HOW TO READ A PAPER

The basics of evidence based medicine

Trisha Greenhalgh

BMJ Books

HOW TO READ A PAPER

2003.3.7

huangzhiman

For www.dnathink.org

HOW TO READ A PAPER

The basics of evidence based medicine

Second edition

TRISHA GREENHALGH
Department of Primary Care and Population Sciences
Royal Free and University College Medical School
London, UK

BMJ
Books

First published in 1997
Second impression 1997
Third impression 1998
Fourth impression 1998
Fifth impression 1999
Sixth impression 2000
Seventh impression 2000
Second Edition 2001
by the BMJ Publishing Group, BMA House, Tavistock Square,
London WC1H 9JR
www.bmjbooks.com

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-7279-1578-9

Cover by Landmark Design, Croydon, Surrey
Typeset by FiSH Books, London
Printed and bound by MPG Books Ltd, Bodmin

Contents

Foreword to the first edition

Not surprisingly, the wide publicity given to what is now called "evidence based medicine" has been greeted with mixed reactions by those who are involved in the provision of patient care. The bulk of the medical profession appears to be slightly hurt by the concept, suggesting as it does that until recently all medical practice was what Lewis Thomas has described as a frivolous and irresponsible kind of human experimentation, based on nothing but trial and error and usually resulting in precisely that sequence. On the other hand, politicians and those who administrate our health services have greeted the notion with enormous glee. They had suspected all along that doctors were totally uncritical and now they had it on paper. Evidence based medicine came as a gift from the gods because, at least as they perceived it, its implied efficiency must inevitably result in cost saving.

The concept of controlled clinical trials and evidence based medicine is not new, however. It is recorded that Frederick II, Emperor of the Romans and King of Sicily and Jerusalem, who lived from 1192 to 1250 AD and who was interested in the effects of exercise on digestion, took two knights and gave them identical meals. One was then sent out hunting and the other ordered to bed. At the end of several hours he killed both and examined the contents of their alimentary canals; digestion had proceeded further in the stomach of the sleeping knight. In the 17th century Jan Baptista van Helmont, a physician and philosopher, became sceptical of the practice of bloodletting. Hence he proposed what was almost certainly the first clinical trial involving large numbers, randomisation, and statistical analysis. This involved taking 200–500 poor people, dividing them into two groups by casting lots and protecting one from phlebotomy while allowing the other to be treated with as much bloodletting as his colleagues thought appropriate. The number of funerals in each group would be used to assess the efficacy of bloodletting. History does not record why this splendid experiment was never carried out.

If modern scientific medicine can be said to have had a beginning, it was in Paris in the mid-19th century where it had its roots in the work and teachings of Pierre Charles Alexandre Louis. Louis introduced statistical analysis to the evaluation of medical treatment and, incidentally, showed that bloodletting was a valueless form of treatment, though this did not change the habits of the physicians of the time or for many years to come. Despite this pioneering work, few clinicians on either side of the Atlantic urged that trials of clinical outcome should be adopted, although the principles of numerically based experimental design were enunciated in the 1920s by the geneticist Ronald Fisher. The field only started to make a major impact on clinical practice after the Second World War following the seminal work of Sir Austin Bradford Hill and the British epidemiologists who followed him, notably Richard Doll and Archie Cochrane.

But although the idea of evidence based medicine is not new, modern disciples like David Sackett and his colleagues are doing a great service to clinical practice, not just by popularising the idea but by bringing home to clinicians the notion that it is not a dry academic subject but more a way of thinking that should permeate every aspect of medical practice. While much of it is based on megatrials and meta-analyses, it should also be used to influence almost everything that a doctor does. After all, the medical profession has been brainwashed for years by examiners in medical schools and Royal Colleges to believe that there is only one way of examining a patient. Our bedside rituals could do with as much critical evaluation as our operations and drug regimes; the same goes for almost every aspect of doctoring.

As clinical practice becomes busier and time for reading and reflection becomes even more precious, the ability effectively to peruse the medical literature and, in the future, to become familiar with a knowledge of best practice from modern communication systems will be essential skills for doctors. In this lively book, Trisha Greenhalgh provides an excellent approach to how to make best use of medical literature and the benefits of evidence based medicine. It should have equal appeal for first-year medical students and grey-haired consultants and deserves to be read widely.

With increasing years, the privilege of being invited to write a foreword to a book by one's ex-students becomes less of a rarity. Trisha Greenhalgh was the kind of medical student who never let her teachers get away with a loose thought and this inquiring attitude seems to have flowered over the years; this is a splendid and timely book and I wish it all the success it deserves. After all, the concept of evidence based medicine is nothing more than the state of mind that every clinical teacher hopes to develop in their students; Dr Greenhalgh's sceptical but constructive approach to medical literature suggests that such a happy outcome is possible at least once in the lifetime of a professor of medicine.

Professor Sir David Weatherall

In November 1995, my friend Ruth Holland, book reviews editor of the *British Medical Journal*, suggested that I write a book to demystify the important but often inaccessible subject of evidence based medicine. She provided invaluable comments on earlier drafts of the manuscript but was tragically killed in a train crash on 8th August 1996. This book is dedicated to her memory.

Preface

When I wrote this book in 1996, evidence based medicine was a bit of an unknown quantity. A handful of academics (including me) were enthusiastic and had already begun running "training the trainers" courses to disseminate what we saw as a highly logical and systematic approach to clinical practice. Others – certainly the majority of clinicians – were convinced that this was a passing fad that was of limited importance and would never catch on. I wrote *How to read a paper* for two reasons. First, students on my own courses were asking for a simple introduction to the principles presented in what was then known as "Dave Sackett's big red book" (Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991) – an outstanding and inspirational volume that was already in its fourth reprint, but which some novices apparently found a hard read. Second, it was clear to me that many of the critics of evidence based medicine didn't really understand what they were dismissing and that until they did, serious debate on the political, ideological, and pedagogical place of evidence based medicine as a discipline could not begin.

I am of course delighted that *How to read a paper* has become a standard reader in many medical and nursing schools and has so far been translated into French, German, Italian, Polish, Japanese, and Russian. I am also delighted that what was so recently a fringe subject in academia has been well and truly mainstreamed in clinical service in the UK. For example, it is now a contractual requirement for all doctors, nurses, and pharmacists to practise (and for managers to manage) according to best research evidence.

In the three and a half years since the first edition of this book was published, evidence based medicine has become a growth industry. Dave Sackett's big red book and Trisha Greenhalgh's little blue book have been joined by some 200 other textbooks and 1500 journal articles offering different angles on the 12 topics covered briefly in the chapters which follow. My biggest task in preparing this second edition has been to update and extend the reference lists to reflect the wide range of excellent material now available to those who wish to go beyond the basics. Nevertheless, there is clearly still room on the bookshelves for a no-frills introductory text so I have generally resisted the temptation to go into greater depth in these pages.

Trisha Greenhalgh

Preface to the first edition: Do you need to read this book?

This book is intended for anyone, whether medically qualified or not, who wishes to find their way into the medical literature, assess the scientific validity and practical relevance of the articles they find, and, where appropriate, put the results into practice. These skills constitute the basics of evidence based medicine.

I hope this book will help you to read and interpret medical papers better. I hope, in addition, to convey a further message, which is this. Many of the descriptions given by cynics of what evidence based medicine is (the glorification of things that can be measured without regard for the usefulness or accuracy of what is measured; the uncritical acceptance of published numerical data; the preparation of all-encompassing guidelines by self-appointed "experts" who are out of touch with real medicine; the debasement of clinical freedom through the imposition of rigid and dogmatic clinical protocols; and the overreliance on simplistic, inappropriate, and often incorrect economic analyses) are actually criticisms of what the evidence based medicine movement is fighting *against,* rather than of what it represents.

Do not, however, think of me as an evangelist for the gospel according to evidence based medicine. I believe that the science of finding, evaluating and implementing the results of medical research can, and often does, make patient care more objective, more logical, and more cost effective. If I didn't believe that, I wouldn't spend so much of my time teaching it and trying, as a general practitioner, to practise it. Nevertheless, I believe that when applied in a vacuum (that is, in the absence of common sense and without regard to the individual circumstances and priorities of the person being offered treatment), the evidence based approach to patient care is a reductionist process with a real potential for harm.

Finally, you should note that I am neither an epidemiologist nor a statistician but a person who reads papers and who has developed a pragmatic (and at times unconventional) system for testing their merits. If you wish to pursue the epidemiological or statistical themes covered in this book, I would encourage you to move on to a more definitive text, references for which you will find at the end of each chapter.

Trisha Greenhalgh

Acknowledgments

I am not by any standards an expert on all the subjects covered in this book (in particular, I am very bad at sums) and I am grateful to the people listed below for help along the way. I am, however, the final author of every chapter and responsibility for any inaccuracies is mine alone.

1. To Professor Dave Sackett and Professor Andy Haines who introduced me to the subject of evidence based medicine and encouraged me to write about it.
2. To Dr Anna Donald, who broadened my outlook through valuable discussions on the implications and uncertainties of this evolving discipline.
3. To the following medical informaticists (previously known as librarians), for vital input into Chapter 2 and the appendices on search strings: Mr Reinhardt Wentz of Charing Cross and Westminster Medical School, London; Ms Jane Rowlands of the BMA library in London; Ms Carol Lefebvre of the UK Cochrane Centre, Summertown Pavilion, Oxford; and Ms Valerie Wildridge of the King's Fund library in London. I strongly recommend Jane Rowlands' Introductory and Advanced Medline courses at the BMA library.
4. To the following expert advisers and proofreaders: Dr Sarah Walters and Dr Jonathan Elford (Chapters 3, 4, and 7), Dr Andrew Herxheimer (Chapter 6), Professor Sir Iain Chalmers (Chapter 8), Professor Brian Hurwitz (Chapter 9), Professor Mike Drummond and Dr Alison Tonks (Chapter 10), Professor Nick Black and Dr Rod Taylor (Chapter 11), and Mr John Dobby (Chapters 5 and 12).
5. To Mr Nick Mole, of Ovid Technologies Ltd, for checking Chapter 2 and providing demonstration software for me to play with.

6. To the many people, too numerous to mention individually, who took time to write in and point out both typographical and factual errors in the first edition. As a result of their contributions, I have learnt a great deal (especially about statistics) and the book has been improved in many ways. Some of the earliest critics of *How to Read a Paper* have subsequently worked with me on my teaching courses in evidence based practice; several have co-authored other papers or book chapters with me, and one or two have become personal friends.

Thanks also to my family for sparing me the time and space to finish this book.

Chapter 1: Why read papers at all?

1.1 Does "evidence based medicine" simply mean "reading medical papers"?

Evidence based medicine is much more than just reading papers. According to the most widely quoted definition, it is "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients".[1] I find this definition very useful but it misses out what for me is a very important aspect of the subject – and that is the use of mathematics. Even if you know almost nothing about evidence based medicine you know it talks a lot about numbers and ratios! Anna Donald and I recently decided to be upfront about this and proposed this alternative definition:

> "Evidence-based medicine is the enhancement of a clinician's traditional skills in diagnosis, treatment, prevention and related areas through the systematic framing of relevant and answerable questions and the use of mathematical estimates of probability and risk".[2]

If you follow an evidence based approach, therefore, all sorts of issues relating to your patients (or, if you work in public health medicine, planning or purchasing issues relating to groups of patients or patient populations) will prompt you to ask questions about scientific evidence, seek answers to those questions in a systematic way, and alter your practice accordingly.

You might ask questions, for example, about a patient's symptoms ("In a 34 year old man with left-sided chest pain, what is the probability that there is a serious heart problem, and if there is, will it show up on a resting ECG?"), about physical or diagnostic signs ("In an otherwise uncomplicated childbirth, does the presence of meconium [indicating fetal bowel movement] in the amniotic fluid indicate significant deterioration in the physiological state of the fetus?"), about the prognosis of an illness ("If a previously well 2 year old has a short fit associated with a high temperature, what is the chance that she will subsequently develop epilepsy?"), about therapy ("In patients with an acute myocardial infarction [heart attack], are the risks associated with thrombolytic drugs [clotbusters] outweighed by the benefits, whatever the patient's age, sex, and ethnic origin?"), about cost effectiveness ("In order to reduce the suicide rate in a health district, is it better to employ more consultant psychiatrists, more community psychiatric nurses or more counsellors?"), and about a host of other aspects of health and health services.

Professor Dave Sackett, in the opening editorial of the very first issue of the journal *evidence based Medicine*,[3] summarised the essential steps in the emerging science of evidence based medicine.

- To convert our information needs into answerable questions (i.e. to formulate the problem).
- To track down, with maximum efficiency, the best evidence with which to answer these questions – which may come from the clinical examination, the diagnostic laboratory, the published literature or other sources.
- To appraise the evidence critically (i.e. weigh it up) to assess its validity (closeness to the truth) and usefulness (clinical applicability).
- To implement the results of this appraisal in our clinical practice.
- To evaluate our performance.

Hence, evidence based medicine requires you not only to read papers but to read the *right* papers at the *right* time and then to alter your behaviour (and, what is often more difficult, the behaviour of other people) in the light of what you have found. I am concerned that the plethora of how-to-do-it courses in evidence based medicine so often concentrate on the third of these five steps (critical appraisal) to the exclusion of all the others. Yet if you have asked the wrong question or sought answers from the wrong sources, you might as well not read any papers at all. Equally, all your training in search techniques and critical appraisal will go to waste if you do not put at least as much effort into implementing valid evidence and measuring progress towards your goals as you do into reading the paper.

If I were to be pedantic about the title of this book, these broader aspects of evidence based medicine should not even get a mention here. But I hope you would have demanded your money back if I had omitted the final section of this chapter (Before you start: formulate the problem), Chapter 2 (Searching the literature), and Chapter 12 (Implementing evidence based findings). Chapters 3–11 describe step three of the evidence based medicine process: critical appraisal, i.e. what you should do when you actually have the paper in front of you.

Incidentally, if you are computer literate and want to explore the subject of evidence based medicine on the Internet, you could try the following websites. If you're not, don't worry (and don't worry either when you discover that there are over 200 websites dedicated to evidence based medicine – they all offer very similar material and you certainly don't need to visit them all).

- **Oxford Centre for evidence based Medicine** A well kept website from Oxford, UK, containing a wealth of resources and links for EBM. http://cebm.jr2.ox.ac.uk
- **POEMs (Patient Oriented Evidence that Matters)** Summaries of evidence that is felt to have a direct impact on patients' choices, compiled by the US *Journal of Family Practice*. http://jfp.msu.edu/jclub/indexes/jcindex.htm
- **SCHARR Auracle** Evidence based, information seeking, well presented links to other evidence based health care sites by the Sheffield Centre for Health and Related Research in the UK. http://panizzi.shef.ac.uk/auracle/aurac.html

1.2 Why do people often groan when you mention evidence based medicine?

Critics of evidence based medicine might define it as: "the increasingly fashionable tendency of a group of young, confident and highly numerate medical academics to belittle the performance of experienced clinicians using a combination of epidemiological jargon and statistical sleight-of-hand" or "the argument, usually presented with near-evangelistic zeal, that no health related action should ever be taken by a doctor, a nurse, a purchaser of health services or a politician unless and until the results of several large and expensive research trials have appeared in print and been approved by a committee of experts".

Others have put their reservations even more strongly.

> "evidence based medicine seems to [replace] original findings with subjectively selected, arbitrarily summarised, laundered, and biased conclusions of indeterminate validity or completeness. It has been carried out by people of unknown ability, experience, and skills using methods whose opacity prevents assessment of the original data".[4]

The palpable resentment amongst many health professionals towards the evidence based medicine movement[5, 6] is mostly a reaction to the implication that doctors (and nurses, midwives, physiotherapists, and other health professionals) were functionally illiterate until they were shown the light and that the few who weren't illiterate wilfully ignored published medical evidence. Anyone who works face to face with patients knows how often it is necessary to seek new information before making a clinical decision. Doctors have spent time in libraries since libraries were invented. We don't put a patient on a new drug without evidence that it is likely to work; apart from anything else, such off licence use of medication is, strictly speaking, illegal. Surely we have all been practising evidence based medicine for years, except when we were deliberately bluffing (using the "placebo" effect for good medical reasons), or when we were ill, overstressed or consciously being lazy?

Well, no, we haven't. There have been a number of surveys on the behaviour of doctors, nurses, and related professionals,[7–10] and most of them reached the same conclusion: clinical decisions are only rarely based on the best available evidence. Estimates in the early 1980s suggested that only around 10–20% of medical interventions (drug therapies, surgical operations, X-rays, blood tests, and so on) were based on sound scientific evidence.[11, 12] These figures have since been disputed, since they were derived by assessing all diagnostic and therapeutic procedures currently in use, so that each procedure, however obscure, carried equal weight in the final fraction. A more recent evaluation using this method classified 21% of health technologies as evidence based.[13]

Surveys which look at the interventions chosen for consecutive series of patients, which reflect the technologies that are actually used rather than simply those that are on the market, have suggested that 60–90% of clinical decisions, depending on the specialty, are "evidence based".[14–18] But as I have argued elsewhere,[19] these studies had methodological limitations. Apart from anything else, they were undertaken in specialised units and looked at the practice of world experts in evidence based medicine; hence, the figures arrived at can hardly be generalised beyond their immediate setting (see section 4.2).

Let's take a look at the various approaches which health professionals use to reach their decisions in reality, all of which are examples of what evidence based medicine *isn't*.

Decision making by anecdote

When I was a medical student, I occasionally joined the retinue of a distinguished professor as he made his daily ward rounds. On seeing a new patient, he would enquire about the patient's symptoms, turn to the massed ranks of juniors around the bed and relate the story of a similar patient encountered 20 or 30 years previously. "Ah, yes. I remember we gave her such-and-such, and she was fine after that." He was cynical, often rightly, about new drugs and technologies and his clinical acumen was second to none. Nevertheless, it had taken him 40 years to accumulate his expertise and the largest medical textbook of all – the collection of cases which were outside his personal experience – was forever closed to him.

Anecdote (storytelling) has an important place in professional learning[20] but the dangers of decision making by anecdote are well illustrated by considering the risk–benefit ratio of drugs and medicines. In my first pregnancy, I developed severe vomiting and was given the anti-sickness drug prochlorperazine (Stemetil). Within minutes, I went into an uncontrollable and very distressing neurological spasm. Two days later, I had recovered fully from this idiosyncratic reaction but I have never prescribed the drug since, even though the estimated prevalence of neurological reactions to prochlorperazine is only one in several thousand cases. Conversely, it is tempting to dismiss the possibility of rare but potentially serious adverse effects from familiar drugs – such as thrombosis on the contraceptive pill – when one has never encountered such problems in oneself or one's patients.

We clinicians would not be human if we ignored our personal clinical experiences, but we would be better advised to base our decisions on the collective experience of thousands of clinicians treating millions of patients, rather than on what we as individuals have seen and felt. Chapter 5 of this book (Statistics for the non-statistician) describes some more objective methods, such as the number needed to treat (NNT) for deciding whether a particular drug (or other intervention) is likely to do a patient significant good or harm.

Decision making by press cutting

For the first 10 years after I qualified, I kept an expanding file of papers which I had ripped out of my medical weeklies before binning the less interesting parts. If an article or editorial seemed to have something new to say, I consciously altered my clinical practice in line with its conclusions. All children with suspected urinary tract infections should be sent for scans of the kidneys to exclude congenital abnormalities, said one article, so I began referring anyone under the age of 16 with urinary symptoms for specialist investigations. The advice was in print and it was recent, so it must surely replace traditional practice – in this case, referring only children below the age of 10 who had had two well documented infections.

This approach to clinical decision making is still very common. How many doctors do you know who justify their approach to a particular clinical problem by citing the results section of a single published study, even though they could not tell you anything at all about the methods used to obtain those results? Was the trial randomised and controlled (see section 3.3)? How many patients, of what age, sex, and disease severity, were involved (see section 4.2)? How many withdrew from ("dropped out of") the study, and why (see section 4.6)? By what criteria were patients judged cured? If the findings of the study appeared to contradict those of other researchers, what attempt was made to validate (confirm) and replicate (repeat) them (see section 7.3)? Were the statistical tests which allegedly proved the authors' point appropriately chosen and correctly performed (see Chapter 5)? Doctors (and nurses, midwives, medical managers, psychologists, medical students, and consumer activists) who like to cite the results of medical research studies have a responsibility to ensure that they first go through a checklist of questions like these (more of which are listed in Appendix 1).

Decision making by expert opinion (eminence based medicine)

An important variant of decision making by press cutting is the use of "off the peg" reviews, editorials, consensus statements, and guidelines. The medical freebies (free medical journals and other "information sheets" sponsored directly or indirectly by the pharmaceutical industry) are replete with potted recommendations and at-a-glance management guides. But who says the advice given in a set of guidelines, a punchy editorial or an amply referenced "overview" is correct?

Professor Cynthia Mulrow, one of the founders of the science of systematic review (see Chapter 8), has shown that experts in a particular clinical field are actually *less* likely to provide an objective review of all the available evidence than a non-expert who approaches the literature with unbiased eyes.[21] In extreme cases, an "expert review" may consist simply of the lifelong bad habits and personal press cuttings of an ageing clinician. Chapter 8 of the book takes you through a checklist for assessing whether a "systematic review" written by someone else really merits the description and Chapter 9 discusses the potential limitations of "off the peg" clinical guidelines.

Decision making by cost minimisation

The general public is usually horrified when it learns that a treatment has been withheld from a patient for reasons of cost. Managers, politicians, and, increasingly, doctors can count on being pilloried by the press when a child with a brain tumour is not sent to a specialist unit in America or a frail old lady is denied indefinite board and lodging on an acute medical ward. Yet in the real world, all health care is provided from a limited budget and it is increasingly recognised that clinical decisions must take into account the economic costs of a given intervention. As Chapter 10 argues, clinical decision making *purely* on the grounds of cost ("cost minimisation" – purchasing the cheapest option with no regard for how effective it is) is usually both senseless and cruel and we are right to object vocally when this occurs.

Expensive interventions should not, however, be justified simply because they are new or because they ought to work in theory or because the only alternative is to do nothing – but because they are very likely to save life or significantly improve its quality. How, though, can the benefits of a hip replacement in a 75 year old be meaningfully compared with those of cholesterol lowering drugs in a middle aged man or infertility investigations for a couple in their 20s? Somewhat counterintuitively, there is no self evident set of ethical principles or analytical tools which we can use to match limited resources to unlimited demand. As you will see in Chapter 10, the much derided quality adjusted life year (QALY) and similar utility based units are simply attempts to lend some objectivity to the illogical but unavoidable comparison of apples with oranges in the field of human suffering.

There is another reason why some people find the term "evidence based medicine" unpalatable. This chapter has argued that evidence based medicine is about coping with change, not about knowing all the answers before you start. In other words, it is not so much about what you have read in the past but about how you go about identifying and meeting your ongoing learning needs and applying your knowledge appropriately and consistently in new clinical situations. Doctors who were brought up in the old school style of never admitting ignorance may find it hard to accept that some aspect of scientific uncertainty is encountered, on average, three times for every two patients seen by experienced teaching hospital consultants[22] (and, no doubt, even more often by their less up to date provincial colleagues). An evidence based approach to ward rounds may turn the traditional medical hierarchy on its head when the staff nurse or junior doctor produces new evidence that challenges what the consultant taught everyone last week. For some senior clinicians, learning the skills of critical appraisal is the least of their problems in adjusting to an evidence based teaching style! If you are interested in reading more about the philosophy and sociology of evidence based medicine, try the references listed at the end of this chapter.[23, 24]

1.3 Before you start: formulate the problem

When I ask my medical students to write me an essay about high blood pressure, they often produce long, scholarly, and essentially correct statements on what high blood pressure is, what causes it, and what the treatment options are. On the day they hand their essays in, most of them know far more about high blood pressure than I do. They are certainly aware that high blood pressure is the single most common cause of stroke and that detecting and treating everyone's high blood pressure would cut the incidence of stroke by almost half. Most of them are aware that stroke, though devastating when it happens, is a fairly rare event and that blood pressure tablets have side effects such as tiredness, dizziness, impotence, and getting "caught short" when a long way from the lavatory.

But when I ask my students a practical question such as "Mrs Jones has developed light-headedness on these blood pressure tablets and she wants to stop all medication; what would you advise her to do?", they are foxed. They sympathise with Mrs Jones' predicament, but they cannot distil from their pages of close written text the one thing that Mrs Jones needs to know. As Richard Smith (paraphrasing T S Eliot) asked a few years ago in a *BMJ* editorial: "Where is the wisdom we have lost in knowledge, and the knowledge we have lost in information?".[25]

Experienced doctors (and many nurses) might think they can answer Mrs Jones' question from their own personal experience. As I argued earlier in this chapter, few of them would be right.[7] And even if they were right on this occasion, they would still need an overall system for converting the ragbag of information about a patient (an ill defined set of symptoms, physical signs, test results, and knowledge of what happened to this patient or a similar patient last time), the particular anxieties and values (utilities) of the patient, and other things that could be relevant (a hunch, a half-remembered article, the opinion of an older and wiser colleague or a paragraph discovered by chance while flicking through a textbook) into a succinct summary of what the problem is and what specific additional items of information we need to solve that problem.

Sackett and colleagues have recently helped us by dissecting the parts of a good clinical question.[26]

- First, define precisely *whom* the question is about (i.e. ask "How would I describe a group of patients similar to this one?").

- Next, define *which* manoeuvre you are considering in this patient or population (for example, a drug treatment) and, if necessary, a comparison manoeuvre (for example, placebo or current standard therapy).

- Finally, define the desired (or undesired) *outcome* (for example, reduced mortality, better quality of life, overall cost savings to the health service, and so on).

The second step may not, in fact, concern a drug treatment, surgical operation or other intervention. The “manoeuvre” could, for example, be the exposure to a putative carcinogen (something that might cause cancer) or the detection of a particular surrogate endpoint in a blood test or other investigation. (A surrogate endpoint, as section 6.3 explains, is something that predicts, or is said to predict, the later development or progression of disease. In reality, there are very few tests which reliably act as crystal balls for patients’ medical future. The statement “The doctor looked at the test results and told me I had six months to live” usually reflects either poor memory or irresponsible doctoring!). In both these cases, the “outcome” would be the development of cancer (or some other disease) several years later. In most clinical problems with individual patients, however, the “manoeuvre” consists of a specific intervention initiated by a health professional.

Thus, in Mrs Jones’ case, we might ask, “In a 68 year old white woman with essential (i.e. common-or-garden) hypertension (high blood pressure), no co-existing illness, and no significant past medical history, do the benefits of continuing therapy with hydrochlorthiazide (chiefly, reduced risk of stroke) outweigh the inconvenience?”. Note that in framing the specific question, we have already established that Mrs Jones has never had a heart attack, stroke or early warning signs such as transient paralysis or loss of vision. If she had, her risk of subsequent stroke would be much higher and we would, rightly, load the risk–benefit equation to reflect this.

In order to answer the question we have posed, we must determine not just the risk of stroke in untreated hypertension but also the likely reduction in that risk which we can expect with drug treatment. This is, in fact, a rephrasing of a more general question (“Do the benefits of treatment in this case outweigh the risks?”) which we should have asked before we prescribed hydrochlorthiazide to Mrs Jones in the first place, and which all doctors should, of course, ask themselves every time they reach for their prescription pad.

Remember that Mrs Jones’ alternative to staying on this particular drug is not necessarily to take no drugs at all; there may be other drugs with equivalent efficacy but less disabling side effects (remember that, as Chapter 6 argues, too many clinical trials of new drugs compare the product with placebo rather than with the best available alternative) or non-medical treatments such as exercise, salt restriction, homeopathy or acupuncture. Not all of these approaches would help Mrs Jones or be acceptable to her, but it would be quite appropriate to seek evidence as to *whether* they might help her.

We will probably find answers to some of these questions in the medical literature and Chapter 2 describes how to search for relevant papers once you have formulated the problem. But before you start, give one last thought to your patient with high blood pressure. In order to determine her personal priorities (how does she value a 10% reduction in her risk of stroke in five years' time compared to the inability to go shopping unaccompanied today?), you will need to approach Mrs Jones, not a blood pressure specialist or the Medline database!

In the early days of evidence based medicine, there was considerable enthusiasm for using a decision tree approach to incorporate the patient's perspective into an evidence based treatment choice.[27, 28] In practice, this often proves impossible, because (I personally would argue) patients' experiences are complex stories that refuse to be reduced to a tree of yes/no decisions.[29] Perhaps the most powerful criticism of evidence based medicine is that it potentially dismisses the patient's own perspective on their illness in favour of an average effect on a population sample or a column of QALYs (see Chapter 10) calculated by a medical statistician.[29–31] In the past few years the evidence based medicine movement has made rapid progress in developing a more practical methodology for incorporating the patient's perspective in clinical decision making, [19, 32] the introduction of evidence based policy,[33] and the design and conduct of research trials.[34, 35] I have attempted to incorporate the patient's perspective into Sackett's five-stage model for evidence based practice;[1] the resulting eight stages, which I have called a context sensitive checklist for evidence based practice, are shown in Appendix 1.

Exercise 1

1. Go back to the fourth paragraph in this chapter, where examples of clinical questions are given. Decide whether each of these is a properly focused question in terms of:

   - the patient or problem
   - the manoeuvre (intervention, prognostic marker, exposure)
   - the comparison manoeuvre, if appropriate
   - the clinical outcome.

2. Now try the following.

   a) A 5 year old child has been on high dose topical steroids for severe eczema since the age of 20 months. The mother believes that the steroids are stunting the child's growth and wishes to change to homeopathic treatment. What information does the dermatologist need to decide (a) whether she is right about the topical steroids and (b) whether homeopathic treatment will help this child?

   b) A woman who is nine weeks pregnant calls out her GP because of abdominal pain and bleeding. A previous ultrasound scan has confirmed that the pregnancy is not ectopic. The GP decides that she might be having a miscarriage and tells her she must go into hospital for a scan and, possibly, an operation to clear out the womb. The woman refuses. What information do they both need in order to establish whether hospital admission is medically necessary?

   c) In the UK, most parents take their babies at the ages of 6 weeks, 8 months, 18 months, and 3 years for developmental checks, where a doctor listens for heart murmurs, feels the abdomen and checks that the testicles are present, and a nurse shakes a rattle and counts how many bricks the infant can build into a tower. Ignoring the social aspects of "well baby clinics", what information would you need to decide whether the service is a good use of health resources?

1 Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. evidence based medicine: what it is and what it isn't. *BMJ* 1996; **312**: 71–2.

2 Donald A, Greenhalgh T. *A hands-on guide to evidence based health care: practice and implementation*. Oxford: Blackwell Science, 2000; in press.

3 Sackett DL, Haynes B. On the need for evidence based medicine. *evidence based Medicine* 1995; **1**: 4–5.

4 James NT. Scientific method and raw data should be considered (letter). *BMJ* 1996; **313**: 169–70.

5 Stradling JR, Davies RJO. The unacceptable face of evidence based medicine. *J Eval Clin Pract* 1997; **3**:99–103.

6 Black D. The limitations to evidence. *J R Coll Physicians Lond* 1998; **32**:23–6.

7 Institute of Medicine. *Guidelines for clinical practice: from development to use*. Washington DC: National Academy Press, 1992.

8 Brook RH, Williams KN, Avery SB. Quality assurance today and tomorrow: forecast for the future. *Ann Intern Med* 1976; **85**: 809–17.

9 Roper WL, Winkenwerde W, Hackbarth GM, Krakauer H. Effectiveness in health care: an initiative to evaluate and improve medical practice. *New Engl J Med* 1988; **319**: 1197–202.

10 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991:305–33.

11 Office of Technology Assessment of the Congress of the United States. *The impact of randomised clinical trials on health policy and medical practice*. Washington DC: US Government Printing Office, 1983.

12 Williamson JW, Goldschmidt PG, Jillson IA. *Medical Practice Information Demonstration Project: final report*. Baltimore, Maryland: Policy Research, 1979.

13 Dubinsky M, Ferguson JH. Analysis of the National Institutes of Health Medicare Coverage Assessment. *Int J Technol Assess Health Care* 1990; **6**: 480–8.

14 Ellis J, Mulligan I, Rowe J, Sackett DL. Inpatient general medicine is evidence based. A-team, Nuffield Department of Clinical Medicine. *Lancet* 1995; **346**: 407–10.

15 Gill P, Dowell AC, Neal RD, Smith N, Heywood P, Wilson AE. Evidence based general practice: a retrospective study of interventions in one training practice. *BMJ* 1996; **312**: 819–21.

16 Geddes J, Game D, Jenkins N, Peterson LA, Pottinger GR, Sackett DL. Inpatient psychiatric treatment is evidence based. *Qual Health Care* 1996; **4**: 215–17.

17 Myles PS, Bain DL, Johnson F, McMahon R. Is anaesthesia evidence based? A survey of anaesthetic practice. *Br J Anaesthesia* 1999; **82**:591–5.

18 Howes N, Chagla L, Thorpe M, McCulloch P. Surgical practice is evidence based. *Br J Surg* 1997; **84**:1220–3.

19 Greenhalgh T. Is my practice evidence based? (editorial) *BMJ* 1996; **313**: 957–8.

20 Macnaughton J. Anecdote in clinical practice. In: Greenhalgh T, Hurwitz B, eds. *Narrative based medicine: dialogue and discourse in clinical practice*. London: *BMJ* Publications, 1999: 202–11.

21 Mulrow C. Rationale for systematic reviews. *BMJ* 1994; **309**: 597–9.

22 Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? *Ann Intern Med* 1985; **103**: 596–9.

23 Tanenbaum SJ. Evidence and expertise: the challenge of the outcomes movement to medical professionalism. *Acad Med* 1999; **74**:757–63.

24 Tonelli MR. The philosophical limits of evidence based medicine. *Acad Med* 1998; **73**:1234–40.

25 Smith R. Where is the wisdom . . . ? *BMJ* 1991; **303**: 798–9.

26 Sackett DL, Richardson WS, Rosenberg WMC, Haynes RB. *evidence based medicine: how to practice and teach EBM*, 2nd edn. London: Churchill Livingstone, 2000.

27 Kassirer JP. Incorporating patients' preferences into medical decisions. *New Engl J Med* 1994; **330**: 1895–6.

28 Dowie J. "Evidence-based", "cost-effective", and "preference-driven" medicine. *J Health Serv Res Policy* 1996; **1**: 104–13.

29 Greenhalgh T. Narrative based medicine in an evidence based world. *BMJ* 1999; **318**: 323–5.

30 Grimley Evans J. evidence based and evidence-biased medicine. *Age Ageing* 1995; **24**: 461–3.

31 Feinstein AR, Horwitz RI. Problems in the "evidence" of "evidence based medicine". *Am J Med* 1997; **103**: 529–35.

32 Greenhalgh T, Young G. Applying the evidence with patients. In: Haines A, Silagy C, eds. *evidence based health care – a guide for general practice*. London: BMJ Publications, 1998.

33 Domenighetti G, Grilli R, Liberati A. Promoting consumers' demand for evidence based medicine. *Int J Technol Assess Health Care* 1998; **14**: 97-105.

34 Fulford KWM, Ersser S, Hope T. *Essential practice in patient-centred care*. Oxford: Blackwell Science, 1996.

35 Entwistle VA, Sheldon TA, Sowden A, Watt IS. Evidence-informed patient choice. Practical issues of involving patients in decisions about health care technologies. *Int J Technol Assess Health Care* 1998; **14**: 212–25.

Chapter 2: Searching the literature

2.1 Reading medical articles

Navigating one's way through the jungle that calls itself the medical literature is no easy task and I make no apology that this chapter is the longest in the book. You can apply all the rules for reading a paper correctly but if you're reading the wrong paper you might as well be doing something else entirely. There are already over 15 million medical articles on our library shelves. Every month, around 5000 medical journals are published worldwide and the number of different journals which now exist solely to summarise the articles in the remainder probably exceeds 250. Only 10–15% of the material which appears in print today will subsequently prove to be of lasting scientific value. A number of research studies have shown that most clinicians are unaware of the extent of the clinical literature and of how to go about accessing it.[1, 2]

Dr David Jewell, writing in the excellent book *Critical reading for primary care*,[3] reminds us that there are three levels of reading.

1. *Browsing*, in which we flick through books and journals looking for anything that might interest us.

2. *Reading for information*, in which we approach the literature looking for answers to a specific question, usually related to a problem we have met in real life.

3. *Reading for research*, in which we seek to gain a comprehensive view of the existing state of knowledge, ignorance, and uncertainty in a defined area.

In practice, most of us get most of our information (and, let's face it, a good deal of pleasure) from browsing. To overapply the rules for critical appraisal which follow in the rest of this book would be to kill the enjoyment of casual reading. Jewell warns us, however, to steer a path between the bland gullibility of believing everything and the strenuous intellectualism of formal critical appraisal.

2.2 The Medline database

If you are browsing (reading for the fun of it), you can read what you like, in whatever order you wish. If reading for information (focused searching) or research (systematic review), you will waste time and miss many valuable articles if you simply search at random. Many (but not all – see section 2.10) medical articles are indexed in the huge Medline database, access to which is almost universal in medical and science libraries in developed countries. Note that if you are looking for a systematic quality checked summary of all the evidence on a particular topic you should probably start with the Cochrane database (see section 2.11) rather than Medline, which uses very similar search principles. However, if you are relatively unfamiliar with both, Medline is probably easier to learn on.

Medline is compiled by the National Library of Medicine of the USA and indexes over 4000 journals published in over 70 countries. Three versions of the information in Medline are available.

- Printed (the *Index Medicus*, a manual index updated every year, from which the electronic version is compiled).
- On-line (the whole database from 1966 to date on a mainframe computer, accessed over the Internet or other electronic server).
- CD-ROM (the whole database on between 10 and 18 CDs, depending on who makes it).

The Medline database is exactly the same, whichever company is selling it, but the commands you need to type in to access it differ according to the CD-ROM software. Commercial vendors of Medline on-line and/or on CD-ROM include Ovid Technologies (OVID), Silver Platter Information Ltd (WinSPIRS), Aries Systems Inc (Knowledge Finder), and PubMed.

