8/99.	0.783

Although the sensitivity and specificity of a test are virtually constant whatever the prevalence of the condition, the positive and negative predictive values depend crucially on prevalence. This is why general practitioners are sceptical of the utility of tests developed exclusively in a secondary care population, and why a good diagnostic test is not necessarily a good screening test.

Question 4: Has workup bias been avoided?

This is easy to check. It simply means, "Did everyone who got the new diagnostic test also get the gold standard, and vice versa?" There is clearly a potential bias in studies where the gold standard test is performed only on people who have **alredy** tested positive for the test being validated.⁷

Question 5: Has expectation bias been avoided?

Expectation bias occurs when pathologists and others who interpret diagnostic specimens are subconsciously influenced by the knowledge of the particular features of the case—for example, the presence of chest pain when interpreting an electrocardiogram. In the context of validating diagnostic tests against a gold standard, all such assessments should be "blind."

Question 6: Was the test shown to be reproducible?

If the same observer performs the same test on two occasions on a subject whose characteristics have not changed, they will get different results in a proportion of cases. Similarly, it is important to confirm that reproducibility between different observers is at an acceptable level.⁹

Question 7: What are the features of the test as derived from this validation study?

All the above standards could have been met, but the test might still be worthless because the sensitivity, specificity, and other crucial features of the test are too low—that is, the test is not valid. What counts as acceptable depends on the condition being screened for. Few of us would quibble about a test for colour blindness that was 95% sensitive and 80% specific, but nobody ever died of colour blindness. The Guthrie heel-prick screening test for congenital hypothyroidism, performed on all babies in Britain soon after birth, is over 99% sensitive but has a positive predictive value of only 6% (it picks up almost all babies with the condition at the expense of a high false positive rate),¹⁰ and rightly so. It is more important to pick up every baby with this treatable condition who would otherwise develop severe mental handicap than to save hundreds the minor stress of a repeat blood test.

Question 8: Were confidence intervals given?

A confidence interval, which can be calculated for virtually every numerical aspect of a set of results, expresses the possible range of results within which the true value will probably lie. If the jury in the first example had found just one more murderer not guilty, the sensitivity of its verdict would have gone down from 67% to 33%, and the positive predictive value of the verdict from 33% to 20%. This enormous (and quite unacceptable) sensitivity to a single case decision is, of course, because we validated the jury's performance on only 10 cases. The larger the sample, the narrower the confidence interval, so it is particularly important to look for confidence intervals if the **paper** you are **reading** reports a study on a relatively small sample.¹¹

Question 9: Has a sensible "normal range" been derived?

If the test gives non-dichotomous (continuous) results—that is, if it gives a numerical value rather than a yes/no result—someone will have to say what values count as abnormal. Defining relative and absolute danger zones for a continuous variable (such as blood pressure) is a complex science, which should take into account the actual likelihood of the adverse outcome which the proposed treatment aims to prevent. This process is made considerably more objective by the use of likelihood ratios (see below).

Question 10: Has this test been placed in the context of other potential tests in the diagnostic sequence?

In general, we treat high blood pressure simply on the basis of a series of resting blood pressure **readings**. Compare this with the sequence we use to diagnose coronary artery stenosis. Firstly, we select patients with a typical history of effort angina. Next, we usually do a resting electrocardiogram, an exercise electrocardiogram, and, in some cases, a radionuclide scan of the heart. Most patients come to a coronary angiogram only after they have produced an abnormal result on these preliminary tests.

If you sent 100 ordinary people for a coronary angiogram, the test might show very different positive and negative predictive values (and even different sensitivity and specificity) than it did in the ill population on which it was originally validated. This means that the various aspects of validity of the coronary angiogram as a diagnostic test are virtually meaningless unless these figures are expressed in terms of

what they contribute to the overall diagnostic work up.

A note on likelihood ratios

Question 9 above described the problem of defining a normal range for a continuous variable. In such circumstances, it can be preferable to express the test result not as "normal" or "abnormal" but in terms of the actual chances of a patient having the target disorder if the test result reaches a particular level. Take, for example, the use of the prostate specific antigen (PSA) test to screen for prostate cancer. Most men will have some detectable antigen in their blood (say, 0.5 ng/ml), and most of those with advanced prostate cancer will have high concentrations (above about 20 ng/ml). But a concentration of, say, 7.4 ng/ml may be found either in a perfectly normal man or in someone with early cancer. There simply is not a clean cutoff between normal and abnormal.¹²

We can, however, use the results of a validation study of this test against a gold standard for prostate cancer (say a biopsy of the prostate gland) to draw up a whole series of two by two tables. Each table would use a different definition of an abnormal test result to classify patients as "normal" or "abnormal." From these tables, we could generate different likelihood ratios associated with an antigen concentration above each different cutoff point. When faced with a test result in the "grey zone" we would at least be able to say, "This test has not proved that the patient has prostate cancer, but it has increased [or decreased] the odds of that diagnosis by a factor of x ."

The likelihood ratio thus has enormous practical value, and it is becoming the preferred way of expressing and comparing the usefulness of different tests.⁶ For example, if a person enters my consulting room with no symptoms at all, I know that they have a 5% chance of having iron deficiency anaemia, since I know that one person in 20 in the population has this condition (in the language of diagnostic tests, the pretest probability of anaemia is 0.05).¹³

- ▲ [Top](#)
- ▲ [Ten men in the...](#)
- ▲ [Validating tests against a...](#)
- ▲ [Does the paper validate...](#)
 - [A note on likelihood...](#)
- ▼ [References](#)

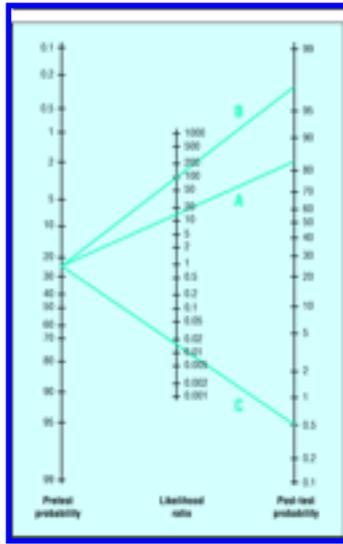


Fig 1 Use of likelihood ratios to calculate post-test probability of someone being a smoker⁶

View larger version (19K):

[\[in this window\]](#)

[\[in a new window\]](#)

Now, if I do a diagnostic test for anaemia, the serum ferritin concentration, the result will usually make the diagnosis of anaemia either more or less likely. A moderately reduced serum ferritin concentration (between 18 and 45 $\mu\text{g/l}$) has a likelihood ratio of 3, so the chances of a patient with this result having iron deficiency anaemia is 0.05×3 —or 0.15 (15%). This value is known as the post-test probability of the serum ferritin test. The likelihood ratio of a very low serum ferritin concentration (below 18 $\mu\text{g/l}$) is 41, making the chances of iron deficiency anaemia in a patient with this result greater than unity. On the other hand, a very high concentration (above 100 $\mu\text{g/l}$; likelihood ratio 0.13) would reduce the chances of the patient being anaemic from 5% to less than 1%.¹³

Figure 1 shows a nomogram, adapted by Sackett and colleagues from an original **paper** by Fagan,¹⁴ for working out post-test probabilities when the pretest probability (prevalence) and likelihood ratio for the test are known. The lines A, B, and C, drawn from a pretest probability of 25% (the prevalence of smoking among British adults), are the trajectories through likelihood ratios of 15, 100, and 0.015, respectively—three different tests for detecting whether someone is a smoker.¹⁵ Actually, test C detects whether the person is a non-smoker, since a positive result in this test leads to a post-test probability of only 0.5%.

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 Publishing Group: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

Acknowledgements

Thanks to Dr Sarah Walters and Dr Jonathan Elford for advice, and in particular to Dr Walters for the jury example.

References

1. WHO Study Group. Diabetes mellitus. *WHO Tech Report Ser* 1985;No 727.
2. McCance DR, Hanson RL, Charles M-A, Jacobsson LTH, Pettitt DJ, Bennett PH, et al. Comparison of tests for glycated haemoglobin and fasting and two-hour plasma glucose concentrations as diagnostic measures for diabetes. *BMJ* 1994;308:1323-8. [[Abstract/Full Text](#)]
3. Andersson DKG, Lundblad E, Svardsudd K. A model for early diagnosis of type 2 diabetes mellitus in primary health care. *Diabet Med* 1993;10:167-73.
4. 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](#)]
5. 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.
6. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology—a basic science for clinical medicine*. London: Little, Brown, 1991:51-68.
7. **Read** MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA* 1995;274:645-51. [[Medline](#)]
8. Mant D. Testing a test: three critical steps. In: Jones R, Kinmonth A-L, eds. *Critical reading for primary care*. Oxford: Oxford University Press, 1995:183-90.
9. Bush B, Shaw S, Cleary P, Delbanco TL, Aronson MD. Screening for alcohol abuse using the CAGE questionnaire. *Am J Med* 1987;82:231-6.

- ▲ [Top](#)
- ▲ [Ten men in the...](#)
- ▲ [Validating tests against a...](#)
- ▲ [Does the paper validate...](#)
- ▲ [A note on likelihood...](#)
- [References](#)

10. Verkerk PH, Derksen-Lubsen G, Vulsma T, Loeber JG, de Vijlder JJ, Verbrugge HP. Evaluation of a decade of neonatal screening for congenital hypothyroidism in the Netherlands. *Ned Tijdschr Geneesk* 1993;137:2199-205.
11. Gardner MJ, Altman DG, eds. *Statistics with confidence: confidence intervals and statistical guidelines*. London: BMJ Books, 1989.
12. Catalona WJ, Hudson MA, Scardino PT, Richie JP, Ahmann FR, Flanigan RC, et al. Selection of optimal prostate specific antigen cutoffs for early diagnosis of prostate cancer: receiver operator characteristic curves. *J Urol* 1994;152:2037-42.
13. Guyatt GH, Patterson C, Ali M, Singer J, Levine M, Turpie I, Meyer R. Diagnosis of iron deficiency anaemia in the elderly. *Am J Med* 1990;88:205-9.
14. Fagan TJ. Nomogram for Bayes' theorem. *N Engl J Med* 1975;293:257-61. [[Medline](#)]
15. How good is that test—using the result. *Bandolier* 1996;3:6-8.

