multiple causes of variation in risk of death that account for the low correlation between the WHO's comprehensive measures of inequality and indices of wealth differences in child mortality.

What is inequality?

Braveman et al believe that health inequalities correlated with factors other than income, social class, and race are not morally important. Citing themselves, they go further and propose that health inequality is defined as the subset of health inequalities correlated with these socioeconomic factors. For a child with an increased risk of death because she lives in a community with a poor immunisation programme and a high prevalence of HIV, it is no solace to know that her risk of death is uncorrelated with income, social class, or race. To most of us, inequality is the state of being unequal. Health inequalities exist when individuals' risks of death and poor health are unequal. The WHO argues that health inequalities should be measured comprehensively. Health scientists can then help determine the causes of inequality and the policies and programmes that can be used to tackle these causes.

Other disciplines such as economics tend to use comprehensive approaches to measuring inequality rather than selective approaches. When economists study income inequality, they do not simply report differences in average income for social class or race groups. Rather, they measure the entire distribution of income across individuals or households and summarise that distribution with measures such as the Gini coefficient. It then becomes a scientific challenge to determine how much is explained by social class or race.

For health, the WHO has adopted the same approach. Firstly, measure the full extent of health inequality in a population. Secondly, use the tools of science to understand what factors explain this inequality. Thirdly, formulate policies that can act on these causes of inequality. Fourthly, monitor and evaluate the impact of these policies on inequality. With this comprehensive approach, an evidence base can be constructed on the causes of health inequality and the policy options available to tackle it.

Competing interests: None declared.

1 Gakidou E, King G. *A framework for measuring health inequality.* World Health Organization, 1999. (Global programme on evidence for health policy discussion paper series No 5.)
2 Wagstaff A. Socioeconomic inequalities in child mortality: comparisons across nine developing countries. *Bull World Health Organ* 2000;78:19-29.
3 Murray CJL, Michaud C, McKenna M, Marks JM. *US patterns of mortality by county and race: 1965-1994.* Cambridge, MA: Harvard School of Public Health and National Center for Disease Prevention and Health Promotion, 1998.
4 Marmot MG, Smith GD, Stansfeld S, Patel C, North F, Head J, et al. Health inequalities among British civil servants: the Whitehall II study. *Lancet* 1991;337:1387-93.

The need for caution in interpreting high quality systematic reviews

Kevork Hopayian

Seahills, Leiston Road, Aldeburgh IP15 5PL
Kevork Hopayian
general practitioner
k.hopayian@btinternet.com

BMJ 2001;323:681–4

The emergence of systematic reviews raised hopes of a new era for the objective appraisal of evidence available on a given topic. Such reviews promised a synthesis of trial results, which could be conflicting, and an escape from the personal bias inherent in traditional reviews and expert opinion.[1] As the discipline of systematic reviews has evolved, however, two new problems have arisen: the quality of reviews is variable[2 3]; and two or more systematic reviews on the same topic may arrive at different conclusions, raising questions on the validity[4–7] or the relevance[8] of the conclusions. Moreover, adherence to a "checklist" system when appraising trials may overlook important clinical details in the original trials and so reduce the validity of the review. I uncovered this last shortcoming when I recently conducted a study of three systematic reviews; the study is reported here.

Summary points

- The discipline of systematic reviews has given clinicians a valuable tool with which to synthesise evidence
- As the methodology of systematic reviews has evolved, the quality of reviews has improved
- Nevertheless, high quality systematic reviews may overlook important clinical details in the papers reviewed, thereby diminishing their validity
- This shortcoming might be avoided if trials were assessed from a clinician's viewpoint as well as from a reviewer's viewpoint

Background

Guidelines have been drawn up to improve the quality of reviews.[9] Differences in the quality of reviews, however, do not always explain discordance. Jadad and McQuay[4] identified six sets of reviews covering six topics in pain research; despite similar quality scores for reviews in each set, four of the sets contained discordant reviews. Jadad et al[8] identified six generic differences between reviews that might lead to discordance: the clinical question asked; the selection and inclusion of studies; data extraction; assessment of study quality; assessment of the ability to combine studies; and statistical methods for data analysis.