The best way to learn to use Medline is to book a session with a trained librarian, informaticist or other experienced user. Unless you are a technophobe, you can pick up the basics in less than an hour. Remember that articles can be traced in two ways.

1. By any word listed on the database including words in the title, abstract, authors' names, and the institution where the research was done (note: the abstract is a short summary of what the article is all about, which you will find on the database as well as at the beginning of the printed article).
2. By a restricted thesaurus of medical titles, known as medical subject heading (MeSH) terms.

To illustrate how Medline works, I have worked through some common problems in searching. The following scenarios have been drawn up using OVID software[4] (because that's what I personally use most often and because it is the version used by the dial up service of the BMA library, to which all BMA members with a modem have free access). I have included notes on WinSPIRS[5] (which many universities use as a preferred system) and PubMed (which is available free on the Internet, comes with ready made search filters which you can insert at the touch of a button, and throws in a search of PreMedline, the database of about to be published and just recently published articles[6]). All these systems (Ovid, WinSPIRS and PubMed) are designed to be used with Boolean logic, i.e. putting in particular words (such as "hypertension", "therapy" and so on) linked by operators (such as "and", "or" and "not", as illustrated on pp 19 and 20). Knowledge Finder[7] is a different Medline software which is marketed as a "fuzzy logic" system; in other words, it is designed to cope with complete questions such as "What is the best therapy for hypertension?" and is said to be more suited to the naïve user (i.e. someone with little or no training). I have certainly found Knowledge Finder's fuzzy logic approach quick and effective and would recommend it as an investment for your organisation if you expect a lot of untrained people to be doing their own searching. The practical exercises included in this chapter are all equally possible with all types of Medline software.

2.3 Problem 1: You are trying to find a particular paper which you know exists

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

This shouldn't take long. You do not need to do a comprehensive subject search. Get into the part of the database which covers the approximate year of the paper's publication (usually the past five years). Selecting this is one of the first things the system asks you to do on the main Medline search screen; if you're already in the main Medline menu, select "database" (Alt-B).

If you know the title of the paper (or the approximate title) 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. Box 2.1 shows some useful OVID field suffixes, most of which are self explanatory. But note the **.ui** suffix, which denotes the unique number which you can use to identify a particular Medline entry. If you find an article which you might wish to call up again, it's often quicker to write down the unique identifier rather than the author, title, journal, and so on.

Box 2.1 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	**1887.yr**

To illustrate the use of field suffixes, let's say you are trying to find a paper called something like "A survey of cervical cancer screening in people with learning disability", which you remember seeing in the *BMJ* a couple of years ago. Make sure you have NOT ticked the box "Map term to subject heading", and then type the following into the computer.

1 cervical cancer.ti

This gives you approximately 750 possible articles in set 1. Now type:

2 survey.ti

This gives you approximately 4500 possible articles in set 2. Now type:

3 learning disability.ti

This gives you approximately 100 possible articles in set 3. Now type:

4 BMJ.jn

This gives you several thousand articles in set 4, i.e. all articles listed in this part of the Medline database for the years you selected from the *BMJ*. Now combine these sets by typing:

5 1 and 2 and 3 and 4

This gives you anything with "cervical cancer" and "survey" and "learning disability" in the title *and* which was published in the *BMJ*: a single article in five steps.[8] Note you can also combine sets in OVID by using the "combine" button at the top of the screen.

You could have done all this in one step using the following command (try it now):

6 (cervical cancer AND survey AND learning disability).ti and BMJ.jn

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

Note that you should not generally use abbreviations for journal titles in OVID, but other software packages may use standard abbreviations. Two important exceptions to this rule in OVID are the *Journal of the American Medical Association (JAMA)* and the *British Medical Journal*, which changed its official title in 1988 to *BMJ*. To search for *BMJ* articles from 1988 to date, you must use *BMJ*; for articles up to and including 1987 you should search under both *British Medical Journal* and *British Medical Journal clinical research ed*. Another important point is that searching for title words will only uncover the *exact* word; for example, this search would have missed an article whose title was about learning *disabilities* rather than disability. To address that problem you need to use a truncation symbol (see p 20).

Often, you don't know the title of a paper but you know who wrote it. Alternatively, you may have been impressed with an article you have read (or lecture you heard) by a particular author and you want to see what else they have published. Clear your previous searches by selecting “edit” from the menu bar at the top of the main search screen, then choosing “delete all”.

Let’s try finding Professor Sir Michael Marmot’s publications over the past five years. The syntax is as follows. Type:

1 marmot-m.au

This gives you all articles on this part of the database in which M Marmot is an author or co-author – approximately 35 papers. But like many authors, Michael is not the only M Marmot in the medical literature and – another problem – he has a middle initial which he uses inconsistently in his publications. Unless you already know his middle initial, you must use a *truncation symbol* to find it out. Type:

2 marmot-m$.au

This gives you about 60 articles, which include the previous 35 you found under M Marmot, plus articles by MA Marmot, MD Marmot and another 25 articles by – we’ve found him – MG Marmot! Note that in OVID, the dollar sign is a truncation symbol meaning “any character or characters”. With Silver Platter search software the equivalent symbol is an asterisk (*). You can use the truncation symbol to search a stem in a textword search; for example, the syntax electric$.tw (in OVID) will uncover articles with “electric”, “electricity”, “electrical”, and so on in the title or abstract.

You could have used the following single line command:

3 (marmot-m or marmot-mg).au

This gives a total of around 60 articles, which you now need to browse by hand to exclude any M Marmots other than Professor Sir Michael!

You may also find it helpful to search by institution field. This will give you all the papers which were produced in a particular research institution. For example, type:

4 (withington hospital and manchester).in

to find all the papers where “Withington Hospital, Manchester” appears in the “institution” field (either as the main address where the research was done or as that of one of the co-authors).

If you can’t remember the title of the article you want but you know some exact key phrases from the abstract, it might be quicker to search under textwords than MeSH terms (which are explained in the next section). The field suffixes you need are **.ti** (title), **.ab** (abstract), and **.tw** (textword = either title or abstract). Let's say you were trying to find an editorial from one of the medical journals (you can't remember which) in 1999 about evidence based medicine. Clear your previous searches, then type:

```
1    evidence based medicine.tw and 1999.yr
```

This gives you a total of about 100 articles. You could now browse the abstracts by hand to identify the one you are looking for. Alternatively, you could refine your search by publication type as follows. Type:

```
2    limit 1 to editorial
```

You could, in fact, have done all this in a single step with the following command:

```
3    evidence based medicine.tw and 1999.yr and
     editorial.pt
```

where **.tw** means "textword" (in title or abstract), **.yr** means "year of publication" and **.pt** means "publication type". (You could also have used the "limit set" button at the top of the screen here and then selected the publication type as "editorial".) Note, however, that this method will *only* pick up articles with the exact string "evidence based medicine" as a textword. It will miss, for example, articles which talk about "evidence based health care" instead of evidence based medicine. For this we need to search under MeSH terms, as explained below, and/or cover all possible variations in the textwords (including different ways of spelling each word).

Exercise 1

1. Try to track down the following articles using as few commands as possible.

   a) A systematic review by Craig and colleagues on the measurement of children's temperature in the axilla compared with the rectum published in a major English language journal in about 2000. (Don't forget that the OVID system needs an initial for the author's name.)

b) A paper by Professor Marsh's team from Oxford on the effect of phenobarbital on the frequency of fits. (Note that you do not need the full address of the institution to search under this field.)

c) A paper describing death rates from different causes in participants in the HOPE (Heart Outcomes Prevention Evaluation) study, by Salim Yusuf and colleagues, published in either the *New England Journal of Medicine* or the *Journal of the American Medical Association* (note that Medline indexes the former under its full name and the latter as *JAMA*).

d) Two articles published in 1995 in the *American Journal of Medical Genetics* on the inheritance of schizophrenia in Israeli subjects. See if you can find them in a single command using field suffixes.

2. Trace the series of ongoing articles published in the *Journal of the American Medical Association* from 1992 to date, entitled "Users' guides to the medical literature". Once you've found them, copy them and keep them. Much of the rest of this book is based on these users' guides.

3. How many articles can you find by Professor David Sackett, who, like Professor Marmot, uses his middle initial inconsistently?

4. Find out how many articles were published by Sandra Goldbeck-Wood in the *BMJ* in 1999. Remember that to restrict your search to a particular year in OVID, use the "limit set" button at the top of the screen and then select "publication year", or, alternatively, use the field suffix **.yr** (e.g. 1994.yr).

2.4 Problem 2: You want to answer a very specific clinical 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 put her on the pill to stop her bones thinning. This seemed a reasonable request, though there were ethical problems to consider. But is there any evidence that going on the pill 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". First, clear your screen by erasing the previous searches. Next, make sure that the "Map text to subject heading" is ticked. Now, type:

1 anorexia nervosa

You have not typed a field suffix (such as **.tw**), so even if you forget to tick the box "Map text to subject heading", the OVID system will automatically do this, i.e. try to match your request to one of its standard medical subject headings (abbreviated MeSH and colloquially known as "mesh terms"). Wait a few seconds, and you should see two options on the screen. The first is "anorexia nervosa" as a MeSH term, and you are offered two additional choices: "Explode" and "Focus". Ignore the "explode" box for now (it is explained on p 24), and consider the "focus" box. 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 as follows. Type:

2 *anorexia nervosa/

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

The other option you were offered was a textword search for the term "anorexia nervosa" in the title or abstract. In other words, whenever you ask OVID to map a term, it will also offer to do a title and abstract search, i.e. find you articles with (say) the words "anorexia nervosa" in the title or abstract even if the article has not been indexed under this MeSH heading. You should tick this box too.

The syntax you see on the screen once the term has been mapped is:

1 anorexia nervosa.mp [title, abstract, registry number word, or MeSH]

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. In this case, 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").

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

2 osteoporosis/

You should get about 3000 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. This can save considerable time. 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. The MeSH term is "contraceptives, oral" (if you had known this you could have used the syntax **contraceptives,oral/** but don't do this, for a reason I'm about to explain).

3 oral contraceptives

OVID maps your set to "contraceptives,oral" and asks you if you want to restrict your set to focus (probably not, so don't tick this box) and if you want to explode the term. 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 using the most specific MeSH terms they can. If you just ask for articles on "asthma" you will miss all the terminal divisions of the branch unless you "explode" the term. (Note, however, that you can only explode a term *down* the MeSH tree, not upwards.)

If you do not tick the "explode" box for "contraceptives,oral", your set will probably only contain around 700 articles, whereas the exploded term contains about 5000! A quick route to explode a topic when you know the MeSH term is:

```
3          exp contraceptives, oral/
```

If you combine these three sets, either by using their set numbers **1 and 2 and 3** or by typing the single line command:

```
4          *anorexia nervosa/ and osteoporosis/ and exp
           contraceptives, oral/
```

you will have searched over 6000 articles and obtained a set of only three references – a letter,[9] an original research study,[10] and a substantive review article.[11] (If you don't find these, check the syntax of your search carefully, then try running the same search through the previous five-year database using the "database" button at the top of the screen.)

Exercise 2

Try to find a set of less than five articles relating to any of the following questions or clinical problems.

1. Is the high incidence of coronary heart disease in certain ethnic Asian groups attributable to differences in insulin levels?
2. The hypothesis linking vitamin C with cure of the common cold is, apparently, something to do with its role as an antioxidant. Is there any (clinical or theoretical) evidence to support this hypothesis?
3. How should thyrotoxicosis be managed in pregnancy?

Make sure you practise finding the MeSH term for each subject, using the asterisk to restrict to focus, and using the slash to denote what you know is a MeSH term. (If the current database disappoints you, re-run your search on previous databases by selecting the "database" button.)

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

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

This is one of the commonest reasons why we approach Medline in real life. We don't have a particular paper in mind or a very specific question to ask and we aren't aiming for an exhaustive overview of the literature. We just want to know, say, what's the latest expert advice on drug treatment for asthma or whether anything new has been written on malaria vaccines.

One method of accomplishing this is to search using MeSH terms and then, if we unearth a large number of articles *but not otherwise*, to use index subheadings. Subheadings are the fine tuning of the Medline indexing system and classify articles on a particular MeSH topic into aetiology, prevention, therapy, and so on. The most useful ones are listed in Box 2.2 (you don't have to memorise these since the OVID mapping process automatically offers you subheadings to tick, but you can truncate the mapping process and therefore save time if you do happen to know the subheading you need). I try not to use subheadings myself, since my librarian colleagues tell me that an estimated 50% of articles in Medline are inadequately or incorrectly classified by subheading.

Box 2.2 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**

Note that the subheading **/th** in Box 2.2 refers to the non-pharmacological therapy of a disease, whereas **/dt** is used for drug therapy. The subheading /**tu** is used exclusively for drugs and means "therapeutic use of". The subheading **/px** is used with non-psychiatric diseases as in this example – **diabetes/px** = psychology of diabetes.

Not all subheadings are used in the indexing system for every topic.

To find the subheadings for a MeSH term such as asthma, type:

1 **sh asthma**

This command will tell you which subheadings are used in the indexing system for this MeSH term. It gives you a number of options, including diagnosis, economics, ethnology, and so on. You should choose **/dt** (drug therapy). You could have typed the single line command:

2 ***asthma/dt**

where * denotes a major focus of the article, / denotes a MeSH term, and **dt** means drug therapy. This will give you around 2000 articles to choose from. You now need to *limit the set*, so start with the frequently used options for limiting a set which are listed as tick boxes below the table on your screen ("human", "reviews", and so on). If you actually want to copy a full article today, select "local holdings". This will restrict your set to journals that are held in the particular library through which you are accessing Medline. If you dial Medline at the BMA library via a computer modem, "local holdings" means journals held at the BMA library, not the library where you are dialling from! Note that options such as "local holdings" reduce your article count in a non-systematic way; there are probably many excellent and relevant articles published in journals that your local library does not take.

If after choosing any relevant options from the frequently used "limit set" boxes, you still have too many articles, now select the "limit set" button at the top of the screen. You must now choose additional options for cutting the set down to a number that you can browse comfortably. 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 software to give you the best of the bunch. In other words, don't overapply the "limit set" commands you find in Box 2.3.

If you are sure you want a review article, select this option. You can get the latest review by selecting first "review articles" and then "latest update". However, given that the very latest update may not be the best overview written in the past year or so, you may be better selecting "publication year" as the current year and trawling through. Remember that only a *systematic review* will have involved, and will include details of, a thorough search of the relevant literature (see Chapter 8).

Box 2.3 Useful "limit set" options

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

The option "AIM journals" denotes all journals listed in the *Abridged Index Medicus*, i.e. 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.[12]

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

3 limit 2 to local holdings

4 limit 3 to human

Exercise 3

Try to find a single paper (by browsing a larger set) to give you a quick answer to the following questions:

1. Is hormone replacement therapy ever indicated in women who have had breast cancer in the past?
2. The North American medical literature often mentions health maintenance organisations. What are these?
3. Imagine that you are a medical journalist who has been asked to write an article on screening for prostate cancer. You want two fairly short review articles, from the mainstream medical literature, to use as your sources.
4. Does watching violence on television lead to violent behaviour in adolescents?

2.6 Problem 4: Your search gives you lots of irrelevant articles

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

Often, a search uncovers dozens of articles that 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. I searched Medline by MeSH terms but I also wanted to search by textwords to pick up articles that the MeSH indexing system had missed (see section 2.7). Unfortunately, my search revealed hundreds of articles I didn't want, all on surrogate motherhood. (Surrogate endpoints are explained in section 6.3 but the point here is that they are nothing to do with surrogate motherhood!) The syntax to exclude the unwanted articles is as follows:

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

2 home adj help.tw

where adj means "adjacent". Similarly, "community adj care", "Macmillan adj nurse". You can even specify the number of words gap between two linked words, as in

3 community adj2 care.tw

which would find "community mental health care" as well as "community child care" and "community care".

Exercise 4

1. Find articles about occupational asthma caused by sugar.
2. The drug chloroquine is most commonly used for the treatment of falciparum malaria. Find out what other uses it has. (Hint: use the subheading **/tu**, which means "therapeutic use of", and remember that malaria is often referred to by its Latin name *plasmodium falciparum*. You should, of course, limit a large search to review articles if you are reading for quick information rather than secondary research).

2.7 Problem 5: Your search gives you no articles at all or not as many as you expected

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

If your carefully constructed search bears little or no fruit, it is possible that there are no relevant articles in the database. More likely, you have missed them. Many important articles are missed not because we constructed a flawed search strategy but because we relied too heavily on a flawed indexing system. I've already talked about the overuse of subheadings (see section 2.5). MeSH terms may also be wrongly assigned or not assigned at all. For this reason, you should adopt a "belt and braces" approach and search under textwords as well as by MeSH. After all, it's 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).

Clear your screen, then consider this example. If you wanted to answer the question: what is the role of aspirin in the prevention and treatment of myocardial infarction?, you could type the single line command:

```
1        (myocardial infarction/pc or myocardial
         infarction/dt) and aspirin/tu
```

which would give you all articles listed in this part of the Medline database which cover the therapeutic use of aspirin and the prevention or treatment of myocardial infarction – 190 or so articles, but no immediate answer to your question. You might be better dropping the subheadings and limiting the set as follows:

```
1        myocardial infarction/ and aspirin/

2        limit 1 to AIM journals

3        limit 2 to review articles
```

a strategy which would give you around 25 review articles, including at least one very useful one which your first search (by subheadings) missed. Now, let's add an extra string to this strategy. Erase the set so far, and work as follows:

```
1     (myocardial infarction and aspirin).mp
2     limit 1 to AIM journals
3     limit 2 to review articles
```

The **.mp** suffix (see p 24) automatically gives you a textword search of the title and abstracts and should give you over 50 articles, most of which look very relevant to your question and some of which were missed when you searched MeSH terms alone.

Another important strategy for preventing incomplete searches is to use the powerful “explode” command. This function is explained on p 24 above and you should use it routinely unless you have good reason not to. Try the following search as an example. We are trying to get hold of a good review article about gonococcal arthritis (a rare type of acute arthritis caused by the gonococcus bacterium). Clear your screen, then type the MeSH term

```
1     *arthritis/
```

This will give you about 1300 articles in which arthritis is the focus. Now search for articles on arthritis in which the word “gonococcal” is mentioned in the title or abstract, by typing:

```
2     gonococcal.tw
3     1 and 2
```

This narrows your search drastically to one or two articles, neither of which offers a comprehensive overview of the subject. And how many have you missed? The answer is quite a few of them, because the MeSH term “arthritis” subdivides into several branches, including “arthritis, infectious”. Try it all again (without erasing the first search) but this time, explode the term “arthritis” before you start and then limit your set to review articles:

```
4     exp arthritis/
5     2 and 4
6     limit 5 to review articles
```

You now have around five articles, including a major overview[13] which your unexploded search missed.You can demonstrate this by typing:

```
7        6 not 3
```

which will show you what the exploded search revealed over and above the unexploded one. Incidentally, if you were also thinking of searching under textwords, the syntax for identifying articles about the problem in men would be **(male not female).tw** and **(men not women).tw**, since the female terms here literally incorporate the male!

2.8 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 a lot 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"). Type:

```
1        ptx stress
```

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

ptx is a useful command when the term you are exploring might be found in several subject areas. If your subject word *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, posttraumatic epilepsy, and so on.

Exercise 5

1. Find where the word "nursing" might appear as part of a MeSH term.
2. Use the tree command to expand the MeSH term "diabetes mellitus".

2.9 Problem 7: Your attempt to limit a set leads to loss of important articles but does not exclude those of low methodological quality

Solution: Apply an evidence based quality filter

What do you do when your closely focused search still gives you several hundred articles to choose from and if applying subheadings or limit set functions seems to lose valuable (and relevant) papers? First, you should consider the possibility that your search wasn't as focused as you thought. But if you can't improve on it, you should try inserting 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 metaanalysis.

These evidence based quality filters, which are listed in Appendices 2 and 3, are highly complex search strategies developed and refined by some of the world's most experienced medical information experts. Carol Lefebvre of the UK Cochrane Centre first introduced me to them and credits their origin to Anne McKibbon who has written extensively on the subject of searching in relation to evidence based practice.[14] You can copy them into your personal computer and save them as cut and paste strategies to be added to your subject searches. Other maximally sensitive search strategies are being developed which will identify cohort studies, case-control studies, and so on, and will soon be available from the UK Cochrane Centre, Summertown Pavilion, Middle Way, Oxford OX2 7LG, email general@cochrane.co.uk.

Exercise 6

1. Search for a good randomised controlled trial of the use of aspirin in the therapy of acute myocardial infarction.
2. Find a systematic review on the risk of gastrointestinal bleeding with non-steroidal antiinflammatory drugs.

2.10 Problem 8: Medline hasn't helped, despite a thorough search

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 librarians who group articles under subheadings and type in the abstracts. 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* References the literature on AIDS and HIV back to 1980. Available via a number of suppliers including Internet Grateful Med (see below).

- *Allied and Complementary Medicine (AMED)* Covers a range of complementary and alternative medicine including homeopathy, chiropractic, acupuncture, and so on. Produced by the British Library, available from a number of suppliers including Silver Platter or OVID. For more details on AMED see http://www.silverplatter.com.catalog/amed.htm

- *Bandolier* Award-winning summary journal with searchable index produced by Andrew Moore and colleagues in Oxford, UK. Includes a range of commissioned review topics for the UK NHS Research and Development Directorate. http://www. jr2.ox.ac.uk:80/Bandolier/

- *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. See text for further details. Abstracts are available free on http://hiru.mcmaster.ca/cochrane/cochrane/revabstr/abidx.htm
- *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. Available from OVID; more details from http://ovid.gwdg. de/ovidweb/fldguide/cc.htm#geni
- *English National Board health care database* A database of journal references primarily of interest to nurses, midwives, and health visitors. http://www.enb.org.uk/hcd.htm
- *Embase* The database of *Excerpta Medica*, which focuses on drugs and pharmacology, but also includes other biomedical specialties. It is more up to date than Medline and with better European coverage. The CD-ROM version is updated monthly. Available via a number of software suppliers including OVID (see reference list).
- *Internet Grateful Med* Public access "front end" to medical databases, offering Medline, HealthSTAR, PreMedline, AIDSLINE, AIDSDRUGS, AIDSTRIALS, and several others. http://igm.nlm.nih.gov/
- *NHS economic evaluation database* Public access database with quality assessed structured abstracts of clinical trials that include an economic evaluation. http://nhscrd.york.ac.uk
- *NHS health technology assessment database* Public access database with quality assessed structured abstracts of clinical trials that include an evaluation of health technology. http://nhscrd.york.ac.uk
- *National Guideline Clearinghouse (US)* A comprehensive database of evidence based clinical practice guidelines and

related documents produced by the US Agency for Health Care Policy and Research (AHCPR), in partnership with the American Medical Association (AMA) and the American Association of Health Plans (AAHP).
http://www.guidelines.gov/index.asp

- *National Research Register (UK)* List of ongoing UK clinical trials by the Medical Research Council and National Research Register.
http://www.update-software.com/National/nrr-frame.html
- *Psyclit* Produced by the American Psychological Association as the computer-searchable version of Psychological Abstracts. It covers psychology, psychiatry and related subjects; journals are included from 1974 and books from 1987 (English language only). Available through several software companies (see reference list) along with Medline.
- *Science Citation Index* Indexes the references cited in articles as well as the usual author, title, abstract and citation of the articles themselves. Useful for finding follow up work done on a key article and for tracking down the addresses of authors. Available (at a charge) from Web of Science at http://wos.mimas.ac.uk/
- *SUMSearch* A new method of searching the Internet for evidence based medical information. Querying a number of key databases, such as Medline, Cochrane abstracts and DARE (see above), SUMSearch aims to select the most appropriate source, format the search query, modify this query if too few or too many hits are found, and return a single document to the clinician. For many queries this is a good first port of call.
http://SUMSearch.uthscsa.edu
- *UNICORN* The main database of the King's Fund, London. Covers a range of journals on health, health management, health economics, and social sciences. Particularly strong on primary health care and the health of Londoners. Accessible at the King's Fund Library, 11–13 Cavendish Square, London W1M 0AN.

2.11 The Cochrane Library

When I wrote the first edition of this book, the Cochrane library was a fairly small and exploratory project but I predicted that by 2000 it would probably have replaced Medline as the medical researcher's first port of call when looking for quality articles and summaries of clinical research. This is indeed the case, and the Cochrane Library now boasts several hundred systematic reviews and hundreds of thousands of peer reviewed summaries of randomised controlled trials. The story behind the Cochrane project is worth telling.

In 1972, epidemiologist Archie Cochrane called for the establishment of a central international register of clinical trials. (It was Cochrane who, as a rebellious young medical student, marched through the streets of London in 1938 bearing a placard which stated, "All effective treatments should be free". His book *Effectiveness and efficiency*[15] caused little reaction at the time but captures the essence of today's evidence based medicine movement.)

Though he never lived to see the eponym, Archie Cochrane's vision of a 100% accurate medical database, the Cochrane Controlled Trials Register, is approaching reality. The Cochrane Library also includes two "metadatabases" (the Cochrane Database of Systematic Reviews and the Database of Abstracts of Reviews of Effectiveness) and a fourth database on the science of research synthesis (the Cochrane Review Methodology Database). This entire library is available on CD-ROM from the BMA bookshop.

Published articles are entered on to the Cochrane databases by members of the Cochrane Collaboration,[16] an international network of (mostly) medically qualified volunteers who each take on the handsearching of a particular clinical journal back to the very first issue. Using strict methodological criteria, the handsearchers classify each article according to publication type (randomised trial, other controlled clinical trial, epidemiological survey, and so on), and prepare structured abstracts in house style. The Collaboration has already identified around 60 000 trials that had not been appropriately tagged in Medline.

All the Cochrane databases are in user friendly Windows style format with a search facility very similar to that used in the common Medline packages. Numerical data in overviews are presented in a standardised graphics way to allow busy clinicians to assess their relevance quickly and objectively. In 1997 some of the founder members of the Cochrane Collaboration published a compilation of articles reflecting on Cochrane's original vision and the projects that have emerged from it. Despite its uninspiring title, *Non-random reflections*... is a fascinating account of one of medicine's most important collaborative initiatives in the 20th century.[17]

Finally, if you are interested in becoming involved with the Cochrane Library projects, contact the Cochrane Library Users Group on http://www.york.ac.uk/inst/crd/clug.htm.

1 Young JM, Ward JE. General practitioners' use of evidence databases. *Med J Australia* 1999; **170**: 56–8.

2 McColl A, Smith H, White P, Field J. General practitioners' perceptions of the route to evidence based medicine: a questionnaire study. *BMJ* 1998; **316**: 361–5.

3 Jones R, Kinmonth A-L. *Critical reading for primary care*. Oxford: Oxford University Press, 1995.

4 For further details on the OVID system of Medline see the company's website http://www.ovid.com.

5 For further details on the WinSPIRS system of Medline see the company's website http://silverplatter.com.

6 The PubMed version of Medline and PreMedline can be accessed free on the Internet on http://www.ncbi.nlm.nih.gov/PubMed/.

7 For further details on the Knowledge Finder system of Medline see the company's website http://www.kfinder.com/newweb/.

8 Stein K, Allen N. Cross sectional survey of cervical cancer screening in women with learning disability. *BMJ* 1999; **318**: 641.

9 Mehler PS. Eating disorders [letter]. *New Engl J Med* 1999; **341**: 614-15.

10 Grinspoon S, Miller K, Coyle C *et al.* Severity of osteopenia in estrogen-deficient women with anorexia nervosa and hypothalamic amenorrhea. *J Clin Endocrinol Metab* 1999; **84**: 2049–55.

11 Grinspoon S, Herzog D, Klibanski A. Mechanisms and treatment options for bone loss in anorexia nervosa. *Psychopharmacol Bull* 1997; **33**: 399–404.

12 Moher D, Fortin P, Jadad AR *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.

13 Angulo JM, Espinoza LR. Gonococcal arthritis. *Comprehensive Therapy* 1999; **25**: 155–62.

14 McKibbon KA. evidence based practice. *Bull Med Library Assoc* 1998; **86**:396–401.

15 Cochrane A. *Effectiveness and efficiency*. London: Nuffield Provincial Hospitals Trust, 1972.

16 Bero L, Rennie D. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995; **274**: 1935–8.

17 Maynard A, Chalmers I, eds. *Non-random reflections on health services research*. London: BMJ Books, 1997.

Chapter 3: Getting your bearings (what is this paper about?)

3.1 The science of "trashing" papers

It usually comes as a surprise to students to learn that some (the purists would say up to 99% of) published articles belong in the bin and should certainly not be used to inform practice. In 1979, the editor of the *British Medical Journal*, Dr Stephen Lock, wrote "Few things are more dispiriting to a medical editor than having to reject a paper based on a good idea but with irremediable flaws in the methods used". Things have improved since then, but not enormously[1] (see Box 3.1).

Most papers appearing in medical journals these days are presented more or less in standard IMRAD format: Introduction (*why* the authors decided to do this particular piece of research), Methods (*how* they did it and how they chose to analyse their results), Results (*what* they found), and Discussion (what they think 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 value of the hypothesis, the nature or potential impact of the results or the speculation in the discussion.

Conversely, bad science is bad science regardless of whether the study addressed an important clinical issue, whether the results are "statistically significant" (see section 5.5), whether things changed in the direction you would have liked them to, and whether, if true, the findings promise immeasurable benefits for patients or savings for the health service. Strictly speaking, *if you are going to trash a paper, you should do so before you even look at the results.*

Box 3.1 Common reasons why papers are rejected for publication

- The study did not examine an important scientific issue (see section 3.2)
- The study was not original – that is, someone else has already done the same or a similar study (see section 4.1)
- The study did not actually test the authors' hypothesis (see section 3.2)
- A different study design should have been used (see section 3.3)
- Practical difficulties (for example, in recruiting subjects) led the authors to compromise on the original study protocol (see section 4.3)
- The sample size was too small (see section 4.6)
- The study was uncontrolled or inadequately controlled (see section 4.4)
- The statistical analysis was incorrect or inappropriate (see chapter 5)
- The authors have drawn unjustified conclusions from their data
- There is a considerable conflict of interest (for example, 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

It is much easier to pick holes in other people's work than to do a methodologically perfect piece of research oneself. When I teach critical appraisal, there is usually someone in the group who finds it profoundly discourteous to criticise research projects into which dedicated scientists have put the best years of their lives. On a more pragmatic note, there may be good practical reasons why the authors of the study have "cut corners" and they know as well as you do that their work would have been more scientifically valid if they hadn't.

Most good scientific journals send papers out to a referee for comments on their scientific validity, originality, and importance before deciding whether to print them. This process is known as *peer review* and much has been written about it.[2] Common defects picked up by referees are listed in Box 3.1.

I recently corresponded with an author whose paper I had refereed (anonymously, though I subsequently declared myself) and recommended that it should not be published. On reading my report, he wrote to the editor and admitted he agreed with my opinion. He described five years of painstaking and unpaid research done mostly in his spare time and the gradual realisation that he had been testing an important hypothesis with the wrong method. He informed the editor that he was "withdrawing the paper with a wry smile and a heavy heart" and pointed out several further weaknesses of his study which I and the other referee had missed. He bears us no grudge and, like Kipling's hero, has now stooped to start anew with worn-out tools. His paper remains unpublished but he is a true (and rare) scientist.

The assessment of methodological quality (critical appraisal) has been covered in detail in many textbooks on evidence based medicine,[3–7] and in Sackett and colleagues' "Users' guides to the medical literature" in the *JAMA*.[8–32] The structured guides produced by these authors on how to read papers on therapy, diagnosis, screening, prognosis, causation, quality of care, economic analysis, and overview are regarded by many as the definitive checklists for critical appraisal. Appendix 1 lists some simpler checklists which I have derived from the users' guides and the other sources cited at the end of this chapter, together with some ideas of my own. If you are an experienced journal reader, these checklists will be largely self explanatory. If, however, you still have difficulty getting started when looking at a medical paper, try asking the preliminary questions in the next section.

3.2 Three preliminary questions to get your bearings

Question 1: Why was the study done and what hypothesis were the authors testing?

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, for example, "Gupta and Brown's prospective survey of grommet insertions demonstrated that...". It is irritatingly common for authors to forget to place their research in context, since the background to the problem is usually clear as daylight to them by the time they reach the writing up stage.

Unless it has already been covered in the introduction, the methods section of the paper should state clearly the hypothesis which the authors have decided to test, such as "This study aimed to determine whether day case hernia surgery was safer and more acceptable to patients than the standard inpatient procedure". Again, this important step may be omitted or, more commonly, buried somewhere mid-paragraph. If the hypothesis is presented in the negative (which it usually is), 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.

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 demonstrate 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.[33]

If you have not discovered what the authors' stated (or unstated) hypothesis was by the time you are halfway through the methods section, you may find it in the first paragraph of the discussion. Remember, however, that not all research studies (even good ones) are set up to test a single definitive hypothesis. *Qualitative* research studies, which are as valid and as necessary as the more conventional quantitative studies, aim to look at particular issues in a broad, open-ended way in order to generate (or modify) hypotheses and prioritise areas to investigate. This type of research is discussed further in Chapter 11. Even quantitative research (which the rest of this book is about) is now seen as more than hypothesis testing. As section 5.5 argues, it is strictly preferable to talk about evaluating the *strength* of evidence around a particular issue than about proving or disproving hypotheses.

Question 2: What type of study was done?

First, decide whether the paper describes a primary or secondary study. Primary studies report research first hand, while secondary (or *integrative*) studies attempt 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.
- *Surveys*, in which something is measured in a group of patients, health professionals, or some other sample of individuals.

The more common types of clinical trials and surveys are discussed in the later sections of this chapter. Make sure you understand any jargon used in describing the study design (see Box 3.2).

Secondary research is composed of:

- *overviews*, considered in Chapter 8, which may be divided into:
  (a) (*non-systematic*) *reviews*, which summarise primary studies
  (b) *systematic reviews*, which do this according to a rigorous and predefined methodology
  (c) *meta-analyses*, which integrate the numerical data from more than one study
- *guidelines*, considered in Chapter 9, which draw conclusions from primary studies about how clinicians should be behaving
- *decision analyses*, which are not discussed in detail in this book but are covered elsewhere;[16, 17, 34–36] these use the results of primary studies to generate probability trees to be used by both health professionals and patients in making choices about clinical management or resource allocation
- *economic analyses*, considered in Chapter 10, 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 broad field of research addressed?

Examples of the sorts of questions that can reasonably be answered by different types of primary research study are given in

Box 3.2 Terms used to describe design features of clinical research studies

Term	*Meaning*
Parallel group comparison	Each group receives a different treatment, with both groups being entered at the same time. In this case, 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 within subject changes
Single blind	Subjects did not know which treatment they were receiving
Double blind	Neither investigators nor subjects knew who was receiving which treatment
Crossover	Each subject received both the intervention and control treatments (in random order) often separated by a *washout* period of 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 that permits investigation of the effects (both separately and combined) of more than one independent variable on a given outcome (for example, a 2 x 2 factorial design tested the effects of placebo, aspirin alone, streptokinase alone or aspirin plus streptokinase in acute heart attack[37])

the sections which follow. One question which frequently cries out to be asked is this: was a randomised controlled trial (see section 3.3 below) the best method of testing this particular hypothesis and if the study was not a randomised controlled trial, should it have been? Before you jump to any conclusions, decide what broad field of research the study covers (see Box 3.3). Then ask whether the right type of study was done to address a question in this field. For more help on this task (which some people find difficult until they have got the hang of it) see the Oxford Centre for EBM website[38] or the journal article by the same group.[39]

Box 3.3 Broad topics of research

Most research studies are concerned with one or more of the following.

- *Therapy* – testing the efficacy of drug treatments, surgical procedures, alternative methods of patient education or other interventions. Preferred study design is randomised controlled trial (see section 3.3 and Chapter 6)
- *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 (see section 3.6 and Chapter 7) in which both the new test and the gold standard test are performed
- *Screening* – demonstrating the value of tests that can be applied to large populations and that pick up disease at a presymptomatic stage. Preferred study design is cross-sectional survey (see section 3.6 and Chapter 7)
- *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 (see section 3.4)
- *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 (see sections 3.4 and 3.5), but case reports (see section 3.7) may also provide crucial information

3.3 Randomised controlled trials

In a randomised controlled trial (RCT), participants in the trial are randomly allocated by a process equivalent to the flip of a coin to either one intervention (such as a drug treatment) or another (such as placebo treatment). Both groups are followed up for a specified time period and analysed in terms of specific outcomes defined at the outset of the study (for example, 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. In reality, however, not every RCT is a bowl of cherries.

Some papers which report trials comparing an intervention with a control group are not, in fact, randomised trials at all. The name for these is *other controlled clinical trials*, a term used to describe comparative studies in which subjects were allocated to intervention or control groups in a non-random manner. This situation may arise, for example, when random allocation would be impossible, impractical or unethical. The problems of non-random allocation are discussed further in section 4.4 in relation to determining whether the two groups in a trial can reasonably be compared with one another on a statistical level.

Some trials count as a sort of halfway house between true randomised trials and non-randomised trials. In these, randomisation is not done truly at random (for example, using sequentially numbered sealed envelopes each with a computer generated random number inside) but by some method which allows the clinician to know which group the patient would be in *before he or she makes a definitive decision to randomise the patient*. This allows subtle biases to creep in, since the clinician might be more (or less) likely to enter a particular patient into the trial if he or she believed that the patient would get active treatment. In particular, patients with more severe disease may be subconsciously withheld from the placebo arm of the trial. Examples of unacceptable methods include randomisation by last digit of date of birth (even numbers to group A, etc.), toss of a coin, sequential allocation (patient A to group 1, patient B to group 2, etc.), and date seen in clinic (all patients seen this week to group A, all those seen next week to group 2, etc.).[40]

Box 3.4 Advantages of the randomised controlled trial design

- Allows rigorous evaluation of a single variable (for example, effect of drug treatment versus placebo) in a precisely defined patient group (for example, menopausal women aged 50–60 years)
- Prospective design (that is, data are collected on events that happen *after* you decide to do the study)
- Uses hypotheticodeductive reasoning (that is, seeks to falsify, rather than confirm, its own hypothesis; see section 3.2)
- Potentially eradicates bias by comparing two otherwise identical groups (but see below and section 4.4)
- Allows for metaanalysis (combining the numerical results of several similar trials) at a later date; see section 8.3

Listed below are examples of clinical questions which would be best answered by a randomised controlled trial but note also the examples in the later sections of this chapter of situations where other types of study could or must be used instead.