This article has been cited by other articles:

- Fallis, D., Fricke, M. (2002). Indicators of Accuracy of Consumer Health Information on the Internet: A Study of Indicators Relating to Information for Managing Fever in Children in the Home. *J. Am. Med. Inform. Assoc.* 9: 73-79 [[Abstract](#)] [[Full text](#)]
- Lysakowski, C., Walder, B., Costanza, M. C., Tramer, M. R. (2001). Transcranial Doppler Versus Angiography in Patients With Vasospasm due to a Ruptured Cerebral Aneurysm: A Systematic Review. *Stroke* 32: 2292-2298 [[Abstract](#)] [[Full text](#)]
- McQueen, M. J. (2001). Overview of Evidence-based Medicine: Challenges for Evidence-based Laboratory Medicine. *Clin Chem* 47: 1536-1546 [[Abstract](#)] [[Full text](#)]
- Holvoet, P., Mertens, A., Verhamme, P., Bogaerts, K., Beyens, G., Verhaeghe, R., Collen, D., Muls, E., Van de Werf, F. (2001). Circulating Oxidized LDL Is a Useful Marker for Identifying Patients With Coronary Artery Disease. *Arterioscler Thromb* 21: 844-848 [[Abstract](#)] [[Full text](#)]
- Kennedy, C. R, HALL, D., DAVIS, A. (2000). Current topic: Neonatal screening for hearing impairment. *Arch. Dis. Child.* 83: 377-383 [[Full text](#)]
- Price, C. P. (2000). Evidence-based Laboratory Medicine: Supporting Decision-Making. *Clin Chem* 46: 1041-1050 [[Abstract](#)] [[Full text](#)]
- Morgan, J. F, Reid, F., Lacey, J H. (2000). The SCOFF questionnaire: a new screening tool for eating disorders. *eWJM* 172: 164-165 [[Full text](#)]
- Santos-Gomes, G., Gomes-Pereira, S., Campino, L., Araújo, M. D. A., Abranches, P. (2000). Performance of Immunoblotting in Diagnosis of Visceral Leishmaniasis in Human

- ▶ [Email this article to a friend](#)
- ▶ [Respond](#) to this article
- ▶ [Read](#) responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by: [Greenhalgh, T.](#)
- ▶ Alert me when: [New articles cite this article](#)

Immunodeficiency Virus-Leishmania sp.-Coinfected Patients. *J. Clin. Microbiol.* 38: 175-178

[\[Abstract\]](#) [\[Full text\]](#)

- Morgan, J. F, Reid, F., Lacey, J H. (1999). The SCOFF questionnaire: assessment of a new screening tool for eating disorders. *BMJ* 319: 1467-1468 [\[Full text\]](#)
- KÜNZLI, N., STUTZ, E. Z., PERRUCHOUD, A. P., BRÄNDLI, O., TSCHOPP, J.-M., BOLOGNINI, G., KARRER, W., SCHINDLER, C., ACKERMANN-LIEBRICH, U., LEUENBERGER, P. (1999). Peak Flow Variability in the SAPALDIA Study and Its Validity in Screening for Asthma-related Conditions. *Am J Respir Crit Care Med* 160: 427-434
[\[Abstract\]](#) [\[Full text\]](#)
- Henderson, A R. (1998). Test accuracy is example of redundant information. *BMJ* 316: 312b-312
[\[Full text\]](#)

Rapid Responses:

Read all [Rapid Responses](#)

Neonatal screening in UK has high sensitivity

Michael Addison

bmj.com, 23 Jan 2001 [\[Full text\]](#)

Diagnostic Tests Revisited

Daan G Uitenbroek

bmj.com, 5 Feb 2002 [\[Full text\]](#)

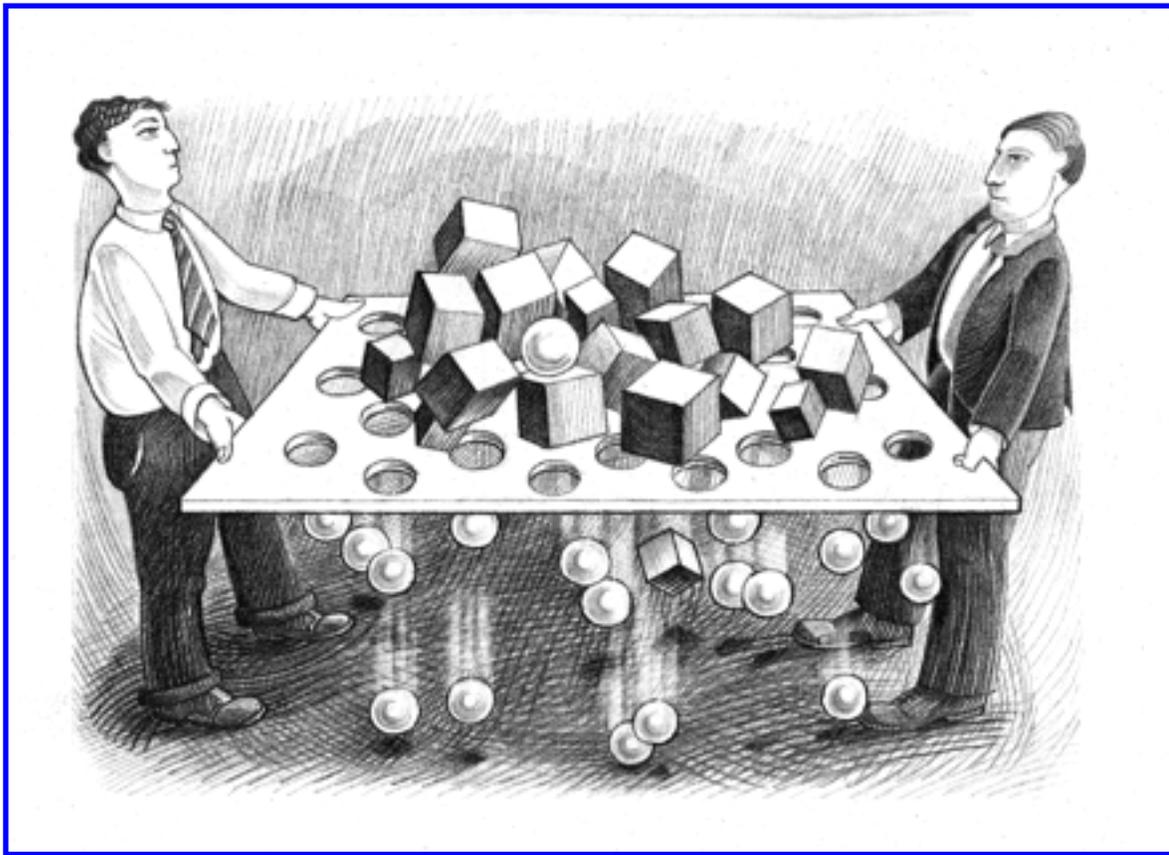
[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



PETER BROWN

[\[View larger version \(179K\)\]](#)

Table 1 Two by two table showing outcome of trial for 10 men accused of murder

Jury verdict	True criminal status	
	Murderer	Not murderer
Guilty	Rightly convicted (2 men)	Wrongly convicted (4 men)
Innocent	Wrongly acquitted (1 man)	Rightly acquitted (3 men)

Table 2 Two by two table notation for expressing the results of validation study for diagnostic or screening test

Result of screening test	Result of gold standard test	
	Disease positive (a+c)	Disease negative (b+d)
Test positive; (a+b)	True positive (a)	False positive (b)
Test negative (c+d)	False negative (c)	True negative (d)

Table 3 Two by two table showing results of validation study of urine glucose testing for diabetes against gold standard³

Result of urine test for glucose	Result of glucose tolerance test	
	Diabetes positive (n=27)	Diabetes negative (n=973)
Glucose present; (n=13)	True positive (n=6)	False positive (n=7)
Glucose absent (n=987)	False negative (n=21)	True negative (n=966)

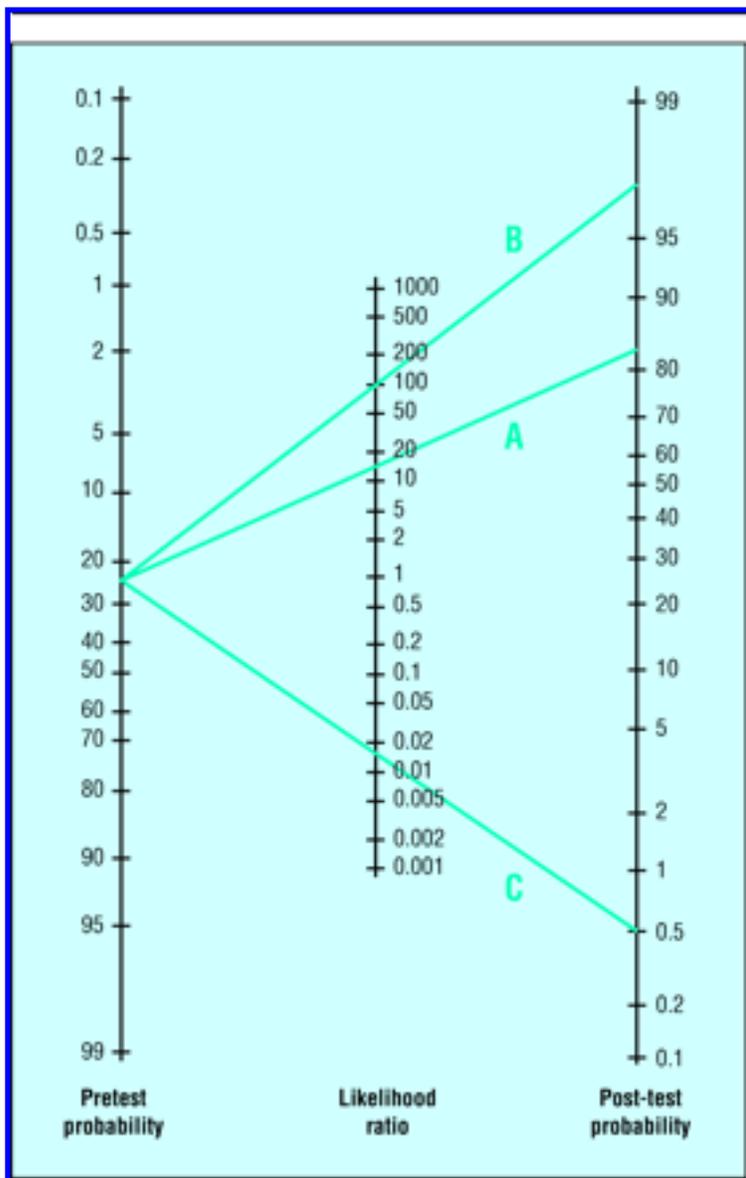


Fig 1 Use of likelihood ratios to calculate post-test probability of someone being a smoker⁶

[\[View larger version \(123K\)\]](#)

BMJ 1997;315:596-599 (6 September)

Education and debate

How to **read** a **paper**: **Papers** that tell you what things cost (economic analyses)

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ What is economic analysis?

An economic analysis can be defined as an analysis that uses analytical techniques to define choices in resource allocation. This article is based largely on a short booklet by Professor Michael Drummond¹ and two of the forerunners to the "Users' Guides to the Medical Literature" series.^{2 3} A recent book, *Elementary Economic Evaluation in Health Care*, is also useful.⁴

- ▲ [Top](#)
- [What is economic analysis?](#)
- ▼ [Measuring costs and benefits...](#)
- ▼ [Ten questions to ask...](#)
- ▼ [References](#)

▶ Measuring costs and benefits of health interventions

Not long ago, I was taken to hospital to have my appendix removed. From the hospital's point of view, the cost of my care included my board and lodging for five days, a proportion of doctors' and nurses' time, drugs and dressings, and investigations (blood tests and a scan). Other direct costs (see [box](#)) included my general practitioner's time

- ▲ [Top](#)
- ▲ [What is economic analysis?](#)
- [Measuring costs and benefits...](#)
- ▼ [Ten questions to ask...](#)
- ▼ [References](#)

for attending me in the middle of the night and the cost of the petrol my husband used when visiting me (not to mention the grapes and flowers).

