Bandolier Professional

Independent evidence-based health care

On quality and validity

If studies are not done properly, any results they produce will be worthless. We call this validity. What constitutes a valid study depends on many factors; there are no absolute hard and fast rules that can cover every clinical eventuality. Validity has a dictionary definition of "sound and defensible".

But the quality of study conduct and reporting is also important, because incorrect conduct can introduce bias. Bias has a dictionary definition of "a one-sided inclination of the mind", and studies with bias may be wrong even if the study is valid.

This essay looks at some of the issues of quality and validity in clinical trials, in meta-analysis, and in epidemiology. It will be in the nature of an overview, because some aspects would require a more detailed examination. The essay will also look at several checklists that can help us examine quality and validity issues in papers we read and want to critically appraise.

Clinical trial quality

Randomisation

We randomise trials to exclude selection bias. Trials are usually performed where there is uncertainty as to whether a treatment works (is better than no treatment or placebo), or whether one treatment is better than another. We start from a position of intellectual equipoise. But trials are often done by believers, and belief, even subconscious belief, might influence our choice of patients to have one treatment or another if we could choose. To avoid this, and to ensure that the two groups of patients are identical, we make the choice randomly. This might be by tossing a coin, or more often by computer-generated randomisation. If we do not randomise we can end up with groups that are not the same, thus invalidating the trial, or with a trial that no-one will believe because trials that are not randomised are often shown to be wrong [1, 2].

For instance, in a systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain, the majority of non-randomised trials concluded that it was effective, while almost all properly randomised trials concluded that it was not effective [2]. Again, in studies of hyperbaric oxygen for multiple sclerosis, randomised studies showed no effect, while one non-randomised comparative study and four case series showed benefit [3].

Table 1: Randomisation and TENS in postoperative pain

	Analgesic result	
Trial design	Positive	Negative
Randomised	2	15
Not properly randomised	17	2

Blinding

We conduct trials blind to minimise observer bias. It's belief again, because even if randomised, we know that Mrs Jones has treatment A and Mr Smith treatment B, our observations may be biased by our belief that Mr Smith overstates his complaint and Mrs Jones understates hers. Only if we have no idea which treatment they received will we be free from a bias that is known to deliver incorrect results (again, see later). Blinding is essential in almost all cases [1].

A meta-analysis of acupuncture for back pain concluded that it was effective [4]. But though the analysis was restricted to randomised trials with valid acupuncture technique independently verified by experts, it included open as well as blind studies. The beneficial effect was found only in the open studies (Table 2).

For clinical trial quality, randomisation and blinding are the two most obvious issues [1], but there are others, including duplication [5], small trials [6], and reporting quality [7,8]. Table 3 shows estimates of how absence of these factors may result in over-estimating treatment effects.

Table 2: Blinding and acupuncture for back pain

	Percent improved with		
Type of study	acupuncture	control	NNT (95%CI)
Blind	57	50	13 (5 to no benefit)
Non-blind	67	38	3.5 (2.4 to 6.5)

Table 3: Factors tending to overestimate treatment effect

Factor	Percentage overestimation of treatment effect
Not randomised	40
Not double-blind	17
Including duplicate information	20
Using only small trials	30
Trials of poor reporting quality	25

Scoring trials for bias

These are the major sources of bias in clinical trials of efficacy. The frequently used Oxford five-point scoring system [9] uses three criteria, full details of which are given in Table 4.

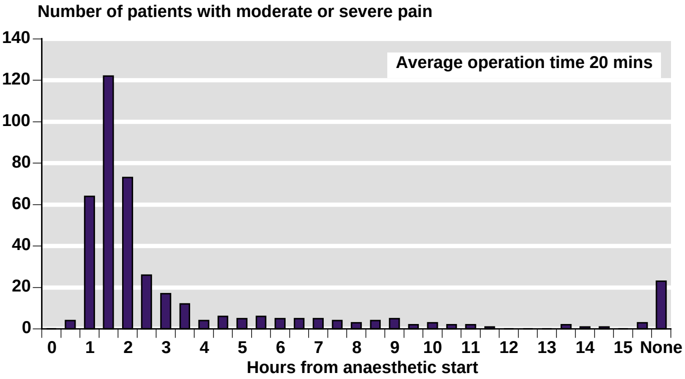
- Is the trial randomised (1 point). Additional point if method is given and appropriate?
- Is the trial double-blind (1 point). Additional point if method given and appropriate?
- Were withdrawals and dropouts described and assigned to different treatments (1 point)?

