BMJ

Assessing the quality of research

Paul Glasziou, Jan Vandenbroucke and Iain Chalmers

BMJ 2004;328;39-41
doi:10.1136/bmj.328.7430.39

Assessing the quality of research

Paul Glasziou, Jan Vandenbroucke, Iain Chalmers

Inflexible use of evidence hierarchies confuses practitioners and irritates researchers. So how can we improve the way we assess research?

The widespread use of hierarchies of evidence that grade research studies according to their quality has helped to raise awareness that some forms of evidence are more trustworthy than others. This is clearly desirable. However, the simplifications involved in creating and applying hierarchies have also led to misconceptions and abuses. In particular, criteria designed to guide inferences about the main effects of treatment have been uncritically applied to questions about aetiology, diagnosis, prognosis, or adverse effects. So should we assess evidence the way Michelin guides assess hotels and restaurants? We believe five issues should be considered in any revision or alternative approach to helping practitioners to find reliable answers to important clinical questions.

Different types of question require different types of evidence

Ever since two American social scientists introduced the concept in the early 1960s,[1] hierarchies have been used almost exclusively to determine the effects of interventions. This initial focus was appropriate but has also engendered confusion. Although interventions are central to clinical decision making, practice relies on answers to a wide variety of types of clinical questions, not just the effect of interventions.[2] Other hierarchies might be necessary to answer questions about aetiology, diagnosis, disease frequency, prognosis, and adverse effects.[3] Thus, although a systematic review of randomised trials would be appropriate for answering questions about the main effects of a treatment, it would be ludicrous to attempt to use it to ascertain the relative accuracy of computerised versus human reading of cervical smears, the natural course of prion diseases in humans, the effect of carriership of a mutation on the risk of venous thrombosis, or the rate of vaginal adenocarcinoma in the daughters of pregnant women given diethylstilboesterol.[4]

To answer their everyday questions, practitioners need to understand the "indications and contraindications" for different types of research evidence.[5] Randomised trials can give good estimates of treatment effects but poor estimates of overall prognosis; comprehensive non-randomised inception cohort studies with prolonged follow up, however, might provide the reverse.

Systematic reviews of research are always preferred

With rare exceptions, no study, whatever the type, should be interpreted in isolation. Systematic reviews are required of the best available type of study for answering the clinical question posed.[6] A systematic review does not necessarily involve quantitative pooling in a meta-analysis.

Although case reports are a less than perfect source of evidence, they are important in alerting us to potential rare harms or benefits of an effective treatment.[7] Standardised reporting is certainly needed,[8] but too few people know about a study showing that more than half of suspected adverse drug reactions were confirmed by subsequent, more detailed research.[9] For reliable evidence on rare harms, therefore, we need a systematic review of case reports rather than a haphazard selection of them.[10] Qualitative studies can also be incorporated in reviews—for example, the systematic compilation of the reasons for non-compliance with hip protectors derived from qualitative research.[11]

References w1-w9 are available on bmj.com

Department of Primary Health Care, University of Oxford, Oxford OX3 7LF
Paul Glasziou *reader*
Leiden University Medical School, Leiden 9600 RC, Netherlands
Jan Vandenbroucke *professor of clinical epidemiology*
James Lind Initiative, Oxford OX2 7LG
Iain Chalmers *editor, James Lind library*
Correspondence to: P Glasziou paul.glasziou@dphpc.ox.ac.uk

BMJ 2004;328:39–41

Level alone should not be used to grade evidence

The first substantial use of a hierarchy of evidence to grade health research was by the Canadian Task Force on the Preventive Health Examination.[12] Although such systems are preferable to ignoring research evidence or failing to provide justification for selecting particular research reports to support recommendations, they have three big disadvantages. Firstly, the definitions of the levels vary within hierarchies so that level 2 will mean different things to different readers. Secondly, novel or hybrid research designs are not accommodated in these hierarchies—for example, reanalysis of individual data from several studies or case crossover studies within cohorts. Thirdly, and perhaps most importantly, hierarchies can lead to anomalous rankings. For example, a statement about one intervention may be graded level 1 on the basis of a systematic review of a few, small, poor quality randomised trials, whereas a statement about an alternative intervention may be graded level 2 on the basis of one large, well conducted, multicentre, randomised trial.