Examples of costs and benefits of health interventions	
Costs	Benefits
<i>Direct:</i>	<i>Economic:</i>
"Board and lodging"	Prevention of illness that is expensive to treat
Drugs, dressings, etc	
Investigations	Avoidance of admission to hospital
Staff salaries	Return to paid work
<i>Indirect:</i>	<i>Clinical:</i>
Work days lost	Postponement of death or disability
Value of "unpaid" work	Relief of pain, nausea, breathlessness, etc
	Improved vision, hearing, muscular strength, etc
<i>Intangible:</i>	<i>Quality of life:</i>
Pain and suffering	Increased mobility and independence
Social stigma	Improved wellbeing
	Release from sick role

In addition to this, there were the indirect costs of my loss in productivity. I was off work for three weeks, and my domestic duties were temporarily carried out by various friends, neighbours, and a hired nanny. Also, from my point of view, there were several intangible costs, such as discomfort, loss of independence, and a cosmetically unsightly scar. As the [box](#) shows, these direct, indirect, and intangible costs constitute one side of the cost-benefit equation. On the benefit side, the operation greatly increased my chances of staying alive and I had a nice rest from work.

In this example, few patients (and even fewer purchasers) would perceive much freedom of choice in deciding to opt for the operation. But most health interventions do not concern definitive procedures for surgical emergencies. At some stage, almost all of us will be forced to decide whether having a routine operation, taking a particular drug, or compromising our lifestyle to treat a chronic but not immediately life threatening condition is "worth it."

It is fine for informed individuals to make choices about their own care by gut reaction ("I'd rather live with my hernia than be cut open," or "I know about the risk of thrombosis but I want to continue to smoke and stay on the pill"). But when the choices are about other people's care, subjective judgments are the last thing that should enter the equation. Most of us would want the planners and policymakers to use objective, explicit, and defensible criteria when making decisions such as "No, this patient may not have a kidney transplant."

One important way of addressing the "what's it worth?" question for a given health state (such as having poorly controlled diabetes or a flare up of rheumatoid arthritis) is to ask someone in that state how they feel. A number of questionnaires have been developed which attempt to measure overall health status, such as the Nottingham health profile, the SF-36 general health questionnaire, and the McMaster health utilities index questionnaire.⁵

Summary points

An economic analysis should be based on a primary study or meta-analysis that is scientifically valid, reliable, and relevant

When deciding whether an economic analysis has been done correctly, you should not simply check the arithmetic but consider whether all direct, indirect, and intangible costs and benefits have been included

In the allocation of limited resources, the comparison of different health states is unavoidable, but instruments for measuring health related quality of life are not as objective as they seem

In some circumstances, disease specific measures of wellbeing are more valid than general measures.⁶ For example, answering "yes" to the question, "Do you get very concerned about the food you are eating?" might indicate anxiety in someone without diabetes but normal self care attitudes in someone with diabetes. There has also been an upsurge of interest in patient specific measures of quality of life, to allow different patients to place different values on particular aspects of their health and wellbeing.⁷ Of course, when quality of life is being analysed from the point of view of the patient, this is a sensible and humane approach. However, the health economist tends to make decisions about groups of patients or populations, in which case patient specific, and even disease specific, measures of quality of life have limited relevance.⁸

The authors of standard instruments (such as the SF-36) for measuring quality of life have often spent

years ensuring they are valid (that they measure what we think they are measuring), reliable (they do so every time), and responsive to change (if an intervention improves or worsens the patient's health, the scale will reflect that). For this reason, you should be highly suspicious of a **paper** that abandons these standard instruments in favour of the authors' own rough and **ready** scale (for example, "functional ability was classified as good, moderate, or poor according to the clinician's overall impression"). Note also that even instruments which have apparently been well validated often do not stand up to rigorous evaluation of their psychometric validity.⁹

Another way of addressing the "what's it worth?" of particular health states is through health state preference values—that is, the value which, in a hypothetical situation, a healthy person would place on a particular deterioration in their health, or which a sick person would place on a return to health. There are three main methods of assigning such values:

- Rating scale measurements—the respondent is asked to make a mark on a fixed line, labelled, for example, "perfect health" at one end and "death" at the other, to indicate where he or she would place the state in question (for example, being confined to a wheelchair by arthritis of the hip);
- Time tradeoff measurements—the respondent is asked to consider a particular health state (for example, infertility) and estimate how many of their remaining years in full health they would sacrifice to be "cured" of the condition;
- Standard gamble measurements—the respondent is asked to consider the choice between living for the rest of their life in a particular health state and taking a "gamble" (such as having an operation) with a given odds of success which would return them to full health if it succeeded but kill them if it failed. The odds are then varied to see at what point the respondent decides the gamble is not worth taking.¹⁰

The quality adjusted life year (QALY) can be calculated by multiplying the preference value for that state with the time the patient is likely to spend in that state. The results of cost-benefit analyses are usually expressed in terms of "cost per QALY," some examples of which are shown in the second [box](#).¹¹

Results of cost-benefit analysis for some medical procedures

Procedure	Cost per QALY (£)
Cholesterol testing and diet therapy	220
Advice to stop smoking from patient's own doctor	270
Hip replacement for arthritis	1 180
Kidney transplant	4 710
Breast cancer screening	5 780
Cholesterol testing and drug therapy if indicated (ages 25-39)	14 150
Neurosurgery for malignant brain tumours	107 780

The use of QALYs is controversial. Any measure of health state preference values is, at best, a reflection of the preferences and prejudices of the individuals who contributed to its development. Indeed, it is possible to come up with different values for QALYs, depending on how the questions from which the health state preference values are derived were posed.¹² Furthermore, it is virtually impossible to combine different QALYs to measure the effect of more than one serious or disabling condition on a patient.¹³ As medical ethicist John Harris has pointed out, QALYs are, like the society that produces them, inherently agist, sexist, racist, and loaded against those with permanent disabilities (since even a complete cure of an unrelated condition would not restore the individual to "perfect health"). Furthermore, QALYs distort our ethical instincts by focusing our minds on years of life rather than people's lives. A disabled premature infant in need of an intensive care cot will, argues Harris, be allocated more resources than it deserves in comparison with a 50 year old woman with cancer, since the infant, were it to survive, would have so many more life years to quality adjust.¹⁴

Other authors have come up with the HYE (healthy years equivalent) measure, which incorporates the individual's likely improvement or deterioration in health status in the future and is said to avoid some, but not all, of the disadvantages of the QALY.¹⁵ Given that the critics of QALYs and HYE have offered no alternative, all encompassing measure of health status, these utility based units are set to remain in the health economist's toolkit for the foreseeable future. For a more detailed discussion of these issues by a multidisciplinary panel, see Anthony Hopkins's booklet *Measures of the Quality of Life*.¹⁶

There is, however, another form of analysis which, although it does not abolish the need to place arbitrary numerical values on life and limb, avoids the buck stopping at the unfortunate health economist. This approach, known as cost-consequences analysis, presents the results of the economic analysis in a disaggregated form. In other words, it expresses different outcomes in terms of their different natural

units (something real such as months of survival, legs amputated, or babies taken home), so that individuals can assign their own values to particular health states before calculating whether the intervention is "worth it."

Ten questions to ask about an economic analysis

The checklist which follows is based on the sources mentioned earlier,^{1 2} as well as suggestions made by a working party set up by the *BMJ* to produce guidelines for journal editors on appraising economic evaluations (M Drummond, personal communication).

- ▲ [Top](#)
- ▲ [What is economic analysis?](#)
- ▲ [Measuring costs and benefits...](#)
 - [Ten questions to ask...](#)
- ▼ [References](#)

Question 1: Is the analysis based on a study that answers a clearly defined clinical question about an economically important issue?

Before pursuing any of the economic arguments, make sure that the trial being analysed is scientifically relevant and capable of giving unbiased and unambiguous answers to the clinical question posed in its introduction.



PETER BROWN

View larger version (128K):

[\[in this window\]](#)

[\[in a new window\]](#)

Question 2: Whose viewpoint are costs and benefits being considered from?

From the Treasury's point of view, the most cost effective health intervention is one which returns all citizens promptly to taxpayer status and, when this status is no longer tenable, causes immediate sudden death. From the drug company's point of view, it would be difficult to imagine a cost-benefit equation

which did not contain one of the company's products, and from a physiotherapist's point of view, the removal of a physiotherapy service would never be cost effective. Almost all economic analyses have some funding, and all have been inspired by someone with a vested interest; the **paper** should say which.

Question 3: Have the interventions being compared been shown to be clinically effective?

In general, the intervention that "works out cheaper" should not be substantially less effective in clinical terms than the one which stands to be rejected on the grounds of cost.

Question 4: Are the interventions sensible and workable in the settings where they are likely to be applied?

Too many research trials look at intervention packages which would be impossible to implement in the non-research setting (they assume, for example, that general practitioners will own a state of the art computer and agree to follow a protocol, that infinite nurse time is available for the taking of blood tests, or that patients will make their personal treatment choices solely on the basis of the trial's conclusions). Remember that standard current practice, which may be to do nothing, should almost certainly be one of the alternatives compared.

Question 5: Which method of analysis was used, and was this appropriate?

This decision can be summarised as follows:

- Cost minimisation analysis would be most appropriate if the interventions produced identical outcomes;
- Cost effectiveness analysis would be most appropriate if the important outcome is unidimensional;
- Cost utility analysis would be most appropriate if the important outcome is multidimensional;
- Cost benefit analysis would be most appropriate if the cost benefit equation for this condition needs to be compared with cost benefit equations for different conditions;
- Cost consequences analysis would be most appropriate if a cost benefit analysis would otherwise be appropriate but the preference values given to different health states are disputed or likely to change.

Question 6: How were costs and benefits measured?

Consider an economic evaluation of a trial comparing the rehabilitation of stroke patients into their own homes, including attendance at a day centre, with a standard alternative intervention (rehabilitation in a long stay hospital). The economic analysis must take into account not just the time of the various professionals involved, the time of the secretaries and administrators who help run the service, "overheads" (such as heating and lighting), and the cost of the food and drugs consumed by the stroke

patients, but also a fraction of the capital cost of building the day centre and maintaining a transport service to and from it.

In a cost effectiveness analysis, changes in health status will be expressed in natural units. But just because the units are natural does not automatically make them appropriate. For example, the economic analysis of the treatment of peptic ulcer by two different drugs might measure outcome as "proportion of ulcers healed after a six week course." Treatments could be compared according to the cost per ulcer healed. However, if the relapse rates on the two drugs were very different, drug A might be falsely deemed "more cost effective" than drug B. A better outcome measure here might be "ulcers that remained healed at one year."

Question 7: Were incremental, rather than absolute, benefits considered?

This question is best illustrated by a simple example. Let's say drug X, at £100 per course, cures 10 out of every 20 patients. Its new competitor, drug Y, costs £120 per course and cures 11 out of 20 patients. The cost per case cured with drug X is £200 (since you spent £2000 curing 10 people), and the cost per case cured with drug Y is £218 (since you spent £2400 curing 11 people).

The incremental cost of drug Y—the extra cost of curing the extra patient—is not £18, but £400, since this is the total amount extra that you have had to pay to achieve an outcome over and above what you would have achieved by giving all patients the cheaper drug. This striking example should be borne in mind the next time a pharmaceutical representative tries to persuade you that his or her product is "more effective and only marginally more expensive."

Question 8: Was the "here and now" given precedence over the distant future?

A bird in the hand is worth two in the bush: in health as well as money terms, we value a benefit today more highly than we value a promise of the same benefit in five years' time. When the costs or benefits of an intervention (or lack of the intervention) will occur some time in the future, their value should be discounted to reflect this. The actual amount of discount that should be allowed for future, as opposed to immediate, health benefit is fairly arbitrary, but most analyses use a figure of around 5% per year.

Question 9: Was a sensitivity analysis performed?