Trials that scored 3 or more were relatively free of bias and could be trusted. Lower scores were shown to be associated with increased treatment effects - they were biased [7,8].

Clinical trial validity

What constitutes trial validity is more difficult. We know it when we see it, or more particularly when our eyes are opened to its absence. Validity is always situation dependent, but criteria might include severity of illness at the start of a trial, the dose or intensity of the intervention, the duration of the intervention or the duration of observation. The first are fairly obvious. Examples of the importance of the duration of an intervention might be seen in a trial of TENS in chronic pain, where a single 10-minute low intensity period was tried, in the face of evidence that it takes months to work. The importance of duration of observations can be seen in, for instance, 5-alpha-reductase inhibitors in benign prostatic hyperplasia, where effects on prostate volume or urine flow continue to improve over several years [10] rather than weeks or months.

Figure 1: Moderate or severe pain after minor orthopaedic surgery



An easy example is that of analgesic trials. In the absence of pain, how do you measure the effect of an analgesic? With difficulty, of course. So, especially in acute pain, patients have to have pain of at least moderate intensity before they can be given a test intervention. Figure 1 shows what happened to 410 patients having minor orthopaedic surgery that lasted about 20 minutes [11]. A trained nursing sister was with them so that when pain was of at least moderate intensity they could enter a clinical trial. Most of them needed an analgesic by about three hours, but some didn't need analgesic until 12 hours or more had elapsed, and 23 (6%) didn't need any analgesia at all. But unless that process was carried out, some patients would have entered analgesic trials who did not have, or would not have had, any pain.

Scoring trials for validity

The Oxford Pain Validity Scale (OPVS) was designed specifically to examine issues regarding validity in pain trials [12] and is described in detail in Table 5. It uses eight criteria (16 points total) to be applied to randomised trials. The criteria include blinding, size, statistics, dropouts, credibility of statistical significance and authors' conclusions, baseline measures and outcomes to examine whether a trial might be considered valid or not.

The scale was applied in a systematic review of back and neck pain. In 13 trials the conclusions of the original authors were found to be incorrect in two cases (and as an aside, this is not an uncommon problem in clinical trial reporting, or even systematic reviews, where frank errors change the results). Using reviewers' conclusion, more valid trials were significantly more likely to have a negative conclusion (Figure 2).

Figure 2: OPVS scale applied to trials of acupuncture for chronic neck and back pain

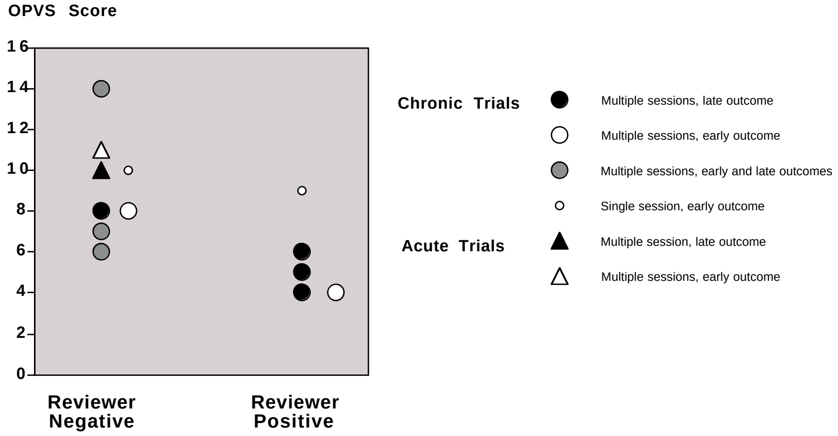


Table 4: The Oxford system for quality scoring controlled trials

Oxford scale for quality scoring controlled trials

This is not the same as being asked to review a paper. It should not take more than 10 minutes to score a report and there are no right or wrong answers.