Table 1 Summary of systematic reviews assessed for validity

Review	Nelemans et al[15]	Koes et al[13]	Watts[12]
Type of review	Qualitative and quantitative—pooled odds ratios	Qualitative—"vote counting" (significant *v* non-significant studies)	Quantitative—pooled odds ratios
Scoring system for assessing quality of methods used	0-100 (following ter Riet[16])	0-100 (following ter Riet[16])	3-9 (following Chalmers[17])
Result	No evidence for effectiveness	No evidence for effectiveness	Evidence for effectiveness

Box 1: Three part focused question

Population—Patients with sciatica

Intervention—Injection of corticosteroid into the epidural space compared with placebo or injection of local anaesthetic

Outcome—Which intervention leads to quicker pain relief?

The case of epidural steroid injection therapy for sciatica is a good illustration of the evolution of reviews. The results of randomised controlled trials of this treatment were inconsistent. Two traditional reviews of these trials appeared—in 1985[10] and 1986.[11] They reached discordant conclusions. A decade later, two systematic reviews—by Watts and Silagy[12] and Koes et al[13]—also reached discordant conclusions. A comparison of these reviews concluded that the difference in their methods—namely, vote counting versus pooling—explained the discordance.[14] A further systematic review (of all types of injection therapies, including epidural) was published by Nelemans et al for the Cochrane Collaboration in 1999.[15] The three systematic reviews overlap in their nature (qualitative versus quantitative), method for assessing the quality of randomised controlled trials (following that of ter Riet et al[16] or Chalmers et al[17]), and conclusions (table 1). I therefore used them to conduct a general study of the validity of systematic reviews.

Assessing the validity of the three reviews

Background and method

My interest in the epidural steroid injection treatment for sciatica stems from a question arising in general practice and a general practice commissioning board. It was framed as a three part, focused question (box 1).[18] I retrieved the relevant trials that were included in all three reviews and critically appraised each individual paper for validity and relevance to this question.[19 20]

I tried to assess the quality of each systematic review using a validated rating scale, the Oxman and Guyatt index.[21] This tool consists of questions about how the review is designed and reported; it does not require knowledge about the trials themselves. It was inappropriate for two reasons, however, to give scores. Firstly, the scale favours trials that combine data and therefore would have discriminated against Koes et al. Secondly, two of the items on the scale relate to aspects of systematic reviews that I am disputing in this article (see box 2 for comments on the criteria used in each review). The final step was the evaluation of the reviews' treatment of the randomised controlled trials against my own appraisals.

Findings

All three reviews were of high quality according to the Oxman and Guyatt index (box 2). Three problems, however, compromised their validity: the relevance of the study population (inclusion of atypical populations); the appropriateness of the intervention

Box 2: Quality of systematic reviews

Criteria	Nelemans et al[15]	Koes et al[13]	Watts and Silagy[12]
Were the search methods used to find evidence (original research) on the primary questions stated?	Yes	Yes	Yes
Was the search for evidence reasonably comprehensive?	The most comprehensive (Medline and Embase, no language restriction)	Reasonably but the least comprehensive (Medline, restricted to English language only)	Medline, no language restriction.
Were the criteria used for deciding which studies to include in the overview reported?	Yes	Yes	Yes
Was bias in the selection of studies avoided?	Yes	Yes	Yes
Were the criteria used for assessing the validity of the included studies reported?	Yes (scale of 0-100, following ter Riet et al[16])	Yes (scale of 0-100 following ter Riet et al[16])	Yes (scale of 3-9 following Chalmers et al[17])
Was the validity of all the studies referred to in the text assessed using appropriate criteria (either in selecting studies for inclusion or in analysing the studies that are cited)?	Not applicable (issue explored in this article)	Not applicable (issue explored in this article)	Not applicable (issue explored in this article)
Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?	Yes	Yes (but see answer to next question)	Yes
Were the findings of the relevant studies combined appropriately, relative to the primary question that the overview addresses?	Partly, but one of the issues explored in this study was whether combination was reasonable	Difficult to say, as combination with pooling was not attempted; results were used for "vote counting"	Partly, but one of the issues explored in this study was whether combination was reasonable
Were the conclusions drawn by the author(s) supported by the data and/or analysis reported in the overview?	Yes (within the review's own terms)	Yes (within the review's own terms)	Yes (within the review's own terms)

These questions on criteria have been taken from Oxman and Guyatt.[21] A further question ("How would you rate the scientific quality of this overview?") asks the rater to give the review a numerical score.