Let's say a cost-benefit analysis comes out as saying that hernia repair by day case surgery costs £1150 per QALY whereas traditional open repair, with its associated hospital stay, costs £1800 per QALY. But, when you look at how the calculations were done, you are surprised at how cheaply the laparoscopic equipment has been costed. If you raise the price of this equipment by 25%, does day case surgery still come out dramatically cheaper? It may, or it may not.

Sensitivity analysis, or exploration of "what ifs," was described earlier in this series in relation to meta-analysis.¹⁷ Exactly the same principles apply here: if adjusting the figures to account for the full range of possible influences gives you a totally different answer, you should not place too much reliance on the

analysis. For a good example of a sensitivity analysis on a topic of both scientific and political importance, see Pharoah and Hollingworth's **paper** on the cost effectiveness of lowering cholesterol (which addresses the difficult issue of who should receive, and who should be denied, effective but expensive drugs to lower cholesterol).¹⁸

Question 10: Were "bottom line" aggregate scores overused?

The notion of cost-consequences analysis, in which the **reader** of the **paper** can attach his or her own values to different utilities, was introduced earlier. In practice, this is an unusual way of presenting an economic analysis, and, more commonly, the **reader** is faced with a cost-utility or cost-benefit analysis which gives a composite score in unfamiliar units which do not translate **readily** into exactly what gains and losses the patient can expect. The situation is analogous to the father who is told "your child's IQ is 115" when he would feel far better informed if he were presented with the disaggregated data: "Johnny can **read**, write, count, and draw pretty well for his age."

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 Publishing Group: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

Acknowledgements

Thanks to Professor Mike Drummond and Dr Alison Tonks for advice on this chapter.

References

1. Drummond M. *Economic analysis alongside controlled trials*. Leeds: Department of Health, 1994. (R&D Directorate, document F51/066 2515 5k.)
2. 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](#)]

- ▲ [Top](#)
- ▲ [What is economic analysis?](#)
- ▲ [Measuring costs and benefits...](#)
- ▲ [Ten questions to ask...](#)
- [References](#)

3. 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](#)]
4. Jefferson T, Demicheli V, Mugford M. *Elementary economic evaluation in health care*. London: BMJ Publishing Group, 1996.
5. Patrick DL, Erikson P. *Health status and health policy*. New York: Oxford University Press, 1993.
6. Fallowfield LJ. Assessment of quality of life in breast cancer. *Acta Oncol* 1995;34:689-94. [[Medline](#)]
7. Hickey AM, Bury G, O'Boyle CA, Bradley F, O'Kelley FD, Shannon W. A new short-form individual quality of life measure (SEIQoL-DW). Application in a cohort of individuals with HIV/AIDS. *BMJ* 1996;313:29-33. [[Full Text](#)]
8. Cairns J. Measuring health outcomes. *BMJ* 1996;313:6. [[Full Text](#)]
9. Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. *JAMA* 1994;272:619-26.
10. Krabbe PFM, Essink-Bot M-L, Bonsel GK. On the equivalence of collectively and individually collected responses: standard-gamble and time-tradeoff judgements of health status. *Med Decis Making* 1996;16:120-32.
11. Ham C. Priority setting in the NHS. *Br J Health Care Manage* 1995;1:27-9.
12. Weinberger M, Oddone EZ, Samsa G, Landsman P. Are health-related quality of life measures affected by the mode of administration? *J Clin Epidemiol* 1996;49:135-40. [[Medline](#)]
13. Richardson J, Hall J, Salkeld G. The measurement of utility in multiphase health states. *Int J Technol Assess Health Care* 1996;12:151-62.
14. Harris J. QALYfying the value of life. *J Med Ethics* 1987;13:117-23. [[Abstract](#)]
15. Mehrez A, Gafni A. Quality-adjusted life years, utility theory and healthy year equivalents. *Med Decis Making* 1989;9:142-9.
16. Hopkins A, ed. *Measures of the quality of life*. London: Royal College of General Practitioners, 1992.
17. **Greenhalgh T. Papers** that summarise other **papers** (systematic reviews and meta-analyses). *BMJ* 1997 (in press).
18. Pharoah PDP, Hollingworth W. Cost-effectiveness of lowering cholesterol concentration with statins in patients with and without pre-existing coronary heart disease: life table method applied to health authority population. *BMJ* 1996;312:1443-8. [[Abstract/Full Text](#)]

This article has been cited by other articles:

- Fanshawe, M., Ellis, C., Habib, S., Konstadt, S. N., Reich, D. L. (2002). A Retrospective Analysis of the Costs and Benefits Related to Alterations in Cardiac Surgery from Routine Intraoperative Transesophageal Echocardiography. *Anesth Analg* 95: 824-827 [[Abstract](#)] [[Full text](#)]

- Figueredo, E., Sadhasivam, S., Saxena, A., Kathirvel, S., Kannan, T. R. (2000). Ondansetron and Evidence-Based Medicine Response. *Anesth Analg* 91: 496-497 [[Full text](#)]

Related letters in BMJ:

Article showed how not to **read** economic evaluations

Jane Henderson, Stavros Petrou, and Tracy Roberts
BMJ 1998 316: 939. [[Letter](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ Related [letters](#) in BMJ
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)



PETER BROWN

[\[View larger version \(196K\)\]](#)

BMJ 1997;315:672-675 (13 September)

Education and debate

How to **read** a **paper**: **Papers** that summarise other **papers** (systematic reviews and meta-analyses)

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

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ **Read** responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by: **[Greenhalgh, T.](#)**
- ▶ Alert me when: [New articles cite this article](#)

▶ Introduction

Remember the essays you used to write as a student? You would browse through the indexes of books and journals until you came across a paragraph that looked relevant, and copied it out. If anything you found did not fit in with the theory you were proposing, you left it out. This, more or less, constitutes the methodology of the journalistic review—an overview of primary studies which have not been identified or analysed in a systematic (standardised and objective) way.

- ▲ [Top](#)
- [Introduction](#)
- ▼ [Evaluating systematic reviews](#)
- ▼ [Meta-analysis for the...](#)
- ▼ [Explaining heterogeneity](#)
- ▼ [References](#)

Summary points

A systematic review is an overview of primary studies that used explicit and reproducible methods

A meta-analysis is a mathematical synthesis of the results of two or more primary studies that addressed the same hypothesis in the same way

Although meta-analysis can increase the precision of a result, it is important to ensure that the methods used for the review were valid and reliable

In contrast, a systematic review is an overview of primary studies which contains an explicit statement of objectives, materials, and methods and has been conducted according to explicit and reproducible methodology (fig 1).

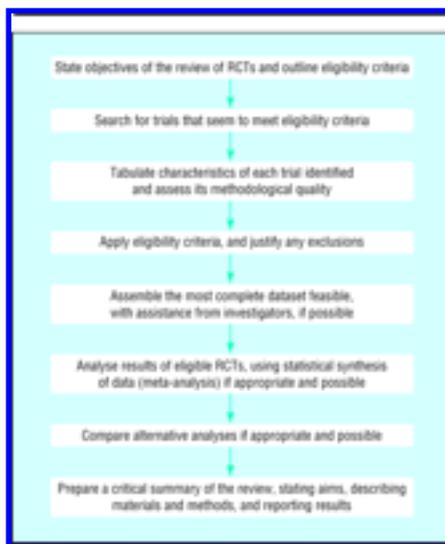


Fig 1 Methodology for a systematic review of randomised controlled trials¹

View larger version (33K):

[\[in this window\]](#)

[\[in a new window\]](#)

Some advantages of the systematic review are given in [box](#). When a systematic review is undertaken, not only must the search for relevant articles be thorough and objective, but the criteria used to reject articles as "flawed" must be explicit and independent of the results of those trials. The most enduring and useful systematic reviews, notably those undertaken by the Cochrane Collaboration, are regularly updated to

incorporate new evidence.²

Box 1: Advantages of systematic reviews³

- Explicit methods limit bias in identifying and rejecting studies
- Conclusions are more reliable and accurate because of methods used
- Large amounts of information can be assimilated quickly by healthcare providers, researchers, and policymakers
- Delay between research discoveries and implementation of effective diagnostic and therapeutic strategies may be reduced
- Results of different studies can be formally compared to establish generalisability of findings and consistency (lack of heterogeneity) of results
- Reasons for heterogeneity (inconsistency in results across studies) can be identified and new hypotheses generated about particular subgroups
- Quantitative systematic reviews (meta-analyses) increase the precision of the overall result

Many, if not most, medical review articles are still written in narrative or journalistic form. Professor Paul Knipschild has described how Nobel prize winning biochemist Linus Pauling used selective quotes from the medical literature to "prove" his theory that vitamin C helps you live longer and feel better.^{3 4} When Knipschild and his colleagues searched the literature systematically for evidence for and against this hypothesis they found that, although one or two trials did strongly suggest that vitamin C could prevent the onset of the common cold, there were far more studies which did not show any beneficial effect.

Experts, who have been steeped in a subject for years and know what the answer "ought" to be, are less able to produce an objective review of the literature in their subject than non-experts.^{5 6} This would be of little consequence if experts' opinions could be relied on to be congruent with the results of independent systematic reviews, but they cannot.⁷

Evaluating systematic reviews

Question 1: Can you find an important clinical question which the review addressed?

The question addressed by a systematic review needs to be defined very precisely, since the reviewer must make a dichotomous (yes/no) decision as to whether each potentially relevant **paper** will be included or, alternatively, rejected as "irrelevant." Thus, for example, the clinical question "Do anticoagulants prevent strokes in patients with atrial fibrillation?" should be refined as an objective: "To assess the effectiveness and safety of warfarin-type anticoagulant therapy in secondary prevention (that is, following a previous stroke or transient ischaemic attack) in patients with non-rheumatic atrial fibrillation: comparison with placebo."⁸

- ▲ [Top](#)
- ▲ [Introduction](#)
- [Evaluating systematic reviews](#)
- ▼ [Meta-analysis for the...](#)
- ▼ [Explaining heterogeneity](#)
- ▼ [References](#)

Question 2: Was a thorough search done of the appropriate databases and were other potentially important sources explored?

Even the best Medline search will miss important **papers**, for which the reviewer must approach other sources.⁹ Looking up references of references often yields useful articles not identified in the initial search,¹⁰ and an exploration of "grey literature" ([box](#)) may be particularly important for subjects outside the medical mainstream, such as physiotherapy or alternative medicine.¹¹ Finally, particularly where a statistical synthesis of results (meta-analysis) is contemplated, it may be necessary to write and ask the authors of the primary studies for raw data on individual patients which was never included in the published review.

Box 2: Checklist of data sources for a systematic review

- Medline database
- Cochrane controlled clinical trials register
- Other medical and paramedical databases
- Foreign language literature
- "Grey literature" (theses, internal reports, non-peer reviewed journals, pharmaceutical industry files)

- References (and references of references, etc) listed in primary sources
- Other unpublished sources known to experts in the field (seek by personal communication)
- Raw data from published trials (seek by personal communication)

Question 3: Was methodological quality assessed and the trials weighted accordingly?

One of the tasks of a systematic reviewer is to draw up a list of criteria, including both generic (common to all research studies) and particular (specific to the field) aspects of quality, against which to judge each trial (see [box](#)). However, care should be taken in developing such scores since there is no gold standard for the "true" methodological quality of a trial¹² and composite quality scores are often neither valid nor reliable in practice.^{13 14} The various Cochrane collaborative review groups are developing topic-specific methodology for assigning quality scores to research studies.¹⁵

Box 3: Assigning weight to trials in a systematic review

Each trial should be evaluated in terms of its:

- Methodological quality—the extent to which the design and conduct are likely to have prevented systematic errors (bias)
- Precision—a measure of the likelihood of random errors (usually depicted as the width of the confidence interval around the result)
- External validity—the extent to which the results are generalisable or applicable to a particular target population

Question 4: How sensitive are the results to the way the review has been done?