Please read the article and try to answer the following questions (see attached instructions):

1 Was the study described as randomised (this includes the use of words such as randomly, random and randomisation)?
2 Was the study described as double-blind?
3 Was there a description of withdrawals and drop outs?

Scoring the items:

Give a score of 1 point for each 'yes' and 0 points for each 'no'. There are no in-between marks.

Give 1 additional point if:

On question 1, the method of randomisation was described and it was appropriate (table of random numbers, computer generated, coin tossing, etc.)
and/or:
If on question 2 the method of double-blinding was described and it was appropriate (identical placebo, active placebo, dummy, etc.)

Deduct 1 point if:
On question 1, the method of randomisation was described and it was inappropriate (patients were allocated alternatively, or according to date of birth, hospital number, etc.)
and/or:
On question 2 the study was described as double-blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy)

Advice on using the scale

1. Randomisation:

If the word randomised or any other related words such as random, randomly, or randomisation are used in the report, but the method of randomisation is not described, give a positive score to this item. A randomisation method will be regarded as appropriate if it allowed each patient to have the same chance of receiving each treatment and the investigators could not predict which treatment was next. Therefore methods of allocation using date of birth, date of admission, hospital numbers or alternation should not be regarded as appropriate.

2. Double-blinding:

A study must be regarded as double-blind if the word double-blind is used (even without description of the method) or if it is implied that neither the caregiver nor the patient could identify the treatment being assessed.

3. Withdrawals and drop outs:

Patients who were included in the study but did not complete the observation period or who were not included in the analysis must be described. The number and the reasons for withdrawal must be stated. If there are no withdrawals, it should be stated in the article. If there is no statement on withdrawals, this item must be given a negative score (0 points).

Table 5: The Oxford pain validity scale (OPVS) combining trial quality and validity

The Oxford Pain Validity Scale (OPVS) should be only be used on trials which are:
*** RANDOMISED**
*** have a start group size ≥10 for all groups relevant to the review question**

ITEM		SCORE (circle one number per item)	COMMENTS
1. Blinding	Was the trial convincingly double-blind?	6	i.e. states double-blind and how this was achieved, eg double-dummy, identical appearance, etc.
	Was the trial convincingly single-blind or unconvincingly double-blind?	3	i.e. states single-blind and how this was achieved, eg observer-blind, patient-blind, etc.
	Was the trial either not blind or the blinding is unclear?	0	
2. Size of trial groups	Was the start group start size ≥40?	3	Not all groups in the trial will necessarily be relevant to the review question. Rate this item using the smallest group that is relevant to the review question.
	Was the start group start size 30 to 39?	2	
	Was the start group start size 20 to 29?	1	
	Was the start group start size 10 to 19?	0	
3. Outcomes	Look at pre hoc list of most desirable outcomes relevant to the review question:		NB: If the trial has not reported the results of any measures relevant to the review question (even if it described them in methods) it should be excluded from the review.
	Did the paper include results for at least one pre hoc desirable outcome, and use the outcome appropriately?	2	
	There were no results for any of the pre hoc desirable outcomes, or, a pre hoc desirable outcome was used inappropriately.	0	
4. Demonstration of internal sensitivity	Look at the baseline levels for the outcomes relevant to the review question:		One way to demonstrate internal sensitivity is by having an additional active control group in the trial which demonstrates a significant difference from placebo (ie the trial design is able to detect a difference). For example, by having an extra group treated with an anlagesic known to be statistically different from placebo, and by demonstrating this difference. Alternatively, internal sensitivity can be demonstrated with a dose response.
	For all treatment groups, baseline levels were sufficient for the trialist to be able to measure a change following the intervention (eg enough baseline pain to detect a difference between baseline and post-treatment levels). Altertatively, did the trial demonstrate internal sensitivity?	1	
	For all treatment groups, baseline levels were insufficient to be able to measure a change following the intervention, or, baseline levels could not be assessed, or internal sensitivity was not demonstrated.	0	
5. Data Analysis	i) Definition of outcomes		There must be at least one outcome measure defined clearly to score 1. This item refers to any outcome measure relevant to the review question, not just pre hoc desirables.
	Did the paper define the relevant outcomes clearly, including where relevant, exactly what 'improved', 'successful treatment', etc represented?	1	
	The paper failed to define the outcomes clearly.	0	
	ii) Data presentation: Location and dispersion		
	Did the paper present either mean data with standard deviations, or dichotomous outcomes, or median with range, or sufficient data to enable extraction of any of these?	1	
	The paper presented none of the above	0	
	iii) Statistical Testing		Corrections for multiple testing must be put in place when a series of tests or measures have been carried out on the same patient group.
	Did the trialist choose an appropriate statistical test, with correction for multiple tests where relevant?	1	
	Inappropriate statistical tests were chosen and/or multiple testing was carried out, but with no correction, or, no statistics were carried out.	0	
	iv) Handling of Dropouts		
	The dropout rate was either ≤10%, or was >10% and includes an intention-to-treat analysis in which dropouts were included appropriately.	1	
	The dropout rate was >10% and dropouts were not included in the analysis, or, it is not possible to calculate a dropout rate from data presented in the paper.	0	
	TOTAL SCORE		