(inclusion of one study with a serious problem in its design); and the adequacy of the outcome measures (inclusion of studies with inappropriate outcome assessments).

Atypical populations

Both the Koes and the Nelemans reviews included atypical populations—notably patients with pain despite or because of spinal surgery.[22 23] One trial had a high proportion of patients with arachnoiditis,[24] which can be a complication of surgery and of epidural injections when the steroid used is methylprednisolone. These populations are clinically and pathologically distinct from patients with back pain or sciatica who are treated by most clinicians and included in all the other trials.

Although the value of "lumping"—that is, the pooling of results from studies with heterogeneous populations—has been cogently defended,[25] guidelines warn against combining studies that are too heterogeneous.[9] The fundamental differences between most of the randomised controlled trials and the atypical ones means that lumping in this case make no clinical sense.

Flawed design

Koes contended that a design could be "fatally" flawed through the use of a checklist system to score randomised controlled trials: "One of the drawbacks of using this list of methodological criteria might be that trials showing a fatal mistake . . . might end up with a high score because of other criteria."[13]

In the trial by Cuckler et al,[26] for example, this did happen. Patients were assessed 24 hours after receiving either epidural steroid or placebo injections; those who had not improved were given active treatment. This led to contamination of the placebo group, so the analysis by intention to treat 13 months later was not really comparing treatment against placebo. Despite this flaw, the trial was included in all three reviews and received a comparatively high rating in all three, and its results were used in pooling by the two quantitative reviews.

That such papers came to be included suggests that problems exist with systems for scoring the quality of the methods used in trials. Application of the score depends on identifying features of the design and conduct of the trial from a checklist but apparently without the substance of the trial being scrutinised. Numbers are bewitching, and it is tempting to see those scores as objective even though they are the product of human judgment. Comparing the scores given by Nelemans and by Koes to the same papers is illuminating. Despite using the same scoring system, Nelemans et al and Koes et al arrived at different scores for the same papers. They came close to agreement (within 10 points) in only four out of seven papers (table 2).

Table 2 Validity scores (on scale of 0-100, following ter Riet et al[16]) awarded by Nelemans et al and Koes et al for included trials

Trial	Nelemans et al[15]	Koes et al[13]
Beliveau[28]	24	45
Breivik*	54	63
Bush*	40	59
Cuckler[26]	57	62
Mathews*	67	67
Rocco[23]	48	49
Serrao*	23	52

*Details not included here.

Inadequate outcome measures

Several validated tools for assessing outcome for musculoskeletal and back pain research are available, measuring pain, disability, or both.[27] Some of the early primary studies used idiosyncratic tools that fell short of the standards we now expect of modern research. There are two consequences for modern reviews: the results of the older trials are less reliable, and their format means they are not comparable with modern studies. The trials by Beliveau et al (1971)[28] and by Snoek et al (1977)[29] (box 3) used idiosyncratic outcome assessments but were included in the reviews by Watts and by Koes. Both Nelemans and Watts included Beliveau (and Cuckler[26]) in their pooling, which casts doubt on their results. As Messerli said in another context: "A meta-analysis is like a Mediterranean bouillabaisse—in concert, all ingredients will enhance its delightful flavour but, no matter how much fresh fish is added, one rotten fish will make it stink."[30]

That such papers were included shows that little weight is given to the measurement of outcomes, something in which clinicians are especially interested; the system used by Nelemans and by Koes et al allots only five out of 100 marks to assessments of outcome.