Carl Counsell and colleagues "proved" (in the Christmas 1994 issue of the *BMJ*) an entirely spurious relationship between the result of shaking a dice and the outcome of an acute stroke.¹⁶ They reported a series of artificial dice rolling experiments in which red, white, and green dice represented different therapies for acute stroke. Overall, the "trials" showed no significant benefit from the three therapies. However, the simulation of a number of perfectly plausible events in the process of meta-analysis—such as the exclusion of several of the "negative" trials through publication bias, a subgroup analysis which

excluded data on red dice therapy (since, on looking back at the results, red dice appeared to be harmful), and other, essentially arbitrary, exclusions on the grounds of "methodological quality"—led to an apparently highly significant benefit of "dice therapy" in acute stroke.

If these simulated results pertained to a genuine medical controversy, how would you spot these subtle biases? You need to work through the "what ifs". What if the authors of the systematic review had changed the inclusion criteria? What if they had excluded unpublished studies? What if their "quality weightings" had been assigned differently? What if trials of lower methodological quality had been included (or excluded)? What if all the patients unaccounted for in a trial were assumed to have died (or been cured)?



PETER BROWN

View larger version (118K):

[\[in this window\]](#)

[\[in a new window\]](#)

An exploration of what ifs is known as a sensitivity analysis. If you find that fiddling with the data in various ways makes little or no difference to the review's overall results, you can assume that the review's conclusions are relatively robust. If, however, the key findings disappear when any of the what ifs changes, the conclusions should be expressed far more cautiously and you should hesitate before changing your practice in the light of them.

Question 5: Have the numerical results been interpreted with common sense and due regard to the broader aspects of the problem?

Any numerical result, however precise, accurate, "significant," or otherwise incontrovertible, must be placed in the context of the painfully simple and often frustratingly general question which the review addressed. The clinician must decide how (if at all) this numerical result, whether significant or not, should influence the care of an individual patient. A particularly important feature to consider when undertaking or appraising a systematic review is the external validity or relevance of the trials that are

included.

▶ Meta-analysis for the non-statistician

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [Evaluating systematic reviews](#)
 - [Meta-analysis for the...](#)
- ▼ [Explaining heterogeneity](#)
- ▼ [References](#)

A good meta-analysis is often easier for the non-statistician to understand than the stack of primary research **papers** from which it was derived. In addition to synthesising the numerical data, part of the meta-analyst's job is to tabulate relevant information on the inclusion criteria, sample size, baseline patient characteristics, withdrawal rate, and results of primary and secondary end points of all the studies included. Although such tables are often visually daunting, they save you having to plough through the methods sections of each **paper** and compare one author's tabulated results with another author's pie chart or histogram.

These days, the results of meta-analyses tend to be presented in a fairly standard form, such as is produced by the computer software MetaView. [3](#) is a pictorial representation (colloquially known as a "forest plot") of the pooled odds ratios of eight randomised controlled trials which each compared coronary artery bypass grafting with percutaneous coronary angioplasty in the treatment of severe angina.¹⁷ The primary (main) outcome in this meta-analysis was death or heart attack within one year.

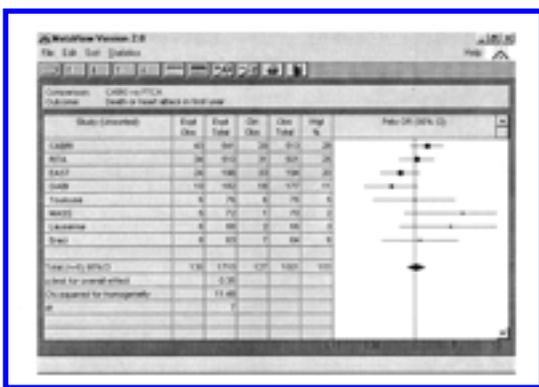


Fig 2 Pooled odds ratios of eight randomised controlled trials of coronary artery bypass grafting against percutaneous coronary angioplasty, shown in MetaView format. Reproduced with authors' permission¹⁷

View larger version (96K):
[\[in this window\]](#)
[\[in a new window\]](#)

The horizontal line corresponding to each of the eight trials shows the relative risk of death or heart

attack at one year in patients randomised to coronary angioplasty compared to patients randomised to bypass surgery. The "blob" in the middle of each line is the point estimate of the difference between the groups (the best single estimate of the benefit in lives saved by offering bypass surgery rather than coronary angioplasty), and the width of the line represents the 95% confidence interval of this estimate. The black line down the middle of the picture is known as the "line of no effect," and in this case is associated with a relative risk of 1.0.

If the confidence interval of the result (the horizontal line) crosses the line of no effect (the vertical line), that can mean either that there is no significant difference between the treatments or that the sample size was too small to allow us to be confident where the true result lies. The various individual studies give point estimates of the relative risk of coronary angioplasty compared with bypass surgery of between about 0.5 and 5.0, and the confidence intervals of some studies are so wide that they do not even fit on the graph. Now look at the tiny diamond below all the horizontal lines. This represents the pooled data from all eight trials (overall relative risk of coronary angioplasty compared with bypass surgery=1.08), with a new, much narrower, confidence interval of this relative risk (0.79 to 1.50). Since the diamond firmly overlaps the line of no effect, we can say that there is probably little to choose between the two treatments in terms of the primary end point (death or heart attack in the first year). Now, in this example, every one of the eight trials also suggested a non-significant effect, but in none of them was the sample size large enough for us to be confident in that negative result.

Note, however, that this neat little diamond does not mean that you might as well offer coronary angioplasty rather than bypass surgery to every patient with angina. It has a much more limited meaning—that the average patient in the trials presented in this meta-analysis is equally likely to have met the primary outcome (death or myocardial infarction within a year), whichever of these two treatments they were randomised to receive. If you **read** the **paper** by Pocock and colleagues¹⁷ you would find important differences in the groups in terms of prevalence of angina and requirement for further operative intervention after the initial procedure.

▶ Explaining heterogeneity

In the language of meta-analysis, homogeneity means that the results of each individual trial are mathematically compatible with the results of any of the others. Homogeneity can be estimated at a glance once the trial results have been presented in the format illustrated in figures [3](#) and [4](#). In [3](#) the lower confidence limit of every trial is below the upper confidence limit of all the others (that is, the horizontal lines all overlap to some extent). Statistically speaking, the trials are homogeneous. Conversely, in [4](#) some lines

- ▲ [Top](#)
- ▲ [Introduction](#)
- ▲ [Evaluating systematic reviews](#)
- ▲ [Meta-analysis for the...](#)
- [Explaining heterogeneity](#)
- ▼ [References](#)

do not overlap at all. These trials may be said to be heterogeneous.

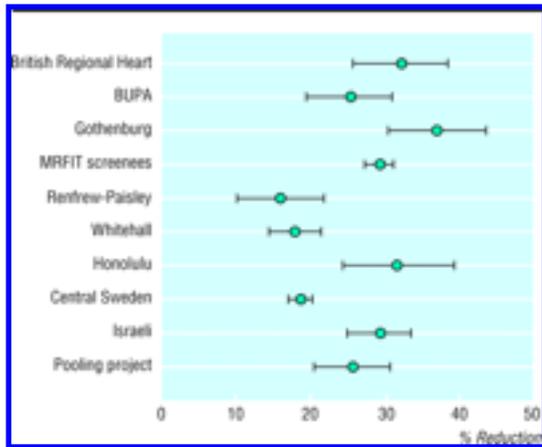


Fig 3 Reduction in risk of heart disease by strategies for lowering cholesterol. Reproduced with permission from Chalmers and Altman¹⁸

View larger version (17K):

[\[in this window\]](#)

[\[in a new window\]](#)

The definitive test for heterogeneity involves a slightly more sophisticated statistical manoeuvre than holding a ruler up against the forest plot. The one most commonly used is a variant of the χ^2 (chi square) test, since the question addressed is whether there is greater variation between the results of the trials than is compatible with the play of chance. Thompson¹⁸ offers the following rule of thumb: a χ^2 statistic has, on average, a value equal to its degrees of freedom (in this case, the number of trials in the meta-analysis minus one), so a χ^2 of 7.0 for a set of eight trials would provide no evidence of statistical heterogeneity. Note that showing statistical heterogeneity is a mathematical exercise and is the job of the statistician, but explaining this heterogeneity (looking for, and accounting for, clinical heterogeneity) is an interpretive exercise and requires imagination, common sense, and hands-on clinical or research experience.

[4](#) shows the results of ten trials of cholesterol lowering strategies. The results are expressed as the percentage reduction in risk of heart disease associated with each reduction of 0.6 mmol/l in serum cholesterol concentration. From the horizontal lines which represent the 95% confidence intervals of each result it is clear, even without knowing the χ^2 statistic of 127, that the trials are highly heterogeneous. Correcting the data for the age of the trial subjects reduced this value to 45. In other words, much of the "incompatibility" in the results of these trials can be explained by the fact that embarking on a strategy which successfully reduces your cholesterol level will be substantially more likely to prevent a heart attack if you are 45 than if you are 85.

Clinical heterogeneity, essentially, is the grievance of Professor Hans Eysenck, who has constructed a

vigorous and entertaining critique of the science of meta-analysis.¹⁹ In a world of lumpers and splitters, Eysenck is a splitter, and it offends his sense of the qualitative and the particular to combine the results of studies which were done on different populations in different places at different times and for different reasons.

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 Publishing Group: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

Eysenck's reservations about meta-analysis are borne out in the infamously discredited meta-analysis which showed (wrongly) that giving intravenous magnesium to people who had had heart attacks was beneficial. A subsequent megatrial involving 58 000 patients (ISIS-4) failed to find any benefit, and the meta-analysts' misleading conclusions were subsequently explained in terms of publication bias, methodological weaknesses in the smaller trials, and clinical heterogeneity.^{20 21}

Acknowledgements

Thanks to Professor Iain Chalmers for advice on this chapter.

References

1. *The Cochrane Centre*. Cochrane Collaboration Handbook [updated 9 December 1996]. The Cochrane Collaboration; issue 1. Oxford: Update Software, 1997.
2. Bero L, Rennie D. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995;274:1935-8.
3. Chalmers I, Altman DG, eds. *Systematic reviews*. London: BMJ Publishing Group, 1995.
4. Pauling L. *How to live longer and feel better*. New York: Freeman, 1986.
5. Oxman AD, Guyatt GH. The science of reviewing research. *Ann NY Acad Sci* 1993; 703: 125-31.
6. Mulrow C. The medical review article: state of the science. *Ann Intern Med* 1987;106: 485-8.