- Is this drug better than placebo or a different drug for a particular disease?
- Is a new surgical procedure better than currently favoured practice?
- Is a leaflet better than verbal advice in helping patients make informed choices about the treatment options for a particular condition?
- Will changing from a margarine high in saturated fats to one high in polyunsaturated fats significantly affect serum cholesterol levels?

RCTs are said to be the gold standard in medical research. Up to a point, this is true (see section 3.8) but only for certain types of clinical question (see Box 3.3 and sections 3.4 to 3.7). The questions which best lend themselves to the RCT design are all about *interventions*, and are mainly concerned with therapy or prevention. It should be remembered, however, that even when we are looking at therapeutic interventions, and especially when we are not, there are a number of important disadvantages associated with randomised trials (see Box 3.5).[41]

Box 3.5 Disadvantages of the randomised controlled trial design

Expensive and time consuming, hence, in practice:

- many trials are either never done, are performed on too few subjects or are undertaken for too short a period (see section 4.6)
- most trials 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 (see section 6.3)

May introduce "hidden bias", especially through:

- imperfect randomisation (see above)
- failure to randomise all eligible patients (clinician offers participation in the trial only to patients he or she considers will respond well to the intervention)
- failure to blind assessors to randomisation status of patients (see section 4.5)

Remember, too, that the results of an RCT may have limited applicability as a result of exclusion criteria (rules about who may not be entered into the study), inclusion bias (selection of trial subjects from a group which is unrepresentative of everyone with the condition (see section 4.2)), refusal of certain patient groups to consent to be included in the trial,[42] analysis of only predefined "objective" endpoints which may exclude important qualitative aspects of the intervention[47] (see Chapter 11), and publication bias (i.e. the selective publication of positive results).[43] Furthermore, RCTs can be well or badly managed,[44] and, once published, their results are open to distortion by an overenthusiastic scientific community or by a public eager for a new wonder drug.[45] Whilst all these problems might also occur with other trial designs, they may be particularly pertinent when a RCT is being sold to you as, methodologically speaking, whiter than white.

There are, in addition, many situations in which RCTs are either unnecessary, impractical or inappropriate.

RCTs are unnecessary

- When a clearly successful intervention for an otherwise fatal condition is discovered.
- When a previous RCT or metaanalysis has given a definitive result (either positive or negative – see section 5.5). Some people would argue that it is actually *unethical* to ask patients to be randomised to a clinical trial without first conducting a systematic literature review to see whether the trial needs to be done at all.

RCTs are impractical

- Where it would be unethical to seek consent to randomise.[46]
- Where the number of patients needed to demonstrate a significant difference between the groups is prohibitively high (see section 4.6).

RCTs are inappropriate

- Where the study is looking at the prognosis of a disease. For this analysis, the appropriate route to best evidence is a longitudinal survey of a properly assembled *inception cohort* (see section 3.4).
- Where the study is looking at the validity of a diagnostic or screening test. For this analysis, the appropriate route to best evidence is a *cross-sectional survey* of patients clinically suspected of harbouring the relevant disorder (see section 3.6 and Chapter 7).
- Where the study is looking at a "quality of care" issue in which the criteria for "success" have not yet been established. For example, an RCT comparing medical versus surgical methods of abortion might assess "success" in terms of number of patients achieving complete evacuation, amount of bleeding, and pain level. The patients, however, might decide that other aspects of the procedure are important, such as knowing in advance how long the procedure will take, not seeing or feeling the abortus come out, and so on. For this analysis, the appropriate route to best evidence is a *qualitative research method*[47] (see Chapter 11).

All these issues have been discussed in great depth by the clinical epidemiologists,[3, 6] who remind us that to turn our noses up at the non-randomised trial may indicate scientific naïveté and not, as many people routinely assume, intellectual rigour. Note also that there is now a recommended format for reporting RCTs in medical journals, which you should try to follow if you are writing one up yourself.[48] For an in-depth discussion of the pros and cons of the RCT, you might like to take a look at the entire issue of the *BMJ* from 31st October 1998 (*BMJ* 1998; **317**: 1167–261), as well as a recent book[49] and journal articles.[50]

3.4 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 medicine 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 RCTs 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.

A special type of cohort study may also be used to determine the prognosis (i.e. what is likely to happen to someone who has it) of a disease. A group of patients who have all been diagnosed as having an early stage of the disease or a positive screening test (see Chapter 7) 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. (Here is a definition which you should commit to memory if you can: *incidence* is the number of new cases of a disease per year, whereas *prevalence* is the overall proportion of the population who suffer from the disease.)

The world's most famous cohort study, which won its two original authors a knighthood, was undertaken by Sir Austen Bradford Hill, Sir Richard Doll and, latterly, Richard Peto. They followed up 40 000 British doctors divided into four cohorts (non-smokers, light, moderate and heavy smokers) using both all cause (any death) and cause specific (death from a particular disease) mortality as outcome measures. Publication of their 10 year interim results in 1964,[51] which showed a substantial excess in both lung cancer mortality and all cause mortality in smokers, with a "dose–response" relationship (i.e. the more you smoke, the worse your chances of getting lung cancer), went a long way to demonstrating that the link between smoking and ill health was causal rather than coincidental. The 20 year[52] and 40 year[53] 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.

Clinical questions which should be addressed by a cohort study include the following.

- Does the contraceptive pill "cause" breast cancer? (Note, once again, that the word "cause" is a loaded and potentially misleading term. As John Guillebaud has argued in his excellent book the Pill,[54] if 1000 women went on the pill tomorrow, some of them would get breast cancer. But some of those would have got it anyway. The question which epidemiologists try to answer through cohort studies is "What is the additional risk of developing breast cancer which this woman would run by taking the pill, over and above her "baseline" risk attributable to her own hormonal balance, family history, diet, alcohol intake, and so on?".)
- Does smoking cause lung cancer?[53]
- 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?

3.5 Case-control studies

In a case-control study, patients with a particular disease or condition (cases) 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 (i.e. what causes it), rather than its treatment. They lie lower down the hierarchy of evidence (see section 3.8) but this design is usually the only option when 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 (see section 4.4). In addition, such a design cannot demonstrate causality; in other words, the *association* of A with B in a case-control study does not prove that A has *caused* B.

Clinical questions which should be addressed by a case-control study include the following.

- Does the prone sleeping position increase the risk of cot death (sudden infant death syndrome)?
- Does whooping cough vaccine cause brain damage? (see section 4.4)
- Do overhead power cables cause leukaemia?

3.6 Cross-sectional surveys

We have probably all been asked to take part in a survey, even if it was only a lady in the street asking us which brand of toothpaste we prefer. Surveys conducted by epidemiologists are run along essentially 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 timepoint but may refer retrospectively to health experiences in the past, such as, for example, the study of patients' casenotes to see how often their blood pressure has been recorded in the past five years.

Clinical questions which should be addressed by a cross-sectional survey include the following.

- What is the "normal" height of a 3 year old child? (This, like other questions about the range of normality, can be answered simply by measuring the height of enough healthy 3 year olds. But such an exercise does not answer the related clinical question "When should an unusually short child be investigated for disease?" since, as in almost all biological measurements, the physiological (normal) overlaps with the pathological (abnormal). This problem is discussed further in section 7.4.)

- What do psychiatric nurses believe about the value of electroconvulsive therapy (ECT) in the treatment of severe depression?
- Is it true that "half of all cases of diabetes are undiagnosed"? (This is an example of the more general question "What is the prevalence (proportion of people with the condition) of this disease in this community?" The only way of finding the answer is to do the definitive diagnostic test on a representative sample of the population.)

3.7 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 2000 . . ."). 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 relatively weak scientific evidence (see section 3.8), a great deal of information can be conveyed in a case report that would be lost in a clinical trial or survey (see Chapter 11). In addition, case reports are immediately understandable by non-academic clinicians and by the lay public. They can, if necessary, be written up and published within days, which gives them a definite edge over meta-analyses (whose gestation period can run into years) or clinical trials (several months). There is certainly a vocal pressure group within the medical profession calling for the reinstatement of the humble case report as a useful and valid contribution to medical science.[55]

Clinical situations in which a case report or case series is an appropriate type of study include the following.

- A doctor notices that two babies born 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.[56] (Anyone who thinks "quick and dirty" case reports are never scientifically justified should remember this example.)

- A patient who has taken two different drugs, terfenadine (for hay fever) and itraconazole (for fungal infection), with no side effects in the past takes them concurrently (i.e. both at the same time) and develops a life-threatening heart rhythm disturbance. The doctors treating him suspect that the two drugs are interacting.[57]

3.8 The traditional 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 order.[20]

1. Systematic reviews and meta-analyses (see Chapter 8).
2. Randomised controlled trials with definitive results (i.e. confidence intervals which do not overlap the threshold clinically significant effect; see section 5.5).
3. Randomised controlled trials with non-definitive results (i.e. a point estimate which suggests a clinically significant effect but with confidence intervals overlapping the threshold for this effect; see section 5.5).
4. Cohort studies.
5. Case-control studies.
6. Cross-sectional surveys.
7. Case reports.

The pinnacle of the hierarchy is, quite properly, reserved for secondary research papers, in which all the primary studies on a particular subject have been hunted out and critically appraised according to rigorous criteria (see Chapter 8). Note, however, that not even the most hard line protagonist of evidence based medicine would place a sloppy metaanalysis or a randomised controlled trial that was seriously methodologically flawed above a large, well designed cohort study. And as Chapter 11 shows, many important and valid studies in the field of qualitative research do not feature in this particular hierarchy of evidence at all. In other words, evaluating the potential contribution of a particular study to medical science requires considerably more effort than is needed to check off its basic design against the six point scale above.

3.9 A note on ethical considerations

When I was a junior doctor, I got a job in a world renowned teaching hospital. One of my humble tasks was seeing the geriatric (elderly) patients in casualty. I was soon invited out to lunch by two charming registrars, who (I later realised) were seeking my help with their research. In return for getting my name on the paper, I was to take a rectal biopsy (i.e. cut out a small piece of tissue from the rectum) on any patient over the age of 90 who had constipation. I asked for a copy of the consent form which patients would be asked to sign. When they assured me that the average 90 year old would hardly notice the procedure, I smelt a rat and refused to cooperate with their project.

I was naïvely unaware of the seriousness of the offence being planned by these doctors. Doing *any* research, particularly that which involves invasive procedures, on vulnerable and sick patients without full consideration of ethical issues is both a criminal offence and potential grounds for a doctor to be "struck off" the medical register. Getting ethical approval for one's research study can be an enormous bureaucratic hurdle,[58] but it is nevertheless a legal requirement (and one which was, until recently, frequently ignored in research into the elderly and those with learning difficulties[59]). Most editors routinely refuse to publish research which has not been approved by the relevant research ethics committee but if you are in doubt about a paper's status, there is nothing to stop you writing to ask the authors for copies of relevant documents.

Note, however, that this hand can be overplayed.[58] Research ethics committees frequently deem research proposals unethical, yet it could be argued that in areas of genuine clinical uncertainty the only ethical option is to allow the informed patient the opportunity to help reduce that uncertainty. The randomised trial which showed that neural tube defects could be prevented by giving folic acid supplements to the mother in early pregnancy[60] is said to have been held back for years because of ethics committee resistance.

1 Altman DG. The scandal of poor medical research. *BMJ* 1994; **308**: 283–4.

2 Jefferson T, Godlee F. *Peer Review in Health Sciences*. London: BMJ Books, 1999.

3 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991.

4 Sackett DL, Richardson WS, Rosenberg WMC, Haynes RB. *evidence based*

medicine: how to practice and teach EBM, 2nd edn. London: Churchill Livingstone, 2000.

5 Crombie IM. *The pocket guide to critical appraisal*. London: BMJ Publications, 1996.

6 Fletcher RH, Fletcher SW, Wagner EH. *Clinical epidemiology: the essentials*, 3rd edn. Baltimore: Williams and Wilkins, 1996.

7 Rose G, Barker DJP. *Epidemiology for the uninitiated*, 4th edn. London: BMJ Publications, 1999.

8 Oxman AD, Sackett DS, Guyatt GH. Users' guides to the medical literature. I. How to get started. *JAMA* 1993; **270**: 2093–5.

9 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.

10 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.

11 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.

12 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.

13 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**:1615–19.

14 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.

15 Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. *JAMA* 1994; **272**: 1367–71.

16 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.

17 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**: 1610–13.

18 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.

19 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.

20 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.

21 Naylor CD, Guyatt GH. Users' guides to the medical literature. X. How to use an article reporting variations in the outcomes of health services. *JAMA* 1996; **275**: 554–8.

22 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.

23 Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R. Cook DJ. Users' guides to the medical literature. XII. How to use articles about health-related quality of life. *JAMA* 1997; **277**: 1232–7.

24 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.

25 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.

26 Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature. XIV. How to decide on the applicability of clinical trial results to your patient. *JAMA* 1998; **279**: 545–9.

27 Richardson WS, Wilson MC, Guyatt GH, Cook DJ, Nishikawa J. Users' guides to the medical literature. XV. How to use an article about disease probability for differential diagnosis. *JAMA* 1999; **281**:1214–19.

28 Guyatt GH, Sinclair J, Cook DJ, Glasziou P. Users' guides to the medical literature. XVI. How to use a treatment recommendation. *JAMA* 1999; **281**: 1836–43.

29 Barratt A, Irwig L, Glasziou P *et al.* Users' guides to the medical literature: XVII. How to use guidelines and recommendations about screening. *JAMA* 1999; **281**: 2029–34.

30 Randolph AG, Haynes RB, Wyatt JC, Cook DJ, Guyatt GH. Users' guides to the medical literature. XVIII. How to use an article evaluating the clinical impact of a computer-based clinical decision support system. *JAMA* 1999; **282**: 67–74.

31 Giacomini MK, Cook DJ. A user's guide to qualitative research in health care. Part I: Are the results of the study valid? *JAMA* 2000; 357–62.

32 Giacomini MK, Cook DJ. A user's guide to qualitative research in health care. Part II: What are the results and how do they help me care for my patients? *JAMA* 2000; 478–82.

33 Popper K. *Conjectures and refutations: the growth of scientific knowledge*. New York: Routledge and Kegan Paul, 1963.

34 Thornton JG, Lilford RJ, Johnson N. Decision analysis in medicine. *BMJ* 1992; **304**: 1099–103.

35 Thornton JG, Lilford RJ. Decision analysis for medical managers. *BMJ* 1995; **310**: 791–4.

36 Dowie J. "Evidence-based", "cost-effective", and "preference-driven" medicine. *J Health Serv Res Policy* 1996; **1**: 104–13.

37 ISIS-2 Collaborative Group. Randomised trial of intravenous streptokinase, aspirin, both, or neither among 17187 cases of suspected acute myocardial infarction: ISIS-2. *Lancet* 1988; **ii**: 349–60.

38 http://cebm.jr2.ox.ac.uk/docs/studies.html.

39 Sackett DL, Wennberg JE. Choosing the best research design for each question. *BMJ* 1997; **315**: 1636.

40 Stewart LA, Parmar MKB. Bias in the analysis and reporting of randomized controlled trials. *Int J Technol Assess Health Care* 1996; **12**: 264–75.

41 Bero LA, Rennie D. Influences on the quality of published drug studies. *Int J Technol Assess Health Care* 1996; **12**: 209–37.

42 MacIntyre IMC. Tribulations for clinical trials. Poor recruitment is hampering research. *BMJ* 1991; **302**: 1099–100.

43 Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ* 1997; **315**: 640–5.

44 Farrell B. Efficient management of randomised controlled trials: nature or nurture. *BMJ* 1998; **317**: 1236–9.

45 McCormack J, Greenhalgh T. Seeing what you want to see in randomised controlled trials: versions and perversions of the UK PDS data. *BMJ* 2000; **320**: 720–3.

46 Lumley J, Bastian H. Competing or complementary: ethical considerations and the quality of randomised trials. *Int J Technol Assess Health Care* 1996; **12**: 247–63.

47 Mays N, Pope C, eds. *Qualitative research in health care,* 2nd edn. London: BMJ Publications, 2000.

48 Altman D. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ* 1996; **313**: 570–1.

49 Jadad AR. *Randomised controlled trials: a user's guide*. London: BMJ Publications, 1998.

50 Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998; **2**: 1–124.

51 Doll R, Hill AB. Mortality in relation to smoking: ten years' observations on British doctors. *BMJ* 1964; **i**: 1399–414, 1460–7.

52 Doll R, Peto R. Mortality in relation to smoking: ten years' observations on British doctors. *BMJ* 1976; **ii**: 1525–36.

53 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.

54 Guillebaud J. *The pill,* 4th edn. Oxford: Oxford University Press, 1991.

55 Macnaughton J. Anecdote in clinical practice. In: Greenhalgh T, Hurwitz B, eds. *Narrative based medicine: dialogue and discourse in clinical practice*. London: BMJ Publications, 1999: 202–11.

56 McBride WG. Thalidomide and congenital abnormalities. *Lancet* 1961; **ii**: 1358.

57 Pohjola-Sintonen S, Viitasalo M, Toivonen L, Neuvonen P. Itraconazole prevents terfenadine metabolism and increases the risk of torsades de pointes ventricular tachycardia. *Eur J Clin Pharm* 1993; **45**: 191–3.

58 Middle C, Johnson A, Petty T, Sims L, Macfarlane A. Ethics approval for a national postal survey: recent experience. *BMJ* 1995; **311**: 659–60.

59 Olde Rickert MGM, ten Have HAMJ, Hoefnagels WHL. Informed consent in biomedical studies on ageing: survey of four journals. *BMJ* 1996; **313**: 1117.

60 MRC Vitamin Research Group. Prevention of neural tube defects. Results of the MRC Vitamin Study. *Lancet* 1991; **338**: 131–7.

Chapter 4: Assessing methodological quality

As I argued in section 3.1, a paper will sink or swim on the strength of its methods section. This chapter considers five essential questions which should form the basis of your decision to "bin" it, suspend judgement or use it to influence your practice.

- Was the study original?
- Who is it about?
- Was it well designed?
- Was systematic bias avoided (i.e. was the study adequately "controlled")?
- Was it large enough and continued for long enough to make the results credible?

These questions are considered in turn below.

4.1 Was the study original?

There is, in theory, no point in testing a scientific question which someone else has already proved one way or the other. But in real life, science is seldom so cut and dried. 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 (if they are methodologically sound) 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 there being several studies in the literature which have addressed the same question in pretty 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 before?" but "Does this new research add to the literature in any way?".

- Is this study bigger, continued for longer or otherwise more substantial than the previous one(s)?
- Are the methods 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 metaanalysis of previous studies?
- Is the population studied different in any way (for example, has the study looked at different ethnic groups, ages or gender than previous studies)?
- Is the clinical issue addressed of sufficient importance, and does there exist 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?

4.2 Who is the study about?

One of the first papers that ever caught my eye was entitled "But will it help *my* patients with myocardial infarction?"[1] I don't remember the details of the article but it opened my eyes to the fact that research on someone else's patients may not have a take home message for my own practice. This is not mere xenophobia. The main reasons why the participants (Sir Iain Chalmers has argued forcefully against calling them "patients")[2] in a clinical trial or survey might differ from patients in "real life" are as follows.

- They were more, or less, ill than the patients you see
- They were from a different ethnic group, or lived a different lifestyle, from your own patients
- They received more (or different) attention during the study than you could ever hope to give your patients
- Unlike most real life patients, they had nothing wrong with them apart from the condition being studied

- None of them smoked, drank alcohol or were taking the contraceptive pill.

Hence, before swallowing the results of any paper whole, 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 putting an advertisement in the local newspaper. However, this method would be a good example of *recruitment bias* since the sample you obtain would be skewed in favour of users who were highly motivated and liked to read newspapers. You would, of course, be better to issue a questionnaire to every user (or to a one in 10 sample of users) who turned up on a particular day.

Who was included in the study?

Many trials in the UK routinely exclude patients with co-existing illness, those who do not speak English, those taking certain other medication, and the 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 all that logical.[3] The results of pharmacokinetic studies of new drugs in 23 year old healthy male volunteers will clearly not be applicable to the average elderly female! This issue, which has been a bugbear of some doctors for some time,[4] has recently been taken up by the patients themselves, most notably in the plea from patient support groups for a broadening of inclusion criteria in trials of anti-AIDS drugs.[5]

Who was excluded from the study?

For example, a RCT 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 who funded the research provide new equipment which would not be available to the ordinary clinician? These factors would not, of course, invalidate the study itself but they may cast doubt on the applicability of its findings to your own practice.

4.3 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. Personally, I assess the basic design of a clinical trial via two questions.

What specific intervention or other manoeuvre was being considered and what was it being compared with?

This is one of the most fundamental questions in appraising any paper. 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 overestimate its originality and potential importance. In the examples in Box 4.1, I have used hypothetical statements so as not to cause offence, but they are all based on similar mistakes seen in print.

What outcome was measured, and how?

If you had an incurable disease for which a pharmaceutical company claimed to have produced a new wonder drug, you would measure the efficacy of the drug in terms of whether it made you live longer (and, perhaps, whether life was *worth* living given your condition and any side effects of the medication). You would not be too interested in the level of some obscure enzyme in your blood which the manufacturer assured you was a reliable indicator of your chances of survival. The use of such *surrogate endpoints* is discussed further in section 6.3.

The measurement of symptomatic (for example, pain), functional (for example, mobility), psychological (for example, anxiety) or social (for example, inconvenience) effects of an intervention is fraught with even more problems. The methodology of developing, administering and interpreting such "soft" outcome measures is beyond the scope of this book. But in general, you

Box 4.1 Examples of problematic descriptions in the methods section of papers

What the authors said	*What thev 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 167 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 got the helpline number	Failure to treat groups equally apart form the specific intervention
"We measured the use of vitamin C in the prevention of the common cold"	A systematic literature search would have found numerous previous studies on this subject (see section 8.1)	Unoriginal study

should always look for evidence in the paper that the outcome measure has been objectively validated, that someone has demonstrated that the scale of anxiety, pain, and so on used in this study has previously been shown to measure 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.[6]

4.4 Was systematic bias avoided or minimised?

Systematic bias is defined by epidemiologists Geoffrey Rose and David Barker as anything which erroneously influences the conclusions about groups and distorts comparisons.[7] 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 like one another 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 using the same outcome measures. Different study designs call for different steps to reduce systematic bias.

Randomised controlled trials

In a RCT, systematic bias is (in theory) avoided by selecting a sample of participants from a particular population and allocating them randomly to the different groups. Section 3.3 describes some ways in which bias can creep into even this gold standard of clinical trial design and Figure 4.1 summarises particular sources to check for.

Non-randomised controlled clinical trials

I recently chaired a seminar in which a multidisciplinary group of students from the medical, nursing, pharmacy, and allied professions were presenting the results of several in-house research studies. All but one of the studies presented were of comparative but non-randomised design – that is, one group of patients (say, hospital outpatients with asthma) had received one intervention (say, an educational leaflet), while another group (say, patients attending GP surgeries with asthma) had received another

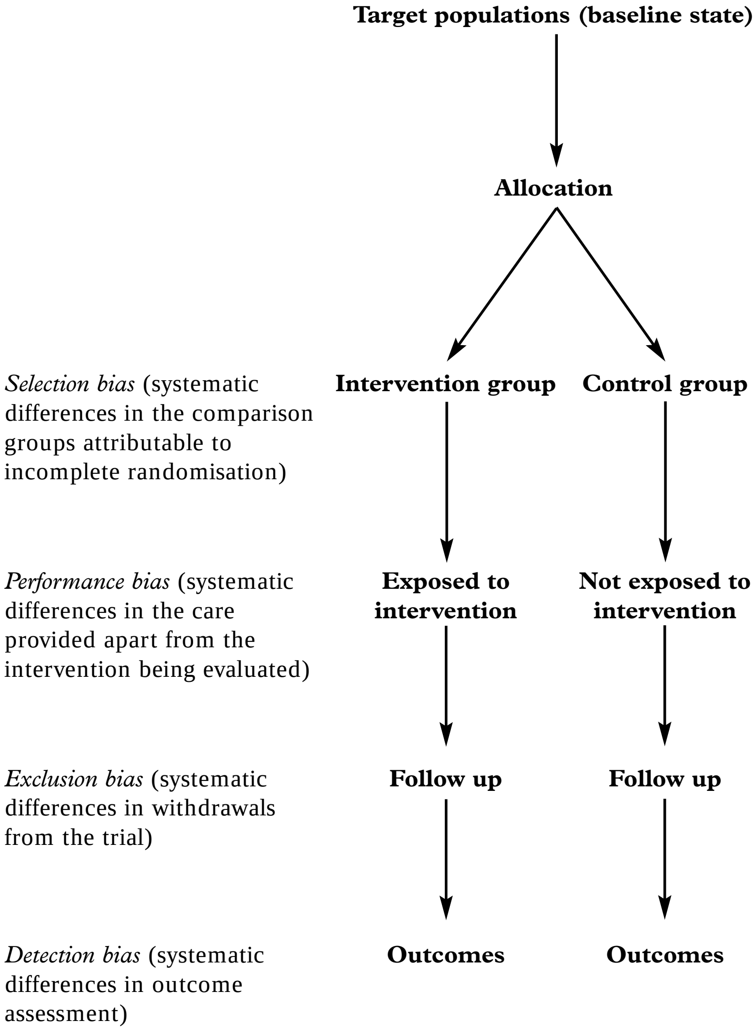

Target populations (baseline state)
Allocation
Intervention group
Control group
Exposed to intervention
Not exposed to intervention
Follow up
Follow up
Outcomes
Outcomes
Selection bias (systematic differences in the comparison groups attributable to incomplete randomisation)
Performance bias (systematic differences in the care provided apart from the intervention being evaluated)
Exclusion bias (systematic differences in withdrawals from the trial)
Detection bias (systematic differences in outcome assessment)

Figure 4.1 Sources of bias to check for in a randomised controlled trial

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 listed in Figure 4.1 for RCTs).

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 .[8, 9] Sometimes, the authors of such a paper will list the important features of each group (such as mean age, sex ratio, markers of disease severity, and so on) in a table to allow you to compare these differences yourself.

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, gender mix, socioeconomic status, presence of co-existing 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 (see section 5.5) will be dangerously misleading.[10]

This problem is illustrated by the various cohort studies on the risks and benefits of alcohol, which have consistently demonstrated a "J-shaped" relationship between alcohol intake and mortality. The best outcome (in terms of premature death) lies with the cohort who are moderate drinkers.[11] Self confessed teetotallers, it seems, are significantly more likely to die young than the average person who drinks three or four alcoholic drinks a day.

But can we assume that teetotallers are, *on average*, identical to moderate drinkers except for the amount they drink? We certainly can't. As we all know, the teetotal population includes those who have been ordered to give up alcohol on health grounds ("sick quitters"), those who, for health or other reasons, have cut out a host of additional items from their diet and lifestyle, those from certain religious or ethnic groups which would be underrepresented in the other cohorts (notably Muslims and Seventh Day Adventists), and those who drink like fish but choose to lie about it.

The details of how these different features of "teetotalism" were controlled for by the epidemiologists are discussed elsewhere.[11] In summary, even when due allowance is made in the analysis for potential confounding variables in subjects who describe themselves as non-drinkers, these subjects' increased risk of premature mortality appears to remain.

Case-control studies

In case-control studies (in which, as I explained in section 3.7, the experiences of individuals with and without a particular disease are analysed retrospectively to identify putative causative events), the process which 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.[12] In order to answer the question "Did the vaccine cause brain damage?", a case-control study had been undertaken in which a "case" was defined as an infant who, previously well, had exhibited fits or other signs suggestive of brain damage within one week of receiving the vaccine. A control was an infant of the same age and sex taken from the same immunisation register, who had received immunisation and who may or may not have developed symptoms at some stage.

New onset of features of brain damage in apparently normal babies is extremely rare but it does happen and the link with recent immunisation could conceivably be coincidental. Furthermore, heightened public anxiety about the issue could have biased the recall of parents and health professionals so that infants whose neurological symptoms predated, or occurred some time after, the administration of pertussis vaccine might be wrongly classified as cases. The judge in the court case ruled that misclassification of three such infants as "cases" rather than controls led to the overestimation of the harm attributable to whooping cough vaccine by a factor of three.[12] Although this ruling has subsequently been challenged, the principle stands – that assignment of "caseness" in a case-control study must be done rigorously and objectively if systematic bias is to be avoided.

4.5 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 you believe that the evaluation of clinical signs and the interpretation of diagnostic tests such as ECGs and X-rays is 100% objective, you haven't been in the game very long.

The chapter "The clinical examination" in Sackett and colleagues' book *Clinical epidemiology – a basic science for clinical medicine*[13] provides substantial evidence that when examining patients, doctors find what they expect and hope to find. It is rare indeed for two competent clinicians to reach agreement beyond what would be expected by chance in more than two cases in every three for any given aspect of the physical examination or interpretation of any diagnostic test.

The level of agreement beyond chance between two observers can be expressed mathematically as the κ (Kappa) score, with a score of 1.0 indicating perfect agreement. κ scores for specialists assessing the height of a patient's jugular venous pressure, classifying diabetic retinopathy from retinal photographs, and interpreting a mammogram X-ray were, respectively, 0.42, 0.55, and 0.67.[13]

The above digression into clinical disagreement should have persuaded you that efforts to keep assessors "blind" (or, to avoid offence to the visually impaired, *masked*) to the group allocation of their patients are far from superfluous. 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 4.1.

An excellent example of controlling for bias by adequate "blinding" was published in the *Lancet* a few years ago.[14] Majeed and colleagues performed a RCT which demonstrated, in contrast with the findings of several previous studies, that the recovery time (days in hospital, days off work, and time to resume full activity) after laparoscopic removal of the gallbladder (the "keyhole surgery" approach) was no quicker than that associated with traditional open operation. The discrepancy between this trial and its predecessors may have been due to Majeed and colleagues' meticulous attempt to reduce bias (see Figure 4.1). The patients were not randomised until after induction of general anaesthesia. Neither the patients nor their carers were aware of which operation had been done, since all patients left the operating theatre with identical dressings (complete with blood stains!). These findings challenge previous authors to ask themselves whether it was expectation bias (see section 7.3) rather than swifter recovery which spurred doctors to discharge the laparoscopic surgery group earlier.

4.6 Were preliminary statistical questions addressed?

As a non-statistician, I tend only to look for three numbers in the methods section of a paper:

1. the size of the sample
2. the duration of follow up
3. the completeness of follow up.

Sample size

One crucial prerequisite before embarking on a clinical trial is to perform a sample size ("power") calculation. 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.[15]

In order to calculate sample size, the clinician must decide two things.

- 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 mmHg and the effect would be a statistically significant lowering of the chances of developing stroke (i.e. the odds are less than 1 in 20 that the reduced incidence occurred by chance).[16] However, if the people being asked to take this drug had only mildly raised blood pressure and no other major risk factors for stroke (i.e. they were relatively young, not diabetic, had normal cholesterol levels, and so on), this level of difference would only prevent around one stroke in every 850 patients treated[17] – a clinical difference in risk which many patients would classify as not worth the effort of taking the tablets.
- What the mean and the standard deviation (abbreviated SD; see section 5.2) of the principal outcome variable is.

If the outcome in question is an event (such as hysterectomy) rather than a quantity (such as blood pressure), the items of data required are the proportion of people experiencing the event in the population and an estimate of what might constitute a clinically significant change in that proportion.

Once these items of data have been ascertained, the minimum sample size can be easily computed using standard formulae, nomograms or tables, which may be obtained from published papers,[15, 18] textbooks,[19] or commercial statistical software packages.[20] Thus, the researchers can, *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 likelihood of detecting a true difference is known as the *power* of the study. It is common for studies to stipulate a power of between 80% and 90%. Hence, when reading a paper about a RCT, you should look for a sentence which reads something like this (which is taken from Majeed and colleagues' cholecystectomy paper described above):

> "For a 90% chance of detecting a difference of one night's stay in hospital using the Mann-Whitney U-test [see Chapter 5, Table 1], 100 patients were needed in each group (assuming SD of 2 nights). This

> gives a power greater than 90% for detecting a difference in operating times of 15 minutes, assuming a SD of 20 minutes."[14]

If the paper you are reading does not give a sample size calculation *and* it appears to show that there is no difference between the intervention and control arms of the trial, you should extract from the paper (or directly from the authors) the information in the two bullet points above and do the calculation yourself. Underpowered studies are ubiquitous in the medical literature, usually because the authors found it harder than they anticipated to recruit their subjects. Such studies typically lead to a type II or ß error, i.e. 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 itself was adequate, a study must be continued for long enough for the effect of the intervention to be reflected in the outcome variable. If the authors were looking at the effect of a new painkiller on the degree of postoperative pain, their study may only have needed a follow up period of 48 hours. On the other hand, if they were looking at the effect of nutritional supplementation in the preschool years on final adult height, follow up should have been measured in decades.

Even if the intervention has demonstrated a significant difference between the groups after, say, six months, that difference may not be sustained. As many dieters know from bitter experience, strategies to reduce obesity often show dramatic results after two or three weeks but if follow up is continued for a year or more, the unfortunate subjects have (more often than not) put most of the weight back on.

Completeness of follow up

It has been shown repeatedly that 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 check ups, and more likely to have experienced side effects on any medication than those who do not withdraw.[13] People who fail to complete questionnaires may feel differently about the issue (and probably less strongly) than those who send them back by return of post. People on a weight reducing programme are more likely to continue coming back if they are actually losing weight.

The reasons why patients withdraw from clinical trials include the following.

- Incorrect entry of patient into trial (i.e. 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 you should never look at the "adverse reaction" rate in the intervention group without comparing it with that on placebo. Inert tablets bring people out in a rash surprisingly frequently!
- Loss of patient motivation ("I don't want to take these tablets any more").
- Withdrawal by clinician for clinical reasons (for example, concurrent illness, pregnancy).
- Loss to follow up (for example, patient moves away).
- Death. Clearly, patients who die will not attend for their outpatient appointments, so unless specifically accounted for they might be misclassified as "dropouts". This is one reason why studies with a low follow up rate (say below 70%) are generally considered invalid.

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 *intent to treat* basis.[21] 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. If you look hard enough in a paper, you will usually find the sentence "Results were analysed on an intent to treat basis", but you should not be reassured until you have checked and confirmed the figures yourself.

There are, in fact, a few situations when intent to treat analysis is, rightly, 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 RCT, for the purposes of the analysis they effectively constitute a cohort study (see section 3.4).

4.7 Summing up

Having worked through the methods section, you should be able to tell yourself in a short paragraph what sort of study was performed, on how many subjects, where the subjects came from, what treatment or other intervention was offered, how long the follow up period was (or, if a survey, what the response rate was), and what outcome measure(s) were used. You should also, at this stage, identify what statistical tests, if any, were used to analyse the results (see Chapter 5). If you are clear about these things before reading the rest of the paper, you will find the results easier to understand, interpret and, if appropriate, reject. You should be able to come up with descriptions such as:

> This paper describes an unblinded randomised trial, concerned with therapy, in 267 hospital outpatients aged between 58 and 93 years, in which four layer compression bandaging was compared with standard single layer dressings in the management of uncomplicated venous leg ulcers. Follow up was six months. Percentage healing of the ulcer was measured from baseline in terms of the surface area of a tracing of the wound taken by the district nurse and calculated by a computer scanning device. Results were analysed using the Wilcoxon matched pairs test.

> This is a questionnaire survey of 963 general practitioners randomly selected from throughout the UK, in which they were asked their year of graduation from medical school and the level at which they would begin treatment for essential hypertension. Response options on the structured questionnaire were '90–99 mmHg', '100-109 mmHg', and '110 mmHg or greater'. Results were analysed using a Chi squared test on a 3 x 2 table to see whether the threshold for treating hypertension was related to whether the doctor graduated from medical school before or after 1975.

> This is a case report of a single patient with a suspected fatal adverse drug reaction to the newly released hypnotic drug Sleepol.

When you have had a little practice in looking at the methods section of research papers along the lines suggested in this chapter, you will find that it is only a short step to start using the checklists in Appendix 1 or the more comprehensive "Users' guides to the medical literature" referenced in Chapter 3. I will return to many of the issues discussed here in Chapter 6, in relation to evaluating papers on drug trials.

1 Mitchell JR. But will it help *my* patients with myocardial infarction? *BMJ* 1982; **285**: 1140–8.
2 Chalmers I. What I want from medical researchers when I am a patient. *BMJ* 1997; **310**: 1315–18.
3 Bero LA, Rennie D. Influences on the quality of published drug studies. *Int J Technol Assess Health Care*1996; **12**: 209–37.
4 Buyse ME. The case for loose inclusion criteria in clinical trials. *Acta Chirurg Belg* 1990; **90**: 129–31.
5 Phillips AN, Davey Smith G, Johnson MA. Will we ever know how to treat HIV infection? *BMJ* 1996; **313**: 608–10.
6 Dunning M, Needham G. *But will it work doctor? Report of conference held in Northampton, 22nd and 23rd May 1996*. London: King's Fund, 1996.
7 Rose G, Barker DJP. *Epidemiology for the uninitiated*, 3rd edn. London: BMJ Publications, 1994.
8 Chalmers TC, Celano P, Sacks HS, Smith H. Bias in treatment assignment in controlled clinical trials. *New Engl J Med* 1983; **309**: 1358–61.
9 Colditz GA, Miller JA, Mosteller JF. How study design affects outcome in comparisons of therapy. I: Medical. *Stat Med* 1989; **8**: 441–54.
10 Brennan P, Croft P. Interpreting the results of observational research: chance is not such a fine thing. *BMJ* 1994; **309**: 727–30.
11 Maclure M. Demonstration of deductive meta-analysis: alcohol intake and risk of myocardial infarction. *Epidemiol Rev* 1993; **15**: 328–51.
12 Bowie C. Lessons from the pertussis vaccine trial. *Lancet* 1990; **335**: 397–9.
13 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991: 19–49.
14 Majeed AW, Troy G, Nicholl JP *et al.* Randomised, prospective, single-blind comparison of laparoscopic versus small-incision cholecystectomy. *Lancet* 1996; **347**: 989–94.
15 Altman D. *Practical statistics for medical research*. London: Chapman and Hall, 1991. The nomogram for calculating sample size or power is on page 456.
16 Medical Research Council Working Party. MRC trial of mild hypertension: principal results. *BMJ* 1985; **291**: 97–104.
17 MacMahon S, Rogers A. The effects of antihypertensive treatment on vascular disease: re-appraisal of the evidence in 1993. *J Vasc Med Biol* 1993; **4**: 265–71.
18 Campbell MJ, Julious SA, Altman DG. Estimating sample size for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995; **311**: 1145–8.
19 Machin D, Campbell MJ, Fayers PM, Pinol APY. *Sample size tables for clinical*

studies, 2nd edn. London: Blackwell Science, 1997.