Conclusion

Does this mean that no conclusions can be drawn from the original randomised controlled trials? Certainly not. Analysis shows that most trials in this field were conducted at a time when trial methodology was less rigorous than it is now. The poor quality of some trials means that we must disregard their findings, or at least resist the temptation to pool them in a meta-analysis. One trial stands out: the trial by Carette et al[31] was, at the time of the Nelemans review, the most recent, largest, and most rigorous. Nelemans awarded it a quality score of 76%. This trial was the best evidence available

Box 3: Outcome assessments

Trial	Examples of outcome assessments used	Comments
Beliveau[28]	Four categories of outcome: completely relieved, improved, unchanged, and worse. Three criteria had to be met for complete recovery: complete disappearance of pain plus full and free lumbar movements plus "greatly improved" straight leg raising	The vagueness of the criteria leaves them open to the subjectivity of the observer. What are full and free lumbar movements? How many degrees constitute "greatly improved" straight leg raising?
Snoek et al[29]	Divided pain into four categories: back pain, radiating pain, impulse pain, and pain that disturbed sleep. For radiating pain, diminished area of radiation was taken as improvement, whereas for all other categories complete disappearance was necessary	It is the degree not the distribution of pain that matters to a patient. Response in most other trials was graded, rather than complete relief or not. Comparison with other trials was thus impossible

at the time, and therefore we should use its results to inform our decisions. To pool it with others of inferior quality is to accept uncritically that a meta-analysis must be better than a single trial. A large, rigorous trial provides better evidence than a non-credible meta-analysis.

Smith et al[32] drew a distinction between the quality and the validity of randomised controlled trials. Quality relates to the conduct of the trial; the scoring systems mentioned above are among several that aim to measure quality. Validity relates to the ability of the trial to answer the question. We can draw a similar distinction in systematic reviews. The quality of the three systematic reviews is high, but their validity is compromised by overlooking important details in the trials themselves. The fact that these oversights occurred in not just one but all three reviews of the same topic suggests that it may be a general rather than an isolated problem. Clinicians were involved in all three reviews, so the oversights did not arise from a lack of involvement by clinicians. Perhaps it was the type of involvement.

This analysis suggests that reading a paper from a clinician's viewpoint is different from reading a paper from the viewpoint of a reviewer, who has a duty to apply a set of criteria from a checklist. Clinicians, whose usefulness up to now has been seen as "content experts" in systematic review teams, may be able to contribute to the future evolution of systematic reviews by exploring these different viewpoints.

Funding: KH holds a primary care enterprise award from the research and development division of the Eastern regional office of the NHS Executive and has been awarded a grant from the Claire Wand Fund.

Competing interests: None declared.