- [▲ Top](#)
- [▲ Introduction](#)
- [▲ Evaluating systematic reviews](#)
- [▲ Meta-analysis for the...](#)
- [▲ Explaining heterogeneity](#)
- References

7. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomised controlled trials and recommendations of clinical experts. *JAMA* 1992;268:240-8.
8. Koudstaal P. Secondary prevention following stroke or TIA in patients with non-rheumatic atrial fibrillation: anticoagulant therapy versus control. *Cochrane Database of Systematic Reviews*. Oxford: Cochrane Collaboration, 1995. (Updated 14 February 1995.)
9. **Greenhalgh** T. Searching the literature. In: *How to read a paper*. London: BMJ Publishing Group, 1997:13-33.
10. Knipschild P. Some examples of systematic reviews. In: Chalmers I, Altman DG. *Systematic reviews*. London: BMJ Publishing Group, 1995:9-16.
11. Knipschild P. Searching for alternatives: loser pays. *Lancet* 1993; 341: 1135-6.
12. Oxman A, ed. Preparing and maintaining systematic reviews. In: *Cochrane Collaboration handbook, section VI*. Oxford: Cochrane Collaboration, 1995. (Updated 14 July 1995.)
13. Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. *Controlled Clin Trials* 1990;11:339-52.
14. Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Health Technol Assess* 1996;12:195-208.
15. Garner P, Hetherington J. Establishing and supporting collaborative review groups. In: *Cochrane Collaboration handbook, section II*. Oxford: Cochrane Collaboration, 1995 (Updated 14 July 1995.)
16. Counsell CE, Clarke MJ, Slattery J, Sandercock PAG. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *BMJ* 1994;309:1677-81. [[Abstract/Full Text](#)]
17. Pocock SJ, Henderson RA, Rickards AF, Hampton JR, Sing SB III, Hamm CW, et al. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet* 1995;346:1184-9.
18. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. In: Chalmers I, Altman DG. *Systematic reviews*. London, BMJ Publishing Group, 1995:48-63.
19. Eysenck HJ. Problems with meta-analysis. In: Chalmers I, Altman DG. *Systematic reviews*. London: BMJ Publishing Group, 1995:64-74.
20. Magnesium, myocardial infarction, meta-analysis and mega-trials. *Drug Ther Bull* 1995;33:25-7.
21. Egger M, Davey Smith G. Misleading meta-analysis: lessons from "an effective, safe, simple" intervention that wasn't. *BMJ* 1995;310:752-4. [[Full Text](#)]

This article has been cited by other articles:

- Stewart, C E, Fielder, A R, Stephens, D A, Moseley, M J (2002). Design of the Monitored Occlusion Treatment of Amblyopia Study (MOTAS). *Br. J. Ophthalmol.* 86: 915-919

[\[Abstract\]](#) [\[Full text\]](#)

- Menz, H. B. (2002). A Retrospective Analysis of JAPMA Publication Patterns, 1991-2000. *J Am Podiatr Med Assoc* 92: 308-313 [\[Abstract\]](#) [\[Full text\]](#)
- 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 [\[Abstract\]](#) [\[Full text\]](#)
- Murphy, C. C., Schei, B., Myhr, T. L., Mont, J. D. (2001). Abuse: A risk factor for low birth weight? A systematic review and meta-analysis. *Can Med Assoc J* 164: 1567-1572 [\[Abstract\]](#) [\[Full text\]](#)
- McQueen, M. J. (2001). Overview of Evidence-based Medicine: Challenges for Evidence-based Laboratory Medicine. *Clin Chem* 47: 1536-1546 [\[Abstract\]](#) [\[Full text\]](#)
- Petrella, R. J (2001). Is exercise effective treatment of osteoarthritis of the knee?. *eWJM* 174: 191-196 [\[Full text\]](#)
- Petrella, R. J (2000). Is exercise effective treatment for osteoarthritis of the knee?. *Br J Sports Med* 34: 326-331 [\[Abstract\]](#) [\[Full text\]](#)
- Sneyd, J.R. (2000). Editorial I: Conflicts of interest: are they a problem for anaesthesia journals? What should we do about them?. *Br J Anaesth* 85: 811-814 [\[Full text\]](#)
- White, P. F., Watcha, M. F. (1999). Has the Use of Meta-Analysis Enhanced Our Understanding of Therapies for Postoperative Nausea and Vomiting?. *Anesth Analg* 88: 1200-1200 [\[Full text\]](#)
- Plotnick, L. H, Ducharme, F. M (1998). Should inhaled anticholinergics be added to beta 2 agonists for treating acute childhood and adolescent asthma? A systematic review. *BMJ* 317: 971-977 [\[Abstract\]](#) [\[Full text\]](#)

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [Read](#) responses to this article
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by: [Greenhalgh, T.](#)
- ▶ Alert me when: [New articles cite this article](#)

Rapid Responses:

Read all [Rapid Responses](#)

Breadth of Useage

Allan White

bmj.com, 14 Aug 1999 [\[Full text\]](#)

[Home](#)

[Help](#)

[Search/Archive](#)

[Feedback](#)

[Search Result](#)

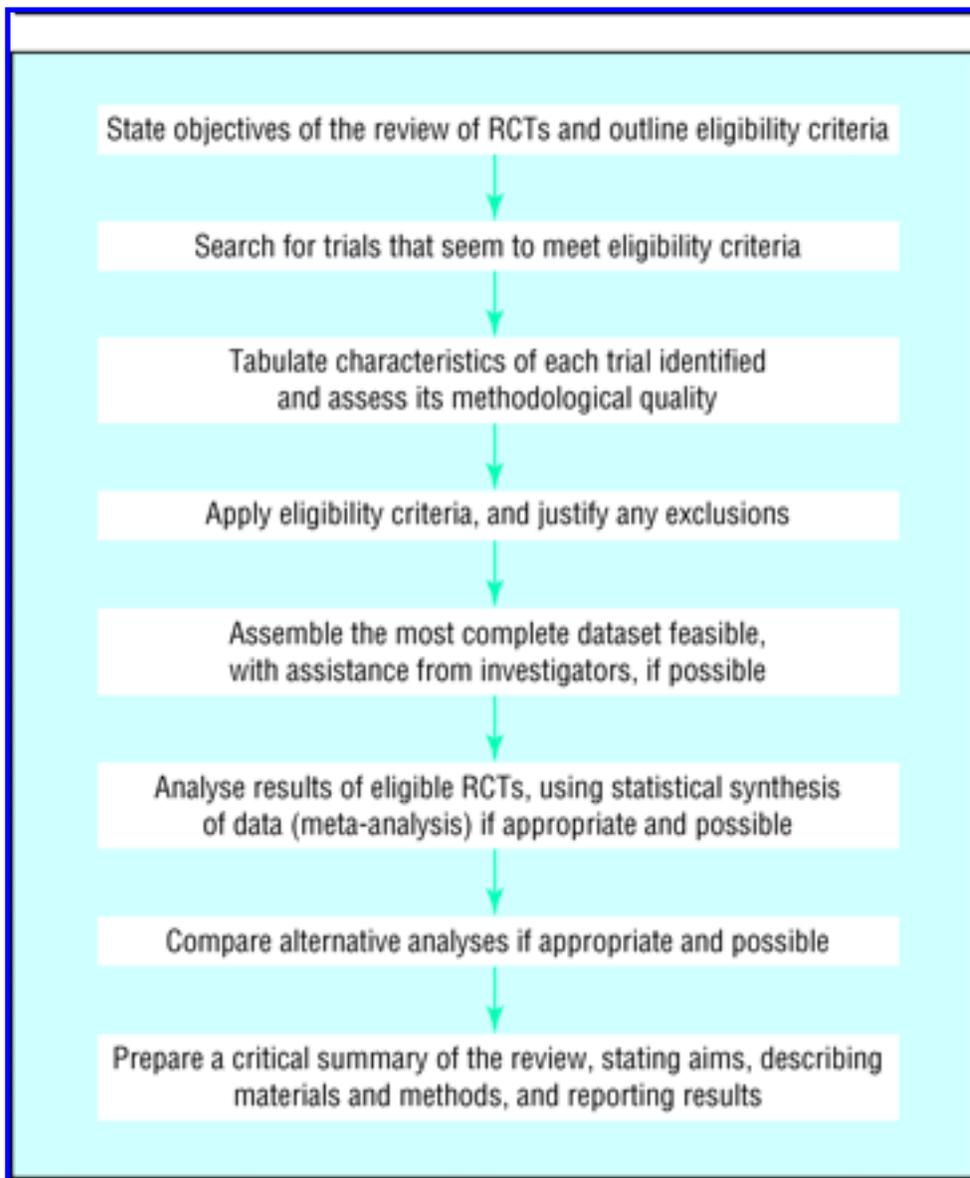


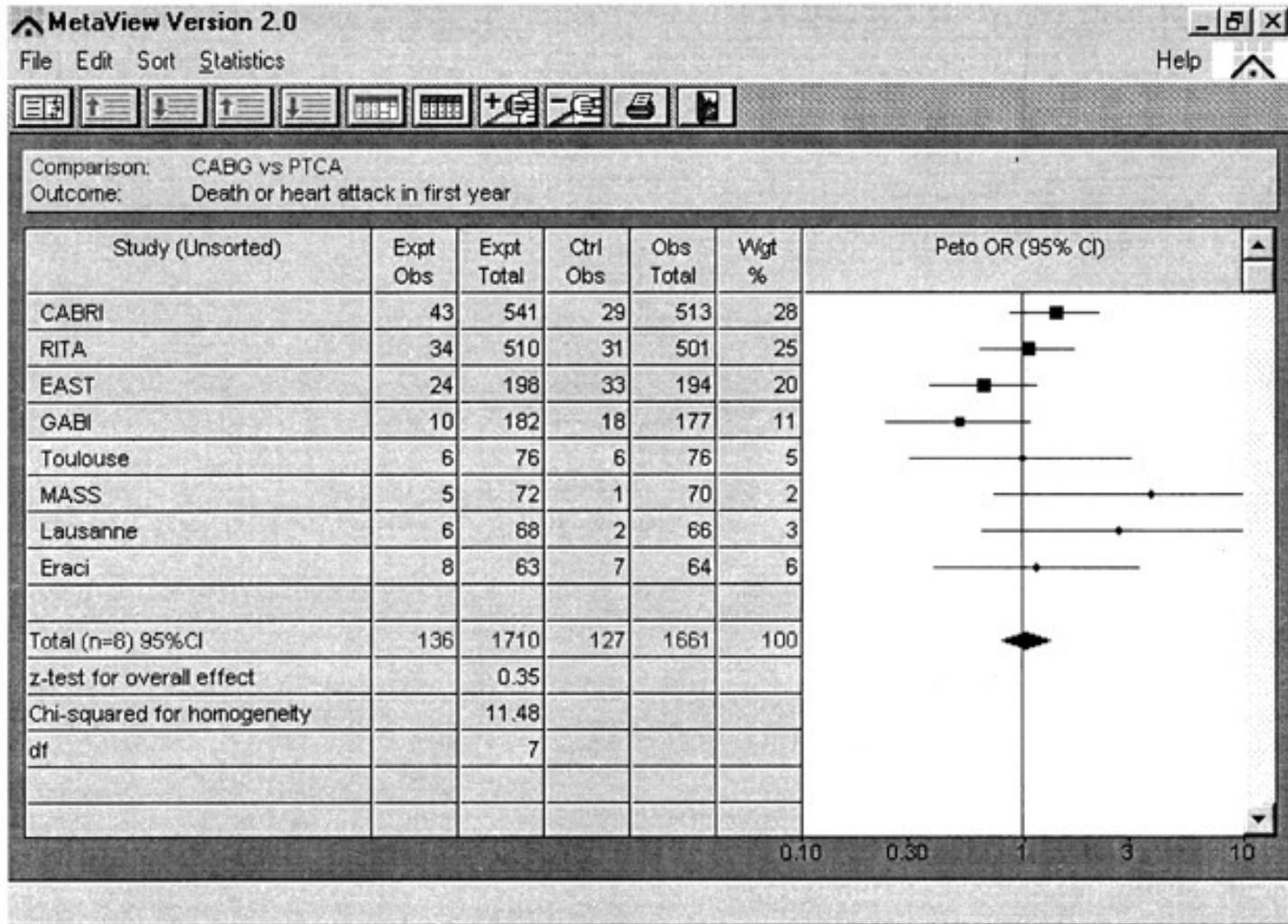
Fig 1 Methodology for a systematic review of randomised controlled trials¹