20 Iwane M, Panesky J, Plante K. A user's review of commercial sample size software for design of biomedical studies using survival data. *Controlled Clin Trials* 1997; **18**: 65–83.

21 Stewart LA, Parmar MKB. Bias in the analysis and reporting of randomized controlled trials. *Int J Technol Assess Health Care* 1996; **12**: 264–75.

Chapter 5: Statistics for the non-statistician

5.1 How can non-statisticians evaluate statistical tests?

In this age where medicine leans increasingly on mathematics, no clinician can afford to leave the statistical aspects of a paper entirely to the "experts". If, like me, you believe yourself to be innumerate, remember that you do not need to be able to build a car in order to drive one. What you do need to know about statistical tests is which is the best test to use for common problems. You need to be able to describe *in words* what the test does and in what circumstances it becomes invalid or inappropriate. Box 5.1 shows some frequently used "tricks of the trade", to which we all need to be alert (in our own as well as other people's practice).

I have found that one of the easiest ways to impress my colleagues is to let slip a comment such as: "Ah, I see these authors have performed a one tailed *F* test. I would have thought a two tailed test would have been more appropriate in these circumstances". As you will see from the notes below, you do not need to be able to perform the *F* test yourself to come up with comments like this, but you do need to understand what its tails mean.

The summary checklist in Appendix 1, explained in detail in the sections below, constitutes my own method for assessing the adequacy of a statistical analysis, which some readers will find too simplistic. If you do, please skip this section and turn either to a more comprehensive presentation for the non-statistician, the "Basic statistics for clinicians" series in the *Canadian Medical Association Journal*,[1–4] or to a more mainstream statistical textbook.[5] If, on the other hand, you find statistics impossibly difficult, take

Box 5.1 Ten ways to cheat on statistical tests when writing up results

- Throw all your data into a computer and report as significant any relationship where "$p < 0.05$" (see section 5.5a).
- If baseline differences between the groups favour the intervention group, remember not to adjust for them (see Section 5.2a).
- Do not test your data to see if they are normally distributed. If you do, you might get stuck with non-parametric tests, which aren't as much fun (see section 5.2b).
- Ignore all withdrawals ("dropouts") and non-responders, so the analysis only concerns subjects who fully complied with treatment (see section 4.6c).
- Always assume that you can plot one set of data against another and calculate an "*r* value" (Pearson correlation coefficient) (see section 5.4a), and that a "significant" *r* value proves causation (see section 5.4b).
- If outliers (points that 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 appear to be spurious results, leave them in (see section 5.3c).
- 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 (see section 5.5b).
- 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 (see section 5.2d).
- 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 females aged 52 to 61 (see section 5.2d).
- 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 (see section 5.2c).

these points one at a time and return to read the next point only when you feel comfortable with the previous ones. None of the points presupposes a detailed knowledge of the actual calculations involved.

The first question to ask, by the way, is "Have the authors used any statistical tests at all?". If they are presenting numbers and claiming that these numbers mean something, without using statistical methods to prove it, they are almost certainly skating on thin ice.

5.2 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 demonstrate that both 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). If there are important differences in these baseline characteristics, even though these may be due to chance, it can pose a challenge to your interpretation of results. In this situation, you can carry out certain adjustments to try to allow for these differences and hence strengthen your argument. To find out how to make such adjustments, see the section on this topic in Douglas Altman's book *Practical statistics for medical research*.[6]

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 a different 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".

All statistical tests are either parametric (i.e. they assume that the data were sampled from a particular form of distribution, such as a normal distribution) or non-parametric (i.e. they do not assume that the data were sampled from a particular type of distribution). In general, parametric tests are more powerful than non-parametric ones and so should be used if at all 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 demonstrate with non-parametric tests and this tempts researchers to use statistics such as the *r* value (see section 5.4) inappropriately. Not only is the *r* value (parametric) easier to calculate than an equivalent non-parametric statistic such as Spearman's σ, but it is also much more likely to give (apparently) significant results. Unfortunately it will also give an entirely spurious and misleading 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 (if present) are given in Table 5.1.

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 5.1 – 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* distribution, as shown in Figure 5.2. (Figure 5.2 in fact shows a negative skew, whereas body weight would be positively skewed. The average adult male body weight is 70 kg and people exist who are 140 kg but nobody weighs less than nothing, so the graph cannot possibly be symmetrical.)

Non-normal (skewed) data can sometimes be *transformed* to give a normal shape graph by plotting the logarithm of the skewed variable or performing some other mathematical transformation (such as square root or reciprocal). Some data, however, cannot be transformed into a smooth pattern and the significance of this is discussed below. For a further, very readable discussion about the normal distribution, see Chapter 7 of Martin Bland's book *An introduction to medical statistics*.[7]

Deciding whether data are normally distributed is not an academic exercise, since it will determine what type of statistical

Table 5.1 Some commonly used statistical tests

Parametric test	Example of equivalent non-parametric test	Purpose of test	Example
Two sample (unpaired) *t* test	Mann-Whitney U test	Compares two independent samples drawn from the same population	To compare girls' heights with boys' heights
One sample (paired) *t* test	Wilcoxon matched pairs test	Compares two sets of observations on a single sample (tests the hypothesis that the mean difference between two measurements is zero)	To compare weight of infants before and after a feed
One way analysis of variance using total sum of squares (e.g. *F* test)	Analysis of variance by ranks (e.g. Kruskall-Wallis test)	Effectively, a generalisation of the paired *t* 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 co-variates	In the above example, to determine if the results differ in males and females
No direct equivalent	χ^2 test	Tests the null hypothesis that the proportions of variables estimated from two (or more) independent samples are the same	To assess whether acceptance into medical school is more likely if the applicant was born in the UK
No direct equivalent	McNemar's test	Tests the null hypothesis that the proportions estimated from a paired sample are the same	To compare the sensitivity and specificity of two different diagnostic tests when applied to the same sample
Product moment correlation coefficient (Pearson's *r*)	Spearman's rank correlation coefficient (σ)	Assesses the *strength* of the straight line association between two continuous variables	To assess whether and to what extent plasma HbA1 level is related to plasma triglyceride level in diabetic patients
Regression by least squares method	No direct equivalent	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	No direct equivalent	Describes the numerical relation between a dependent variable and several predictor variables (co-variates)	To determine whether and to what extent a person's age, body fat, and sodium intake determine their blood pressure

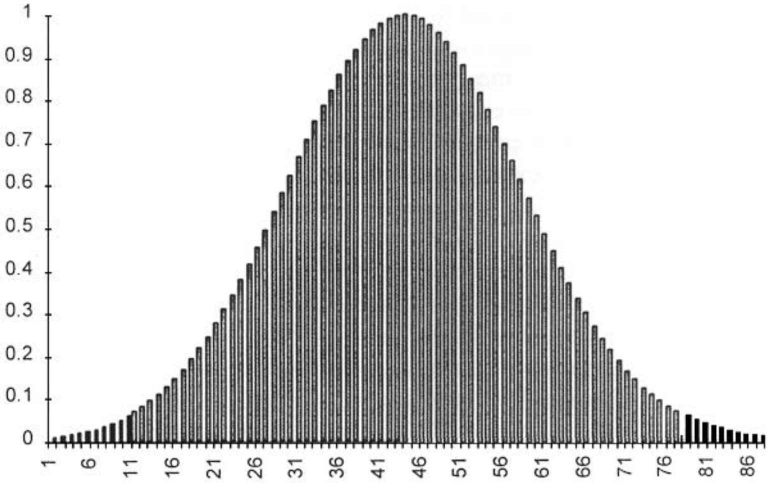


Figure 5.1 Example of a normal curve

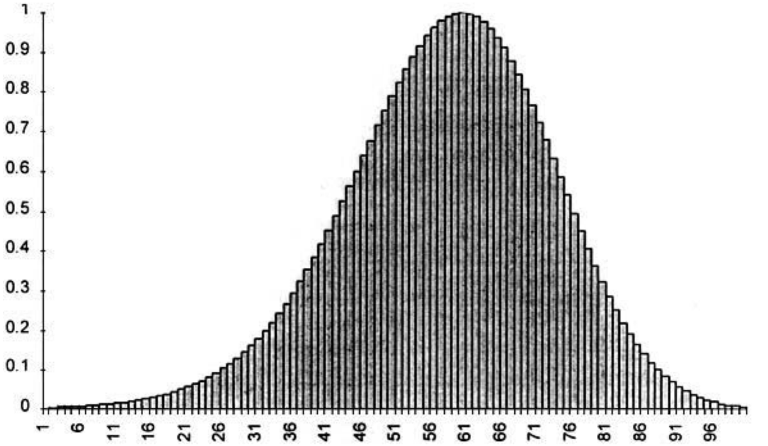


Figure 5.2 Example of a skew curve

tests to use. For example, linear regression (see section 5.4) will give misleading results unless the points on the scatter graph form a particular distribution about the regression line, i.e. 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 is very definitely cheating.

If the statistical tests in the paper are obscure, why have the authors chosen to use them and have they included a reference?

There sometimes seems to be an infinite number of possible statistical tests. In fact, most statisticians could survive with a formulary of about a dozen. The rest are small print and should be reserved for special indications. If the paper you are reading appears to describe a standard set of data wich have been collected in a standard way, but the test used is unpronounceable and 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.

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

Even if you are not interested in the statistical justification, common sense should tell you why points 8 and 9 in Box 5.1 amount to serious cheating. If you trawl for long enough you will inevitably find some category of patient which appears to have done particularly well or badly. However, each time you look to see if a particular subgroup is different from the rest you greatly increase the likelihood that you will eventually find one which appears to be so, even though the difference is entirely due to chance.

Similarly, 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.[8] Terminating an intervention trial prematurely for ethical reasons when subjects in one arm are faring particularly badly is different, and is discussed elsewhere.[8]

Going back and raking over your data to look for "interesting results" (retrospective subgroup analysis or, more colloquially, data dredging) can lead to false conclusions.[9] In an early study on the use of aspirin in the prevention of stroke in predisposed patients, the results showed a significant effect in both sexes combined and a retrospective subgroup analysis appeared to show that the effect was confined to males.[10] This conclusion led to aspirin being withheld from women for many years until the results of other studies (including a large metaanalysis[11]) showed this subgroup effect to be spurious.

This and other examples are given in a paper by Oxman and Guyatt, "A consumer's guide to subgroup analysis", which reproduces a useful checklist for deciding whether apparent differences in subgroup response are real.[12]

5.3 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, in fact, no great mystery about this. If you measure something twice on each subject (for example, lying and standing blood pressure), you will probably be interested not just in the average difference in 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 having 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 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 concept of a test with tails always has me thinking of devils or snakes, which I guess just reflects my aversion to statistics. In fact, the term "tail" refers to the extremes of the distribution – the dark areas in Figure 5.1. Let's say that that graph represents the diastolic blood pressures of a group of individuals 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 hand "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 put it *up*? Even if there are valid physiological reasons why that might be the case 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 and 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 (for example, faulty equipment), errors in interpretation (for example, misreading a meter reading), or errors in calculation (for example, 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. A few years ago, while doing a research project, I measured a number of different hormone levels 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) is quite a sophisticated statistical manoeuvre. If you are interested, try the relevant section in Douglas Altman's book.[13]

5.4 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.[14]

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

- The data (or, strictly, the population from which the data are drawn) should be normally distributed. If they are not, non-parametric tests of correlation should be used instead (see Table 5.1).
- The two variables should be structurally independent (that is, one should not be forced to vary with the other). If they are not, a paired t or other paired test should be used instead.
- Only a single pair of measurements should be made on each subject, since the measurements made on successive subjects need to be statistically independent of each other if we are to end up with unbiased estimates of the population parameters of interest.[14]
- 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 (see section 5.5), or a confidence interval, which expresses the range within which the "true" R value is likely to lie (see section 5.5). (Note that lower case "r" represents the correlation coefficient of the sample, whereas upper case "R" represents the correlation coefficient of the entire population.)

Remember, too, that even if the *r* value is an appropriate value to calculate from a set of data, it does not tell you whether the relationship, however strong, is causal (see below).

What, then, is regression? The term "regression" refers to a mathematical equation which 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 which calculated it) allows the target variable to be predicted from two or more independent variables (often known as *co-variables*).

The simplest regression equation, which you may remember from your schooldays, 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 that a quadratic rather than a linear regression would probably be more appropriate).

Of course, even when you have fed sufficient height–weight data into a computer for it to calculate the regression equation which 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 level into the computer and ask it how much each of these co-variables contributes to the overall equation (or model).

The elementary principles described here, particularly the points on the previous page, 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 in Martin Bland's textbook[14] and in the fourth article in the "Basic statistics for clinicians" series.[4]

Box 5.2 Tests for causality

- Is there evidence from true experiments in humans?
- Is the association strong?
- Is the association consistent from study to study?
- Is the temporal relationship appropriate (i.e. did the postulated cause precede the postulated effect)?
- Is there a dose-response gradient (i.e. 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 proven causal association?

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. In order to demonstrate 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. Box 5.2 gives some criteria, originally developed by Sir Austen Bradford Hill, which should be met before assuming causality.[15]

5.5 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 one in 20 (expressed as $p < 0.05$ and equivalent to a betting odds of 20 to one) as "statistically significant" and a p value of less than one in 100 ($p < 0.01$) as "statistically highly significant".

By definition, then, one chance association in 20 (this must be around one major published result per journal issue) will appear to be significant when it isn't, and one in 100 will appear highly significant when it is really what my children call a "fluke". Hence, if you *must* analyse multiple outcomes from your dataset, you need to make a correction to try to allow for this (some authors recommend the Bonferoni method[16, 17]).

A result in the statistically significant range (p < 0.05 or p < 0.01 depending on what you have chosen as the cutoff) suggests that the authors should reject the null hypothesis (i.e. the hypothesis that there is no real difference between two groups). But as I have argued earlier (see section 4.6), a p value in the non-significant range tells you that *either* there is no difference between the groups *or* there were too few subjects to demonstrate such a difference if it existed. It does not tell you which.

The p value has a further limitation. Gordon 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 a 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?".[1]

For this, we need confidence intervals, which are considered next.

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 which show a statistically significant difference between two arms of the trial) and "negative" ones (those which appear to show no difference), whether the strength of the evidence is *strong* or *weak* and whether the study is *definitive* (i.e. obviates the need for further similar studies). The calculation of confidence intervals has been covered with great clarity in Gardner and Altman's book *Statistics with confidence*[18] and their interpretation has been covered by Guyatt and colleagues.[2]

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 only conducted 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 – + 12) higher survival".

In this particular example, the 95% confidence interval overlaps zero difference and, if we were expressing the result as a dichotomy (i.e. Is the hypothesis "proven" or "disproven"?), 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 that the strength of that inference is weak".[2]

As section 8.3 argues, 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 "would a much larger trial be likely to show a significant benefit?". To answer this question, look at the *upper* 95% confidence interval of the result. There is only one chance in 40 (i.e. a 2½% chance, since the other 2½% of extreme results will lie below the *lower* 95% confidence interval) that the real result will be this much or more. Now ask yourself "Would this level of difference be *clinically* significant?" and if it wouldn't, you can classify the trial as not only negative but also definitive. If, on the other hand, the upper 95% confidence interval represented a clinically significant level of difference between the groups, the trial may be negative but it is also non-definitive.

Until recently, the use of confidence intervals was relatively uncommon in medical papers. In one survey of 100 articles from three top journals (*The New England Journal of Medicine*, *Annals of Internal Medicine*, and *Canadian Medical Association Journal*), only 43% reported any confidence intervals at all, whereas 66% gave a *p* value.[1] The figure is now probably somewhat higher but even so, many authors do not interpret their confidence intervals correctly. You should check carefully in the discussion section to see whether the authors have correctly concluded (a) whether and to what extent their trial supported their hypothesis and (b) whether any further studies need to be done.

5.6 The bottom line (quantifying the risk of benefit and harm)

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 (and I promise you they *are* simple: 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 which means something to the non-statistician. The 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, let me tell you about a survey which Tom Fahey and his colleagues conducted in 1995.[19] They wrote to 182 board members of district health authorities in England (all of whom would be in some way responsible for making important health service decisions) and put the following data to them about four different rehabilitation programmes for heart attack victims.

They asked which one they would prefer to fund.

- Programme A – which reduced the rate of deaths by 20%.
- Programme B – which produced an absolute reduction in deaths of 3%.
- Programme C – which increased patients' survival rate from 84% to 87%.
- Programme D – which meant that 31 people needed to enter the programme to avoid one death.

Of the 140 board members who responded, only three spotted that all four "programmes" in fact related to the same set of results. The other 137 all selected one of the programmes in preference to one of the others, thus revealing (as well as their own ignorance) the need for better basic training in epidemiology for health authority board members.

Table 5.2 Effect of coronary artery bypass graft on survival

	Outcome at 10 years		Total number of patients randomised in each group
Treatment	Dead	Alive	
Medical therapy	404	921	1324
CABG	350	974	1325

Let's continue with the example in Table 5.2, which Fahey and colleagues reproduced from a study by Salim Yusuf and colleagues.[20] I have expressed the figures as a 2 x 2 table giving details of which treatment the patients received in their randomised trial and whether they were dead or alive 10 years later.

Simple maths tells you that patients on medical therapy have a 404/1324 = 0.305 or 30.5% chance of being dead at 10 years (and a 0.695 or 69.5% chance of still being alive). Let's call the risk of death CER (control event rate). Patients randomised to CABG have a 350/1325 = 0.264 or 26.4% chance of being dead at 10 years (and a 0.736 or 73.6% chance of still being alive). Let's call their risk of death the EER (experimental event rate).

The *relative risk* of death, i.e. the risk in CABG patients compared with controls, is CER/EER or 0.264/0.305 = 0.87 (87%).

The *relative risk reduction*, i.e. the amount by which the risk of death is reduced by CABG, (CER-EER)/CER = (0.305–0.264)/ 0.305 = 0.041/0.305 = 13%.

The *absolute risk reduction* (or risk difference), i.e. the absolute amount by which CABG reduces the risk of death at 10 years, is 0.305−0.264 = 0.041 (41%).

The *number needed to treat*, i.e. how many patients need a CABG in order to prevent, on average, one additional death by 10 years, is the reciprocal of the absolute risk reduction, 1/ARR = 1/0.041 = 24.

The final way of expressing the effect of treatment which I want to introduce here is the *odds ratio*. Look back at Table 5.2 and you will see that the "odds" of dying compared to the "odds" of surviving for patients in the medical treatment group are 404/921 = 0.44, and for patients in the CABG group are 350/974 = 0.36. The *ratio* of these odds will be 0.44/0.36 = 1.22, which is another way of expressing the fact that in this study, patients in the CABG group did better.

The general formulae for calculating these "bottom line" effects of an intervention are reproduced in Appendix 4 and for a discussion on which of these values is most useful in which circumstances, see Jaenschke and colleagues' article in the "Basic statistics for clinicians" series[3] or Chapter 7 (Deciding on the best therapy) of Sackett *et al*'s clinical epidemiology textbook.[21]

5.7 Summary

It is possible to be seriously misled by taking the statistical competence (and/or the intellectual honesty) of authors for granted. Statistics can be an intimidating science and understanding its finer points often calls for expert help. But I hope that this chapter has shown you that the statistics used in most medical research papers can be evaluated by the non-expert using a simple checklist such as that in Appendix 1. In addition, you might like to check the paper you are reading (or writing) against the common errors given in Box 5.1.

[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. 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: 461–2.

7 Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 1987: 112–29.

8 Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Stat Med* 1987; **7**: 1231–42.

9 Stewart LA, Parmar MKB. Bias in the analysis and reporting of randomized controlled trials. *Int J Technol Assess Health Care* 1996; **12**: 264–75.

10 Canadian Cooperative Stroke Group. A randomised trial of aspirin and sulfinpyrazone in threatened stroke. *New Engl J Med* 1978; **299**: 53–9.

11 Antiplatelet Triallists Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. *BMJ* 1988; **296**: 320–1.

12 Oxman AD, Guyatt GH. A consumer's guide to subgroup analysis. *Ann Intern Med* 1992; **116**: 79–84.

13 Altman D. *Practical statistics for medical research*. London: Chapman and Hall, 1995: 126–30.

14 Bland M. *An introduction to medical statistics*. Oxford: Oxford University Press, 1987: 188–215.

15 Bradford Hill A. The environment and disease: association or causation? *Proc R Soc Med* 1965; **58**: 295–300. Adapted version is reproduced with permission from Haines A. Multi-practice research: a cohort study. In: Jones R, Kinmonth A-L, eds. *Critical reading for primary care*. Oxford: Oxford University Press, 1995: 124.

16 Altman D. *Practical statistics for medical research*. London: Chapman and Hall, 1995: 210–12.

17 Pocock SJ, Geller XPL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**: 487–98.

18 Gardner MJ, Altman DG, eds. *Statistics with confidence: confidence intervals and statistical guidelines*. London: BMJ Publications, 1989.

19 Fahey T, Griffiths S, Peters TJ. evidence based purchasing: understanding the results of clinical trials and systematic reviews. *BMJ* 1995; **311**: 1050–60.

20 Yusuf S, Zucker D, Peduzzi P *et al*. Effect of coronary artery bypass surgery on survival: overview of ten year results from randomized trials by the Coronary Artery Surgery Triallists Collaboration. *Lancet* 1994; **344**: 563–70.

21 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991: 187–248.

Chapter 6: Papers that report drug trials

6.1 "Evidence" and marketing

If you are a clinical doctor or nurse practitioner (i.e. if you prescribe drugs), the pharmaceutical industry is interested in you and spends a proportion of its multimillion pound annual advertising budget trying to influence you (see Box 6.1). The most effective way of changing the prescribing habits of a clinician is via a personal representative (known to most of us in the UK as the "drug rep" and to our North American colleagues as the "detailer"), who travels round with a briefcase full of "evidence" in support of his or her wares.[1] Indeed, as Chapter 12 discusses in more detail, the evidence based medicine movement has learnt a lot from the drug industry in recent years about changing the behaviour of physicians and now uses the same sophisticated techniques of persuasion in what is known as "academic detailing" of individual health professionals.[2]

Before you agree to meet a "rep", remind yourself of some basic rules of research methodology. As sections 3.4 and 3.6 argued, questions about the benefits of therapy should ideally be addressed with randomised controlled trials. But preliminary questions about pharmacokinetics (i.e. how the drug behaves while it is getting to its site of action), particularly those relating to bioavailability, require a straight dosing experiment in healthy (and, if ethical and practicable, sick) volunteers.

Common (and hopefully trivial) adverse drug reactions may be picked up, and their incidence quantified, in the RCTs undertaken to demonstrate the drug's efficacy. But rare (and usually more serious) adverse drug reactions require both pharmacovigilance surveys (collection of data prospectively on patients receiving a newly licensed drug) and case-control studies (see section 3.5) to establish association. Ideally, individual rechallenge experiments (where the patient who has had a reaction considered to be caused by the drug is given the drug again in carefully supervised circumstances) should be performed to establish causation.[3]

Box 6.1 Ten tips for the pharmaceutical industry: how to present your product in the best light

- Think up a plausible physiological mechanism why the drug works and become slick at presenting it. Preferably, find a surrogate endpoint that is heavily influenced by the drug, though it may not be strictly valid (see section 6.3)
- When designing clinical trials, select a patient population, clinical features, and trial length that reflect the maximum possible response to the drug
- If possible, compare your product only with placebos. If you must compare it with a competitor, make sure the latter is given at subtherapeutic dose
- Include the results of pilot studies in the figures for definitive studies ("Russian doll publication"), so it looks like more patients have been randomised than is actually the case
- Omit mention of any trial that had a fatality or serious adverse drug reaction in the treatment group. If possible, don't publish such studies
- Get your graphics department to maximise the visual impact of your message. It helps not to label the axes of graphs or say whether scales are linear or logarithmic. Make sure you do not show individual patient data or confidence intervals
- Become master of the hanging comparative ("better" – but better than what?)
- Invert the standard hierarchy of evidence so that anecdote takes precedence over randomised trials and metaanalyses
- Name at least three local opinion leaders who use the drug and offer "starter packs" for the doctor to try
- Present a "cost effectiveness" analysis that shows that your product, even though more expensive than its competitor, "actually works out cheaper" (see section 10.1)

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.[4] 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.[5] Reference back to section 3.6 and a look at the classic *Lancet* series on placebo effects[6–12] or the more recent overview from the UK Health Technology Assessment Programme[13] should remind you why uncontrolled before and after studies are the stuff of teenage magazines, not hard science.

Dr Andrew Herxheimer, who edited *Drug and Therapeutics Bulletin* for many years, once undertook a survey of "references" cited in advertisements for pharmaceutical products in the leading UK medical journals. He tells me that a high proportion of such references cite "data on file" and many more refer to publications written, edited, and published entirely by the industry. Evidence from these sources has sometimes (though by no means invariably) been shown to be of lower scientific quality than that which appears in independent, peer reviewed journals.[5] And let's face it, if you worked for a drug company which had made a major scientific breakthrough you would probably submit your findings to a publication such as the *Lancet* or the *New England Journal of Medicine* before publishing them in-house. In other words, you don't need to "trash" papers about drug trials *because* of where they have been published, but you do need to look closely at the methods and statistical analysis of such trials.

6.2 Making decisions about therapy

Sackett and colleagues, in their book *Clinical epidemiology – a basic science for clinical medicine*,[14] argue that before starting a patient on a drug, the doctor should:

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

For example, in the treatment of high blood pressure, the doctor or nurse practitioner 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 comparative trials, as well as between non-drug treatments such as salt restriction
- the *treatment target* might be a phase V diastolic blood pressure (right arm, sitting) of less than 90 mmHg 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. In a veiled slight on surrogate endpoints, Sackett and his team remind us that the choice of specific therapy should be determined by evidence of what *does* work and not on what *seems* to work or *ought* to work. "Today's therapy", they warn (p 188), "when derived from biologic facts or uncontrolled clinical experience, may become tomorrow's bad joke."[14]

6.3 Surrogate endpoints

I have not included this section solely because it is a particular hobby horse of mine. If you are a practising (and non-academic) clinician, your main contact with published papers may well be through what gets fed to you by a drug rep. The pharmaceutical industry is a slick player at the surrogate endpoint game and I make no apology for labouring the point that such outcome measures must be evaluated very carefully.

I will define a surrogate endpoint as "*a variable which is relatively easily measured and which predicts a rare or distant outcome of either a toxic stimulus (for example, pollutant) or a therapeutic intervention (for example, drug, surgical procedure, piece of advice), but which is not itself a direct measure of either harm or clinical benefit*". The growing interest in surrogate endpoints in medical research reflects two important features of their use.

- They can considerably reduce the *sample size, duration* and, therefore, *cost* of clinical trials.

- 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 endpoints include:

- pharmacokinetic measurements (for example, concentration-time curves of a drug or its active metabolite in the bloodstream)
- in vitro (i.e. laboratory) measures such as the mean inhibitory concentration (MIC) 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) "biological markers of disease" (for example, microalbuminuria in the measurement of diabetic kidney disease[15])
- radiological appearance (for example, shadowing on a chest X-ray).

Surrogate endpoints have a number of drawbacks. First, a change in the surrogate endpoint 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?". Second, the surrogate endpoint may not closely reflect the treatment target; in other words, it may not be valid or reliable. Third, the use of a surrogate endpoint has the same limitations as the use of any other *single* measure of the success or failure of therapy – it ignores all the other measures! Overreliance on a single surrogate endpoint as a measure of therapeutic success usually reflects a narrow or naïve clinical perspective.

Finally, surrogate endpoints 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 liable to be invalid.[16–18]

- In animal studies, the population being studied has fairly uniform biological characteristics and may be genetically inbred.

- Both the tissue and the disease being studied may vary in important characteristics (for example, susceptibility to the pathogen, rate of cell replication) from the parallel condition in human subjects.
- The animals are kept in a controlled environment which minimises the influence of lifestyle variables (for example, diet, exercise, stress) and concomitant medication.
- Giving high doses of chemicals to experimental animals may distort the usual metabolic pathways and thereby give misleading results. Animal species best suited to serve as a surrogate for humans vary for different chemicals.

Box 6.2 Ideal features of a surrogate endpoint

- The surrogate endpoint should be reliable, reproducible, clinically available, easily quantifiable, affordable, and exhibit a "dose–response" effect (that is, the higher the level of the surrogate endpoint, 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 co-variable. The relation between the surrogate endpoint and the disease should have a biologically plausible explanation
- It should be sensitive – that is, a "positive" result for the surrogate endpoint should pick up all or most patients at increased risk of adverse outcome
- It should be specific – that is, a "negative" result should exclude all or most of those without increased risk of adverse outcome
- There should be a precise cutoff between normal and abnormal values
- It should have an acceptable positive predictive value – that is, a "positive" result should always or usually mean that the patient thus identified is at increased risk of adverse outcome (see section 7.2)
- It should have an acceptable negative predictive value – that is, a "negative" result should always or usually mean that the patient thus identified is not at increased risk of adverse outcome (see section 7.2)
- It should be amenable to quality control monitoring
- Changes in the surrogate endpoint should rapidly and accurately reflect the response to treatment – in particular, levels should normalise in states of remission or cure

The ideal features of a surrogate endpoint are shown in Box 6.2 – and microalbuminuria in diabetic kidney disease is a good example of a marker that fulfils most if not all of these criteria.[15] If the rep who is trying to persuade you of the value of the drug cannot justify the endpoints used, you should challenge him or her to produce additional evidence.

One important example of the invalid use of a surrogate endpoint is the CD4 cell count (a measure of one type of white blood cell which, when I was at medical school, was known as the "T-helper cell") in monitoring progression to AIDS in HIV positive subjects. The CONCORDE trial[19] was a RCT comparing early versus late initiation of zidovudine therapy in patients who were HIV positive but clinically asymptomatic. Previous studies had shown that early initiation of therapy 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 while 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 issued earlier by authors suspicious of the validity of this endpoint.[20] Subsequent research in this field has attempted to identify a surrogate endpoint that correlates with real therapeutic benefit, i.e. progression of asymptomatic HIV infection to clinical AIDS and survival time after the onset of AIDS. A recent review summarises the story and concludes that a combination of several markers (including percentage of CD4 C29 cells, degree of fatigue, age, and haemoglobin level) predicts progression much better than the CD4 count.[21]

If you think this is an isolated example of the world's best scientists all barking up the wrong tree in pursuit of a bogus endpoint, check out the literature on using ventricular premature beats (a minor irregularity of the heartbeat) to predict death from serious heart rhythm disturbance,[22, 23] blood levels of antibiotics to predict clinical cure of infection,[24] plaques on an MRI scan to chart the progression of multiple sclerosis,[25] and the use of the prostate specific antigen (PSA) test to measure the response to therapy in prostate cancer (see p 116).[26, 27] You might also like to see the fascinating literature on the development of valid and relevant surrogate endpoints in the field of cancer prevention.[28]

Clinicians are increasingly sceptical of arguments for using new drugs, or old drugs in new indications, that are not justified by direct evidence of effectiveness. Before surrogate endpoints can be used in the marketing of pharmaceuticals, those in the industry must justify the utility of these measures by demonstrating a plausible and consistent link between the endpoint and the development or progression of disease.

It would be wrong to suggest that the pharmaceutical industry develops surrogate endpoints with the deliberate intention to mislead the licensing authorities and health professionals. Surrogate endpoints, as I argued in section 6.1, have both ethical and economic imperatives. However, the industry does have a vested interest in overstating its case on the strength of these endpoints. Given that much of the data relating to the validation of surrogate endpoints 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 the setting up of a data archive which would pool data across studies.[29] If, like me, you continually find yourself questioning the validity of surrogate endpoints, you might like to read more about the subject in a recent review.[30]

6.4 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 antiinflammatory drug will recognise the gastric erosion example. The question to ask him or her is not "What is the incidence of gastric erosion on your drug?" but "What is the incidence of potentially life-threatening gastric bleeding?". Other questions to ask drug reps, reproduced from an article in *Drug and Therapeutics Bulletin*[31] and other sources,[1, 5, 14] are listed below.

1. See representatives only by appointment. Choose to see only those whose product interests you and confine the interview to that product.

2. Take charge of the interview. Do not hear out a rehearsed sales routine but ask directly for the information below.

3. Request independent published evidence from reputable peer reviewed journals.

4. Do not look at promotional brochures, which often contain unpublished material, misleading graphs, and selective quotations.

5. Ignore anecdotal "evidence" such as the fact that a medical celebrity is prescribing the product.

6. Using the "STEP" acronym, ask for evidence in four specific areas.

   - *Safety* – i.e. 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*, which is best measured by comparing the pooled withdrawal rates between the drug and its most significant competitor
   - *Efficacy*, of which the most relevant dimension is how the product compares with your current favourite
   - *Price*, which should take into account indirect as well as direct costs (see section 10.3).

7. Evaluate the evidence stringently, paying particular attention to the power (sample size) and methodological quality of clinical trials and the use of surrogate endpoints. Do not accept theoretical arguments in the drug's favour (for example, "longer half life") without direct evidence that this translates into clinical benefit.

8. Do not accept the newness of a product as an argument for changing to it. Indeed, there are good scientific arguments for doing the opposite.[32]

9. Decline to try the product via starter packs or by participating in small scale, uncontrolled "research" studies.

10. Record in writing the content of the interview and return to these notes if the rep requests another audience.

[1] Shaughnessy AF, Slawson DC. Pharmaceutical representatives. *BMJ* 1996; **312**: 1494-5.

[2] Thomson O'Brien MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Educational outreach visits: effects on professional practice and health care outcomes. In: *The Cochrane Library*, issue 1. Oxford: Update Software, 2000.

[3] Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991: 297–301.

4 Bardelay D. Visits from medical representatives: fine principles, poor practice. *Prescriber Int* 1995; **4**: 120–2.

5 Bero LA, Rennie D. Influences on the quality of published drug studies. *Int J Technol Assess Health Care* 1996; **12**: 209–37.

6 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.

7 Joyce CR. Placebo and complementary medicine. *Lancet* 1994; **344**: 1279–81.

8 Laporte JR, Figueras A. Placebo effects in psychiatry. *Lancet* 1994; **344**:1206–9.

9 Johnson AG. Surgery as a placebo. *Lancet* 1994; **344**: 1140–2.

10 Thomas KB. The placebo in general practice. *Lancet* 1994; **344**:1066–7.

11 Chaput de Saintonge DM, Herxheimer A. Harnessing placebo effects in health care. *Lancet* 1994; **344**: 995–8.

12 Gotzsche PC. Is there logic in the placebo? *Lancet* 1994; **344**: 925–6.

13 Crow R, Gage H, Hampson S, Hart J, Kimber A, Thomas H. The role of expectancies in the placebo effect and their use in the delivery of health care: a systematic review. *Health Technol Assess* 1999; **3**(3). Available in full text on http://www.hta.nhsweb.nhs.uk/

14 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991: 187–248.

15 Epstein M, Parving HH, Ruilope LM. Surrogate endpoints and renal protection: focus on microalbuminuria. *Blood Pressure* 1997; **2** (suppl): 52–7.

16 Gotzsche P, Liberati A, Torri V, Rosetti L. Beware of surrogate outcome measures. *Int J Technol Assess Health Care* 1996; **12**: 238–46.

17 Lipkin M. Summary of recommendations for colonic biomarker studies of candidate chemopreventive compounds in phase II clinical trials. *J Cellular Biochem* 1994; **19** (suppl): 94-8.

18 Kimbrough RD. Determining acceptable risks: experimental and epidemiological issues. *Clin Chem* 1994; **40**: 1448–53.

19 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.

20 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.

21 Hughes MD, Daniels MJ, Fischl MA, Kim S, Schooley RT. CD4 cell count as a surrogate endpoint in HIV clinical trials: a meta-analysis of studies of the AIDS Clinical Trials Group. *AIDS* 1998; **12**:1823–32.

22 Epstein AE, Hallstrom AO, Rogers WJ *et al*. Mortality following ventricular arrhythmia suppression by encainide, flecainide and moricizine after myocardial infarction. *JAMA* 1993; **270**: 2451–5.

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 Pharmacokinet* 1995; **28**: 143–60.

25 Anonymous. Interferon beta-1b – hope or hype? *Drug Therapeut Bull* 1996; **34**: 9–11.

26 Carducci MA, DeWeese TL, Nelson JB. Prostate-specific antigen and other markers of therapeutic response. *Urologic Clin North Am* 2000; **26**: 291–302.

27 Schroder FH, Kranse R, Barbet N, Hop WC, Kandra A, Lassus M. Prostate-specific antigen: a surrogate endpoint for screening new agents against prostate cancer? *Prostate* 2000; **42**: 107–15.