This ranking problem arises because of the objective of collapsing the multiple dimensions of quality (design, conduct, size, relevance, etc) into a single grade. For example, randomisation is a key methodological feature in research into interventions,[13] but reducing the quality of evidence to a single level reflecting proper randomisation ignores other important dimensions of randomised clinical trials. These might include:

- Other design elements, such as the validity of measurements and blinding of outcome assessments
- Quality of the conduct of the study, such as loss to follow up and success of blinding
- Absolute and relative size of any effects seen
- Confidence intervals around the point estimates of effects.

None of the current hierarchies of evidence includes all these dimensions, and recent methodological research suggests that it may be difficult for them to do so.[14] Moreover, some dimensions are more important for some clinical problems and outcomes than for others, which necessitates a tailored approach to appraising evidence.[15] Thus, for important recommendations, it may be preferable to present a brief summary of the central evidence (such as "double-blind randomised controlled trials with a high degree of follow up over three years showed that …"), coupled with a brief appraisal of why particular quality dimensions are important. This broader approach to the assessment of evidence applies not only to randomised trials but also to observational studies. In the final recommendations, there will also be a role for other types of scientific evidence—for example, on aetiological and pathophysiological mechanisms—because concordance between theoretical models and the results of empirical investigations will increase confidence in the causal inferences.[16 17]

What to do when systematic reviews are not available

Although hierarchies can be misleading as a grading system, they can help practitioners find the best relevant evidence among a plethora of studies of diverse quality. For example, to answer a therapeutic question, the hierarchy would suggest first looking for a systematic review of randomised controlled trials. However, only a fraction of the hundreds of thousands of reports of randomised trials have been considered for possible inclusion in systematic reviews.[18] So when there is no existing review, a busy clinician might next try to identify the best of several randomised trials. If the search fails to identify any randomised trials, non-randomised cohort studies might be informative. For non-therapeutic questions, however, search strategies should accommodate the need for observational designs that answer questions about aetiology, prognosis, or adverse effects.[19] Whatever evidence is found, this should be clearly described rather than simply assigned to a level. Such considerations have led the authors of the *BMJ*'s *Clinical Evidence* to use a hierarchy for finding evidence but to forgo grading evidence into levels. Instead, they make explicit the type of evidence on which their conclusions are based.

Balanced assessments should draw on a variety of types of research

For interventions, the best available evidence for each outcome of potential importance to patients is needed.[20] Often this will require systematic reviews of several different types of study. As an example, consider a woman interested in oral contraceptives. Evidence is available from controlled trials showing their contraceptive effectiveness. Although contraception is the main intended beneficial effect, some women will also be interested in the effects of oral contraceptives on acne or dysmenorrhoea. These may have been assessed in short term randomised controlled trials comparing different contraceptives. Any beneficial intended effect needs to be weighed against possible harms, such as increases in thromboembolism and breast cancer. The best evidence for such potential harms is likely to come from non-randomised cohort studies or case-control studies. For example, fears about negative consequences on fertility after long term use of oral contraceptives were allayed by such non-randomised studies. The figure gives an example of how all this information might be amalgamated into a balance sheet.[21 22]

Sometimes, rare, dramatic adverse effects detected with case reports or case control studies prompt further investigation and follow up of existing randomised cohorts to detect related but less severe adverse effects. For example, the case reports and case-control studies showing that intrauterine exposure to diethylstilboestrol could cause vaginal adenocarcinoma led to further investigation and follow up of the mothers and children (male as well as female) who had participated in the relevant randomised trials. These investigations showed several less serious but more frequent adverse effects of diethylstilboestrol that would have otherwise been difficult to detect.[4]

Conclusions

Given the flaws in evidence hierarchies that we have described, how should we proceed? We suggest that there are two broad options: firstly, to extend, improve, and standardise current evidence hierarchies[22]; and, secondly, to abolish the notion of evidence hierarchies and levels of evidence, and concentrate instead on teaching practitioners general principles of research so

that they can use these principles to appraise the quality and relevance of particular studies.[5]

We have been unable to reach a consensus on which of these approaches is likely to serve the current needs of practitioners more effectively. Practitioners who seek immediate answers cannot embark on a systematic review every time a new question arises in their practice. Clinical guidelines are increasingly prepared professionally—for example, by organisations of general practitioners and of specialist physicians or the NHS National Institute for Clinical Excellence—and this work draws on the results of systematic reviews of research evidence. Such organisations might find it useful to reconsider their approach to evidence and broaden the type of problems that they examine, especially when they need to balance risks and benefits. Most importantly, however, the practitioners who use their products should understand the approach used and be able to judge easily whether a review or a guideline has been prepared reliably.