[\[View larger version \(196K\)\]](#)



PETER BROWN

[\[View larger version \(184K\)\]](#)



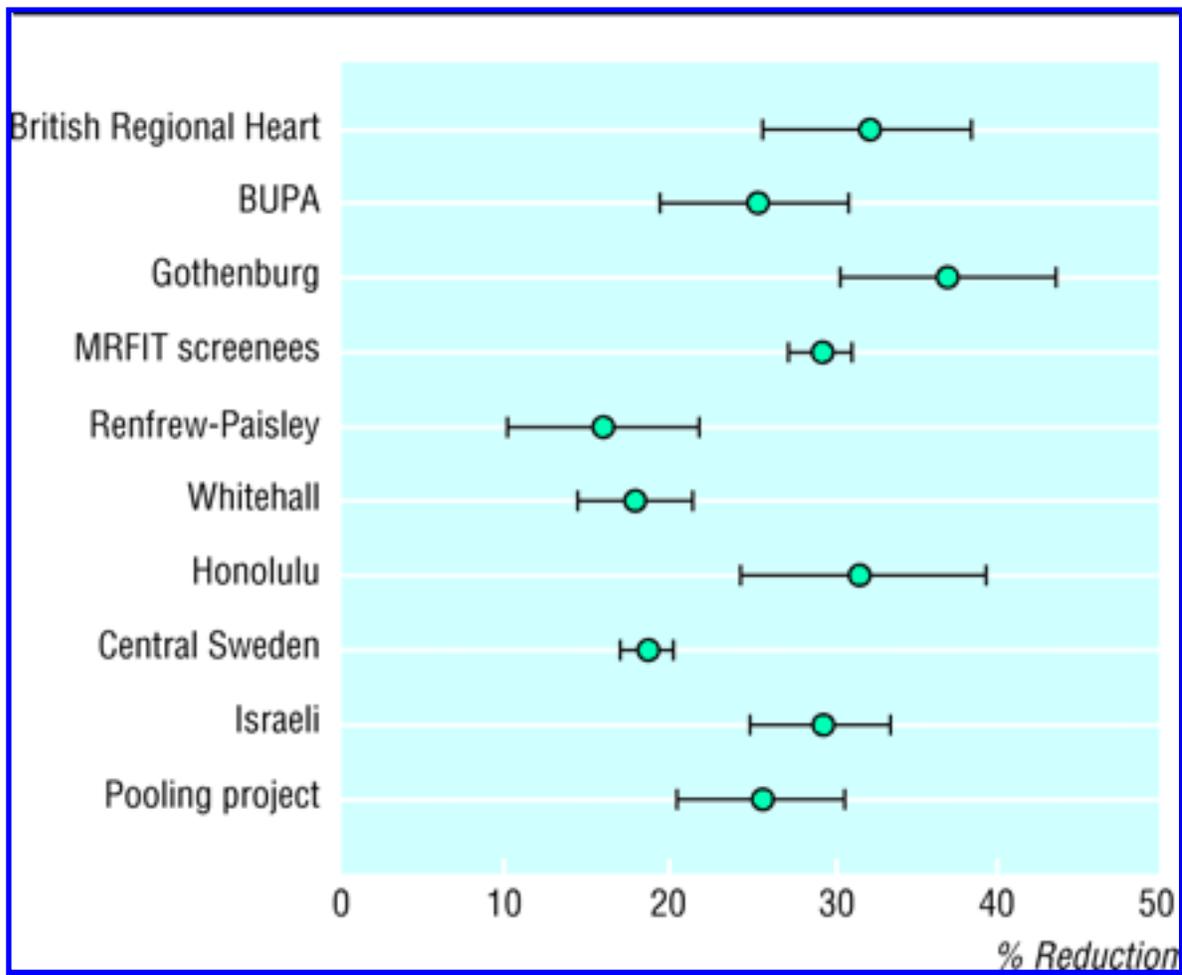


Fig 3 Reduction in risk of heart disease by strategies for lowering cholesterol. Reproduced with permission from Chalmers and Altman¹⁸

[\[View larger version \(107K\)\]](#)

BMJ 1997;315:740-743 (20 September)

Education and debate

How to read a paper: Papers that go beyond numbers (qualitative research)

Trisha **Greenhalgh**, senior lecturer,^a Rod Taylor, senior lecturer^b

^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, ^b Exeter and Devon Research and Development Support Unit, Postgraduate Medical School, Wonford, Exeter EX2 5EQ

Correspondence to: Dr **Greenhalgh** p.greenhalgh@ucl.ac.uk

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ This article has been cited by [other articles](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#) || [Taylor, R.](#)
- ▶ Alert me when:
[New articles cite this article](#)

▶ What is qualitative research?

Epidemiologist Nick Black has argued that a finding or a result is more likely to be accepted as a fact if it is quantified (expressed in numbers) than if it is not.¹ There is little or no scientific evidence, for example, to support the well known "facts" that one couple in 10 is infertile, or that one man in 10 is homosexual. Yet, observes Black, most of us are happy to accept uncritically such simplified, reductionist, and blatantly incorrect statements so long as they contain at least one number.

Researchers who use qualitative methods seek a deeper truth. They aim to "study things in their natural setting, attempting to make sense of, or interpret, phenomena in terms of the meanings people bring to them,"² and they use "a holistic perspective which preserves the complexities of human behaviour."¹

- ▲ [Top](#)
- [What is qualitative research?](#)
- ▼ [Evaluating papers that describe...](#)
- ▼ [Conclusion](#)
- ▼ [References](#)

Summary points

Qualitative methods aim to make sense of, or interpret, phenomena in terms of the meanings people bring to them

Qualitative research may define preliminary questions which can then be addressed in quantitative studies

A good qualitative study will address a clinical problem through a clearly formulated question and using more than one research method (triangulation)

Analysis of qualitative data can and should be done using explicit, systematic, and reproducible methods

Questions such as "How many parents would consult their general practitioner when their child has a mild temperature?" or "What proportion of smokers have tried to give up?" clearly need answering through quantitative methods. But questions like "Why do parents worry so much about their children's temperature?" and "What stops people giving up smoking?" cannot and should not be answered by leaping in and measuring the first aspect of the problem that we (the outsiders) think might be important. Rather, we need to listen to what people have to say, and we should explore the ideas and concerns which the subjects themselves come up with. After a while, we may notice a pattern emerging, which may prompt us to make our observations in a different way. We may start with one of the methods shown in box [box](#), and go on to use a selection of others.

Box 1: Examples of qualitative research methods

Documents—Study of documentary accounts of events, such as meetings

Passive observation—Systematic watching of behaviour and talk in natural occurring settings

Participant observation—Observation in which the researcher also occupies a role or part in the setting, in addition to observing

In depth interviews—Face to face conversation with the purpose of exploring issues or topics in detail. Does not use preset questions, but is shaped by a defined set of topics

Focus groups—Method of group interview which explicitly includes and uses the group interaction to generate data

Box [box](#) summarises (indeed, overstates) the differences between the qualitative and quantitative approaches to research. In reality, there is a great deal of overlap between them, the importance of which is increasingly being recognised.⁴

Box 2 Qualitative versus quantitative research—the overstated dichotomy

	Qualitative	Quantitative
Social theory	Action	Structure
Methods	Observation, interview	Experiment, survey
Question	What is X? (classification)	How many Xs? (enumeration)
Reasoning	Inductive	Deductive
Sampling method	Theoretical	Statistical
Strength	Validity	Reliability

Reproduced with permission from Mays and Pope, *Qualitative Research in Health Care*³

Quantitative research should begin with an idea (usually articulated as a hypothesis), which then, through measurement, generates data and, by deduction, allows a conclusion to be drawn. Qualitative research, in contrast, begins with an intention to explore a particular area, collects "data" (observations and interviews), and generates ideas and hypotheses from these data largely through what is known as inductive reasoning.³ The strength of the quantitative approach lies in its reliability (repeatability)—that is, the same measurements should yield the same results time after time. The strength of qualitative research lies in validity (closeness to the truth)—that is, good qualitative research, using a selection of data collection methods, really should touch the core of what is going on rather than just skimming the surface. The validity of qualitative methods is greatly improved by using a combination of research methods, a process known as triangulation, and by independent analysis of the data by more than one researcher.

The so called iterative approach (altering the research methods and the hypothesis as the study progresses, in the light of information gleaned along the way) used by qualitative researchers shows a commendable sensitivity to the richness and variability of the subject matter. Failure to recognise the legitimacy of this approach has, in the past, led critics to accuse qualitative researchers of continually moving their own goalposts. Though these criticisms are often misguided, there is, as Nicky Britten and colleagues have observed, a real danger "that the flexibility [of the iterative approach] will slide into sloppiness as the researcher ceases to be clear about what it is (s)he is investigating."⁵ These authors warn that qualitative researchers must, therefore, allow periods away from their fieldwork for reflection, planning, and consultation with colleagues.

Evaluating **papers** that describe qualitative research

By its very nature, qualitative research is non-standard, unconfined, and dependent on the subjective experience of both the researcher and the researched. It explores what needs to be explored and cuts its cloth accordingly. It is debatable, therefore, whether an all-encompassing critical appraisal checklist along the lines of the Users' Guides to the Medical Literature^{6 7 8 9 10 11 12 13 14 15 16 17 18 19} could ever be developed. Our own view, and that of a number of individuals who have attempted, or are currently working on, this very task,^{3 5} is that such a checklist may not be as exhaustive or as universally applicable as the various guides for appraising quantitative research, but that it is certainly possible to set some ground rules. The list which follows has been distilled from the published work cited earlier,^{2 3 5} and also from our own research and teaching experiences. You should note, however, that there is a great deal of disagreement and debate about the appropriate criteria for critical appraisal of qualitative research, and the ones given here are likely to be modified in the future.

- ▲ [Top](#)
- ▲ [What is qualitative research?](#)
- Evaluating **papers** that describe...
- ▼ [Conclusion](#)
- ▼ [References](#)

Question 1: Did the **paper describe an important clinical problem addressed via a clearly formulated question?**

A previous article in this series explained that one of the first things you should look for in any research **paper** is a statement of why the research was done and what specific question it addressed.²⁰ Qualitative **papers** are no exception to this rule: there is absolutely no scientific value in interviewing or observing people just for the sake of it. **Papers** that cannot define their topic of research more closely than "We decided to interview 20 patients with epilepsy" inspire little confidence that the researchers really knew what they were studying or why.

You might be more inclined to **read** on if the **paper** stated in its introduction something like, "Epilepsy is a common and potentially disabling condition, and up to 20% of patients do not remain free of fits while

taking medication. Antiepileptic medication is known to have unpleasant side effects, and several studies have shown that a high proportion of patients do not take their tablets regularly. We therefore decided to explore patients' beliefs about epilepsy and their perceived reasons for not taking their medication."

Question 2: Was a qualitative approach appropriate?