28 See entire issue of *Journal of Cellular Biochemistry* 1994; **19** (suppl).
29 Aickin M. If there is gold in the labelling index hills, are we digging in the right place? *J Cellular Biochem* 1994; **19** (suppl): 91–3.
30 Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998; **54**: 1014–29.
31 Anonymous. Getting good value from drug reps. *Drug Therapeut Bull* 1983; **21**: 13–15.
32 Ferner RE. Newly licensed drugs. *BMJ* 1996; **313**: 1157–8.

Chapter 7: Papers that report diagnostic or screening tests

7.1 Ten men in the dock

If you are new to the concept of validating diagnostic tests and if algebraic explanations ("Let's call this value *x*...") leave you cold, the following example may help you. Ten men are awaiting trial for murder. Only three of them actually committed a murder; the other seven 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.

This information can be expressed in what is known as a 2 x 2 table (Table 7.1). Note that the "truth" (i.e. 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 title row.

Table 7.1 2 x 2 table showing outcome of trial for 10 men accused of murder

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

You should be able to see that these figures, if they are typical, reflect a number of 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 he is actually a murderer.
- If this jury found a person innocent, he has a three in four chance of actually being innocent.
- In five cases out of every 10 the jury gets the verdict 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 chapter considers these five features applied to diagnostic (or screening) tests when compared with a "true" diagnosis or gold standard. Section 7.4 also introduces a sixth, slightly more complicated (but very useful) feature of a diagnostic test – the likelihood ratio. (After you have read the rest of this chapter, look back at this section. By then, you should be able to work out that the likelihood ratio of a positive jury verdict in the above example is 1.17, and that of a negative one 0.78. If you can't, don't worry – many eminent clinicians have no idea what a likelihood ratio is.)

7.2 Validating diagnostic tests against a gold standard

Our window-cleaner told me the other day 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 general practitioner's surgery had asked him to produce a urine specimen and dipped a special stick in it. The stick stayed green, which meant, apparently, that there was no sugar (glucose) in his urine. This, the nurse had said, meant that he did not have diabetes.

I had trouble explaining to the window-cleaner that the test result did not necessarily mean this at all, 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 7 mmol/l in the fasting state, or above 11.1 mmol/l two hours after a 100 g oral glucose load (the much-dreaded "glucose tolerance test", where the subject has to glug down every last drop of a sickly glucose drink and wait two hours for a blood test). These values must be achieved on two separate occasions if the person has no symptoms, but on only one occasion if they have typical symptoms of diabetes (thirst, passing large amounts of urine, and so on).

These stringent criteria can be termed the *gold standard* for diagnosing diabetes. In other words, if you fulfil the WHO criteria you can call yourself diabetic and if you don't, you can't (although note that experts rightly challenge categorical statements such as this and indeed, since the first edition of this book was published the cutoff values in the gold standard test for diabetes using blood glucose levels have all changed[1]). The same cannot be said for dipping a stick into a random urine specimen. For one thing, you might be a true diabetic but have a high renal threshold; that is, your kidneys conserve glucose much better than most people's, so your blood glucose level would have to be much higher than most people's for any glucose to appear in your urine. Alternatively, you may be an otherwise normal individual with a *low* renal threshold, so glucose leaks into your urine even when there isn't any excess in your blood. In fact, as anyone with diabetes will tell you, diabetes is very often associated with a negative test for urine glucose.

There are, however, many advantages in using a urine dipstick rather than the full blown glucose tolerance test to "screen" people for diabetes. The test is cheap, convenient, easy to perform and interpret, acceptable to patients, and gives an instant yes/no result. In real life, people like my window-cleaner may decline to take an oral glucose tolerance test. Even if he was prepared to go ahead with it, his general practitioner might decide that the window-cleaner's symptoms did not merit the expense of this relatively sophisticated investigation. I hope you can see that even though the urine test cannot say for sure if someone is diabetic, it has a definite practical edge over the gold standard. That, of course, is why we use it!

Table 7.2 2 x 2 table notation for expressing the results of a validation study for a diagnostic or screening test

	Result of gold standard test	
Result of screening 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 7.3 Features of a diagnostic test which can be calculated by comparing it with a gold standard in a validation study

Feature of the test	Alternative name	Question which the feature addresses	Formula (see Table 7.2)
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 (s)he has the condition?	a/a+b
Negative predictive value	Indicates the post-test probability of a negative test*	If a person tests negative, what is the probability that (s)he does not have the condition?	d/c+d
Accuracy		What proportion of all tests have given the correct result (i.e. 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, as opposed to without, the condition?	Sensitivity/ (1- specificity)

* The post-test probability of a negative test is (1 - NPV)

In order to assess objectively just how useful the urine glucose 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. Such an exercise is known as a *validation study*. We could express the results of the validation study in a 2 x 2 table (also known as a 2 x 2 matrix) as in Table 7.2, and calculate various features of the test as in Table 7.3, just as we did for the features of the jury in section 7.1.

If the values for the various features of a test (such as sensitivity and specificity) fell within reasonable limits, we would be able to say that the test was *valid* (see question 7 below). The validity of urine testing for glucose in diagnosing diabetes has been looked at by Andersson and colleagues,[2] whose data I have used in the example in Table 7.4. In fact, the original study was performed on 3268 subjects, of whom 67 either refused to produce a specimen or, for some other reason, were not adequately tested. For simplicity's sake, I have ignored these irregularities and expressed the results in terms of a denominator (total number tested) of 1000 subjects.

Table 7.4 2 x 2 table showing results of validation study of urine glucose testing for diabetes against gold standard of glucose tolerance test (based on reference 2)

		Result of glucose tolerance test	
		Diabetes positive **27 subjects**	Diabetes negative **973 subjects**
Result of urine test for glucose	Glucose present **13 subjects**	True positive **6**	False positive **7**
	Glucose absent **987 subjects**	False negative **21**	True negative **966**

In actual fact, these data came from an epidemiological survey to detect the prevalence of diabetes in a population; the validation of urine testing was a side issue to the main study. If the validation had been the main aim of the study, the subjects selected would have included far more diabetic individuals, as question 2 in section 7.3 below will show. If you look up the original paper, you will also find that the gold standard for diagnosing true diabetes was not the oral glucose tolerance test but a more unconventional series of observations. Nevertheless, the example serves its purpose, since it provides us with some figures to put through the equations listed in the last column of Table 7.3. We can calculate the important features of the urine test for diabetes as follows.

- 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 = 966/987 = 97.9%
- Accuracy = (a+d)/(a+b+c+d) = 972/1000 = 97.2%
- Likelihood ratio of a positive test = sensitivity/(1 – specificity) = 22.2/0.7 = 32
- Likelihood ratio of a negative test = (1 – sensitivity)/specificity = 77.8/99.3 = 0.78

From these features, you can probably 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 true diabetics. In the presence of classic symptoms and a family history, the window-cleaner's baseline odds (pretest likelihood) of having the condition are pretty high and they are only reduced to about four-fifths of this (the likelihood ratio of a negative test, 0.78; see section 7.4) after a single negative urine test. In view of his symptoms, this man clearly needs to undergo a more definitive test for diabetes. Note that as the definitions in Table 7.3 show, if the test had been positive the window-cleaner would have good reason to be concerned, since even though the test is not very *sensitive* (i.e. it is not good at picking up people with the disease), it is pretty *specific* (i.e. it *is* good at excluding people without the disease).

Students often get mixed up about the sensitivity/specificity dimension of a test and the positive/negative predictive value dimension. As a rule of thumb, the sensitivity or specificity tells you about the *test in general*, whereas the predictive value tells you about *what a particular test result means for the patient in front of you*. Hence, sensitivity and specificity are generally used more by epidemiologists and public health specialists whose day to day work involves making decisions about *populations*.

A screening mammogram (breast X-ray) might have an 80% sensitivity and a 90% specificity for detecting breast cancer, which means that the test will pick up 80% of cancers and exclude 90% of women without cancer. But imagine you were a GP or practice nurse and a patient comes to see you for the result of her mammogram. The question she will want answered is (if the test has come back positive), "What is the chance that I've got cancer?" or (if it has come back negative) "What is the chance that I can now forget about the possibility of cancer?". Many patients (and far too many health professionals) assume that the negative predictive value of a test is 100%, i.e. if the test is "normal" or "clear" they think there is no chance of the disease being present – and you only need to read the confessional stories in women's magazines ("I was told I had cancer but tests later proved the doctors wrong") to find examples of women who have assumed that the positive predictive value of a test is 100%.

7.3 Ten questions to ask about a paper which claims to validate a diagnostic or screening test

In preparing the tips below, I have drawn on three main published sources: the "Users' guides to the medical literature"[3, 4] and the book by the same authors;[5] a more recent article in the *JAMA*,[6] and David Mant's simple and pragmatic guidelines for "testing a test".[7]

Question 1 Is this test potentially relevant to my practice?

This is the "so what?" question which Sackett and colleagues call the *utility* of the test.[5] 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? If the answers to these questions are all "no", you may be able to reject the paper without reading further than the abstract or introduction.

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

You need to ask, first, whether the test has been compared with anything at all! Papers have occasionally been written (and, in the past, published) in which nothing has been done except perform the new test on a few dozen subjects. This exercise may give a range of possible results for the test, but it certainly does not confirm that the "high" results indicate that the target disorder (the disease you are looking for) is present or that the "low" results indicate that it isn't.

Next, you should verify that the "gold standard" test used in the survey merits the term. A good way of assessing a gold standard is to use the "so what?" questions listed above. For many conditions, there is no absolute gold standard diagnostic test which will say for certain if it is present or not. Unsurprisingly, these tend to be the very 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 here (or a variant of it) is not being used to contribute to the definition of the gold standard.

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

If you validated a new test for cholesterol in 100 healthy male medical students, you would not be able to say how the test would perform in women, children, older people, those with diseases that seriously raise the cholesterol level or even those who had never been to medical school! Although few people would be naïve enough to select quite such a biased sample for their validation study, one paper found that only 27% of published studies explicitly defined the spectrum of subjects tested in terms of age, sex, symptoms and/or disease severity, and specific eligibility criteria .[6]

Defining both the range of participants and the spectrum of disease to be included is essential if the values for the different features of the test are to be worth quoting, i.e. if they are to be transferable to other settings. A particular diagnostic test may, conceivably, be more sensitive in female subjects than males or in younger rather than older subjects. For the same reasons, as Sackett and colleagues stipulate, the subjects on which any test is verified should include those with both mild and severe disease, treated and untreated, and those with different but commonly confused conditions.[5]

Whilst the sensitivity and specificity of a test are virtually constant whatever the prevalence of the condition, the positive and negative predictive values are crucially dependent on prevalence. This is why general practitioners are, often rightly, sceptical of the utility of tests developed exclusively in a secondary care population, where the severity of disease tends to be greater (see section 4.2), and why a good *diagnostic* test (generally used when the patient has some symptoms suggestive of the disease in question) is not necessarily a good *screening* test (generally used in people without symptoms, who are drawn from a population with a much lower prevalence of the disease).

Question 4 Has work up 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?". I hope you have no problem spotting the potential bias in studies where the gold standard test is only performed on people who have already tested positive for the test being validated. There are, in addition, a number of more subtle aspects of work up bias which are beyond the scope of this book. If you are interested, you could follow the discussion on this subject in Read and colleagues' paper.[6]

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 ECG. In the context of validating diagnostic tests against a gold standard, the question means "Did the people who interpreted one of the tests know what result the other test had shown on each particular subject?". As I explained in section 4.5, all assessments should be "blind" – that is, the person interpreting the test should not be given any inkling of what the result is expected to be in any particular case.

Question 6 Was the test shown to be reproducible both within and between observers?

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. All tests show this feature to some extent, but a test with a reproducibility of 99% is clearly in a different league from one with a reproducibility of 50%. A number of factors may contribute to the poor reproducibility of a diagnostic test: the technical precision of the equipment, observer variability (for example, in comparing a colour with a reference chart), arithmetical errors, and so on.

Look back again at section 4.5 to remind yourself of the problem of interobserver agreement. Given the same result to interpret, two people will agree in only a proportion of cases, generally expressed as the κ score. If the test in question gives results in terms of numbers (such as the blood cholesterol level in mmol/l), interobserver agreement is hardly an issue. If, however, the test involves reading X-rays (such as the mammogram example in section 4.5) or asking a person questions about their drinking habits,[8] it is important to confirm that reproducibility between observers is at an acceptable level.

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 test itself is not valid, i.e. its sensitivity, specificity, and other crucial features are too low. That is arguably the case for using urine glucose as a screening test for diabetes (see section 7.2 above). After all, if a test has a false negative rate of nearly 80%, it is more likely to mislead the clinician than assist the diagnosis if the target disorder is actually present.

There are no absolutes for the validity of a screening test, since 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 the UK soon after birth, is over 99% sensitive but has a positive predictive value of only 6% (in other words, it picks up almost all babies with the condition at the expense of a high false positive rate),[9] and rightly so. It is far more important to pick up every single baby with this treatable condition who would otherwise develop severe mental handicap than to save hundreds of parents the relatively minor stress of a repeat blood test on their baby.

Question 8 Were confidence intervals given for sensitivity, specificity, and other features of the test?

As section 5.5 explained, 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 lie. Go back to the jury example in section 7.1. If they had found just one more murderer not guilty, the sensitivity of their 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 only validated the jury's performance on 10 cases. The confidence intervals for the features of this jury are so wide that my computer programme refuses to calculate them! Remember, the larger the sample size, 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. If you would like the formula for calculating confidence intervals for diagnostic test features, see Gardner and Altman's textbook *Statistics with confidence*.[10]

Question 9 Has a sensible "normal range" been derived from these results?

If the test gives non-dichotomous (continuous) results – i.e. if it gives a numerical value rather than a yes/no result – someone will have to say at what value the test result will count as abnormal. Many of us have been there with our own blood pressure reading. We want to know if our result is "okay" or not, but the doctor insists on giving us a value such as "142/92". If 140/90 were chosen as the cutoff for high blood pressure, we would be placed in the "abnormal" category, even though our risk of problems from our blood pressure is very little different from that of a person with a blood pressure of 138/88. Quite sensibly, many practising doctors advise their patients, "Your blood pressure isn't quite right, but it doesn't fall into the danger zone. Come back in three months for another check". Nevertheless, the doctor must at some stage make the decision that *this* blood pressure needs treating with tablets but *that* one does not.

Defining relative and absolute danger zones for a continuous physiological or pathological variable 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 section 7.4). For an entertaining discussion on the different possible meanings of the word "normal" in diagnostic investigations, see Sackett and colleagues' textbook,[5] page 59.

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

In general, we treat high blood pressure simply on the basis of the blood pressure reading alone (although we tend to rely on a series of readings rather than a single value). Compare this with the sequence we use to diagnose stenosis ("hardening") of the coronary arteries. First, we select patients with a typical history of effort angina (chest pain on exercise). Next, we usually do a resting ECG, an exercise ECG, and, in some cases, a radionuclide scan of the heart to look for areas short of oxygen. Most patients only come to a coronary angiogram (the definitive investigation for coronary artery stenosis) *after* they have produced an abnormal result on these preliminary tests.

If you took 100 people off the street and sent them straight for a coronary angiogram, the test might display very different positive and negative predictive values (and even different sensitivity and specificity) than it did in the sicker 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.

7.4 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 PSA in their blood (say, 0.5 ng/ml) and most of those with advanced prostate cancer will have very high levels of PSA (above about 20 ng/ml). But a PSA level of, say, 7.4 ng/ml may be found in either a perfectly normal man or in someone with early cancer. There simply is not a clean cutoff between normal and abnormal.[11]

We can, however, use the results of a validation study of the PSA test against a gold standard for prostate cancer (say a biopsy) to draw up a whole series of 2 x 2 tables. Each table would use a different definition of an abnormal PSA result to classify patients as "normal" or "abnormal". From these tables, we could generate different likelihood ratios associated with a PSA level above each different cutoff point. Then, when faced with a PSA 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*". (In fact, as I mentioned in section 6.3, the PSA test is not a terribly good discriminator between the presence and absence of cancer, whatever cutoff value is used. In other words, there is no value for PSA that gives a particularly high likelihood ratio in cancer detection.)

Although the likelihood ratio is one of the more complicated aspects of a diagnostic test to calculate, it has enormous practical value and it is becoming the preferred way of expressing and comparing the usefulness of different tests. As Sackett and colleagues explain at great length in their textbook,[5] the likelihood ratio can be used directly in ruling a particular diagnosis in or out. 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 has this condition (in the language of diagnostic tests, this means that the pretest probability of anaemia, equivalent to the prevalence of the condition, is 0.05).[12]

Now, if I do a diagnostic test for anaemia, the serum ferritin level, the result will usually make the diagnosis of anaemia either more or less likely. A moderately reduced serum ferritin level (between 18 and 45 (µg/l) has a likelihood ratio of 3, so the chance of a patient with this result having iron deficiency anaemia is generally calculated as 0.05 x 3 or 0.15 (15%). This value is known as the post-test probability of the serum ferritin test. (Strictly speaking, likelihood ratios should be used on odds rather than on probabilities, but the simpler method shown here gives a good approximation when the pretest probability is low. In this example, a pretest probability of 5% is equal to a pre-test odds of 0.05/0.95

or 0.053; a positive test with a likelihood ratio of 3 gives a post-test odds of 0.158, which is equal to a post-test probability of 14%[12]).

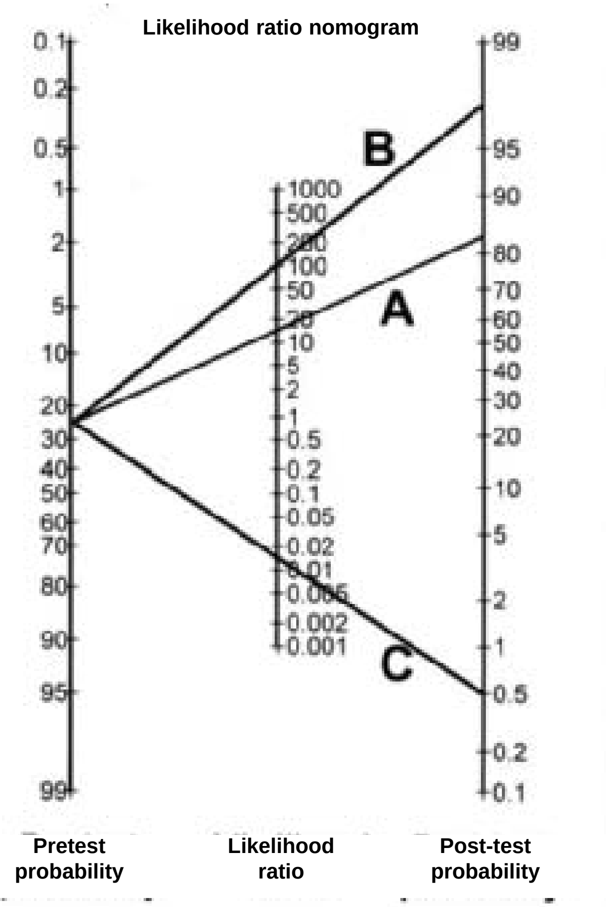


Figure 7.1 Using likelihood ratios to calculate the post-test probability of someone being a smoker

Figure 7.1 shows a nomogram, adapted by Sackett and colleagues from an original paper by Fagan,[13] 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 amongst British adults) are respectively the trajectories through likelihood ratios of 15, 100, and 0.015 – three different tests for detecting whether someone is a smoker.[14] 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%.

In summary, as I said at the beginning of this chapter, you can get a long way with diagnostic tests without referring to likelihood ratios. I avoided them myself for years. But if you put aside an afternoon to get to grips with this aspect of clinical epidemiology, I predict that your time will have been well spent.

1 Puavilai G, Chanprasertyotin S, Sriphrapradaeng A. Diagnostic criteria for diabetes mellitus and other categories of glucose intolerance: 1997 criteria by the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus (ADA), 1998 WHO consultation criteria, and 1985 WHO criteria. *Diabetes Res Clin Pract* 1999; **44**: 21–6.

2 Andersson DKG, Lundblad E, Svardsudd K. A model for early diagnosis of Type 2 diabetes mellitus in primary health care. *Diabetic Med* 1993; **10**: 167–73.

3 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.

4 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.

5 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology – a basic science for clinical medicine*. London: Little, Brown, 1991: 51–68.

6 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.

7 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.

8 Bush B, Shaw S, Cleary P *et al.* Screening for alcohol abuse using the CAGE questionnaire. *Am J Med* 1987; **82**: 231–6.

9 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.

10 Gardner MJ, Altman DG, eds. *Statistics with confidence: confidence intervals and statistical guidelines*. London: BMJ Publications, 1989.

11 Catalona WJ, Hudson MA, Scardino PT *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.

12 Guyatt GH, Patterson C, Ali M *et al.* Diagnosis of iron deficiency anaemia in the elderly. *Am J Med* 1990; **88**: 205–9.

13 Fagan TJ. Nomogram for Bayes' theorem. *New Engl J Med* 1975; **293**: 257–61.

14 Anonymous. How good is that test – using the result. *Bandolier* 1996; **3**: 6-8.

Chapter 8: Papers that summarise other papers (systematic reviews and meta-analyses)

8.1 When is a review systematic?

Remember the essays you used to write when you first started college? You would mooch round the library, browsing through the indexes of books and journals. When you came across a paragraph that looked relevant you copied it out and 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 (i.e. standardised and objective) way. Journalists get paid according to how much they write rather than how much they read or how critically they process it, which explains why most of the "new scientific breakthroughs" you read about in your newspaper today will probably be discredited before the month is out.

In contrast, a *systematic review* is an overview of primary studies which:

- contains a statement of objectives, materials, and methods
- has been conducted according to explicit and reproducible methodology (see Figure 8.1).[1]

The most enduring and useful systematic reviews, notably those undertaken by the Cochrane Collaboration (see section 2.11), are regularly updated to incorporate new evidence.

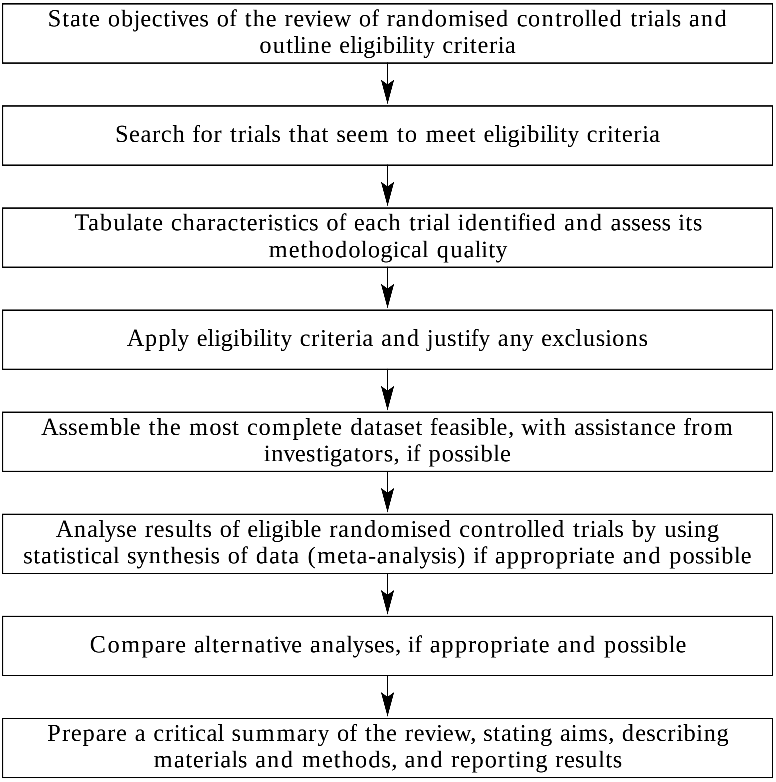

State objectives of the review of randomised controlled trials and outline eligibility criteria
Search for trials that seem to meet eligibility criteria
Tabulate characteristics of each trial identified and assess its methodological quality
Apply eligibility criteria and justify any exclusions
Assemble the most complete dataset feasible, with assistance from investigators, if possible
Analyse results of eligible randomised controlled trials by using statistical synthesis of data (meta-analysis) if appropriate and possible
Compare alternative analyses, if appropriate and possible
Prepare a critical summary of the review, stating aims, describing materials and methods, and reporting results

Figure 8.1 Method for a systematic review of randomised controlled trials

Many, if not most, medical review articles are still written in journalistic form. Professor Paul Knipschild, in Iain Chalmers' and Doug Altman's excellent book *Systematic reviews*,[2] describes 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] 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.

Linus Pauling probably did not deliberately intend to deceive his readers, but since his enthusiasm for his espoused cause outweighed his scientific objectivity, he was unaware of the *selection bias* influencing his choice of papers. Much work has been done, most notably by Professor Cynthia Mulrow of the University of Texas Health Science Center, USA, which confirms the sneaky feeling that were you or I to attempt what Pauling did – i.e. hunt through the medical literature for "evidence" to support our pet theory – we would make an equally idiosyncratic and unscientific job of it.[4, 5] Mulrow, along with Iain Chalmers at the UK Cochrane Centre and Peter Gøtzsche and Andy Oxman of the Nordic Cochrane Centre (see section 2.11), deserves much of the credit for persuading the rest of the medical community that flawed secondary research, exemplified by the journalistic review, is as scientifically dangerous as flawed primary research. Some advantages of the systematic review are given in Box 8.1.

Box 8.1 Advantages of systematic reviews[3]

- Explicit methods *limit bias* in identifying and rejecting studies
- Conclusions are hence more *reliable* and *accurate*
- Large amounts of *information* can be assimilated quickly by health care providers, researchers, and policymakers
- Delay between research discoveries and *implementation* of effective diagnostic and therapeutic strategies is potentially reduced (see Chapter 12)
- Results of different studies can be formally compared to establish *generalisability* of findings and *consistency* (lack of heterogeneity) of results (see section 8.4)
- Reasons for *heterogeneity* (inconsistency in results across studies) can be identified and new hypotheses generated about particular subgroups (see section 8.4)
- Quantitative systematic reviews (metaanalyses) increase the *precision* of the overall result (see sections 4.6 and 8.3)

Experts, who have been steeped in a subject for years and know what the answer "ought" to be, were once shown to be significantly less able to produce an objective review of the literature in their subject than non-experts.[6] This would have been of little consequence if experts' opinions could be relied upon to be congruent with the results of independent systematic reviews, but at the time they most certainly couldn't.[7] These condemning studies are still widely quoted by people who would replace all subject experts (such as cardiologists) with search and appraisal experts (people who specialise in finding and criticising papers on any subject). But no one in more recent years has replicated the findings; in other words, perhaps we should credit today's experts with more of a tendency to base their recommendations on a thorough assessment of the evidence! As a general rule, however, if you are going to pay someone to seek out the best objective evidence of the benefits of anticoagulants in atrial fibrillation, you should ask someone who is an expert in systematic reviews to work alongside an expert in atrial fibrillation.

To be fair to Linus Pauling, he did mention a number of trials whose results seriously challenged his theory that vitamin C prevents the common cold,[3] but he described all such trials as "methodologically flawed". As Knipschild reminds us, so were many of the trials which Pauling *did* include in his analysis but because their results were consistent with his theory, Pauling was, perhaps subconsciously, less critical of weaknesses in their design.

I mention this to illustrate the point that, when undertaking a systematic review, 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. In other words, you don't trash a trial because all other trials in this area showed something different (see section 8.4); you trash it because, *whatever the results showed*, the trial's objectives or methods did not meet your inclusion criteria (see section 3.1).

8.2 Evaluating systematic reviews

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

Look back to Chapter 3, in which I explained the importance of defining the question when reading a paper about a clinical trial or other form of primary research. I called this "getting your bearings" since one sure way to be confused about a paper is to fail to ascertain what it is about. The definition of a specific answerable question is, if anything, even more important (and even more frequently omitted!) when preparing an overview of primary studies. If you have ever tried to pull together the findings of a dozen or more clinical papers into an essay, editorial or summary notes for an examination, you will know that it is all too easy to meander into aspects of the subject which you never intended to cover.

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". The question "Do anticoagulants prevent strokes in patients with atrial fibrillation?" sounds pretty specific, until you start looking through the list of possible studies to include. Does "atrial fibrillation" include both rheumatic and non-rheumatic forms (which are known to be associated with very different risks of stroke) and does it include intermittent atrial fibrillation (my grandfather, for example, used to go into this arrhythmia for a few hours whenever he drank coffee and would have counted as a "grey case" in any trial)?

Does "stroke" include both ischaemic stroke (caused by a *blocked* blood vessel in the brain) and haemorrhagic stroke (caused by a *burst* blood vessel)? And, talking of burst blood vessels, shouldn't we be weighing the side effects of anticoagulants against their possible benefits? Should true anticoagulants such as heparin and warfarin be compared with placebo or with other drugs that reduce the clotting tendency of the blood, such as aspirin and related products? Finally, should the review cover trials on patients who have already had a previous stroke or transient ischaemic attack (a mild stroke which gets better within 24 hours) or should it be limited to trials on patients without these major risk factors for a further stroke? The "simple" question posed earlier is becoming unanswerable, and we must refine it as follows.

> "To assess the effectiveness and safety of warfarin-type anticoagulant therapy in secondary prevention (i.e. following a previous stroke or transient ischaemic attack) in patients with non-rheumatic atrial fibrillation: comparison with placebo."[8]

Question 2 Was a thorough search done of the appropriate database(s) and were other potentially important sources explored?

As Figure 8.1 illustrates, one of the benefits of a systematic review is that, unlike a narrative or journalistic review, the author is required to tell you where the information in it came from and how it was processed. As I explained in Chapter 2, searching the

Medline database for relevant articles is a very sophisticated science and even the best Medline search will miss important papers, for which the reviewer must approach the other databases listed in section 2.10.

In the search for trials to include in a review, the scrupulous avoidance of linguistic imperialism is a scientific as well as a political imperative. As much weight must be given, for example, to the expressions "Eine Placebo-kontrolierte Doppel-blindstudie" and "une étude randomisée a double insu face au placebo" as to "a double blind, randomised controlled trial"![9] Furthermore, particularly where a statistical synthesis of results (metaanalysis) is contemplated, it may be necessary to write and ask the authors of the primary studies for raw data on individual patients which were never included in the published review (see section 8.3).

Even when all this has been done, the systematic reviewer's search for material has hardly begun. As Paul Knipschild and his colleagues showed when they searched for trials on vitamin C and cold prevention, their electronic databases only gave them 22 of their final total of 61 trials. Another 39 trials were uncovered by handsearching the manual *Index Medicus* database (14 trials not identified previously), searching the references of the trials identified in Medline (15 more trials), the references of the references (nine further trials), and the references of the references of the references (one additional trial not identified by any of the previous searches).

Box 8.2 Checklist of data sources for a systematic review

- Medline database
- Cochrane controlled clinical trials register (see section 2.11)
- Other medical and paramedical databases (see section 2.10)
- Foreign language literature
- "Grey literature" (theses, internal reports, non-peer reviewed journals, pharmaceutical industry files)
- References (and references of references, etc.) cited in primary sources
- Other unpublished sources known to experts in the specialty (seek by personal communication)
- Raw data from published trials (seek by personal communication)

Do not be too hard on a reviewer, however, if he or she has not followed this counsel of perfection to the letter. After all, Knipschild and his team found that only one of the trials not identified in Medline met stringent criteria for methodological quality and ultimately contributed to their systematic review of vitamin C in cold prevention.[9] An exploration of "grey literature" (see Box 8.2) may be of greater relative importance when looking at trials outside the medical mainstream such as physiotherapy or alternative medicine.[10]

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

Chapters 3 and 4 and Appendix 1 of this book provide some checklists for assessing whether a paper should be rejected outright on methodological grounds. But given that only around 1% of clinical trials are said to be beyond criticism in terms of methodology, the practical question is how to ensure that a "small but perfectly formed" study is given the weight it deserves in relation to a larger study whose methods are adequate but more open to criticism.

Box 8.3 Assigning weight to trials in a systematic review

Each trial should be evaluated in terms of its

- *Methodological quality*—that is, extent to which the design and conduct are likely to have prevented systematic errors (bias) (see section 4.4)
- *Precision*—that is, a measure of the likelihood of random errors (usually depicted as the width of the confidence interval around the result)
- *External validity*—that is, the extent to which the results are generalisable or applicable to a particular target population

(Additional aspects of "quality" such as scientific importance, clinical importance, and literary quality, are rightly given great weight by peer reviewers and journal editors but are less relevant to the systematic reviewer once the question to be examined has been defined)

Methodological shortcomings which invalidate the results of trials are often generic (i.e. they are independent of the subject matter of the study; see Appendix 1), but there may also be particular methodological features which distinguish between good, medium, and poor quality in a particular field. Hence, one of the tasks of a systematic reviewer is to draw up a list of criteria, including both generic and particular aspects of quality, against which to judge each trial. In theory, a composite numerical score could be calculated which would reflect "overall methodological quality". In reality, however, care should be taken in developing such scores since there is no gold standard for the "true" methodological quality of a trial[11] and such composite scores are probably neither valid nor reliable in practice.[12, 13] The various Cochrane Collaborative review groups are in the process of developing both general and topic specific methods for assigning quality scores to research studies.[14–16] Currently, less than half of all published meta-analyses contain reproducible criteria for assessing the quality of the trials that were included and excluded.[17]

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

If you don't understand what this question means, look up the tongue in cheek paper by Carl Counsell and colleagues in the Christmas 1994 issue of the *BMJ*, which "proved" an entirely spurious relationship between the result of shaking a dice and the outcome of an acute stroke.[18] The authors report a series of artificial dice rolling experiments in which red, white, and green dice respectively 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 metaanalysis – such as the exclusion of several of the "negative" trials through publication bias (see section 3.3), 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.

You cannot, of course, cure anyone of a stroke by rolling a dice, but if these simulated results pertained to a genuine medical controversy (such as which groups of postmenopausal women should take hormone replacement therapy or whether breech babies should routinely be delivered by caesarean section), how would you spot these subtle biases? The answer is that 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 unaccounted for patients in a trial were assumed to have died (or been cured)?

An exploration of what ifs is known as a *sensitivity analysis*. If you find that fiddling with the data like this 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 change, 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?

As the next section shows, it is easy to be "phased" by the figures and graphs in a systematic review. But 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 of included trials (see Box 8.3). A trial may be of high methodological quality and have a precise and numerically impressive result but it may, for example, have been conducted on participants under the age of 60 and hence may not be valid for people over 75. The inclusion in systematic reviews of irrelevant studies is guaranteed to lead to absurdities and reduce the credibility of secondary research, as Professor Sir John Grimley Evans argued (see section 9.1).[19]

8.3 Metaanalysis for the non-statistician

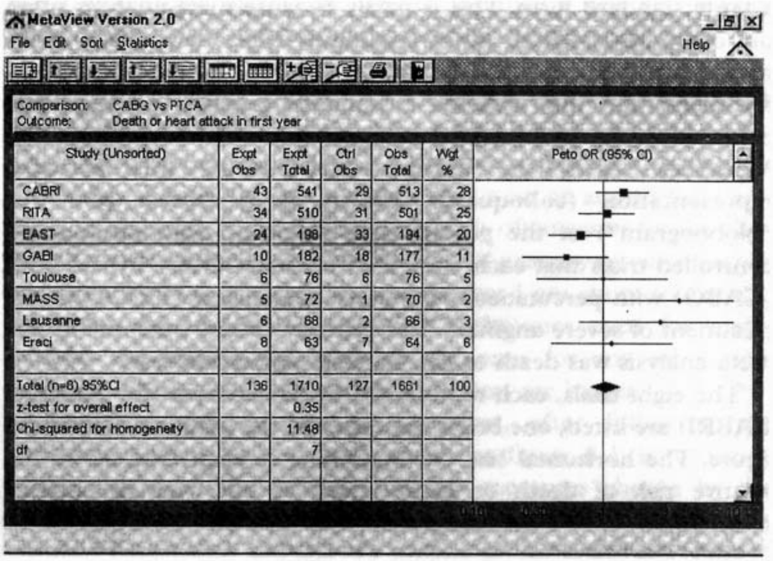
If I had to pick one word which exemplifies the fear and loathing felt by so many students, clinicians, and consumers towards evidence based medicine, that word would be "metaanalysis". The metaanalysis, defined as a *statistical synthesis of the numerical results of several trials which all addressed the same question*, is the statisticians' chance to pull a double whammy on you. First, they phase you with all the statistical tests in the individual papers and then they use a whole new battery of tests to produce a new set of odds ratios, confidence intervals, and values for significance.

As I confessed in Chapter 5, I too tend to go into panic mode at the sight of ratios, square root signs, and half-forgotten Greek letters. But before you consign metaanalysis to the set of newfangled techniques which you will never understand, remember two things. First, the metaanalyst may wear an anorak but he or she is *on your side*. A good metaanalysis is often easier for the non-statistician to understand than the stack of primary research papers from which it was derived, for reasons which I am about to explain. Second, the underlying statistical techniques used for metaanalysis are exactly the same as the ones for any other data analysis – it's just that some of the numbers are bigger. Helpfully, an international advisory group have come up with a standard format for meta-analyses (the QUOROM statement,[20] analogous to the CONSORT format for randomised controlled trials I mentioned in Chapter 4).

The first task of the metaanalyst, after following the preliminary steps for systematic review in Figure 8.1, is to decide which out of all the various outcome measures chosen by the authors of the primary studies is the best one (or ones) to use in the overall synthesis. In trials of a particular chemotherapy regimen for breast cancer, for example, some authors will have published cumulative mortality figures (i.e. the total number of people who have died to date) at cutoff points of three and 12 months, whereas other trials will have published six month, 12 month, and five year cumulative mortality. The metaanalyst might decide to concentrate on 12 month mortality because this result can be easily extracted from all the papers. He or she may, however, decide that three month mortality is a clinically important endpoint and would need to write to the authors of the remaining trials asking for the raw data from which to calculate these figures.