Evidence hierarchies with the randomised trial at the apex have been pivotal in the ascendancy of numerical reasoning in medicine over the past quarter century.[17] Now that this principle is widely appreciated, however, we believe that it is time to broaden the scope by which evidence is assessed, so that the principles of other types of research, addressing questions on aetiology, diagnosis, prognosis, and unexpected effects of treatment, will become equally widely understood. Indeed, maybe we do have something to learn from Michelin guides: they have separate grading systems for hotels and restaurants, provide the details of the several quality dimensions behind each star rating, and add a qualitative commentary (www.viamichelin.com).

	Short term outcomes	Long term outcomes
Benefits		
Contraception	☺Effective (2 controlled trials[w1 w2])	Return to normal fertility soon after cessation (nested case-control and cohort studies[w3])
Dysmenorrhoea	☺Possible reduction in pain and work absence (systematic review of 5 poor quality RCTs[w4])	Not applicable
Harms		
Breast cancer	☹Increased risk: relative risk 1.24 (95% CI 1.15 to 1.33) for current users (individual patient data analysis of 54 observational studies[w5])	No increased risk detected 10 years after cessation (systematic review of 45 observational studies[w5])
Venous thromboembolism	☹Increased risk: 2.5-fold to 6-fold increase in relative risk (systematic review of 5 non-randomised studies) and relative risk 1.1 (0.4 to 2.9 in one RCT with 9898 participants[w6])	Return to background risk after cessation[w6 w7]
Minimal or uncertain effects		
Weight gain	☺No weight gain (3 placebo controlled RCTs of 4-9 months[w8])	Unknown
Heavy menstrual bleeding	☺Insufficient evidence (one, 3 armed RCT with 43 participants[w9])	Not applicable

Example of possible evidence table for short and long term effects of oral contraceptives. (Absolute effects will vary with age and other risk factors such as smoking and blood pressure. RCT = randomised controlled trial)

Summary points

- Different types of research are needed to answer different types of clinical questions
- Irrespective of the type of research, systematic reviews are necessary
- Adequate grading of quality of evidence goes beyond the categorisation of research design
- Risk-benefit assessments should draw on a variety of types of research
- Clinicians need efficient search strategies for identifying reliable clinical research

We thank Andy Oxman and Mike Rawlins for helpful suggestions.

Contributors and sources: As a general practitioner, PG uses the his own and others' evidence assessments, and as a teacher of evidence based medicine helps others find and appraise research. JV is an internist and epidemiologist by training; he has extensively collaborated in clinical research, which made him strongly aware of the diverse types of evidence that clinicians use and need. IC's interest in these issues arose from witnessing the harm done to patients from eminence based medicine.

Competing interests: None declared.