1 Mulrow C. The medical review article; state of the science. *Ann Intern Med* 1987;106:485-8.
2 Jadad A, Moher M, Browman G, Booker L, Sigouin C, Fuentes M, et al. Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. *BMJ* 2000;320:537-40.
3 Furlan A, Clarke J, Esmail R, Sinclair S, Irvin E, Bombardier C. A critical review of reviews on the treatment of chronic low back pain. *Spine* 2001;26:E155-62.
4 Jadad A, McQuay HJ. Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology. *J Clin Epidemiol* 1996;49:235-43.
5 Prins J, Buller H. Meta-analysis: the final answer, or even more confusion? *Lancet* 1996;348:199.
6 Petticrew M, Kennedy S. Detecting the effects of thromboprophylaxis: the case of the rogue reviews. *BMJ* 1997;315:665-8.
7 Lindback M, Hjortdahl P. How do two meta-analyses of similar data reach opposite conclusions? *BMJ* 1999;318:873-4.
8 Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. *Can Med Assoc J* 1997;156:1411-6.
9 NHS Centre for Reviews and Dissemination. *Undertaking systematic reviews of research on effectiveness. Guidelines for those carrying out or commissioning reviews.* York: NHS Centre for Reviews and Dissemination, University of York, 2001.
10 Kepes E, Duncalf D. Treatment of backache with spinal injections of local anesthetics, spinal and systemic steroids. A review. *Pain* 1985;22:33-47.
11 Benzon H. Epidural steroid injections for low back pain and lumbosacral radiculopathy. *Pain* 1986;224:277-95.
12 Watts R, Silagy C. A meta-analysis on the efficacy of epidural corticosteroids in the treatment of sciatica. *Anaesth Intens Care* 1995;23:564-9.
13 Koes B, Scholten R, Mens J, Bouter L. Efficacy of epidural injections for low-back pain and sciatica: a systematic review of randomized clinical trials. *Pain* 1995;63:279-88.
14 Hopayian K, Mugford M. Conflicting conclusions from two systematic reviews of epidural steroid injections for sciatica: which evidence should general practitioners heed? *Br J Gen Pract* 1999;49(Jan):57-61.
15 Nelemans P, Bie RA de, Vet HCW de, Sturmans F. Injection therapy for subacute and chronic benign low back pain. In: *Cochrane Database of Syst Rev* 2001;(3):CD001824.
16 Ter Riet G, Kleijnen J, Knipschild P. Acupuncture and chronic pain: a criteria based meta-analysis. *J Clin Epidemiol* 1990;43:1191-9.
17 Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroede B, Reitman D, et al. A method for assessing the quality of a randomized control trial. *Controlled Clinical Trials* 1981;2(1):31-49.
18 Richardson W, Wilson M, Nishikawa J, Hayward R. The well-built clinical question: a key to evidence based decisions. *ACP Journal Club* 1995;123:A12-3.
19 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.
20 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;271:59-6.
21 Oxman A, Guyatt G. Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991;44:91-8.
22 Dallas T, Lin R, Wu W, Wolskee P. Epidural morphine and methylprednisolone for low-back pain. *Anesthesiology* 1987;67:408-11.
23 Rocco A, Frank E, Kaul A, Lipson S, Gallo J. Epidural steroids, epidural morphine and epidural steroids combined with morphine in the treatment of post-laminectomy syndrome. *Pain* 1989;36:29 7-303.
24 Glynn C, Dawson D, Sanders R. A double-blind comparison between epidural morphine and epidural clonidine in patients with chronic non-cancer pain. *Pain* 1988;34:123-8.
25 Gøtzsche P. Why we need a broad perspective on meta-analysis. *BMJ* 2000;321:585-6.
26 Cuckler JM, Bernini PA, Wiesel SW, Booth RE Jr, Rothman RH, Pickens GT. The use of epidural steroids in the treatment of lumbar radicular pain. A prospecitive, randomized, double-blind study. *J Bone Joint Surg Am* 1985;67(1):63-6.
27 Ruta D, Garratt A, Wardlaw D, Russell I. Developing a valid and reliable measure for health outcome for patients with low back pain. *Pain* 1994;19:1187-96.
28 Beliveau P. A comparison between epidural anaesthesia with and without corticosteroid in the treatment of sciatica. *Rheum Phys Med* 1971;11:40-3.
29 Snoek W, Weber H, Jørgensen B. Double blind evaluation of extradural methyl prednisolone for herniated lumbar discs. *Acta Orthop Scand* 1977;48:635-41.
30 Messerli F. Meta-analysis. Are calcium antagonists safe? *Lancet* 1985:767-8.
31 Carette S, Leclaire R, Marcoux S, Morin F, Blaise G, St Pierre A, et al. Epidural corticosteroid injections for sciatica due to herniated nucleus pulposus. *N Engl J Med* 1997;336:1634-40.
32 Smith AS, Oldman A, McQuay H, Moore R. Teasing apart quality and validity in systematic reviews: an example from acupuncture trials in chronic neck and back pain. *Pain* 2000;86:119-32.

(Accepted 25 June 2001)

Corrections and clarifications

Implementing clinical governance: turning vision into reality

Unfortunate results can ensue when even a single vowel is mistyped. We assigned nine competing interests to Aidan Halligan and Liam Donaldson, the authors of this article (9 June, pp 1413-7). We can reassure readers (especially those who noticed this unlikely declaration), however, that both authors completed and signed our rigorous competing interests form saying "None declared."

Influence of variation in birth weight within normal range and within sibships on IQ at age 7 years: cohort study

In this article by Thomas D Matte and colleagues (11 August, pp 310-4) the same editing error slipped into two tables and persisted to final publication. The value 1.0, which appeared four times in table 3 and twice in table 4, was wrong. The entry in each case should read "Reference 0."

Birmingham trust criticised

In shortening this news article by Cherrill Hicks (4 August, p 249), we inadvertently deleted a couple of important words, leading to a shift in meaning. Referring to a report of the clinical governance review at Northern Birmingham Mental Health NHS Trust, we said that this was the first report [rather than one of the first reports] to cover mental health services. It is in fact the second report—the first was of the Wrightington, Wigan and Leigh Trust.