In addition to crunching the numbers, part of the metaanalyst's job description is to tabulate relevant information on the inclusion criteria, sample size, baseline patient characteristics, withdrawal ("dropout") rate, and results of primary and secondary endpoints of all the studies included. If this task has been done properly, you will be able to compare both the methods and the results of two trials whose authors wrote up their research in different ways. 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. This is partly because metaanalysts often use computer software to do the calculations for them,[21] and this software includes a standard graphics package which presents results as illustrated in Figure 8.2. I have reproduced in the format of one commonly used software package (with the authors' permission) this pictorial representation (colloquially known as a "forest plot" or "blobbogram") of the pooled odds ratios of eight RCTs which each compared coronary artery bypass graft (CABG) with percutaneous coronary angioplasty (PTCA) in the treatment of severe angina.[22] The primary (main) outcome in this meta-analysis was death or heart attack within one year.



Comparison: CABG vs PTCA
Outcome: Death or heart attack in first year

Study (Unsorted)	Expt Obs	Expt Total	Ctrl Obs	Obs Total	Wgt %	Peto OR (95% CI)
CABRI	43	541	29	513	28	
RITA	34	510	31	501	25	
EAST	24	198	33	194	20	
GABI	10	182	18	177	11	
Toulouse	6	76	6	76	5	
MASS	5	72	1	70	2	
Lausanne	6	68	2	66	3	
Eraci	8	63	7	64	6	
Total (n=8) 95%CI	136	1710	127	1661	100	
z-test for overall effect		0.35				
Chi-squared for homogeneity		11.48				
df		7				

Figure 8.2 Pooled odds ratios of eight randomised controlled trials of coronary artery bypass graft against percutaneous coronary angioplasty shown in MetaView format

The eight trials, each represented by its acronym (for example, CABRI), are listed one below the other on the left hand side of the figure. The horizontal line corresponding to each trial shows the relative risk of death or heart attack at one year in patients randomised to PTCA compared to patients randomised to CABG. 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 CABG rather than PTCA) and the width of the line represents the 95% confidence interval of this estimate (see section 5.5). 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 (RR) of 1.0. In other words, if the horizontal line for any trial does not cross the line of no effect, there is a 95% chance that there is a "real" difference between the groups.

As sections 4.6 and 5.5 argued, if the confidence interval of the result (the horizontal line) *does* cross the line of no effect (i.e. the vertical line), that can mean *either* that there is no significant difference between the treatments *and/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 PTCA compared to CABG of between about 0.5 and 5.0, and the confidence intervals of some studies are so wide that they don't even fit on the graph.

Now, here comes the fun of metaanalysis. Look at the tiny diamond below all the horizontal lines. This represents the *pooled* data from all eight trials (overall relative risk PTCA:CABG = 1.08), with a new, much narrower, confidence interval of this relative risk (0.79–1.50). 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 endpoint (death or heart attack in the first year). Now, in this example, every single 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 a PTCA rather than a CABG to every patient with angina. It has a much more limited meaning – that the *average* patient in the trials presented in this metaanalysis is equally likely to have met the primary outcome (death or heart attack

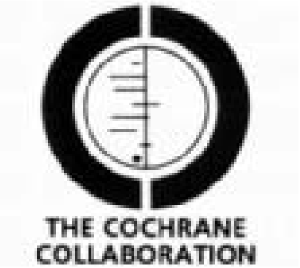


Figure 8.3 Cochrane Collaboration Logo

within a year) whichever of these two treatments they were randomised to receive. If you read the paper by Pocock and colleagues,[22] you would find important differences in the groups in terms of prevalence of angina and requirement for further operative intervention after the initial procedure. The choice of treatment should also, of course, take into account how the patient feels about undergoing major heart surgery (CABG) as opposed to the relatively minor procedure of PTCA.

In many meta-analyses, “non-significant” trials (i.e. ones which, on their own, did not demonstrate a significant difference between treatment and control groups) contribute to a pooled result which *is* statistically significant. The most famous example of this, which the Cochrane Collaboration adopted as its logo (Figure 8.3), is the metaanalysis of seven trials of the effect of giving steroids to mothers who were expected to give birth prematurely. Only two of the seven trials showed a statistically significant benefit (in terms of survival of the infant) but the improvement in precision (i.e. the narrowing of confidence intervals) in the pooled results, shown by the narrower width of the diamond compared with the individual lines, demonstrates the strength of the evidence in favour of this intervention. This metaanalysis showed that infants of steroid treated mothers were 30–50% less likely to die than infants of control mothers. This example is discussed further in section 12.1 in relation to changing clinicians’ behaviour.

If you have followed the arguments on metaanalysis of published trial results this far, you might like to read up on the more sophisticated technique of metaanalysis of individual patient data, which provides a more accurate and precise figure for the point estimate of effect.[23] You might also like to seek out the excellent review series on metaanalysis published in the *BMJ* a few years ago,[24–29] together with a special supplement to that series on the different software packages now available for metaanalysis, which was only published on the Web.[21]

8.4 Explaining heterogeneity

In everyday language, "homogeneous" means "of uniform composition", and "heterogeneous" means "many different ingredients". In the language of metaanalysis, homogeneity means that the results of each individual trial are 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 8.2 and 8.4. In Figure 8.2, the lower confidence interval of every trial is below the upper confidence interval of all the others (i.e. the horizontal lines all overlap to some extent). Statistically speaking, the trials are homogeneous. Conversely, in Figure 8.4, there are some trials whose lower confidence interval is above the upper confidence interval of one or more other trials (i.e. some lines do not overlap at all). These trials may be said to be heterogeneous.

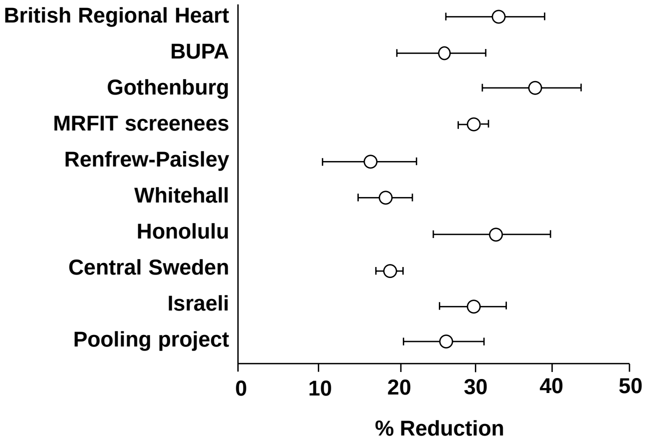


Figure 8.4 Reduction in risk of heart disease by strategies to lower cholesterol concentration[30]

You may have spotted by now (particularly if you have already read section 5.5 on confidence intervals) that pronouncing a set of trials heterogeneous on the basis of whether their confidence intervals overlap is somewhat arbitrary, since the confidence interval itself is arbitrary (it can be set at 90%, 95%, 99% or indeed any other value). The definitive test involves a slightly more sophisticated statistical manoeuvre than holding a ruler up against the blobbogram. The one most commonly used is a variant of the chi square (χ^2) test (see Table 5.1), since the question addressed is, "Is there greater variation between the results of the trials than is compatible with the play of chance?".

The χ^2 statistic for heterogeneity is explained in more detail by Simon Thompson,[30] who offers the following useful 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 metaanalysis minus one), so a χ^2 of 7.0 for a set of eight trials would provide no evidence of statistical heterogeneity. (In fact, it would not prove that the trials were homogeneous either, particularly since the χ^2 test has low power [see section 4.6] to detect small but important levels of heterogeneity.)

A χ^2 value much greater than the number of trials in a meta-analysis tells us that the trials which contributed to the analysis are different in some important way from one another. There may, for example, be known differences in methodology (for example, authors may have used different questionnaires to assess the symptoms of depression) or known clinical differences in the trial participants (for example, one centre might have been a tertiary referral hospital to which all the sickest patients were referred). There may, however, be unknown or unrecorded differences between the trials which the metaanalyst can only speculate upon until he or she has extracted further details from the trials' authors. Remember: demonstrating statistical heterogeneity is a mathematical exercise and is the job of the statistician but explaining this heterogeneity (i.e. looking for, and accounting for, *clinical* heterogeneity) is an interpretive exercise and requires imagination, common sense, and hands on clinical or research experience.

Figure 8.4, which is reproduced with permission from Simon Thompson's chapter on the subject,[30] shows the results of 10 trials of cholesterol lowering strategies. The results are expressed as the percentage reduction in heart disease risk associated with each 0.6 mmol/l reduction in serum cholesterol level. The horizontal lines represent the 95% confidence intervals of each result and it is clear, even without being told the χ^2 statistic of 127, that the trials are highly heterogeneous.

To simply "average out" the results of the trials in Figure 8.4 would be very misleading. The metaanalyst must return to his or her primary sources and ask, "In what way was trial A different from trial B, and what do trials C, D and H have in common which makes their results cluster at one extreme of the figure?". In this example, a correction for the age of the trial subjects reduced χ^2 from 127 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 (such as a special diet) 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.

This, essentially, is the basis of the grievance of Professor Hans Eysenck, who has constructed a vigorous and entertaining critique of the science of metaanalysis.[31] In a world of lumpers and splitters, Eysenck is a splitter and it offends his sense of the qualitative and the particular (see Chapter 11) to combine the results of studies which were done on different populations in different places at different times and for different reasons.

Eysenck's reservations about metaanalysis are borne out in the infamously discredited metaanalysis which demonstrated (wrongly) that there was significant benefit to be had from giving intravenous magnesium to heart attack victims. A subsequent megatrial involving 58000 patients (ISIS-4) failed to find any benefit whatsoever and the metaanalysts' misleading conclusions were subsequently explained in terms of publication bias, methodological weaknesses in the smaller trials, and clinical heterogeneity.[32, 33] (Incidentally, for more debate on the pros and cons of metaanalysis versus megatrials, see LeLorier and colleagues' *Lancet* article.[34])

Although Eysenck's mathematical naïveté is embarrassing ("If a medical treatment has an effect so recondite and obscure as to require a metaanalysis to establish it, I would not be happy to have it used on me"), I have a great deal of sympathy for the body of his argument. As one who tends to side with the splitters, I would put Eysenck's misgivings about metaanalysis high on the list of required reading for the aspiring systematic reviewer. Indeed, I recently threw my own hat into the ring when Simon Griffin published a metaanalysis of primary studies into the management of diabetes by primary health care teams.[35] Although I have a high regard for Simon as a scientist, I felt strongly that he had not been justified in performing a mathematical summation of what I believed were very different studies all addressing slightly different questions. As I said in my commentary on his article, "Four apples and five oranges makes four apples and five oranges, not nine appleoranges".[36] But Simon numbers himself among the lumpers and there are plenty of people cleverer than me who have argued that he was entirely correct to analyse his data as he did. Fortunately, the two of us have agreed to differ – and on a personal level we remain friends.

For an authoritative review of the technicalities of integrating heterogeneous pieces of evidence into systematic reviews, see the article by Cindy Mulrow and colleagues.[37]

1 Reproduced from The Cochrane Centre brochure, UK Cochrane Centre, Summertown Pavilion, Middle Way, Oxford OX2 7LG, UK.

2 Chalmers I, Altman DG, eds. *Systematic reviews*. London: BMJ Publications, 1995.

3 Pauling L. *How to live longer and feel better*. New York: Freeman, 1986.

4 Mulrow C. The medical review article: state of the science. *Ann Intern Med* 1987; **106**: 485–8.

5 Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997; **126**: 376–80.

6 Oxman AD, Guyatt GH. The science of reviewing research. *Ann NY Acad Sci* 1993; **703**: 125–31.

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*, updated 14 February 1995. Oxford: The Cochrane Library, Issue 2, 2000.

9 Knipschild P. Some examples of systematic reviews. In: Chalmers I, Altman DG, eds. *Systematic reviews*. London: BMJ Publications, 1995: 9–16.

10 Knipschild P. Searching for alternatives: loser pays. *Lancet* 1993; **341**: 1135–6.

11 Oxman A, ed. Preparing and maintaining systematic reviews. In: *Cochrane Collaboration Handbook*, section VI. Oxford: Update Software, 2000.

12 Emerson JD, Burdick E, Hoaglin DC *et al.* 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.

13 Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. *Int J Technol Assess Health Care* 1996; **12**: 195–208.

14 Bero L, Rennie D. The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *JAMA* 1995: **274**: 1935–8.

15 Garner P, Hetherington J. Establishing and supporting collaborative review groups. In: *Cochrane Collaboration Handbook*, section II. Oxford: Update Software, 2000.

16 Verhagen AP, de Vet HC, de Bie RA *et al*. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. *J Clin Epidemiol* 1998; **51**: 1235–41.

17 Moher D, Cook DJ, Jadad AR *et al*. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. *Health Technol Assess* 1999; **3**(12). Available in full text on http://www.hta.nhsweb.nhs.uk/.

18 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.

19 Grimley Evans J. evidence based and evidence-biased medicine. *Age Ageing* 1995; **24**: 461–3.

20 Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, for the QUOROM Group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. *Lancet* 1999; **354**: 1896-900. Available in full text on http://www.thelancet.com/newlancet/eprint/2/index.html

21 Egger M, Davey Smith G. Meta-analysis software. Electronic publication by BMJ to supplement series on meta-analysis (see references 24–29 below). http://bmj.com/archive/7126/7126ed9.htm

22 Pocock SJ, Henderson RA, Rickards AF *et al*. Meta-analysis of randomised trials comparing coronary angioplasty with bypass surgery. *Lancet* 1995; **346**: 1184–9.

23 Clarke MJ, Stewart LA. Obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analyses? In: Chalmers I, Altman DG, eds. *Systematic reviews*. London: BMJ Publications, 1995: 37–47.

24 Egger M, Davey Smith G. Meta-analysis: potentials and promise. *BMJ* 1997; **315**: 1371–4.

25 Davey Smith G, Egger M, Phillips A. Meta-analysis: principles and procedures. *BMJ* 1997; **315**: 1533–7.

26 Davey Smith G, Egger M, Phillips A. Meta-analysis: beyond the grand mean? *BMJ* 1997; **315**: 1610–14.

27 Egger M, Davey Smith G. Meta-analysis: bias in location and selection of studies. *BMJ* 1998; **316**: 61–6.

28 Egger M, Schneider M, Davey Smith G. Meta-analysis: spurious precision? Meta-analysis of observational studies. *BMJ* 1998; **316**: 140–4.

29 Davey Smith G, Egger M. Meta-analysis: unresolved issues and future developments. *BMJ* 1998; **316**: 221–5.

30 Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. In: Chalmers I, Altman DG, eds. *Systematic reviews*. London: BMJ Publications, 1995: 48–63.

31 Eysenck HJ. Problems with meta-analysis. In: Chalmers I, Altman DG, eds. *Systematic reviews*. London: BMJ Publications, 1995: 64–74.

32 Anon. Magnesium, myocardial infarction, meta-analysis and mega-trials. *Drug Therapeut Bull* 1995; **33**: 25–7.

33 Egger M, Davey Smith G. Misleading meta-analysis: lessons from "an effective, safe, simple" intervention that wasn't. *BMJ* 1995; **310**: 752–4.

34 LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analysis and subsequent large randomised controlled trials. *New*

Engl J Med 1997; **337**: 536–42.
35 Griffin S. Diabetes care in general practice: meta-analysis of randomised controlled trials. *BMJ* 1998; **317**: 390–5.
36 Greenhalgh T. Meta-analysis is a blunt and potentially misleading instrument for analysing methods of service delivery. *BMJ* 1998; **317**: 395–6.
37 Mulrow C, Langhorne P, Grimshaw J. Integrating heterogeneous pieces of evidence in systematic reviews. *Ann Intern Med* 1997; **127**: 989–95.

Chapter 9: Papers that tell you what to do (guidelines)

9.1 The great guidelines debate

Never was the chasm between old fashioned clinicians and old style health service managers wider than in their respective attitudes to clinical guidelines. Managers (in which I include politicians and all those who implement, administer, evaluate, and finance the actions of clinicians but who do not themselves see patients) tend to love guidelines. Clinicians, save for the important minority who actually write them, usually have a strong aversion to guidelines.

Before we carry this political hot potato any further, we need a definition of guidelines, for which the following will suffice.

Box 9.1 Purpose of guidelines

- To make evidence based standards explicit and accessible (but see below – few guidelines currently in circulation are truly evidence based)
- To make decision making in the clinic and at the bedside easier and more objective
- To provide a yardstick for assessing professional performance
- To delineate the division of labour (for example, between GPs and consultants)
- To educate patients and professionals about current best practice
- To improve the cost effectiveness of health services
- To serve as a tool for external control

> "Guidelines are systematically developed statements to assist practitioner decisions about appropriate health care for specific clinical circumstances."[1]

The purposes which guidelines serve are given in Box 9.1. The image of the medical buffoon blundering blithely through the outpatient clinic still diagnosing the same illnesses and prescribing the same drugs he (or she) learnt about at medical school 40 years previously, and never having read a paper since, knocks the "clinical freedom" argument (i.e. that a doctor's professional opinion is beyond reproach) right out of the arena. Such hypothetical situations are grist to the mill of those who would impose "expert guidelines" on most if not all medical practice and hold to account all those who fail to keep in step.

Box 9.2 Drawbacks of guidelines (real and perceived)

- Guidelines may be intellectually suspect and reflect "expert opinion", which may formalise unsound practice
- By reducing medical practice variation they may standardise to "average" rather than best practice
- They inhibit innovation and prevent individual cases from being dealt with discretely and sensitively
- They could, theoretically, be used medicolegally (both in and out of context) to dictate what a competent practitioner "would have done" in particular circumstances
- Guidelines developed at national or regional level may not reflect local needs or have the ownership of local practitioners
- Guidelines developed in secondary care may not reflect demographic, clinical or practical differences between this sector and the primary care setting
- Guidelines may produce undesirable shifts in the balance of power between different professional groups (for example, between clinicians and academics or purchasers and providers); hence, guideline development may be perceived as a political act

But the counter argument to the excessive use, and particularly the compulsive imposition, of clinical guidelines is a powerful one and it has been expressed very eloquently by Professor John Grimley Evans.

> "There is a fear that in the absence of evidence clearly applicable to the case in hand a clinician might be forced by guidelines to make use of evidence which is only doubtfully relevant, generated perhaps in a different grouping of patients in another country at some other time and using a similar but not identical treatment. This is evidence-*biased* medicine; it is to use evidence in the manner of the fabled drunkard who searched under the street lamp for his door key because that is where the light was, even though he had dropped the key somewhere else."[2]

Grimley Evans' fear, which every practising clinician shares but few can articulate, is that politicians and health service managers who have jumped on the evidence based medicine bandwagon will use guidelines to decree the treatment of diseases rather than of patients. They will, it is feared, make judgements about people and their illnesses subservient to published evidence that an intervention is effective "on average". This, and other real and perceived disadvantages of guidelines, are given in Box 9.2, which has been compiled from a number of sources.[3–10]

The mushrooming guidelines industry owes its success at least in part to a growing "accountability culture" that is now (many argue) being set in statute in many countries. In the UK National Health Service, all doctors, nurses, pharmacists, and other health professionals now have a contractual duty to provide clinical care based on best available research evidence.[11] Officially produced or sanctioned guidelines are a way of both supporting and policing that laudable goal. Whilst the medicolegal implications of "official"guidelines have rarely been tested in the UK,[12] US courts have ruled that guideline developers can be held liable for faulty guidelines and that doctors cannot pass off their liability for poor clinical performance by claiming that adherence to guidelines corrupted their judgement.[4]

9.2 Do guidelines change clinicians' behaviour?

An early systematic review of randomised trials and "other robust designs" by Grimshaw and Russell[13] demonstrated that, *in the research setting* (in which participants were probably highly selected and evaluation was an explicit part of guideline introduction), all but four of 59 published studies demonstrated improvements – i.e. changes in line with the guideline recommendations – in the process of care (i.e. what doctors did), and all but two of the 11 studies which measured outcome (i.e. what happened to the patients) reported significant improvements in the group using guidelines compared to the "usual care" group. Grimshaw subsequently set up a special subgroup of the Cochrane Collaboration (see section 2.11) to review and summarise emerging research on the use of guidelines and other related issues in improving professional practice. You can find details of the Effective Practice and Organisation of Care (EPOC) Group on the Cochrane website.[14]

EPOC and other groups who research the effectiveness of guidelines would probably be the first to emphasise that, despite the broadly positive findings of the research studies, guidelines do not *necessarily* improve either performance or outcome. Both Grimshaw and Russell[13] and others[15, 16] found wide variation in the size of the improvements in performance achieved by clinical guidelines. The former authors concluded that the probability of a guideline being effective depended on three factors which are summarised in Table 9.1: the development strategy (where and how the guidelines were produced), the dissemination strategy (how they were brought to the attention of clinicians), and the implementation strategy (how the clinician was prompted to follow them).

Table 9.1 Classification of clinical guidelines in terms of probability of being effective (after Grimshaw and Russell[13])

Probability of being effective	Development strategy	Dissemination strategy	Implementation strategy
High	Internal	Specific educational intervention (for example, problem based learning package)	Patient specific reminder at time of consultation
Above average	Intermediate	Continuing education (for example, lecture)	Patient specific feedback
Below average	External, local	Mailing targeted groups	General feedback
Low	External, national	Publication in journal	General reminder

Table 9.1, in a nutshell, tells us that the most effective guidelines are developed locally by the people who are going to use them, introduced as part of a specific educational intervention, and implemented via a patient specific prompt that appears at the time of the consultation. Grimshaw's conclusions were initially misinterpreted by some people as implying that there was no place for nationally developed guidelines since only locally developed ones had any impact. In fact, whilst local adoption and ownership is undoubtedly crucial to the success of a guideline programme, local teams would be foolish not to draw on the range of expensively produced resources of evidence based national and international recommendations.[17]

Subsequent publications have identified a number of barriers to the adoption of guidelines in practice.[6, 7, 8, 18] These include:

- (apparent) disagreements amongst experts about the quality of evidence ("*Well, if they can't agree among themselves . . .*")
- lack of appreciation of evidence by practitioners ("*That's all very well, but when I trained we were always taught to hold back on steroids for asthma*")
- defensive medicine ("*I'll check all the tests anyway – belt and braces*")
- strategic and cost constraints ("*We can't afford to replace the equipment*")
- specific practical constraints ("*Where on earth did I put those guidelines?*")
- failure of patients to accept procedures ("*Mrs Brown insists she only needs a smear every five years*")
- competing influences of other non-medical factors ("*When we get the new computer system up and running . . .*")
- lack of appropriate, patient specific feedback on performance ("*I seem to be treating this condition OK*").

For a more detailed discussion on the barriers to implementing guidelines, see Grimshaw and Russell's comprehensive discussion of the subject,[19] the review on developing[17] and using[20] guidelines from the *BMJ*'s 1999 series on guidelines, and original research by other writers.[15, 21] In a nutshell, the successful introduction of guidelines needs "careful attention to the principles of change management: in particular, . . . leadership, energy, avoidance of unnecessary uncertainty, good communication, and, above all, time".[8]

Box 9.3 Proposed format for structured abstracts of clinical practice guidelines[22]

- *Objective* – The primary objective of the guideline, including the health problem and the targeted patients, providers, and settings
- *Options* – The clinical practice options considered in formulating the guideline
- *Outcomes* – Significant health and economic outcomes considered in comparing alternative practices
- *Evidence* – How and when evidence was gathered, selected, and synthesised
- *Values* – Disclosure of how values were assigned to potential outcomes of practice options and who participated in the process
- *Benefits, harms, and costs* – The type and magnitude of benefits, harms, and costs expected for patients from guideline implementation
- *Recommendations* – Summary of key recommendations
- *Validation* – Report of any external review, comparison with other guidelines or clinical testing of guideline use
- *Sponsors* – Disclosure of the people who developed, funded or endorsed the guideline

9.3 Questions to ask about a set of guidelines

Like all published articles, guidelines would be easier to evaluate if they were presented in a standardised format. Box 9.3 reproduces a suggested structured abstract for clinical guidelines[22] but since few published guidelines currently follow such a format, you will probably have to scan the full text for answers to the questions below. In preparing the list which follows, I have drawn on a number of previously published checklists and discussion documents.[8, 15, 18, 20, 23, 24, 25]

Question 1 Did the preparation and publication of these guidelines involve a significant conflict of interest?

I will resist labouring the point, but a drug company that makes hormone replacement therapy or a research professor whose life's work has been spent perfecting this treatment might be tempted to recommend it for wider indications than the average clinician.

Question 2 Are the guidelines concerned with an appropriate topic, and do they state clearly the goal of ideal treatment in terms of health and/or cost outcome?

Key questions in relation to choice of topic, reproduced from an article in the *BMJ*,[26] are given in Box 9.4.

A guideline which says "do this" without telling the practitioner why such an action is desirable is bad psychology as well as slack science. The intended outcome if the guideline is followed might be better patient survival, lower complication rates, increased patient satisfaction or savings in direct or indirect costs (see section 10.2). Whatever it is, it would be nice to know.

Question 3 Was the guideline development panel headed by a leading expert in the field and was a specialist in the methods of secondary research (e.g. metaanalyst, health economist) involved?

If a set of guidelines has been prepared entirely by a panel of internal "experts", you should, paradoxically, look at them particularly critically since researchers have been shown to be less objective in appraising evidence in their own field of expertise than in someone else's.[27] The involvement of an outsider (an expert in guideline development rather than in the particular clinical topic)

Box 9.4 Key questions on choice of topic for guideline development[26]

- Is the topic high volume, high risk, high cost?
- Are there large or unexplained variations in practice?
- Is the topic important in terms of the process and outcome of patient care?
- Is there potential for improvement?
- Is the investment of time and money likely to be repaid?
- Is the topic likely to hold the interest of team members?
- Is consensus likely?
- Will change benefit patients?
- Can change be implemented?

to act as arbiter and methodological adviser will, hopefully, make the process more objective.

Question 4 Have all the relevant data been scrutinised and do the guidelines' conclusions appear to be in keeping with the data?

On the most basic level, was the literature analysed at all or are these guidelines simply a statement of the preferred practice of a selected panel of experts (i.e. consensus guidelines)? If the literature was looked at, was a systematic search done and if so, did it follow the methodology described in section 8.2? Were all papers unearthed by the search included or was an explicit scoring system used to reject those of poor methodological quality and give those of high quality the extra weight they deserved?

Of course, up to date systematic reviews should ideally be the raw material for guideline development.[28] But in many cases, a search for rigorous and relevant research on which to base guidelines proves fruitless and the authors, unavoidably, resort to "best available" evidence or expert opinion. Given that in many clinical areas, the opinion of experts is still the best "evidence" around, guideline developers should adopt rigorous methods to ensure that it isn't just the voice of the expert who talks for longest in the meetings that drives the recommendations. Paul Shekelle from the RAND Corporation in the USA has undertaken some exciting research into methods for improving the rigour of consensus recommendations so as to ensure, for example, that an appropriate mix of experts is chosen, everyone reads the available research evidence, everyone gets an equal vote, all points of contention (raised anonymously) are fully discussed, and the resulting recommendations indicate the extent of agreement and dissent between the panel.[29, 30, 31] The UK Health Technology Assessment Programme has produced a valuable overview of the strengths and limitations of consensus methods which is available in full text on the Internet.[32]

Question 5 Do the guidelines address variations in medical practice and other controversial areas (e.g. optimum care in response to genuine or perceived underfunding)?

It would be foolish to make dogmatic statements about ideal practice without reference to what actually goes on in the real world. There are many instances where some practitioners are marching to an altogether different tune from the rest of us (see section 1.2) and a good guideline should face such realities head on rather than hoping that the misguided minority will fall into step by default.

Another thorny issue which guidelines should tackle head on is where essential compromises should be made if financial constraints preclude "ideal" practice. If the ideal, for example, is to offer all patients with significant coronary artery disease a bypass operation (at the time of writing it isn't, but never mind), and the health service can only afford to fund 20% of such procedures, who should be pushed to the front of the queue?

Question 6 Are the guidelines valid and reliable?

In other words, can you trust them, and if a different guideline development panel addressed the same question, would they come up with the same guidelines? These, of course, are the two $64 000 questions. The academic validity of guidelines depends on whether they are supported by high quality research studies and how strong the evidence from those studies is. In particular, issues of probability and confidence should have been dealt with acceptably (see section 4.6).

Question 7 Are the guidelines clinically relevant, comprehensive, and flexible?

In other words, are they written from the perspective of the practising doctor, nurse, midwife, physiotherapist, and so on and do they take account of the type of patients he or she is likely to see, and in what circumstances? Perhaps the most frequent source of trouble here is when guidelines developed in secondary care and intended for use in hospital outpatients (who tend to be at the sicker end of the clinical spectrum) are passed on to the primary health care team to be used in the primary care setting where, in general, patients are less ill and may well need fewer investigations and less aggressive management. This issue is discussed in section 7.2 in relation to the different utility of diagnostic and screening tests in different populations.

Guidelines should cover all, or most, clinical eventualities. What if the patient is intolerant of the recommended medication? What if you can't send off all the recommended blood tests? What if the patient is very young, very old or suffers from a co-existing illness?

These, after all, are the patients who prompt most of us to reach for our guidelines, while the more "typical" patient tends to be managed without recourse to written instructions.

Flexibility is a particularly important consideration for national and regional bodies who set themselves up to develop guidelines. It has been repeatedly demonstrated that ownership of guidelines by the people who are intended to use them locally is crucial to whether or not the guidelines are actually used.[6, 7, 13] If there is no free rein for practitioners to adapt them to meet local needs and priorities, a set of guidelines will probably never get taken out of the drawer.

Question 8 Do the guidelines take into account what is acceptable to, affordable by, and practically possible for patients?

There is an apocryphal story of a physician in the 1940s (a time when no effective medicines for high blood pressure were available) who discovered that restricting the diet of hypertensive patients to plain, boiled, unsalted rice dramatically reduced their blood pressure and also reduced the risk of stroke. The story goes, however, that the diet made the patients so miserable that a lot of them committed suicide.

This is an extreme example but I have seen guidelines for treating constipation in the elderly which offered no alternative to the combined insults of large amounts of bran and twice daily suppositories. Small wonder that the district nurses who were issued with them (for whom I have a good deal of respect) have gone back to giving castor oil.

For a further discussion on how to incorporate the needs and priorities of patients in guideline development, see a 1995 report from the College of Health.[33]

Question 9 Did the guidelines include recommendations for their own dissemination, implementation, and regular review?

Given the well documented gap between what is known to be good practice and what actually happens,[15, 19, 34] and the barriers to the successful implementation of guidelines discussed in section 9.2, it would be in the interests of those who develop guidelines to suggest methods of maximising their use. If this objective were included as standard in the "Guidelines for good guidelines", the guideline writers' output would probably include fewer ivory tower recommendations and more that are plausible, possible, and capable of being explained to patients.

1 Field, MJ, Lohr KN. *Clinical practice guidelines: direction of a new agency*. Washington DC: Institute of Medicine, 1990.

2 Grimley Evans J. evidence based and evidence-biased medicine. *Age Ageing*. 1995; **24**: 461–3.

3 Edwards P, Jones S, Shale D, Thursz M. *Shared care – a model for clinical management*. Oxford: Radcliffe Medical Press, 1996.

4 Hurwitz B. Clinical guidelines and the law: advice, guidance or regulation? *J Eval Clin Pract* 1995; **1**: 49–60.

5 Chalmers I. Why are opinions about the effects of health care so often wrong? *Medicolegal J* 1993; **62**: 116–30.

6 Delamothe T. Wanted: guidelines that doctors will follow. *BMJ* 1993; **307**: 218.

7 Greenhalgh PM. *Shared care for diabetes – a systematic review*. London: Royal College of General Practitioners, 1994 (Occasional Paper 67).

8 Ayers P, Renvoize T, Robinson M. Clinical guidelines: key decisions for acute service providers. *Br J Health Care Manage* 1995; **1**: 547–51.

9 Newton J, Knight D, Woolhead G. General practitioners and clinical guidelines: a survey of knowledge, use and beliefs. *Br J Gen Pract* 1996; **46**: 513–17.

10 Woolf SH, Grol R, Hutchinson A, Eccles M, Grimshaw J. Clinical guidelines: potential benefits, limitations, and harms of clinical guidelines. *BMJ* 1999; **318**: 527–30.

11 Department of Health. *A first class service: quality in the new NHS*. London: The Stationery Office, 1998.

12 Hurwitz B. Clinical guidelines: legal and political considerations of clinical practice guidelines. *BMJ* 1999; **318**: 661–4.

13 Grimshaw JM, Russell IT. Effect of clinical guidelines on medical practice. A systematic review of rigorous evaluations. *Lancet* 1993; **342**: 1317–22.

14 EPOC abstracts can be found via the public-access site listing all Cochrane abstracts on http://hiru.mcmaster.ca/cochrane/cochrane/revabstr/abidx.htm.

15 Lomas J, Haynes RB. A taxonomy and critical review of tested strategies for the application of clinical practice recommendations. From "official" to "individual" clinical policy. *Am J Prevent Med* 1987; **4**: 77–94.

16 Thomas L, Cullum N, McColl E, Rousseau N, Soutter J, Steen N. Guidelines in professions allied to medicine (Cochrane Review). In: *The Cochrane Library*, Issue 2. Oxford: Update Software, 2000.

17 Shekelle PG, Woolf SH, Eccles M, Grimshaw J. Clinical guidelines: developing guidelines. *BMJ* 1999; **318**: 593–6.

18 Report from General Practice 26. *The development and implementation of clinical guidelines*. London: Royal College of General Practitioners, 1995.

19 Grimshaw JM, Russell IT. Achieving health gain through guidelines II: ensuring guidelines change medical practice. *Qual Health Care* 1994; **3**: 45–52.

20 Feder G, Eccles M, Grol R, Griffiths C, Grimshaw J. Clinical guidelines: using clinical guidelines. *BMJ* 1999; **318**: 728–30.

21 Oxman A. *No magic bullets: a systematic review of 102 trials of interventions to help health professionals deliver services more effectively and efficiently*. London: North East Thames Regional Health Authority, 1994.

22 Hayward RSA, Wilson MC, Tunis SR, Bass EB, Rubin HR, Haynes RB. More informative abstracts of articles describing clinical practice guidelines. *Ann Intern Med* 1993; **118**: 731–7.

23 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.

24 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.

25 Effective Health Care Bulletin. *Implementing clinical guidelines: can guidelines be used to improve clinical practice?* Leeds: University of Leeds, 1994.

26 Thomson R, Lavender M, Madhok R. How to ensure that guidelines are effective. *BMJ* 1995; **311**: 237–42.

27 Cook DJ, Mulrow CD, Haynes RB. Systematic reviews: synthesis of best evidence for clinical decisions. *Ann Intern Med* 1997; **126**: 376–80.

28 Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR. The relation between systematic reviews and practice guidelines. *Ann Intern Med* 1997; **127**: 210–16.

29 Shekelle PG, Kahan JP, Bernstein SJ, Leape LL, Kamberg CJ, Park RE. The reproducibility of a method to identify the overuse and underuse of medical procedures. *New Engl J Med* 1998; **338**: 1888–95.

30 Shekelle PG, Roland M. Measuring quality in the NHS: lessons from across the Atlantic. *Lancet* 1998; **352**: 163–4.

31 Campbell SM, Hann M, Roland MO, Quayle JA, Shekelle PG. The effect of panel membership and feedback on ratings in a two-round Delphi survey: results of a randomized controlled trial. *Med Care* 1999; **37**: 964–8.

32 Murphy MK, Black NA, Lamping DL *et al.* Consensus development methods and their use in clinical guideline development. *Health Technol Assess* 1998; **2**(3). Available in full text on http://www.hta.nhsweb.nhs.uk/

33 Kelson M. *Consumer involvement initiatives in clinical audit and outcomes. A review of developments and issues in the identification of good practice*. London: College of Health, 1995.

34 Haines AP. The science of perpetual change. *Br J Gen Pract* 1996; **46**: 115–19.

Chapter 10: Papers that tell you what things cost (economic analyses)

10.1 What is economic analysis?

An economic analysis can be defined as *one that involves the use of analytical techniques to define choices in resource allocation*. Most of what I have to say on this subject comes from advice prepared by Professor Michael Drummond's team for authors and reviewers of economic analyses[1] and three of the "Users' guides to the medical literature" series,[2, 3, 4] as well as the excellent pocket sized summary by Jefferson and colleagues,[5] all of which emphasise the importance of setting the economic questions about a paper in the context of the overall quality and relevance of the study (see section 10.3).

The first economic evaluation I ever remember was a TV advertisement in which the pop singer Cliff Richard tried to persuade a housewife that the most expensive brand of washing up liquid on the market "actually works out cheaper". It was, apparently, stronger on stains, softer on the hands, and produced more bubbles per penny than "a typical cheap liquid". Although I was only nine at the time, I was unconvinced. Which "typical cheap liquid" was the product being compared with? How much stronger on stains was it? Why should the effectiveness of a washing up liquid be measured in terms of bubbles produced rather than plates cleaned?

Forgive me for sticking with this trivial example but I'd like to use it to illustrate the four main types of economic evaluation which you will find in the literature (see Table 10.1 for the conventional definitions).

Table 10.1 Types of economic analysis

Type of analysis	Outcome measure	Conditions of use	Example
Cost minimisation analysis	No outcome measure	Used when the effect of both interventions is known (or may be assumed) to be identical	Comparing the price of a brand name drug with that of its generic equivalent when bioequivalence has been demonstrated
Cost effectiveness analysis	Natural units (for example, life years gained)	Used when the effect of the interventions can be expressed in terms of one main variable	Comparing two preventive treatments for an otherwise fatal condition
Cost utility analysis	Utility units (for example, quality adjusted life years)	Used when the effect of the interventions on health status has two or more important dimensions (for example, benefits and side effects of drugs)	Comparing the benefits of two treatments for varicose veins in terms of surgical result, cosmetic appearance, and risk of serious adverse event (for example, pulmonary embolus)
Cost benefit analysis	Monetary units (for example, estimated cost of loss in productivity)	Used when it is desirable to compare an intervention for one condition with an intervention for a different condition	For a purchasing authority, to decide whether to fund a heart transplantation programme or a stroke rehabilitation ward

- *Cost minimisation analysis*. Sudso costs 47p per bottle whereas Jiffo costs 63p per bottle.
- *Cost effectiveness analysis*. Sudso gives you 15 more clean plates per wash than Jiffo.
- *Cost utility analysis*. In terms of quality adjusted housewife hours (a composite score reflecting time and effort needed to scrub plates clean and hand roughness caused by the liquid), Sudso provides 29 units per pound spent whereas Jiffo provides 23 units.
- *Cost benefit analysis*. The net overall cost (reflecting direct cost of the product, indirect cost of time spent washing up, and estimated financial value of a clean plate relative to a slightly grubby one) of Sudso per day is 7.17p, while that of Jiffo is 9.32p.