Equivalence trials

Studies of analgesics of an A versus B design are notoriously difficult to interpret, but we have guidance of what to expect from equivalence trials, with useful guides about what features of equivalence trials are important in determining their validity [13]. The intellectual problem with equivalence (A versus B) trials is that the same result is consistent with three conclusions:

- Both A and B are equally effective
- Both A and B are equally ineffective
- Trials inadequate to detect differences between A and B

To combat the problems posed by the latter two conclusions, McAlister & Sackett [13] suggest several criteria in addition to those used for superiority trials (A and/or B versus placebo). These are shown in Table 6, with the following expansion:

Control shown previously to be effective?

Ideally documented in a systematic review of placebo controlled trials with benefits exceeding a clinically important effect. Without this information both may be equally ineffective.

Patients and outcomes similar to original trials?

Obvious, this one. If they are not, then any conclusion about equivalence is doomed. Beware, though, trials designed to show equivalent efficacy being used to demonstrate differences in harm or toxicity, for which they were not powered.

Regimens applied in identical fashion?

The most common example is that of choosing the best dose of A versus an ineffective dose of B (no names, no pack drill, but no prizes for picking out numerous examples especially from pharmaceutical company sponsored trials showing "our drug is better than yours"). Should be OK if licensed doses are chosen.

Other pitfalls to look out for are low compliance or frequent treatment changes, incomplete follow up, disproportionate use of cointerventions and lack of blinding.

Appropriate statistical analysis?

Equivalence trials are designed to rule out meaningful differences between two treatments. Often one-sided tests of difference are used. Lack of significant superiority is not necessarily the same as defining an appropriate level of equivalence and testing for it.

Intention to treat analysis confers the risk of making a false-negative conclusion that treatments have the same efficacy when they do not. In equivalence trials the conservative approach may be to compare patients actually on treatment. Both analyses should probably be used.

Prespecified equivalence margin?

How different is different? Equivalence trials should have a prior definition of how big a difference is a difference, and justify it. Even more than that, they have to convince you that the lack of that difference means that treatments would, in fact, be equivalent.

Size?

Most equivalence trials do not have enough power to detect even a 50% difference between treatments, and a 1994 review [14] found that 84% were too small to detect a 25% difference. Size is everything when we want to show no difference, and the smaller the difference that is important, the larger the trial has to be.