1 Campbell DT, Stanley JC. *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally College, 1963.
2 Sackett DL, Wennberg JE. Choosing the best research design for each question. *BMJ* 1997;315:1636.
3 Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, et al. Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. *JAMA* 2000;284:1290-6.
4 Grant A, Chalmers I. Some research strategies for investigating aetiology and assessing the effects of clinical practice. In: Macdonald RR, ed. *Scientific basis of obstetrics and gynaecology*. 3rd ed. London: Churchill Livingstone, 1985:49-84.
5 Vandenbroucke JP. Observational research and evidence-based medicine: What should we teach young physicians? *J Clin Epidemiol* 1998;51:467-72.
6 Centre for Evidence-Based Medicine. Levels of evidence. www.cebm.net/levels_of_evidence.asp (accessed 13 Nov 2003).
7 Vandenbroucke JP. In defense of case reports and case series. *Ann Intern Med* 2001;134:330-4.
8 Aronson JK. Anecdotes as evidence. *BMJ* 2003;326:1346.
9 Venning GR. Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. *BMJ* 1982;284:249-254.
10 Jenicek M. *Clinical case reporting in evidence-based medicine*. Oxford: Butterworth-Heinemann, 1999.
11 Van Schoor NM, Deville WL, Bouter LM, Lips P. Acceptance and compliance with external hip protectors: a systematic review of the literature. *Osteoporos Int* 2002;13:917-24.
12 Canadian Task Force on the Periodic Health Examination. The periodic health examination. *CMAJ* 1979;121:1193-254.
13 Kunz R, Vist G, Oxman AD. Randomisation to protect against selection bias in healthcare trials (Cochrane methodology review). In: *Cochrane Library*. Issue 2. Oxford: Update Software, 2003.
14 Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282:1054-60.
15 Sterne JA, Juni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in "meta-epidemiological" research. *Stat Med* 2002;21:1513-24.
16 Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295-300.
17 Vandenbroucke JP, de Craen AJ. Alternative medicine: a "mirror image" for scientific reasoning in conventional medicine. *Ann Intern Med* 2001;135:507-13.
18 Mallett S, Clarke M. How many Cochrane reviews are needed to cover existing evidence on the effects of health care interventions. *ACP J Club* 2003;139:A11.
19 Rosendaal FR. Bridging case-control studies and randomized trials. *Curr Control Trials Cardiovasc Med* 2001;2:109-110.
20 Glasziou PP, Irwig LI, An evidence based approach to individualising treatment. *BMJ* 1995;311:1356-9.
21 Skegg DCG. Oral contraception and health. *BMJ* 1999;318:69-70.
22 Schünemann HJ, Best D, Vist G, Oxman AD, for the GRADE Working Group. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. *CMAJ* 2003;169:677-80

(Accepted 20 October 2003)

Summary points

- Infection with herpes simplex virus type 2 is mostly asymptomatic and cannot be cured
- Prevalence by age differs between populations and geographical areas
- Serology tests are commercially available with acceptable sensitivity and specificity
- Use of the tests to screen low prevalence groups would give high rates of false positive results
- Potential biotechnical, medical, epidemiological, psychosocial and ethical advantages and disadvantages must be balanced at both the individual and public health level
- Screening cannot currently be ethically justified

their sexual relationships and also prevent transmission. Concern about infecting a partner is common among those diagnosed, although relationship issues and not infection control seem to be the main cause for this.[w1]

The chosen ethical principles for guidance should be intellectually and emotionally acceptable in the affected society, in our case primarily patients at an STD clinic, their partners, the clinic staff, and policy officials. We do not know whether this is the case, and more information is needed from social science research.

In our opinion, justice as solidarity (see bmj.com) should be paired with autonomy in ethical deliberations of preventive health interventions. If the goal is solidarity rather than conformity, patients must be free to decide what they think is right, because that is what moral responsibility is all about. Without professional truthfulness—the basic tenet of patient involvement in clinical decisions—solidarity could never be accepted as an argument by itself. Patients must understand and feel comfortable with the messages from health institutions. They must also be convinced that reasonable societal support will be available and affordable for those infected with HSV-2 as well as for their partners.

We thank the seminar group in medical ethics, Lund, for helpful comments.

Contributors and sources: IK is a specialist in infectious diseases and professor in epidemiology and public health with an interest in the ethics of public health. G-BL, IK, and BMA are part of a herpes research network in western Sweden. TN is professor in medical ethics. IK and TN had the idea for this paper, and IK wrote the first draft. All authors contributed to the final version from their special fields of competence. The article is based on sources from the Pub Med and medical ethics research literature. They are all guarantors.

Competing interests: None declared.