You should be able to see immediately that the most sensible analysis to use in this example is cost effectiveness analysis. Cost minimisation analysis (see Table 10.1) is inappropriate since Sudso and Jiffo do not have identical effectiveness. Cost utility analysis is unnecessary since, in this example, we are interested in very little else apart from the number of plates cleaned per unit of washing up liquid; in other words, our outcome has only one important dimension. Cost benefit analysis is, in this example, an absurdly complicated way of telling you that Sudso cleans more plates per penny.

There are, however, many situations where health professionals, particularly those who purchase health care from real cash limited budgets, must choose between interventions for a host of different conditions whose outcomes (such as cases of measles prevented, increased mobility after a hip replacement, reduced risk of death from heart attack or likelihood of giving birth to a live baby) cannot be directly compared with one another. Controversy surrounds not just how these comparisons should be made (see section 10.2) but also who should make them, and to whom the decision makers for the "rationing" of health care should be accountable. These essential, fascinating, and frustrating questions are beyond the scope of this book but if you are interested I would recommend you look up the references listed at the end of this chapter.[6–14]

10.2 Measuring the 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 10.1) included my general practitioner's time 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).

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 divided between various friends, neighbours, and a nice young girl from a nanny agency. And, from my point of view, there were several *intangible* costs, such as discomfort, loss of independence, the allergic rash I developed on the medication, and the cosmetically unsightly scar which I now carry on my abdomen.

As Box 10.1 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. In addition, I had a nice rest from work and, to be honest, I rather enjoyed all the attention and sympathy. (Note that the "social stigma" of appendicitis can be a positive one. I would be less likely to brag about my experience if my hospital admission had been precipitated by, say, an epileptic fit or a nervous breakdown, which have negative social stigmata.)

Box 10.1 Examples of costs and benefits of health interventions

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

In the appendicitis 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 acutely life threatening diseases. Most of us can count on developing at least one chronic, disabling, and progressive condition such as ischaemic heart disease, high blood pressure, arthritis, chronic bronchitis, cancer, rheumatism, prostatic hypertrophy or diabetes. At some stage, almost all of us will be forced to decide whether having a routine operation, taking a particular drug or making a compromise in our lifestyle (reducing our alcohol intake or sticking to a low-fat diet) 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 judgements 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, Mrs Brown 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 asthma) 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 (widely used in the UK) and the McMaster Health Utilities Index Questionnaire (popular in North America).[15]

In some circumstances, disease specific measures of wellbeing are more valid than general measures. 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.[16] 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. 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. If you would like to get up to speed in the ongoing debate on how to measure health related quality of life, take time to look up some of the references listed at the end of this chapter.[15, 17–25]

The authors of standard instruments for measuring quality of life (such as the SF-36) have often spent years ensuring they are valid (i.e. they measure what we think they are measuring), reliable (they do so every time), and responsive to change (i.e. 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 which eschews these standard instruments in favour of the authors' own rough and ready scale ("Functional ability was classified as good, moderate or poor according to the clinician's overall impression" or "We asked patients to score both their pain and their overall energy level from one to ten, and added the results together"). Note also that even instruments which have apparently been well validated often do not stand up to rigorous evaluation of their psychometric validity.[17]

Another way of addressing the "What's it worth?" of particular health states is through *health state preference values*, i.e. 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.[26]

- *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 wheelchair bound from arthritis of the hip).
- *Time trade-off 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" (for example, 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.

The quality adjusted life year or 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 Box 10.2.[27]

I find it almost impossible to discuss QALYs without my blood starting to boil (and I am not alone in feeling this way[28]). Any measure of health state preference values is, at best, a reflection of the preferences and prejudices of the individuals who contributed

Box 10.2 Cost per QALY (1990 figures)

Cholesterol testing and diet therapy	£220
Advice to stop smoking from patient's own doctor	£270
Hip replacement for arthritis	£1180
Kidney transplant	£4710
Breast cancer screening	£5780
Cholesterol testing and drug treatment if indicated (ages 25–39)	£14 150
Neurosurgery for malignant brain tumours	£107 780

to its development. Indeed, it is possible to come up with different values for QALYs depending on how the questions from which health state preference values are derived were posed.[29]

As medical ethicist John Harris has pointed out, QALYs are, like the society which produces them, inherently ageist, 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 life years 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.[30]

There is an increasingly confusing array of alternatives to the QALY. Some of the ones that were in vogue when this book went to press include the following.

- *Healthy years equivalent* or HYE, a QALY type measure that incorporates the individual's likely improvement or deterioration in health status in the future.[31]
- *Willingness to pay* (WTP) or *willingness to accept* (WTA) – measures of how much people would be prepared to pay to gain certain benefits or avoid certain problems.[5]
- *Healthy life years* or HeLY – which incorporates the risk of mortality (death) and morbidity (sickness) into a single number.[32]

- *Disability adjusted life year* or DALY – used mainly in the developing world to assess the overall burden of chronic disease and deprivation,[33, 34] an increasingly used measure that is not without its critics.[35]
- *TWiST* (time spent without symptoms of disease and toxicity of treatment) and *Q-TWiST* (quality adjusted TWiST)![36]

My personal advice on all these measures is to look carefully at what goes into the number that is supposed to be an "objective" indicator of a person's (or population's) health status and at how the different measures might differ according to different disease states. In my view, they all have potential uses but none of them is an absolute or incontrovertible measure of health or illness! (Note, also, that I do not claim to be an expert on any of these measures or on how to calculate them, which is why I have offered a generous list of additional references at the end of this chapter.)

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 with 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 (i.e. something real such as months of survival, legs amputated or take home babies), so that individuals can assign their own values to particular health states before comparing two quite different interventions (for example, infertility treatment versus cholesterol lowering, as in the example I mentioned in Chapter 1). Cost consequences analysis allows for the health state preference values of both individuals and society to change with time and is particularly useful when these are disputed or likely to change. This approach may also allow the analysis to be used by different groups or societies from the ones on which the original trial was performed.

10.3 Ten questions to ask about an economic analysis

The elementary checklist which follows is based largely on the sources mentioned in the first paragraph of this chapter. I strongly recommend that for a more definitive list, you check out these sources, especially the official recommendations by the BMJ working group.[1]

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

Before you attempt to digest what a paper says about costs, quality of life scales or utilities, 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 (see Chapter 4). Furthermore, if there is clearly little to choose between the interventions in terms of either costs or benefits, a detailed economic analysis is probably pointless.

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

From the patient's point of view, he or she generally wants to get better as quickly as possible. From the Treasury's point of view, the most cost effective health intervention is one that 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. There is no such thing as an economic analysis which is devoid of perspective. Most assume the perspective of the health care system itself, although some take into account the hidden costs to the patient and society (for example, due to work days lost). There is no "right" perspective for an economic evaluation but the paper should say clearly whose costs and whose benefits have been counted "in" and "out".

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

Nobody wants cheap treatment if it doesn't work. The paper you are reading may simply be an economic analysis, in which case it will be based on a previously published clinical trial, or it will be an economic evaluation of a new trial whose clinical results are presented in the same paper. Either way, you must make sure that the intervention that "works out cheaper" is not substantially less effective in clinical terms than the one that stands to be rejected on the grounds of cost. (Note, however, that in a resource limited health care system, it is often very sensible to use treatments that are a little less effective when they are a lot less expensive than the best on offer!)

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

A research trial that compares one obscure and unaffordable intervention with another will have little impact on medical practice. Remember that standard current practice (which may be "doing nothing") should almost certainly be one of the alternatives compared. 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).

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

This decision can be summarised as follows (see section 10.2).

- If the interventions produced identical outcomes ⇨ cost minimisation analysis
- If the important outcome is unidimensional ⇨ cost effectiveness analysis
- If the important outcome is multidimensional ⇨ cost utility analysis
- If the outcomes can be expressed meaningfully in monetary terms (i.e. if it is possible to weigh the cost–benefit equation for this condition against the cost–benefit equation for another condition) ⇨ cost benefit analysis
- If a cost benefit analysis would otherwise be appropriate but the preference values given to different health states are disputed or likely to change ⇨ cost consequences analysis

Question 6 How were costs and benefits measured?

Look back at section 10.2, where I outlined some of the costs associated with my appendix operation. Now imagine a more complicated example – the rehabilitation of stroke patients into their own homes with attendance at a day centre compared 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, 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.

There are no hard and fast rules for deciding which costs to include. If calculating "cost per case" from first principles, remember that someone has to pay for heating, lighting, personnel support, and even the accountants' bills of the institution. In general terms, these "hidden costs" are known as overheads and generally add an extra 30–60% onto the cost of a project. The task of costing things like operations and outpatient visits in the UK is easier than it used to be because these experiences are now bought and sold within the NHS at a price which reflects (or should reflect) all overheads involved. Be warned, however, that unit costs of health interventions calculated in one country often bear no relation to those of the same intervention elsewhere, even when these costs are expressed as a proportion of GNP.

Benefits such as earlier return to work for a particular individual can, on the face of it, be measured in terms of the cost of employing that person at his or her usual daily rate. This approach has the unfortunate and politically unacceptable consequence of valuing the health of professional people higher than that of manual workers, homemakers or the unemployed and that of the white majority higher than that of (generally) lower paid minority ethnic groups. It might therefore be preferable to derive the cost of sick days from the average national wage.

In a cost effectiveness analysis, changes in health status will be expressed in natural units (see section 10.2). 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 which remained healed at one year".

In cost benefit analysis, where health status is expressed in utility units such as QALYs, you would, if you were being really rigorous about evaluating the paper, look back at how the particular utilities used in the analysis were derived (see section 10.2). In particular, you will want to know whose health preference values were used – those of patients, doctors, health economists or the government.

For a more detailed and surprisingly readable account of how to "cost" different health care interventions, see the report from the UK Health Technology Assessment programme.[37]

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, i.e. 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 pretty 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 in section 8.2 in relation to metaanalysis. 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 cholesterol lowering drugs).[38]

Question 10 Were "bottom line" aggregate scores overused?

In section 10.2, I introduced the notion of cost consequences analysis, in which the reader of the paper can attach his or her own values to different utilities. 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 intelligence quotient 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".

10.4 Conclusion

I hope this chapter has shown that the critical appraisal of an economic analysis rests as crucially on asking questions such as "Where did those numbers come from?" and "Have any numbers been left out?" as on checking that the sums themselves were correct. Whilst few papers will fulfil all the criteria listed in section 10.3 and summarised in Appendix 1, you should, after reading the chapter, be able to distinguish an economic analysis of moderate or good methodological quality from one which slips "throwaway costings" ("drug X costs less than drug Y; therefore it is more cost effective") into its results or discussion section.

1 Drummond MF, Jefferson TO on behalf of the BMJ Economic Evaluation Working Party. Guidelines for authors and peer reviewers of economic submissions for the BMJ. *BMJ* 1996; **313**: 275–83.

2 Guyatt GH, Naylor CD, Juniper E, Heyland DK, Jaeschke R, Cook DJ. Users' guides to the medical literature. XII. How to use articles about health-related quality of life. *JAMA* 1997; **277**: 1232–7.

3 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.

4 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.

5 Jefferson T, Demicheli V, Mugford M. *Elementary economic evaluation in health care*. London: BMJ Publications, 1996.

6 New B. The rationing debate. Defining a package of healthcare services the NHS is responsible for. The case for. *BMJ* 1997; **314**: 503–5.

7 Klein R. The rationing debate. Defining a package in healthcare services the NHS is responsible for. The case against. *BMJ* 1997; **314**: 506–9.

8 Culyer AJ. The rationing debate: maximising the health of the whole community. The case for. *BMJ* 1997; **314**: 667–9.

9 Harris J. The rationing debate: maximising the health of the whole community. The case against: what the principal objective of the NHS should really be. *BMJ* 1997; **314**: 669–72.

10 Williams A, Evans JG. The rationing debate. Rationing health care by age. *BMJ* 1997; **314**: 820–5.

11 Lenaghan J. The rationing debate. Central government should have a greater role in rationing decisions. The case for. *BMJ* 1997; **314**: 967–70.

12 Harrison S. The rationing debate. Central government should have a greater role in rationing decisions. The case against. *BMJ* 1997; **314**: 970–3.

13 Doyal L. The rationing debate. Rationing within the NHS should be explicit. The case for. *BMJ* 1997; **314**: 1114–18.

14 Coast J. The rationing debate. Rationing within the NHS should be explicit. The case against. *BMJ* 1997; **314**: 1118–22.

15 Bowling A. Measuring health. Milton Keynes: Open University Press, 1997.

16 Bradley C, ed. *Handbook of psychology and diabetes*. London: Harwood Academic Publishers, 1994.

17 Gill TM, Feinstein AR. A critical appraisal of quality-of-life measurements. *JAMA* 1994; **272**: 619–26.

18 Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* 1995; **273**: 59–65.

19 Fallowfield LJ. Assessment of quality of life in breast cancer. *Acta Oncol* 1995; **34**: 689–94.

20 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.

21 Laupacis A, Wong C, Churchill D. The use of generic and specific quality-of-life measures in hemodialysis patients treated with erythropoietin. *Controlled Clin Trials* 1991; **12** (suppl): 168S–179S.

22 Tugwell P, Bombardier C, Buchanan WW *et al.* Methotrexate in rheumatoid arthritis. Impact on quality of life assessed by traditional standard-item and individualized patient preference health status questionnaires. *Arch Intern Med* 1990; **150**: 59–62.

23 Cairns J. Measuring health outcomes. *BMJ* 1996; **313**: 6.

24 Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. *JAMA* 1994; **272**: 619–26.

25 Guyatt GH, Cook DJ. Health status, quality of life, and the individual patient. A commentary on: Gill TM, Feinstein AR. A critical appraisal of the quality of quality of life measurements. *JAMA* 1994; **272**: 630–1.

26 Brazier J, Deverill M, Green C, Harper R, Booth A. A review of the use of health status measures in economic evaluation. *Health Technol Assess* 1999; **3**(9). Available in full text on http://www.hta.nhsweb.nhs.uk/

27 Ham C. Priority setting in the NHS. *Br J Health Care Manage* 1995; **1**: 27–9.

28 Naylor CD. Cost-effectiveness analysis: are the outputs worth the inputs? *ACP Journal Club* 1996; **124**: a12–14.

29 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.

30 Harris J. QALYfying the value of life. *J Med Ethics* 1987; **13**: 117–23.

31 Gafni A. Alternatives to the QALY measure for economic evaluations. *Supportive Care Cancer* 1997; **5**: 105–11.

32 Hyder AA, Rotllant G, Morrow RH. Measuring the burden of disease: healthy life-years. *Am J Pub Health* 1998; **88**: 196–202.

33 Ustun TB, Rehm J, Chatterji S *et al*. Multiple-informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *Lancet* 1999; **354**: 111–15.

34 Gwatkin DR, Guillot M, Heuveline P. The burden of disease among the global poor. *Lancet* 1999; **354**: 586–9.

35 Arnesen T, Nord E. The value of DALY life: problems with ethics and validity of disability adjusted life years. *BMJ* 1999; **319**: 1423–5.

36 Billingham LJ, Abrams KR, Jones DR. Methods for the analysis of quality-of-life and survival data in health technology assessment. *Health Technol Assess* 1999; **3**(10). Available in full text on http://www.hta.nhsweb.nhs.uk/

37 Johnston K, Buxton MJ, Jones DR, Fitzpatrick R. Assessing the costs of healthcare technologies in clinical trials. *Health Technol Assess* 1999; **3**(6). Available in full text on http://www.hta.nhsweb.nhs.uk/

38 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.

Chapter 11: Papers that go beyond numbers (qualitative research)

11.1 What is qualitative research?

The pendulum is swinging. Fifteen years ago, when I took up my first research post, a work weary colleague advised me: "Find something to measure, and keep on measuring it until you've got a boxful of data. Then stop measuring and start writing up". "But what should I measure?", I asked. "That", he said cynically, "doesn't much matter."

This true example illustrates the limitations of an exclusively quantitative (counting and measuring) perspective in 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, one man in 10 is homosexual, and the incidence of coronary heart disease was rising but is now falling. 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.

Qualitative researchers 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]

Interpretive or qualitative research was for years the territory of the social scientists. It is now increasingly recognised as being not just complementary to but, in many cases, a prerequisite for the quantitative research with which most of us who trained in the biomedical sciences are more familiar. Certainly, the view that the two approaches are mutually exclusive has itself become "unscientific" and it is currently rather trendy, particularly in the fields of primary care and health services research, to say that you are doing some qualitative research – and since the first edition of this book was published, qualitative research has even become mainstream within the evidence based medicine movement.[3, 4, 5]

Dr Cecil Helman, author of a leading textbook on the anthropological aspects of health and illness,[6] told me the following story to illustrate the qualitative quantitative dichotomy. A small child runs in from the garden and says, excitedly, "Mummy, the leaves are falling off the trees". "Tell me more," says his mother. "Well, five leaves fell in the first hour, then ten leaves fell in the second hour . . ." That child will become a quantitative researcher.

A second child, when asked "tell me more", might reply, "Well, the leaves are big and flat, and mostly yellow or red, and they seem to be falling off some trees but not others. And mummy, why did no leaves fall last month?" That child will become a qualitative researcher.

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 hang out, listen to what people have to say, and 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 11.1, and go on to use a selection of others.

Box 11.2, which is reproduced with permission from Nick Mays and Catherine Pope's introductory book *Qualitative research in health care*,[7] 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.[8, 9]

Box 11.1 Examples of qualitative research methods

Documents	Study of documentary accounts of events, such as meetings
Passive observation	Systematic watching of behaviour and talk in naturally 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 that explicitly includes and uses the group interaction to generate data

Box 11.2 Qualitative versus quantitative research – the overstated dichotomy[7]

	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

As section 3.2 explains, 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 is different. It begins with an intention to explore a particular area, collects "data" (i.e. observations and interviews), and generates ideas and hypotheses from these data largely through what is known as *inductive reasoning*.[7] The strength of the quantitative approach lies in its *reliability* (repeatability), i.e. the same measurements should yield the same results time after time. The strength of qualitative research lies in *validity* (closeness to the truth), i.e. 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 the use of more than one method (see Box 11.1) in combination, a process known as *triangulation*, and by more than one researcher analysing the same data independently.

Those who are ignorant about qualitative research often believe that it constitutes little more than hanging out and watching leaves fall. It is beyond the scope of this book to take you through the substantial literature on how to (and how not to) proceed when observing, interviewing, leading a focus group, and so on. But sophisticated methods for all these techniques certainly exist and if you are interested I suggest you try the introductory[7, 10, 11] or more detailed[2, 12] texts listed at the end of this chapter.

Qualitative methods really come into their own when researching uncharted territory, i.e. where the variables of greatest concern are poorly understood, ill defined, and cannot be controlled.[1, 13] In such circumstances, the definitive hypothesis may not be arrived at until the study is well under way. But it is in precisely these circumstances that the qualitative researcher must ensure that (s)he has, at the outset, carefully delineated a particular focus of research and identified some specific questions to try to answer (see Question 1 in section 11.2 below). The methods of qualitative research allow for and even encourage[2] modification of the research question in the light of findings generated along the way. (In contrast, as section 5.2 showed, sneaking a look at the interim results of a quantitative study is statistically invalid!)

The so-called *iterative* approach (altering the research methods and the hypothesis as you go along) employed 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. Whilst 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".[13] They warn that qualitative researchers must, therefore, allow periods away from their fieldwork for reflection, planning, and consultation with colleagues.

11.2 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" (see references 8–32 in Chapter 3) could ever be developed. My own view, and that of a number of individuals who have attempted or are currently working on this very task,[7, 12, 13, 14] 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, 7, 13] and also from discussions with Dr Rod Taylor of Exeter University, who has worked with the CASP Project on a more detailed and extensive critical appraisal guide for qualitative papers.

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

In section 3.2, I 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. 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 which 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 fit free on 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".

As I explained in section 11.1, the iterative nature of qualitative research is such that the definitive research question may not be clearly focused at the outset of the study but, as Britten and colleagues point out, it should certainly have been formulated by the time the report is written!

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), qualitative methods are clearly inappropriate! If you think a case-control, cohort study or randomised trial would have been better suited to the research question posed in the paper than the qualitative methods that were actually used, you might like to compare that question with the examples in section 3.3 to confirm your hunch.

Question 3 How were (a) the setting and (b) the subjects selected?

Look back at Box 11.2, which contrasts the *statistical* sampling methods of quantitative research with *theoretical* ones of qualitative research. Let me explain what this means. In the earlier chapters of this book, particularly section 4.2, I emphasised the importance, in quantitative research, of ensuring that a truly random sample of subjects is recruited. A random sample will ensure 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 and 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 interpreter/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.

We would also wish to select some women who had had shared antenatal care between an obstetrician and their general practitioner and some women who had been cared for by community midwives throughout the pregnancy. In this example, it might be particularly instructive to find women who had had their care provided by male doctors, even though this would be a relatively unusual situation. Finally, we might choose to study patients who gave birth in the setting of a large, modern, "high-tech" maternity unit as well as some who did so in a small community hospital. Of course, all these specifications will give us "biased" samples but that is exactly what we want.

Watch out for qualitative research where the sample has been selected (or appears to have been selected) purely on the basis of convenience. In the above example, taking the first dozen Punjabi patients to pass through the nearest labour ward would be the easiest way to notch up interviews, but the information obtained may be considerably less helpful.

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

Given that qualitative research is necessarily grounded in real life experience, a paper describing such research should not be "trashed" simply because the researchers have declared a particular cultural perspective or personal involvement with the subjects of the research. Quite the reverse: they should be congratulated for doing just that. 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 (see Box 11.1) 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 and Bloggs [reference to Bloggs and Bloggs’ paper on how to measure serum rhubarb]”.

I now spend quite a lot of my time doing qualitative research and I can confirm that it’s infinitely more fun. I and my research team spent an interesting few years devising a unique combination of techniques to measure the beliefs, hopes, fears, and attitudes of diabetic patients from a particular minority ethnic group (British Sylhetis). We had to develop, for example, a valid way of simultaneously translating and transcribing interviews which were conducted in Sylheti, a complex dialect of Bengali which has no written form. We found that patients’ attitudes appear to be heavily influenced by the presence in the room of certain of their relatives, so we contrived to interview some patients in both the presence and the absence of these key relatives.

I could go on describing the methods we devised to address this particular research issue[15] but I have probably made my point: 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 via *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 tailormade 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?".

A good qualitative research paper may 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. In analysing my own research into health beliefs in diabetic patients, three of us looked in turn at a typed interview transcript and assigned codings to particular statements. We then compared our decisions and argued (sometimes heatedly) about our disagreements. Our analysis revealed differences in the interpretation of certain statements which we were unable to fully resolve; in other words, our inability to "wrap up" all aspects of the interpretation was itself an important item of data.[15] All this is legitimate methodology for analysing qualitative data. What is *not* legitimate is to assume that there is a single "right" way to interpret the data. Having said that, there are some researchers who feel strongly that only the person most immersed in the fieldwork has genuine insight into the meaning of the data; in other words, interpretation should not be "triangulated" by all and sundry simply to give a show of improving validity.

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

We obviously cannot assess the credibility of qualitative results via the precision and accuracy of measuring devices, nor their significance via confidence intervals and numbers needed to treat. It 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, presented in standard IMRAD format (see section 3.1), should clearly distinguish the study's results (usually a set of numbers) from the interpretation of those results. 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 imply 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?
- How well does the explanation cohere with what we already know?[16]

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

One of the most common 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 I described in Question 3. 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.

11.3 Conclusion

Doctors have traditionally placed high value on number based 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 with the evidence.[4, 5]

In 1993, Catherine Pope and Nicky Britten presented at a conference a paper entitled "Barriers to qualitative methods in the medical mindset", in which they showed their collection of rejection letters from biomedical journals.[17] 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 appeared to be incapable of distinguishing good qualitative research from bad.

Somewhat ironically, poor quality qualitative papers now appear regularly in some medical journals, who appear to have undergone an about face in editorial policy since Pope and Britten's exposure of the "medical mindset". I hope, therefore, that the questions listed above, and the references below, will assist reviewers in both camps: those who continue to reject qualitative papers for the wrong reasons and those who have climbed on the qualitative bandwagon and are now *accepting* such papers for the wrong reasons! 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.

1 Black N. Why we need qualitative research. *J Epidemiol Commun Health* 1994; **48**: 425–6.

2 Denzin NK, Lincoln YS, eds. *Handbook of qualitative research*. London: Sage Publications, 1994.

3 Green J, Britten N. Qualitative research and evidence based medicine. *BMJ* 1998; **316**: 1230–2.

4 Giacomini MK, Cook DJ. A user's guide to qualitative research in health care. Part I. Are the results of the study valid? *JAMA* 2000; 357–62.

5 Giacomini MK, Cook DJ. A user's guide to qualitative research in health care: Part II. What are the results and how do they help me care for my patients? *JAMA* 2000; 478–82.

6 Helman C. *Culture, health and illness*, 4th edn. London: Butterworth Heinemann, 2000.

7 Mays N, Pope C, eds. *Qualitative research in health care*, 2nd edn. London: BMJ Publications, 2000.

8 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.

9 Bradley F, Wiles R, Kinmonth AL, Mant D, Gantley M. Development and evaluation of complex interventions in health services research: case study of the Southampton Heart Integrated Care Project (SHIP). The SHIP Collaborative Group. *BMJ* 1999; **318**: 711–15.

10 Pope C, Mays N. Qualitative research in health care: assessing quality in qualitative research. *BMJ* 2000; **320**: 50–2.

11 Pope C, Ziebland S, Mays N. Qualitative research in health care: analysing qualitative data. *BMJ* 2000; **320**: 114–16.

12 Murphy E, Dingwall R, Greatbatch D, Parker S, Watson P. Qualitative research

methods in health technology assessment: a review of the literature. *Health Technol Assess* 1998; **2**(16). Available in full text on http://www.hta.nhsweb.nhs.uk/.

13 Britten N, Jones R, Murphy E, Stacy R. Qualitative research methods in general practice and primary care. *Fam Pract* 1995; **12**: 104–14.

14 Taylor R, personal communication.

15 Greenhalgh T, Chowdhury AM, Helman C. Health beliefs and folk models of diabetes in British Bangladeshis: a qualitative study. *BMJ* 1998; **316**: 978–83.

16 Mays N, Pope C, eds. *Qualitative research in health care,* 2nd edn. London: BMJ Publications, 2000: 15.

17 Pope C, Britten N. The quality of rejection: barriers to qualitative methods in the medical mindset. Paper presented at BSA Medical Sociology Group annual conference, September 1993.

Chapter 12: Implementing evidence based findings

12.1 Surfactants versus steroids: a case study in adopting evidence based practice

Health professionals' failure to practise in accordance with the best available evidence cannot be attributed entirely to ignorance or stubbornness. Consultant paediatrician Dr Vivienne van Someren has described an example that illustrates many of the additional barriers to getting research evidence into practice: the prevention of neonatal respiratory distress syndrome in premature babies.[1]

It was discovered back in 1957 that babies born more than six weeks early may get into severe breathing difficulties because of lack of a substance called surfactant (which lowers the surface tension within the lung alveoli and reduces resistance to expansion) in their lungs. Pharmaceutical companies began research in the 1960s to develop an artificial surfactant that could be given to the infant to prevent the life threatening syndrome developing but it was not until the mid-1980s that an effective product was developed.

By the late 1980s a number of randomised trials had taken place and a metaanalysis published in 1990 suggested that the benefits of artificial surfactant greatly outweighed its risks. In 1990 a 6000 patient trial (OSIRIS) was begun which involved almost all the major neonatal intensive care units in the UK. The manufacturer was awarded a product licence in 1990 and by 1993, practically every eligible premature infant in the UK was receiving artificial surfactant.

Another treatment had also been shown a generation ago to prevent neonatal respiratory distress syndrome: administration of the steroid drug dexamethasone to mothers in premature labour. Dexamethasone worked by accelerating the rate at which the fetal lung reached maturity. Its efficacy had been demonstrated in experimental animals in 1969 and in clinical trials on humans, published in the prestigious journal *Pediatrics*, as early as 1972. Yet despite a significant beneficial effect being confirmed in a number of further trials and a metaanalysis published in 1990, the take up of this technology was astonishingly slow. It was estimated in 1995 that only 12–18% of eligible mothers currently received this treatment in the USA.[2]

The quality of the evidence and the magnitude of the effect were similar for both these interventions.[3,4] Why were the paediatricians so much quicker than the obstetricians at implementing an intervention which prevented avoidable deaths? Dr van Someren has considered a number of factors, listed in Table 12.1.[1] The effect of artificial surfactant is virtually immediate and the doctor administering it witnesses directly the "cure" of a terminally sick baby. Pharmaceutical industry support for a large (and, arguably, scientifically unnecessary) trial ensured that few consultant paediatricians appointed in the early 1990s would have escaped being introduced to the new technology.

In contrast, steroids, particularly for pregnant women, were unfashionable and perceived by patients to be "bad for you". In doctors' eyes, dexamethasone treatment was old hat for a host of unglamorous diseases, notably cancer, and the scientific mechanism for its effect on fetal lungs was not readily understood. Most poignantly of all, an obstetrician would rarely get a chance to witness directly the life saving effect on an individual patient.

The above example is far from isolated. Effective health care strategies frequently take years to catch on,[5] even amongst the experts who should be at the cutting edge of practice.[6] It would appear that for a new technology to be adopted readily by individual health professionals, a number of conditions must be satisfied. The evidence should be unequivocal and of high quality (preferably from large RCTs with well defined, clinically important endpoints); the user of the technology must personally believe that it is effective; he or she should have the opportunity to try out the intervention in controlled circumstances; possible adverse effects of the technology should be placed in proportion to the likely benefits; and clinical conflicts of interest (for example, an obstetrician's divided loyalty between two patients) should be identified and explored.

Table 12.1 Factors influencing implementation of evidence to prevent neonatal respiratory distress syndrome (van Someren, personal communication)

	Surfactant treatment	Prenatal steroid treatment
Perception of mechanism	Corrects a surfactant deficiency disease	Ill defined effect on developing lung tissue
Timing of effect	Minutes	Days
Impact on prescriber	Views effect directly (has to stand by ventilator)	Sees effect as statistic in annual report
Perception of side effects	Perceived as minimal	Clinicians' and patients' anxiety disproportionate to actual risk
Conflict between two patients	No (paediatrician's patient will benefit directly)	Yes (obstetrician's patient will not benefit directly)
Pharmaceutical industry interest	High (patented product; huge potential revenue)	Low (product out of patent; small potential revenue)
Trial technology	"New" (developed in late 1980s)	"Old" (developed in early 1970s)
Widespread involvement of clinicians in trials	Yes	No

12.2 Changing health professionals' behaviour: evidence from studies on individuals

The Cochrane Effective Practice and Organisation of Care Group (EPOC, described in Chapter 9, page 142) has done an excellent job of summarising the literature accumulated from research trials on what is and is not effective in changing professional practice, both in promoting effective innovations and in encouraging professionals to resist "innovations" that are ineffective or harmful. A major EPOC review was recently published in an excellent book edited by Andy Haines and Anna Donald.[7] Its main findings were that:

- *Consistently effective* methods include educational outreach visits (academic detailing), reminders or prompts (manual or computerised) issued at the time of the consultation, multifaceted interventions (a "belt and braces" combination of two or more methods), and interactive educational meetings.
- *Sometimes effective* methods included audit and feedback (any summary of clinical performance given back to individual clinicians), local opinion leaders (see below), and patient mediated interventions (such as information leaflets or patient held prompts).
- *Little or no effect* was found with didactic educational meetings or the distribution of printed guidelines.

The different individual methods are considered in more detail below.

Ineffective: didactic education

Until relatively recently, *education* (at least in relation to the training of doctors) was more or less synonymous with the didactic talk and chalk sessions that most of us remember from school and college. The "bums on seats" approach to postgraduate education (filling lecture theatres up with doctors or nurses and wheeling on an "expert" to impart pearls of wisdom) is relatively cheap and convenient for the educators but is largely ineffective in producing sustained behaviour change in practice.[8, 9] Indeed, one study demonstrated that the number of reported CME (continuing medical education) hours attended was *inversely* correlated with doctors' competence![10]

Mostly effective: interactive, hands on education

Encouragingly, the most powerful drive to learn amongst health professionals is probably not external sticks and carrots but the desire to be more competent in treating patients.[11] The major changes currently under way in medical and nursing education in the UK (and many other countries) strongly support the use of hands on methods structured around real clinical problems ("problem based learning")[12] and strategies to link professionals' learning with the needs of the service,[13] improved teamwork,[14] and organisational development.[15]

Ineffective: "standard issue" guidelines and protocols

Another overview by EPOC supports the common sense impression that producing and disseminating written guidelines, protocols, and policies does not generally change behaviour unless accompanied by other measures as well.[16] The underlying reasons for the gap between evidence and practice have been fairly well described,[16–18] and include:

- lack of knowledge (the guideline or protocol remains unread or misunderstood)
- lack of confidence (the clinician does not "believe in" the recommendation)
- fear of legal or patient pressure or loss of income
- lack of a physical skill (e.g. inability to inject a joint or operate via an endoscope)
- inadequate resources (e.g. a limited budget for drug prescribing, non-availability of key equipment such as an MRI scanner)
- failure to remember to carry out a procedure or task, due to pressure of work or old habits, which tend to die hard.

In one large study, the main factors associated with successfully following a guideline or protocol were the practitioner's perception that it was uncontroversial (68% compliance vs 35% if it was perceived to be controversial), evidence based (71% vs 57% if not), contained explicit recommendations (67% vs 36% if the recommendations were vague), and required no change to existing routines (67% vs 44% if a major change was recommended).[17]

Mostly effective: locally produced and owned protocols

The importance of ownership (i.e. the feeling of those being asked to play by new rules that they have been involved in drawing up those rules) was emphasised in section 9.2 (see in particular Table 9.1, page 142) in relation to clinical guidelines. There is an extensive management theory literature to support the common sense notion that professionals will oppose changes that they perceive as threatening to their livelihood (i.e. income), self-esteem, sense of competence or autonomy. It stands to reason, therefore, that involving health professionals in setting the standards against which they are going to be judged generally produces greater changes in patient outcomes than occur if they are not involved.[18]

Other studies have shown that nationally produced recommendations are more likely to be adopted locally if they are adapted to reflect local needs and priorities (for example, by deleting sections not applicable locally and adding information about relevant local services) and made accessible (for example, via a wallchart or by inserting a summary in a widely used handbook such as a house officers' formulary). The integration of evidence based recommendations with the realities of local services can often be successfully achieved via written, multiprofessional "care pathways" for particular conditions which state not only what intervention is needed at different stages in the course of the condition but also whose responsibility it is to undertake the task and to follow up if it gets missed.[19, 20]

Mostly effective: high quality decision support

There is a growing literature on the use of high technology strategies such as computerised *decision support systems* that incorporate the research evidence and can be accessed by the busy practitioner at the touch of a button. Dozens of these systems are currently being developed and since the last edition of this book, a major systematic review has been published,[21] which is summarised in a book chapter by Paul Taylor and Jeremy Wyatt.[22] Briefly, the review found that overall, around two thirds of studies of computerised decision support demonstrated improved performance, with the best results being in drug dosing and active clinical care (for example, management of asthma) and the worst results in diagnosis. An important criticism of decision support systems, however, is that the information on which their "advice" is based may not necessarily be evidence based![22]

Computerised "prompts" that appear during the consultation or at other crucial points in the decision making sequence are one of the most effective methods for achieving behaviour change. But anecdotally, clinicians frequently complain that the computer systems they work with are cumbersome, address the wrong questions or provide unworkable recommendations. As Taylor and Wyatt suggest, "Poor design and a failure to consider the practicalities of clinical settings have perhaps hindered the take up of decision support systems, but such systems could never be designed to fit seamlessly into existing ways of working".[22] In other words, given the evidence of improved patient care with good decision support systems, we clinicians should probably make more effort than we currently do to accommodate them within our practice.