Table 6: Criteria for validity in superiority and active-control equivalence trials

Superiority trials	Active-control equivalence trials
Randomised allocation	Randomised allocation
Randomisation concealed	Randomisation concealed
All patients randomised accounted for	All patients randomised accounted for
Intention to treat analysis	Intention to treat analysis **and on-treatment analysis**
Clinicians and patients blinded to treatment received	Clinicians and patients blinded to treatment received
Groups treated equally	Groups treated equally
Groups identical at baseline	Groups identical at baseline
Clinically important outcomes	Clinically important outcomes
	Active control previously shown to be effective
	Patients and outcomes similar to trials previously showing efficacy
	Both regimens applied in an optimal fashion
	Appropriate null hypothesis tested
	Equivalence margin pre-specified
Trial of sufficient size	Trial of sufficient size

When can we say that drugs have a "class effect"?

Class (noun); "*any set of people or things grouped together or differentiated from others*". An increasingly asked question is that of whether a set of drugs forms a class, and whether there is a 'class effect'. Class effect is usually taken to mean similar therapeutic effects and similar adverse effects, both in nature and extent. If such a 'class effect' exists, then it makes decision-making easy: you choose the cheapest.

Criteria for drugs to be grouped together as a class involve some or all of the following:

- Drugs with similar chemical structure
- Drugs with similar mechanism of action
- Drugs with similar pharmacological effects

Declaring a class effect requires a bit of thought, though. How much thought, and of what type, has been considered in one of that brilliant JAMA series on users guides to the medical literature [15]. No one should declare a class effect and choose the cheapest without reference to the rules of evidence set out in this paper.

Class: levels of evidence for efficacy

These are shown in Table 7, though if it comes down to levels 3 and 4 evidence for efficacy, the ground is pretty shaky. Level 1 evidence is what we always want and almost always never get, the large randomised head to head comparison. By the time there are enough compounds around to form a class, there is almost no organisation interested in funding expensive, new, trials to test whether A is truly better than B.

Most of the time we will be dealing with randomised trials of A versus placebo or standard treatment and B versus placebo or standard treatment. This will be level 2 evidence based on clinically important outcomes (a healing event) or validated surrogate outcomes (reduction of cholesterol with a statin). So establishing a class effect will likely involve quality systematic review or meta-analysis of quality randomised trials.

What constitutes quality in general is captured in Table 7, though there will be some situation dependent factors. The one thing missing from consideration in Table 7 is size. There probably needs to be some prior estimate of how many patients or events constitutes a reasonable number for analysis.

Class: levels of evidence for safety

These are shown in Table 8. There are always going to be problems concerning rare, but serious, adverse events. The inverse rule of three tells us that if we have seen no serious adverse events in 1500 exposed patients, then we can be 95% sure that they do not occur more frequently than 1 in 500 patients.

Randomised trials of efficacy will usually be underpowered to detect rate, serious adverse events, and we will usually have to use other study designs. In practice the difficulty will be that soon after new treatments are introduced there will be a paucity of data for these other types of study. Only rarely will randomised trials powered to detect rare adverse events be conducted.

Most new treatments are introduced after being tested on perhaps a few thousand patients in controlled trials. Caution in treatments for chronic conditions are especially difficult if trials are only short-term, and where other diseases and treatments are likely.

Table 7: Levels of evidence for safgety for class effect

Level	Type of study	Advantages	Criteria for validity
1	RCT	Only design that permits detection of adverse effects when the adverse effect is similar to the event the treatment is trying to prevent	Underpowered for detecting adverse events unless specifically designed to do so
2	Cohort	Prospective data collection, defined cohort	Critically depends on follow up, classification and measurement accuracy
3	Case-control	Cheap and usually fast to perform	Selection and recall bias may provide problems, and temporal relationships may not be clear.
4	Phase 4 studies	Can detect rare but serious adverse events if large enough	No control or unmatched control Critically depends on follow up, classification and measurement accuracy
5	Case series	Cheap and usually fast	Often small sample size, selection bias may be a problem, no control group
6	Case report(s)	Cheap and usually fast	Often small sample size, selection bias may be a problem, no control group