1 Koutsky LA, Ashley RL, Holmes KK, et al. The frequency of unrecognized type 2 herpes simplex virus infection among women. *Sex Transm Dis* 1990;17(2):90-4.
2 Smith JS, Robinson NJ. Age-specific prevalence of infection with herpes simplex virus types 2 and 1: a global review. *J Infect Dis* 2002;186(suppl 1):3-28.
3 Wald A, Zeh J, Selke S, Warren T, Ryncarz AJ, Ashley R, et al. Reactivation of genital herpes simplex virus type 2 infection in asymptomatic seropositive persons. *N Engl J Med* 2000;342:844-50.
4 Brown ZA, Wald A, Morrow RA, Selke S, Zeh J, Corey L. Effect of serological status and cesarean delivery on transmission rates of herpes simplex virus from mother to infant. *JAMA* 2003;289:203-9
5 Rouse DW, Stringer JSA. An appraisal of screening for maternal type-specific herpes simplex virus antibodies to prevent neonatal herpes. *Am J Obstet Gynecol* 2000;183:400-5.
6 Scoular A. Using the evidence base on genital herpes: optimising the use of diagnostic tests and information provision. *Sex Transm Infect* 2002;78:160-5.
7 Ashley RL. Sorting out the new HSV type specific antibody tests. *Sex Transm Infect* 2001;77:232-7.
8 Fleming DT, McQuillan GM, Johnson JR, Nahmias, Aral OS, Lee FK, et al. Herpes simplex virus type 2 in the United States, 1976 to 1994. *N Engl J Med* 1997;337:1105-11.
9 Roest RW, van der Meijden WI, van Dijk G, Groen J, Mulder PG, Verjans GM, et al. Prevalence and association between herpes simplex virus types 1 and 2-specific antibodies in attendees at a sexually transmitted disease clinic. *Int J Epidemiol* 2001;30:580-8.
10 Nilstun T. *Forskningsetik i vård och medicin.* [Research ethics in care and medicine]. 2nd ed. Lund: Studentlitteratur, 1994.
11 Francoeur RT. *Biomedical ethics. A guide to decision making.* Canada: J Wiley, 1983.
12 Persson K, Månsson A, Jonsson E, Nordenfelt E. Decline of herpes simplex virus type 2 and Chlamydia trachomatis infections from 1970 to 1993 indicated by a similar change in antibody pattern. *Scand J Infect Dis* 1995;27:195-9.
13 Gillon R. Medical ethics: four principles plus attention to scope. *BMJ* 1994;309:184.
14 Beauchamp TL, Childress JF. *Principles of biomedical ethics.* Oxford: Oxford University Press, 2001.
15 Narouz N, Allan PS, Wade AH. Genital herpes: general practitioners' knowledge and opinions. *Sex Transm Infect* 2002;78:198-200.
16 Tranöy KE. *Medicinsk etik i vår tid* [Medical ethics in our time]. Lund: Studentlitteratur, 1993.
17 Fleming DT, Wasserheit JN. From epidemiological synergy to public health policy and practice: the contribution of other sexually transmitted diseases to sexual transmission of HIV infection. *Sex Transm Infect* 1999;75:3-17.
18 Corey L, Handsfield HH. Genital herpes and public health. Addressing a global problem. *JAMA* 2000;283:791-4.
19 Johnson RE, Nahmias AJ, Magder LS, Lee FK, Brooks CA, Snowden CB. A seroepidemiologic survey of the prevalence of herpes simplex type 2 infection in the United States. *N Engl J Med* 1989;321:7-12.

(Accepted 28 June 2004)

Corrections and clarifications

Scandals have eroded US public's confidence in drug industry

In this news article by Jeanne Lenzer (31 July, p 247) we wrongly said that one of the authors of a study on the sources of lead in children living near a smelter failed to acknowledge that her husband owned the smelter in question. In fact, her husband didn't own the smelter; he owned the consulting firm that advised the owner of the smelter.

How protective is the working time directive?

We appeared to slash the salaries of new doctors in this editorial by Rhona MacDonald (7 August, pp 301-2). In the fifth paragraph we dropped a zero by mistake: 12 550 additional UK doctors would cost up to £780m [not £78m] ($1420m; €1150).

Assessing the quality of research

One of the authors of this Education and Debate article by Paul Glasziou and colleagues has advised us that his published name was missing his middle initial (*BMJ* 2004;328:39-41). Jan Vandenbroucke is in fact Jan P Vandenbroucke.

UK health minister under pressure to ban smoking in public

In this news article by Claire Laurent (14 August, p 368), we said that a *BMJ* paper reinforced the findings of earlier studies linking passive smoking to coronary heart disease. Unfortunately, we inadvertently cited the wrong paper. The correct citation is: *BMJ* 2004;329:200-5. The web version and link have been amended.