Sometimes effective: audit and feedback

Audit is a cycle of setting standards, measuring performance, changing practice to work towards the standard, and measuring the change. Feedback consists of telling the physician how his or her performance (for example, number of hysterectomies performed per head of population, total cost of drugs prescribed per month, and so on) over time compares either with a group norm (such as the levels achieved by fellow physicians) or with an external standard (such as expert consensus). Studies in this area have been directed mainly at reducing the number of inappropriate diagnostic tests ordered by junior doctors or improving drug prescribing in either general practice or hospitals.[7]

Most studies of clinical audit provide modest but undramatic evidence of improvement in clinician performance with this method[23] but one wonders how much publication bias there is in this finding (I personally would never publish an audit that showed I'd got worse despite trying to improve!). The EPOC review is highly critical of the quality of the studies and hardly any of the primary studies published so far have measured the impact on patients.[23]

Studies of feedback to clinicians on their performance suggest that this method is only effective in changing practice if:

- the health professional already accepts that his or her practice needs to change
- the health professional has the resources and authority to implement the required changes
- the feedback is offered in "real time" (i.e. at the time when the practice is being implemented) rather than retrospectively.[7, 23]

Both audit and feedback work better as part of a multifaceted intervention, for example, when combined with an interactive education programme.[7]

Sometimes effective: social influence strategies

One method of "education" which the pharmaceutical industry has shown to be highly effective is one to one contact between doctors and company representatives (known in the USA as detailers), whose influence on clinical behaviour may be so dramatic that they have been dubbed the "stealth bombers" of medicine.[24] In the USA, this tactic has been harnessed by the government in what is known as *academic detailing*: the educator books in to see the physician in the same way as industry representatives but in this case the "rep" provides objective, complete and comparative information about a range of different drugs and encourages the clinician to adopt a critical approach to the evidence. Such a strategy can achieve dramatic short term changes in practice[25] but may be ineffective if the educator attends only briefly and fails to ascertain the perspective of the clinician before attempting to influence it.[26]

Academic detailing is one example of a general behaviour change policy known as *social influence*, in which the practitioner is persuaded that his or her current practice is out of step with that of colleagues or experts.[27, 28] Social influence policies also include use of the mass media, processes where members of a group or network influence one another,[29] and those where local opinion leaders – officially defined as individuals nominated by their colleagues as educationally influential – are used as a mouthpiece for policy change.[30] A recent EPOC review confirms that such individuals are often, but not always, effective in acting as the vehicle for evidence based change in practice.[31] A systematic review by Andy Oxman and colleagues provides some overall support for social influence methods but gives several examples of so-called social influence policies that, in reality, failed to influence.[32]

Sometimes effective: patient led strategies

One important method of initiating change in the behaviour of health professionals is *pressure from patients and the general public*. A number of organisations now produce evidence based information leaflets for patients; for example, the "Effectiveness matters" series based on the *Effective health care bulletin* from the Centre for

Reviews and Dissemination in York,[33] the MIDIRS booklet *Through the maze,*[34] which is based, among other sources, on the Cochrane Pregnancy and Childbirth Database, or the leaflet by the charity Diabetes UK "What diabetic care to expect".[35] A number of electronic information aids on particular conditions are now available either as interactive video discs[36] or over the Internet[37] but the evidence that such tools can significantly improve the proportion of decisions that are evidence based (as opposed to just understood and agreed by the patient) is still in doubt.[38]

The power of informed patient choice can also be harnessed more directively; for example, through the "prompted care" model of diabetes care, where patients are sent a structured checklist of tasks (such as blood pressure monitoring and inspection of feet) every six months and advised to ask their doctor to complete it.[39] For an overview of the patient's perspective in evidence based health care and examples of how the informed user can shape the behaviour of professionals, see the chapter by Sandy Oliver and colleagues[38] and the book by Fulford and colleagues *Essential practice in patient-centred care*.[40]

Sometimes effective: rules and incentives

Administrative strategies for influencing clinicians' behaviour include, at one extreme, changes in the law (for example, withdrawing the product licence for a drug) or institutional policy (such as the imposition of a restricted formulary of drugs and equipment). More commonly, they involve introducing barriers to undesired practice (such as requiring the approval of a specialist when ordering certain tests) or reducing barriers to desired practice (such as altering order forms to reflect the preferred dosing interval for antibiotics).[41]

Financial incentives may be set up to prompt health professionals to perform more of a desired intervention (such as the UK "target" system for cervical smears by general practitioners[42]) or less of an undesired one (such as changing a fee for service remuneration policy for clinic doctors to a flat rate salary[43]).

Such strategies, however, may run counter to the philosophy of involving professionals in setting standards and gaining their ownership of the changes. In addition, whilst the evidence that administrative and financial strategies achieve changes in behaviour is strong, these changes may generate much resented "hassle" and are not always translated into desired patient outcomes.[7] A restrictive policy to minimise “unnecessary” drug prescribing in the elderly, for example, achieved its limited objective of reducing expenditure on medication but was associated with increased rates of admission to nursing homes.[44] This illustrates the important point that implementing evidence based practice is not an all or none, unidimensional achievement, as I have argued elsewhere.[45]

In summary, there is no shortage of strategies for changing the behaviour of health professionals, an increasing number of which have now been rigorously evaluated. As Oxman and colleagues concluded after reviewing 102 published studies on different ways to influence the behaviour of clinicians, “There are no ‘magic bullets’ for improving the quality of health care, but there is a wide range of interventions available that, if used appropriately, could lead to important improvements in professional practice and patient outcomes.”[32] Although many new studies and systematic reviews have been published since Oxman penned these words, they still reflect the gist of the evidence on this fascinating topic.

12.3 Managing change for effective clinical practice: evidence from studies on organisational change

A number of projects in the UK have looked systematically at the implementation of evidence based findings in health care organisations. These include:

- GRiPP (Getting Research into Practice and Purchasing), led by the Anglia and Oxford Regional Health Authority[46]
- PACE (Promoting Action on Clinical Effectiveness), led by the King’s Fund[47]
- PLIP (Purchaser Led Implementation Projects), led by the North Thames Regional Office.[48]

All these projects were founded on the view that it was not sensible to rely on one course of action – such as producing a set of clinical guidelines – but rather that the overall process had to be managed using a wide variety of ways and means including incentives, educational programmes, academic detailing, local guidelines development, and so on. A number of separate projects were initiated through these overall programmes. GRiPP included the use of steroids in preterm delivery, the management of services for stroke patients, the use of dilatation and curettage (D & C) in women with heavy periods, and insertion of grommets into children with glue ear. The 12 PACE projects included initiatives to improve hypertension management by GPs and leg ulcer care within an acute hospital. The PLIP projects were similarly topic based initiatives, including promoting secondary prevention of coronary heart disease in primary care and introduction of guidelines for the eradication of the ulcer causing bacterium *H. pylori*.

The lessons in Boxes 12.1 (from GRiPP), 12.2 (from PACE) and 12.3 (from PLIP) were the result of a rigorous evaluation process undertaken by the different participating groups.[47–49]

Box 12.1 Lessons from the GRiPP (Getting Research into Practice and Purchasing) project[49]

1. Prerequisites for implementing changes in clinical practice are *nationally available research evidence* and *clear*, *robust* and *local justification for change*.
2. There should be *consultation and involvement* of all interested parties, led by a respected product champion.
3. The *knock-on effect* of change in one sector (e.g. acute services) onto others (e.g. general practice or community care) should be addressed.
4. *Information* about current practice and the effect of change needs to be available.
5. *Relationships* between purchasers and providers need to be good.
6. Contracts (e.g. between purchasers and providers) are best used to *summarise agreement* that has already been negotiated elsewhere, not to table points for discussion.
7. Implementing evidence *may not save money*.
8. Implementing evidence *takes more time than is usually anticipated*.

12.4 The evidence based organisation: a question of culture

A publication by the UK National Association of Health Authorities and Trusts (NAHAT) entitled “Acting on the evidence” observes that:

Box 12.2 Barriers to change identified in the PACE (Promoting Action on Clinical Effectiveness) projects[47]

1. *Lack of perception of relevance*. Practitioners believe there is no need to change and/or that their practice is already evidence based.
2. *Lack of resources*. Stakeholders feel they do not have the time or money to become involved.
3. *Short term outlook*. Stakeholders are driven by short term incentives, e.g. an annual contracting round, which may conflict with the timescale needed for effecting change.
4. *Conflicting priorities*. Stakeholders have other demands on their energies, such as reducing waiting lists or dealing with specific complaints.
5. *Difficulty in measuring outcomes*. Health outcomes are notoriously difficult to measure, yet many stakeholders mistakenly seek to measure the success of the project in terms of bottom line health gains.
6. *Lack of necessary skills*. Unfamiliar skills may be needed for effective clinical practice, such as those for searching and critical appraisal of research.
7. *No history of multidisciplinary working*. Members of different disciplines may not be used to working together in a collaborative fashion.
8. *Inadequate or ambiguous evidence*. If the validity or relevance of the research literature itself is open to question, change will (perhaps rightly) be more difficult to achieve.
9. *Perverse incentives*. Stakeholders may be pulled in a different direction from that required for clinical effectiveness, e.g. by a drug company "research" project or because of an outdated item of service payment scheme.
10. *Intensity of contribution required*. Changing practice requires a lot of enthusiasm, hard work, and long term vision on the part of the project leaders.

"However hard the organisations responsible for producing... effectiveness information try, they cannot change practice themselves. Only health authorities and trusts, and the managers and clinicians who work within them, have the power (and the responsibility) to translate the evidence into real, meaningful and lasting improvements in patient care."[49]

Box 12.3 Lessons from PLIP (Purchaser Led Implementation Projects)[48]

1. *Find organisations and individuals that are ready*. The "soil" must be fertile for the seeds to germinate. Much effort will be wasted, and project workers will become demoralised, if organisations are offered an idea whose time has not yet come.
2. *Pick a good topic*. The ideal topic for a change programme is locally relevant, based on sound evidence, and able to demonstrate tangible benefits in a short time.
3. *Appoint the right project worker and give them protected time*. Drive, personality, motivation, enthusiasm, and non-threatening style are necessary (but not sufficient) characteristics for success. An overworked project worker with conflicting demands on their time and a limited contract is likely to get distracted and start looking for another job.
4. *Locate the project worker within the sphere you want to influence*. If the project is located, for example, in primary care, the project worker needs to be based there. Being seen to be independent of statutory bodies and commercial companies can add credibility and increase goodwill.
5. *Build on existing structures and systems*. If bodies (such as audit advisory groups or educational consortia) already exist and are arranging events, plug into these rather than setting up a separate programme.
6. *Get key stakeholders on board*. Genuine commitment from the "movers and shakers", including funders, opinion leaders, and those in "political" positions, is crucial.
7. *Engage in a constant process of review*. Taking time out to reflect on questions such as "What worked?" "What didn't?" "What have we learnt?", "Where do we go now?", "What would happen if we do X?" and so on is difficult but greatly helps to keep the project on the road.
8. *Tailor your approach*. Flexibility and responsiveness are particularly important when things seem to be going badly; for example, when people say they have insufficient time or resources to deliver on a task. Think of ways of doing things differently, extend deadlines, compromise on the task, offer an extra pair of hands, and so on.
9. *Promote teamwork*. If a project rests entirely on the enthusiasm of a key individual, it will almost certainly flounder when that individual moves on. Getting at least two people to take responsibility, and building a wider team who know and care about the project, will help ensure sustainability.

The report recognises that the task of educating and empowering managers and clinical professionals to use evidence as part of their everyday decision making is massive and offers advice for any organisation wishing to promote the principles of evidence based practice. An action checklist for health care organisations working towards an evidence based culture for clinical and purchasing decisions, listed at the end of Appendix 1, is reproduced from the NAHAT report.[49]

First and foremost, key players within the organisation, particularly chief executives, board members, and senior clinicians, must create an evidence based culture where decision making is *expected* to be based on the best available evidence. High quality, up to date information sources (such as the Cochrane electronic library and the Medline database) should be available in every office and staff given protected time to access them. Ideally, users should only have to deal with a single access point for all available sources. Information on the clinical and cost effectiveness of particular technologies should be produced, disseminated and used together. Individuals who collate and disseminate this information within the organisation need to be aware of who will use it and how it will be applied and tailor their presentation accordingly. They should also set standards for, and evaluate, the quality of the evidence they are circulating.

Individuals on the organisation's internal mailing list for effectiveness information need training and support if they are to make the best use of this information. The projects described in section 12.3 provided important practical lessons, but there is much still to learn about the practicalities of implementing evidence within large (and small) organisations. As the NAHAT report emphasises, separately funded pilot projects on specific clinical issues such as those in GRiPP, PACE or PLIP are useful for demonstrating that change is possible and offering on the job training in implementing evidence, but health authorities and trusts must now move on from this experimental phase and work towards a culture in which clinical and cost effectiveness are part of the routine dialogue between purchasers and providers, and between managers and clinicians.

But changing organisational culture is no simple task. One key step is to create an environment in which the enthusiasm and expertise of the change agents can be harnessed rather than stifled in the organisation as a whole. As Davies and Nutley have pointed out, "Learning is something achieved by individuals, but 'learning organisations' can configure themselves to maximise, mobilise, and retain this learning potential".[50] Drawing on the work of Senge,[51] they offer five key features of a learning organisation.

1. Encouragement to move beyond traditional professional or departmental boundaries (an approach Senge called "open systems thinking").
2. Attention to individuals' personal learning needs.
3. Learning in teams, since it is largely through teams that organisations achieve their objectives.
4. Changing the way people conceptualise issues, hence allowing new, creative approaches to old problems.
5. A shared vision with coherent values and clear strategic direction, so that staff willingly pull together towards a common goal.

12.5 Theories of change

So much of the literature on implementing evidence in the real world of clinical practice is in the form of practical checklists or the "Ten tips for success" type format. Checklists and tips, as you can see from Boxes 12.1, 12.2 and 12.3 and Appendix 1, can be enormously useful, but such lists can leave you hungry for some coherent conceptual models on which to hang your own real life experiences, especially since, as the other chapters in this book suggest, the messages of evidence based medicine are themselves founded on a highly coherent (some would argue *too* coherent!) underpinning theory. As a conference delegate once said to me, "We need a Dave Sackett of change management".

But the management literature offers not one but several dozen different conceptual frameworks for looking at change, leaving the non-expert confused about where to start. It was my attempt to make sense of this multiplicity of theories that led me to write a series of six articles published recently in the *British Journal of General Practice* and entitled "Theories of change". In these articles, I explored six different models of professional and organisational change in relation to effective clinical practice.

1. *Adult learning theory* – the notion that adults learn via a cycle of thinking and doing. This explains why instructional education is so consistently ineffective (see p 182) and why hands on practical experience with the opportunity to reflect and discuss with colleagues is the fundamental basis for both learning and change.[52]
2. *Psychoanalytic theory* – Freud's famous concept of the unconscious, which influences (and sometimes overrides) our conscious, rational self. People's resistance to change can sometimes have powerful and deeprooted emotional explanations.[53]
3. *Group relations theory* – based on studies by specialists at London's Tavistock Clinic on how teams operate (or fail to operate) in the work environment. Relationships both within the team and between the team and its wider environment can act as barriers to (or catalysts of) change.[54]
4. *Anthropological theory* – the notion that organisations have cultures, i.e. ways of doing things and of thinking about problems that are, in general, highly resistant to change. A relatively minor proposed change towards evidence based practice (such as requiring consultants to look up evidence routinely on the Cochrane database) may in reality be highly threatening to the culture of the organisation (in which, for example, the "consultant opinion" has traditionally carried an almost priestly status).[55]
5. *Organisational strategy* – the notion that "mainstreaming" a change within an organisation requires more than one person's desire for it to happen. The vision for change must be shared amongst a critical mass of staff and must usually be accompanied by systematic changes to the visible structures of the organisation, to the roles and responsibilities of key individuals, and to information and communication systems.[56]
6. *Complexity theory* – the notion that large organisations (such as the UK National Health Service) depend crucially on the dynamic, evolving, and local relationships and communication systems between individuals. Supporting key interpersonal relationships and improving the quality and timeliness of information available locally are often more crucial factors in achieving sustained change than "top down" directives or overarching national or regional programmes.[57]

I am not alone in my search for useful theories to explain and promote change. One of the problems in the “instructional” approach to professional behaviour change is the assumption that people behave in a particular way *because (and only because) they lack knowledge* and that imparting knowledge will therefore change behaviour. Theresa Marteau and colleagues’ short and authoritative critique shows that this model has neither theoretical coherence nor empirical support.[58] Information, they conclude, may be *necessary* for professional behaviour change but it is rarely if ever *sufficient*. Psychological theories that might inform the design of more effective educational strategies, suggest Marteau and colleagues, include:

- *behavioural learning* – the notion that behaviour is more likely to be repeated if it is associated with rewards and less likely if it is punished
- *social cognition* – when planning an action, individuals ask themselves “Is it worth the cost?”, “What do other people think about this?” and “Am I capable of achieving it?”
- *stages of change models* – in which all individuals are considered to lie somewhere on a continuum of readiness to change from no awareness that there is a need to change through to sustained implementation of the desired behaviour.[58]

There are, as I have said, many additional theories that might come in useful when identifying barriers and bridges to achieving best clinical practice. The most important advice might be “Don’t try to explain or predict a complex universe using only one of them!”.

12.6 Priorities for further research on the implementation process

Following the success of GRiPP, the UK Department of Health identified further studies into the implementation of evidence based policy as a major spoke of its research and development strategy[59] and made training in research methodology a statutory requirement for all doctors undergoing higher professional training.[60] The 20 priority areas in which new research proposals were specifically invited by the NHS Central Research and Development Committee include the following questions.[61]

- *Who are the players in the implementation process?* Research is needed to examine the relative roles of individuals, professionals, purchasers, providers, the public, the media, commercial organisations, and policymakers in the implementation process.

- *What are the levers of change and barriers to change?* Research could investigate the effectiveness in achieving change of contracts (as used in the UK internal market), financial incentives, professional and regulatory pressures, organisational incentives and disincentives, and structural issues.
- *What interventions can be used to bring about change?* A range of interventions could be explored, including the use of guidelines, clinical audit, feedback, outreach visits, consensus building processes, opinion leaders, patient pressures, process redesign, and decision support or reminder systems.
- *How does the nature of the evidence affect the implementation process?* Additional studies are required into the nature of the evidence underlying current and proposed clinical practices, the availability, strength, and relevance of RCT evidence, the use of observational information, qualitative data, and other non-RCT evidence, the integration of evidence from disparate sources, and the transfer of evidence from one setting to another.

Studies to address these issues are currently ongoing in the UK.

1 Van Someren V. Changing clinical practice in the light of the evidence: two contrasting stories from perinatology. In: Haines A, Donald A, eds. *Getting research findings into practice*. London: BMJ Publications, 1998: 143–51.

2 Anonymous. Effect of corticosteroids for fetal maturation on perinatal outcomes. NIH Consensus Development Panel on the Effect of Corticosteroids for Fetal Maturation on Perinatal Outcomes. *JAMA* 1995; **273**: 413–8.

3 Crowley P. *Corticosteroids prior to preterm delivery* (updated January 1996). Cochrane Database of Systematic Reviews. London: BMJ Publications, 1996.

4 Halliday HL. Overview of clinical trials comparing natural and synthetic surfactants. *Biol Neonate* 1995; **67** (suppl 1): 32–47.

5 Haines A, Donald A. Looking forward: getting research findings into practice: making better use of research findings. *BMJ* 1998; **317**: 72–5.

6 Antmann EM, Lau J, Kupelnick B *et al*. A comparison of the results of meta-analyses of randomised controlled trials and recommendations of clinical experts. *JAMA* 1992; **268**: 240–8.

7 Bero L, Grilli R, Grimshaw J, Harvey E, Oxman A, Thomson MA. Closing the gap between research and practice. In: Haines A, Donald A, eds. *Getting research findings into practice*. London: BMJ Publications, 1998: 27–35.

8 Davis DA, Thomson MA, Oxman AD. Changing physician performance: a systematic review of the effect of CME strategies. *JAMA* 1995; **274**: 700–5.

9 Stanton F, Grant J. *The effectiveness of continuing professional development*. London: Joint Centre for Medical Education, Open University, 1997.

10 Caulford PG, Lamb SB, Kaigas TB, Hanna E, Norman GR, Davis DA. Physician incompetence: specific problems and predictors. *Acad Med* 1993; **270** (suppl): 16–18.

11 Fox RD. *Changing and learning in the lives of physicians*. NewYork: Praeger, 1989.

12 Vernon DT, Blake RL. Does problem-based learning work? A meta-analysis of evaluative research. *Acad Med* 1993; **68**: 550–63.

13 Bashook PG, Parboosingh J. Continuing medical education: recertification and the maintenance of competence. BMJ 1998; **316**: 545–8.

14 Elwyn G. Professional and practice development plans for primary care teams. *BMJ* 1998; **316**: 1619–20. See also ensuing correspondence in *BMJ* 1998; **317**: 1454–5.

15 Koeck C. Time for organisational development in healthcare organisations. *BMJ* 1998; **317**: 1267–8.

16 Freemantle N, Harvey EL, Wolf F, Grimshaw JM, Grilli R, Bero LA. Printed educational materials: effects on professional practice and health care outcomes. In: *The Cochrane Library,* Issue 1. Oxford: Update Software, 2000.

17 Grol R, Dalhuijsen J, Thomas S, Veld C, Rutten G, Mokkink H. Attributes of clinical guidelines that influence use in general practice: observational study. *BMJ* 1998; **317**: 858–61.

18 Report from General Practice 26. *The development and implementation of clinical guidelines*. London: Royal College of General Practitioners, 1995.

19 Robins A, Gallagher A, Rossiter MA, Lloyd BW. Evaluation of joint medical and nursing notes with pre-printed prompts. *Qual Health Care* 1997; **6**: 192–3.

20 Riley K. Care pathways: paving the way. *Health Services J* 1998; **26/3/98**: 30–1.

21 Hunt DL, Haynes RB, Hanna SE, Smith K. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *JAMA* 1998; **280**: 1339–46.

22 Taylor P, Wyatt J. Decision support. In: Haines A, Donald A, eds. *Getting research findings into practice*. London: BMJ Publications 1998: 86–98.

23 Thompson O'Brien MA, Oxman AD, Davis DA, Haynes RB, Freemantle N, Harvey EL. Audit and feedback: effects on professional practice and health care outcomes. In: *The Cochrane Library,* Issue 1. Oxford: Update Software, 2000.

24 Shaughnessy AF, Slawson DC. Pharmaceutical representatives. *BMJ* 1996; **312**: 1494–5.

25 Avorn J, Soumerai SB. A new approach to reducing suboptimal drug use. *JAMA* 1983; **250**: 1728–32.

26 Ray WA. Reducing antipsychotic drug prescribing for nursing-home patients: a controlled trial of the effect of an educational visit. *Am J Pub Health* 1987; **77**: 1448–50.

27 Wyatt J, Paterson-Brown S, Johanson R, Altman DG, Bradburn M, Fisk N. Trial of outreach visits to enhance use of systematic reviews in 25 obstetric units. *BMJ* 1998; **317**: 1041–6.

28 Thompson O'Brien MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Educational outreach visits: effects on professional practice and health care outcomes. In: *The Cochrane Library,* Issue 1. Oxford: Update Software, 2000.

29 Barnes RD, Bell S. Interpractice visits by general practitioners. *Aust Fam Physician* 1994; **23**: 1922–8.

30 Lomas J, Enkin M, Anderson GM, Hannah WJ, Vadya E, Singer J. Opinion leaders vs audit and feedback to implement practice guidelines: delivery after previous cesarean section. *JAMA* 1991; **265**: 2202–7.

31 Thompson O'Brien MA, Oxman AD, Haynes RB, Davis DA, Freemantle N, Harvey EL. Local opinion leaders: effects on professional practice and health care outcomes. In: *The Cochrane Library*, Issue 1. Oxford: Update Software, 2000.

32 Oxman A, Davis D, Haynes RB, Thomson MA. No magic bullets: a systematic review of 102 trials of interventions to help health professionals deliver services more effectively or efficiently. *Can Med Assoc J* 1995; **153**: 1423–43.

33 Effective Health Care Bulletins. Published by Universities of Leeds and York. Nuffield Institute for Health, 71-75 Clarendon Rd, Leeds LS2 9PL.

34 National Childbirth Trust, King's Fund. *Through the maze: a comprehensive guide to sources of research-based information on pregnancy, birth and post-natal care*. Obtainable from the National Childbirth Trust, Alexandra House, Oldham Terrace, Acton, London W3 6HN, price £3. Website http://www.nct-online.org.

35 British Diabetic Association. *What diabetic care to expect*. Obtainable from British Diabetic Association, 10 Queen Anne St, London WC1.

36 Kasper J, Mulley A, Wennberg J. Developing shared decision-making programmes to improve the quality of health care. *Qual Rev Bull* 1992; **18**: 182–90.

37 Coiera E. The Internet's challenge to health care provision. *BMJ* 1996; **312**: 3–4.

38 Oliver S, Entwhistle V, Hodnett E. Roles for lay people in the implementation of health care research. In: Haines A, Donald A, eds. *Getting research findings into practice*. London: BMJ Publications 1998: 43–51.

39 Hurwitz B, Goodman C, Yudkin J. Prompting the care of non-insulin dependent (type II) diabetic patients in an inner city area: one model of community care. *BMJ* 1993; **306**: 624–30.

40 Fulford KWM, Ersser S, Hope T. *Essential practice in patient-centred care*. Oxford: Blackwell Science, 1996.

41 Avorn J, Soumerai SB, Taylor W, Wessels MR, Janousek J, Weiner M. Reduction of incorrect antibiotic dosing through a structured educational order form. *Arch Intern Med* 1988; **148**: 1720–4.

42 Ridsdale L. *evidence based general practice: a critical reader*. London: WB Saunders, 1995: 59–76.

43 Hickson GB, Altemeier WA, Perrin JM. Physician reimbursement by salary or fee-for-service: effect on physician practice behaviour in a randomised prospective trial. *Pediatrics* 1987; **80**: 344–50.

44 Soumerai SB, Ross-Degnan D, Avorn J, McLaughlin TJ, Choodnovskiy I. Effects of Medicaid drug-payment limits on admission to hospitals and nursing homes. *New Engl J Med* 1991; **325**: 1072–7.

45 Greenhalgh T. "Is my practice evidence based?" (editorial) *BMJ* 1996; **313**: 957–8.

46 Dunning M, McQuay H, Milne R. Getting a GRiPP. *Health Services J* 1994; **104**: 18–20.

47 Dunning M, Abi-aad G, Gilbert D, Gillam S, Livett H. *Turning evidence into everyday practice*. London: King's Fund, 1999.

48 Evans D, Haines A, eds. *Implementing evidence based changes in healthcare*. Oxford: Radcliffe Medical Press, 2000.

49 Appleby J, Walshe K, Ham C. *Acting on the evidence: a review of clinical effectiveness: sources of information, dissemination and implementation*. Birmingham: NAHAT, 1995.

50 Davies HTO, Nutley SM. Developing learning organisations in the new NHS. *BMJ* 2000; **320**: 998–1001.

51 Senge P. *The fifth discipline: the art and practice of the learning organisation*. New York: Doubleday, 1994.

52 Greenhalgh T. Change and the individual 1: adult learning theory. *Br J Gen Pract* 2000; **50**: 76–7.

53 Greenhalgh T. Change and the individual 2: psychoanalytic theory. *Br J Gen Pract* 2000; **50**:164–5.

54 Greenhalgh T. Change and the team: group relations theory. *Br J Gen Pract* 2000; **50**: 252–3.

55 Greenhalgh T. Change and the organisation 1: culture and context. *Br J Gen Pract* 2000; **50**: 340–1.

56 Greenhalgh T. Change and the organisation 2: strategy. *Br J Gen Pract* 2000; **50**: 424–5.

57 Greenhalgh T. Change and complexity: the rich picture. *Br J Gen Pract* 2000; **50**: 514–15.

58 Marteau T, Snowden A, Armstrong D. Implementing research findings in practice: beyond the information deficit model. In: Haines A, Donald A, eds. *Getting research findings into practice*. London: BMJ Publications, 1998: 36–42

59 Department of Health. *Research for health*. London: HMSO, 1993.

60 Research and Development Task Force. *Supporting research and development in the NHS* (Culyer Report). London: HMSO, 1994.

61 Advisory Group to the NHS Central Research and Development Committee. *An agenda for the evaluation of methods to promote the implementation of research findings in the NHS*. Leeds: Department of Health, 1995.

Appendix 1: Checklists for finding, appraising, and implementing evidence

Unless otherwise stated, these checklists can be applied to randomised controlled trials, other controlled clinical trials, cohort studies, case-control studies or any other research evidence.

Is my practice evidence based? A context sensitive checklist for individual clinical encounters (see Chapter 1)

1. Have I identified and prioritised the clinical, psychological, social, and other problem(s), taking into account the patient's perspective?
2. Have I performed a sufficiently competent and complete examination to establish the likelihood of competing diagnoses?
3. Have I considered additional problems and risk factors which may need opportunistic attention?
4. Have I, where necessary, sought evidence (from systematic reviews, guidelines, clinical trials, and other sources) pertaining to the problems?
5. Have I assessed and taken into account the completeness, quality, and strength of the evidence?
6. Have I applied valid and relevant evidence to this particular set of problems in a way that is both scientifically justified and intuitively sensible?

7. Have I presented the pros and cons of different options to the patient in a way they can understand, and incorporated the patient's utilities into the final recommendation?
8. Have I arranged review, recall, referral or other further care as necessary?

Checklist for searching Medline or the Cochrane library (see Chapter 2)

1. To look for an article you know exists, search by textwords (in title, abstract or both) or use field suffixes for author, title, institution, journal, and publication year.
2. For a maximally sensitive search on a subject, search under both MeSH headings (exploded) and textwords (title and abstract), then combine the two using the Boolean operator "or".
3. For a focused (specific) search on a clear-cut topic, perform two or more sensitive searches as in (2), and combine them using the Boolean operator "and".
4. To find articles which are likely to be of high methodological quality, insert an evidence based qualifying string for therapeutic interventions, aetiology, diagnostic procedures or epidemiology (see Appendix 2) and/or use maximally sensitive search strategies for randomised trials, systematic reviews, and meta-analyses (see Appendix 3).
5. Refine your search as you go along. For example, to exclude irrelevant material, use the Boolean operator "not".
6. Use subheadings only when this is the only practicable way of limiting your search, since manual indexers are fallible and misclassifications are common.
7. When limiting a large set, browse through the last 50 or so abstracts yourself rather than expecting the software to pick the best half dozen.

Checklist to determine what a paper is about (see Chapter 3)

1. Why was the study done (what clinical question did it address)?
2. What type of study was done?

- Primary research (experiment, randomised controlled trial, other controlled clinical trial, cohort study, case-control study, cross-sectional survey, longitudinal survey, case report or case series)?
- Secondary research (simple overview, systematic review, metaanalysis, decision analysis, guideline development, economic analysis)?

3. Was the study design appropriate to the broad field of research addressed (therapy, diagnosis, screening, prognosis, causation)?
4. Was the study ethical?

Checklist for the methods section of a paper (see Chapter 4)

1. Was the study original?
2. Who is the study about?
   - How were subjects recruited?
   - Who was included in, and who was excluded from, the study?
   - Were the subjects studied in "real life" circumstances?
3. Was the design of the study sensible?
   - What intervention or other manoeuvre was being considered?
   - What outcome(s) were measured, and how?
4. Was the study adequately controlled?
   - If a "randomised trial", was randomisation truly random?
   - If a cohort, case-control or other non-randomised comparative study, were the controls appropriate?
   - Were the groups comparable in all important aspects except for the variable being studied?
   - Was assessment of outcome (or, in a case-control study, allocation of caseness) "blind"?
5. Was the study large enough, and continued for long enough, and was follow-up complete enough, to make the results credible?

Checklist for the statistical aspects of a paper (see Chapter 5)

1. Have the authors set the scene correctly?
   - Have they determined whether their groups are comparable and, if necessary, adjusted for baseline differences?
   - What sort of data have they got and have they used appropriate statistical tests?
   - If the statistical tests in the paper are obscure, why have the authors chosen to use them?
   - Have the data been analysed according to the original study protocol?
2. Paired data, tails, and outliers
   - Were paired tests performed on paired data?
   - Was a two tailed test performed whenever the effect of an intervention could conceivably be a negative one?
   - Were outliers analysed with both common sense and appropriate statistical adjustments?
3. Correlation, regression, and causation
   - Has correlation been distinguished from regression and has the correlation coefficient ("*r* value") been calculated and interpreted correctly?
   - Have assumptions been made about the nature and direction of causality?
4. Probability and confidence
   - Have "*p* values" been calculated and interpreted appropriately?
   - Have confidence intervals been calculated and do the authors' conclusions reflect them?
5. Have the authors expressed their results in terms of the likely harm or benefit which an individual patient can expect, such as:
   - relative risk reduction?
   - absolute risk reduction?
   - number needed to treat?
   - odds ratio?

Checklist for material provided by a pharmaceutical company representative (see Chapter 6)

1. Does this material cover a subject which interests me and is clinically important in my practice?
2. Has this material been published in independent peer reviewed journals? Has any significant evidence been omitted from this presentation or withheld from publication?
3. 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?
4. 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?
5. Has each trial or metaanalysis defined the condition to be treated, the patients to be included, the interventions to be compared, and the outcomes to be examined?
6. Does the material provide direct evidence that the drug will help my patients live a longer, healthier, more productive, and symptom-free life?
7. 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?
8. 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?
9. If large amounts of material have been provided by the representative, which three papers provide the strongest evidence for the company's claims?

Checklist for a paper which claims to validate a diagnostic or screening test (see Chapter 7)

1. Is this test potentially relevant to my practice?
2. Has the test been compared with a true gold standard?
3. Did this validation study include an appropriate spectrum of subjects?
4. Has work up bias been avoided?
5. Has observer bias been avoided?
6. Was the test shown to be reproducible both within and between observers?
7. What are the features of the test as derived from this validation study?
8. Were confidence intervals given for sensitivity, specificity, and other features of the test?
9. Has a sensible "normal range" been derived from these results?
10. Has this test been placed in the context of other potential tests in the diagnostic sequence for the condition?

Checklist for a systematic review or metaanalysis (see Chapter 8)

1. Did the review address an important clinical question?
2. Was a thorough search done of the appropriate database(s) and were other potentially important sources explored?
3. Was methodological quality assessed and the trials weighted accordingly?
4. How sensitive are the results to the way the review has been done?
5. Have the numerical results been interpreted with common sense and due regard to the broader aspects of the problem?

Checklist for a set of clinical guidelines (see Chapter 9)

1. Did the preparation and publication of these guidelines involve a significant conflict of interest?
2. Are the guidelines concerned with an appropriate topic and do they state clearly the goal of ideal treatment in terms of health and/or cost outcome?
3. Was the guideline development panel headed by a leading expert in the field (ideally it should not be) and was a specialist in the methodology of secondary research (e.g. metaanalyst, health economist) involved?
4. Have all the relevant data been scrutinised and do the guidelines' conclusions appear to be in keeping with the data?
5. Do they address variations in medical practice and other controversial areas (e.g. optimum care in response to genuine or perceived underfunding)?
6. Are the guidelines valid and reliable?
7. Are they clinically relevant, comprehensive, and flexible?
8. Do they take into account what is acceptable to, affordable by, and practically possible for patients?
9. Do they include recommendations for their own dissemination, implementation, and periodic review?

Checklist for an economic analysis (see Chapter 10)

1. Is the analysis based on a study which answers a clearly defined clinical question about an economically important issue?
2. Whose viewpoint are costs and benefits being considered from?
3. Have the interventions being compared been shown to be clinically effective?
4. Are the interventions sensible and workable in the settings where they are likely to be applied?
5. Which method of economic analysis was used and was this appropriate?

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

6. How were costs and benefits measured?
7. Were incremental, rather than absolute, benefits compared?
8. Was health status in the "here and now" given precedence over health status in the distant future?
9. Was a sensitivity analysis performed?
10. Were "bottom line" aggregate scores overused?

Checklist for a qualitative research paper (see Chapter 11)

1. Did the article describe an important clinical problem addressed via a clearly formulated question?
2. Was a qualitative approach appropriate?
3. How were (a) the setting and (b) the subjects selected?
4. What was the researcher's perspective and has this been taken into account?
5. What methods did the researcher use for collecting data and are these described in enough detail?
6. What methods did the researcher use to analyse the data and what quality control measures were implemented?
7. Are the results credible and, if so, are they clinically important?

8. What conclusions were drawn and are they justified by the results?
9. Are the findings of the study transferable to other clinical settings?

Checklist for health care organisations working towards an evidence based culture for clinical and purchasing decisions (see Chapter 12)

1. *Leadership*. How often has effectiveness information or evidence based medicine been discussed at board meetings in the last 12 months? Has the board taken time out to learn about developments in clinical and cost effectiveness?
2. *Investment*. What resources is the organisation investing in finding and using clinical effectiveness information? Is there a planned approach to promoting evidence based medicine which is properly resourced and staffed?
3. *Using available resources*. What action has been taken by the organisation in response to official directives requiring organisational support for evidence based practice? What has changed in the organisation as a result?
4. *Implementation*. Who is responsible for receiving, acting on, and monitoring the implementation of Effective Health Care bulletins? What action has been taken on each of the bulletins issued to date?
5. *Clinical guideline*s. Who is responsible for receiving, acting on, and monitoring clinical practice guidelines? Do those arrangements ensure that both managers and clinicians play their part in guideline development and implementation?
6. *Training*. Has any training been provided to staff within the organisation (both clinical and non-clinical) on appraising and using evidence of effectiveness to influence clinical practice?
7. *Contracts*. How often does clinical and cost effectiveness information form an important part of contract negotiation and agreement? How many contracts contain terms which set out how effectiveness information is to be used?

8. *Incentives.* What incentives – both individual and organisational – exist to encourage the practice of evidence based medicine? What disincentives exist to discourage inappropriate practice and unjustified variations in clinical decision making?

9. *Information systems.* Is the potential of existing information systems to monitor clinical effectiveness being used to the full? Is there a business case for new information systems to address the task and is this issue being considered when IT purchasing decisions are made?

10. *Clinical audit.* Is there an effective clinical audit programme throughout the organisation, capable of addressing issues of clinical effectiveness and bringing about appropriate changes in practice?

Appendix 2: Evidence based quality filters for everyday use

1 Therapeutic interventions (What works?)

1 exp clinical trials
2 exp research design
3 randomised 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

2 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

3 Diagnostic procedures

1 exp "sensitivity and specificity"

2 exp diagnostic errors

3 exp mass screening

4 1 or 2 or 3

4 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 3: Maximally sensitive search strings (to be used mainly for research)

1 Maximally sensitive qualifying string for randomised controlled trials

1 RANDOMISED CONTROLLED TRIAL.pt

2 CONTROLLED CLINICAL TRIAL.pt

3 RANDOMISED 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 27 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 seems to be taking an unacceptably long time to run.

2 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 an unacceptably long time to run)

Appendix 4: Assessing the effects of an intervention

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

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

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

Relative risk = CER/EER

Absolute risk reduction (ARR) = CER – EER

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

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

Odds ratio for a particular outcome event =

$$\frac{\text{odds of outcome event vs odds of no event in control group}}{\text{odds of outcome event vs odds of no outcome event in experimental group}}$$

= (a/b)/(c/d)

= ad/bc

*The outcome event can be desirable (e.g. cure) or undesirable (e.g. an adverse drug reaction). In the latter case, it is semantically preferable to refer to the relative or absolute risk *increase*.

Index