If the objective of the research was to explore, interpret, or obtain a deeper understanding of a particular clinical issue, qualitative methods were almost certainly the most appropriate ones to use. If, however, the research aimed to achieve some other goal (such as determining the incidence of a disease or the frequency of an adverse drug reaction, testing a cause and effect hypothesis, or showing that one drug has a better risk-benefit ratio than another), a case-control study, cohort study, or randomised trial may have been better suited to the research question.¹⁹

Question 3: How were the setting and the subjects selected?

The second [box](#) contrasts the statistical sampling methods of quantitative research with theoretical methods of qualitative research. In quantitative research, it is vital to ensure that a truly random sample of subjects is recruited so that the results reflect, on average, the condition of the population from which that sample was drawn.

In qualitative research, however, we are not interested in an "on average" view of a patient population. We want to gain an in depth understanding of the experience of particular individuals or groups; we should therefore deliberately seek out individuals or groups who fit the bill. If, for example, we wished to study the experience of non-English speaking British Punjabi women when they gave birth in hospital (with a view to tailoring the interpreting or advocacy service more closely to the needs of this patient group), we would be perfectly justified in going out of our way to find women who had had a range of different birth experiences—an induced delivery, an emergency caesarean section, a delivery by a medical student, a late miscarriage, and so on—rather than a "random" sample of British Punjabi mothers.

Question 4: What was the researcher's perspective, and has this been taken into account?



PETER BROWN

View larger version (131K):

[\[in this window\]](#)

[\[in a new window\]](#)

It is important to recognise that there is no way of abolishing, or fully controlling for, observer bias in qualitative research. This is most obviously the case when participant observation is used, but it is also true for other forms of data collection and of data analysis. If, for example, the research concerns the experience of asthmatic adults living in damp and overcrowded housing and the perceived effect of these surroundings on their health, the data generated by techniques such as focus groups or semistructured interviews are likely to be heavily influenced by what the interviewer believes about this subject and by whether he or she is employed by the hospital chest clinic, the social work department of the local authority, or an environmental pressure group. But since it is inconceivable that the interviews could have been conducted by someone with no views at all and no ideological or cultural perspective, the most that can be required of the researchers is that they describe in detail where they are coming from so that the results can be interpreted accordingly.

Question 5: What methods did the researcher use for collecting data—and are these described in enough detail?

I once spent two years doing highly quantitative, laboratory based experimental research in which around 15 hours of every week were spent filling or emptying test tubes. There was a standard way to fill the test tubes, a standard way to spin them in the centrifuge, and even a standard way to wash them up. When I finally published my research, some 900 hours of drudgery was summed up in a single sentence: "Patients' serum rhubarb levels were measured according to the method described by Bloggs et al [reference to Bloggs et al's published **paper**]."

The methods section of a qualitative **paper** often cannot be written in shorthand or dismissed by reference to someone else's research techniques. It may have to be lengthy and discursive since it is telling a unique story without which the results cannot be interpreted. As with the sampling strategy, there are no hard and fast rules about exactly what details should be included in this section of the **paper**. You should simply ask, "have I been given enough information about the methods used?", and, if you

have, use your common sense to assess, "are these methods a sensible and adequate way of addressing the research question?"

Question 6: What methods did the researcher use to analyse the data—and what quality control measures were implemented?

The data analysis section of a qualitative research **paper** is where sense can most **readily** be distinguished from nonsense. Having amassed a thick pile of completed interview transcripts or field notes, the genuine qualitative researcher has hardly begun. It is simply not good enough to flick through the text looking for "interesting quotes" which support a particular theory. The researcher must find a systematic way of analysing his or her data, and, in particular, must seek examples of cases which appear to contradict or challenge the theories derived from the majority.

One way of doing this is by content analysis: drawing up a list of coded categories and "cutting and pasting" each segment of transcribed data into one of these categories. This can be done either manually or, if large amounts of data are to be analysed, via a tailor-made computer database. The statements made by all the subjects on a particular topic can then be compared with one another, and more sophisticated comparisons can be made such as "did people who made statement A also tend to make statement B?"

In theory, the **paper** will show evidence of "quality control"—that is, the data (or at least, a sample of them) will have been analysed by more than one researcher to confirm that they are both assigning the same meaning to them, although in practice this is often difficult to achieve. Indeed, when researching this article, we could find no data on the interobserver reliability of any qualitative study to illustrate this point.

Question 7: Are the results credible, and if so, are they clinically important?

We obviously cannot assess the credibility of qualitative results through the precision and accuracy of measuring devices, nor their significance via confidence intervals and numbers needed to treat. It usually takes little more than plain common sense to determine whether the results are sensible and believable, and whether they matter in practice.

One important aspect of the results section to check is whether the authors cite actual data. Claims such as "general practitioners did not usually recognise the value of audit" would be infinitely more credible if one or two verbatim quotes from the interviewees were reproduced to illustrate them. The results should be independently and objectively verifiable—after all, a subject either made a particular statement or (s)he did not—and all quotes and examples should be indexed so that they can be traced back to an identifiable subject and setting.

Question 8: What conclusions were drawn, and are they justified by the results?

A quantitative research **paper** should clearly distinguish the study's results (usually a set of numbers) from the interpretation of those results (the discussion). The **reader** should have no difficulty separating

what the researchers *found* from what they think it *means*. In qualitative research, however, such a distinction is rarely possible, since the results are by definition an interpretation of the data.

It is therefore necessary, when assessing the validity of qualitative research, to ask whether the interpretation placed on the data accords with common sense and is relatively untainted with personal or cultural perspective. This can be a difficult exercise, because the language we use to describe things tends to impugn meanings and motives which the subjects themselves may not share. Compare, for example, the two statements, "three women went to the well to get water" and "three women met at the well and each was carrying a pitcher."

It is becoming a cliché that the conclusions of qualitative studies, like those of all research, should be "grounded in evidence"—that is, that they should flow from what the researchers found in the field. Mays and Pope suggest three useful questions for determining whether the conclusions of a qualitative study are valid:

- how well does this analysis explain why people behave in the way they do?
- how comprehensible would this explanation be to a thoughtful participant in the setting?; and
- how well does the explanation cohere with what we **already** know?³

Question 9: Are the findings of the study transferable to other clinical settings?

One of the commonest criticisms of qualitative research is that the findings of any qualitative study pertain only to the limited setting in which they were obtained. In fact, this is not necessarily any truer of qualitative research than of quantitative research. Look back at the example of British Punjabi women described above. You should be able to see that the use of a true *theoretical* sampling frame greatly increases the transferability of the results over a "convenience" sample.

▶ Conclusion

Doctors have traditionally placed high value on numerical data, which may in reality be misleading, reductionist, and irrelevant to the real issues. The increasing popularity of qualitative research in the biomedical sciences has arisen largely because quantitative methods provided either no answers or the wrong answers to important questions in both clinical care and service delivery.¹ If you still feel that qualitative research is necessarily second rate by virtue of being a "soft" science, you should be aware that you are out of step

- ▲ [Top](#)
- ▲ [What is qualitative research?](#)
- ▲ [Evaluating papers that describe...](#)
- [Conclusion](#)
- ▼ [References](#)

with the evidence.

In 1993, Pope and Britten presented a **paper** to the BSA Medical Sociology Group conference entitled "Barriers to qualitative methods in the medical mindset," in which they showed their collection of rejection letters from biomedical journals. The letters revealed a striking ignorance of qualitative methodology on the part of reviewers. In other words, the people who had rejected the **papers** often seemed to be incapable of distinguishing good qualitative research from bad. Somewhat ironically, qualitative **papers** of poor quality now appear regularly in some medical journals, whose editors have climbed on the qualitative bandwagon without gaining an ability to appraise such **papers**. Note, however, that the critical appraisal of qualitative research is a relatively underdeveloped science, and the questions posed in this chapter are still being refined.

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 Publishing Group: tel 0171 383 6185/6245; fax 0171 383 6662. Price £13.95 UK members, £14.95 non-members.

Acknowledgements

Thanks to Professor Nick Black for advice on this article.

References

1. Black N. Why we need qualitative research. *J Epidemiol Community Health* 1994;48:425-6. [[Medline](#)]
2. Denkin NK, Lincoln YS, eds. *Handbook of qualitative research*. London: Sage, 1994.
3. Mays N, Pope C, eds. *Qualitative research in health care*. London: BMJ Publishing Group, 1996.
4. Abell P. Methodological achievements in sociology over the past few decades with specific reference to the interplay of qualitative and quantitative methods. In: Bryant C, Becker H, eds. *What has sociology achieved?* London: Macmillan, 1990.

- ▲ [Top](#)
- ▲ [What is qualitative research?](#)
- ▲ [Evaluating papers that describe...](#)
- ▲ [Conclusion](#)
- [References](#)

5. Britten N, Jones R, Murphy E, Stacy R. Qualitative research methods in general practice and primary care. *Fam Pract* 1995;12:104-14.
6. Oxman AD, Sackett DL, Guyatt GH. Users' guides to the medical literature. I. How to get started. *JAMA* 1993;270:2093-5.
7. 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.
8. 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. [\[Medline\]](#)
9. 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. [\[Medline\]](#)
10. 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. [\[Medline\]](#)
11. 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. [\[Medline\]](#)
12. 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. [\[Medline\]](#)
13. Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. *JAMA* 1994;272:1367-71. [\[Medline\]](#)
14. 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. [\[Medline\]](#)
15. 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.
16. 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.
17. 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. [\[Medline\]](#)
18. 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. [\[Medline\]](#)
19. 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. [\[Medline\]](#)
20. **Greenhalgh** T. Getting your bearings (deciding what the **paper** is about). *BMJ* 1997;315:243-6. [\[Full Text\]](#)
21. Kinmonth A-L. Understanding and meaning in research and practice. *Fam Pract* 1995;12:1-2.

This article has been cited by other articles:

- Walker, S., McGeer, A., Simor, A. E., Armstrong-Evans, M., Loeb, M. (2000). Why are antibiotics prescribed for asymptomatic bacteriuria in institutionalized elderly people?: A qualitative study of physicians' and nurses' perceptions. *Can Med Assoc J* 163: 273-277 [[Abstract](#)] [[Full text](#)]
- Rychetnik, L, Frommer, M, Hawe, P, Shiell, A (2002). Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health* 56: 119-127 [[Abstract](#)] [[Full text](#)]
- Cook, D. J., Meade, M. O., Perry, A. G. (2001). Qualitative Studies on the Patient's Experience of Weaning From Mechanical Ventilation. *Chest* 120: 469S-473 [[Abstract](#)] [[Full text](#)]
- Cranney, M., Warren, E., Barton, S., Gardner, K., Walley, T. (2001). Why do GPs not implement evidence-based guidelines? A descriptive study. *Fam. Pract.* 18: 359-363 [[Abstract](#)] [[Full text](#)]
- Lloyd, G., Skarratts, D., Robinson, N., Reid, C. (2000). Communication skills training for emergency department senior house officers--a qualitative study. *Emerg Med J* 17: 246-250 [[Abstract](#)] [[Full text](#)]

- ▶ [Email this article to a friend](#)
- ▶ [Respond to this article](#)
- ▶ [PubMed citation](#)
- ▶ [Related articles in PubMed](#)
- ▶ [Download to Citation Manager](#)
- ▶ Search Medline for articles by:
[Greenhalgh, T.](#) || [Taylor, R.](#)
- ▶ Alert me when:
[New articles cite this article](#)

[Home](#)
[Help](#)
[Search/Archive](#)
[Feedback](#)
[Search Result](#)



PETER BROWN

[\[View larger version \(221K\)\]](#)