Table 8: The Oxman & Guyatt scoring system for systematic review

The Oxman & Guyatt index of scientific quality

1	Were the search methods used to find evidence on the primary question(s) stated?	No	Partially	Yes
2	Was the search for evidence reasonably comprehensive?	No	Can't tell	Yes
3	Were the criteria used for deciding which studies to include in the overview reported?	No	Partially	Yes
4	Was bias in the selection of studies avoided?	No	Can't tell	Yes
5	Were the criteria used for assessing the validity of the included studies reported?	No	Partially	Yes
6	Was the validity of all the studies referred to in the text assessed using appropriate criteria ?	No	Can't tell	Yes
7	Were the methods used to combine the findings of the relevant studies (to reach a conclusion) reported?	No	Partially	Yes
8	Were the findings of the relevant studies combined appropriately relative to the primary question of the overview ?	No	Can't tell	Yes
9	Were the conclusions made by the author(s) supported by the data and/or analysis reported in the overview?	No	Partially	Yes
10	How would you rate the scientific quality of this overview?			

Flaws

Extensive		Major		Minor		Minimal
1	**2**	**3**	**4**	**5**	**6**	**7**

Scoring system for reviews

The purpose is to evaluate the scientific quality (i.e., adherence to scientific principles) of research overviews (review articles) published in the medical literature. It is not intended to measure literary quality, importance, relevance, originality, or other attributes of overviews.

The index ([16], Table 8) is for assessing overviews of primary ("original") research on pragmatic questions regarding causation, diagnosis, prognosis, therapy, or prevention. A research overview is a survey of research. The same principles that apply to epidemiological surveys apply to overviews; a question must be clearly specified, a target population identified and accessed, appropriate information obtained from that population in an unbiased fashion, and conclusions derived, sometimes with the help of formal statistical analysis, as is done in "meta-analyses". The fundamental difference between overviews and epidemiological surveys is the unit of analysis, not scientific issues that the questions in this index address.

Since most published overviews do not include a methods section it is difficult to answer some of the questions in the index. Base your answers, as much as possible, on information provided in the overview. If the methods that were used are reported incompletely relative to a specific item, score that item as "partially". Similarly, if there is no information provided regarding what was done relative to a particular question, score it as "can't tell", unless there is information in the overview to suggest either that the criterion was or was not met.

For Question 8, if no attempt has been made to combine findings, and no statement is made regarding the inappropriateness of combining findings, check "no". If a summary (general) estimate is given anywhere in the abstract, the discussion, or the summary section of the paper, and it is not reported how that estimate was derived, mark "no" even if there is a statement regarding the limitations of combining the findings of the studies reviewed. If in doubt mark "can't tell".

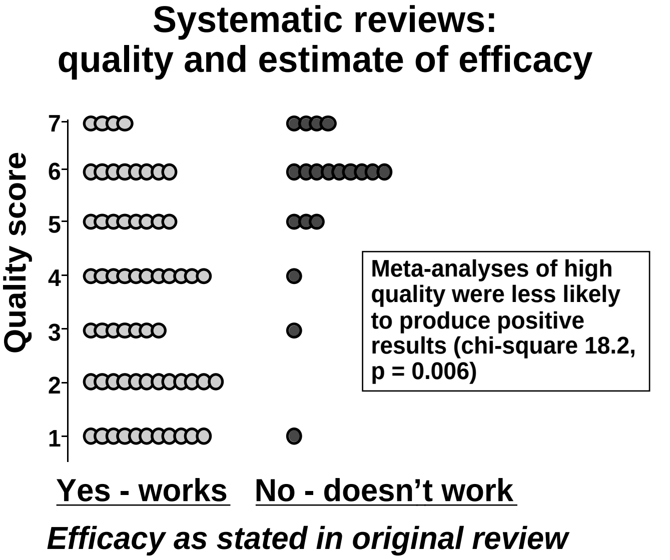
For an overview to be scored as "yes" on Question 9, data (not just citations) must be reported that support the main conclusions regarding the primary question(s) that the overview addresses.

The score for Question 10, the overall scientific quality, should be based on your answers to the first nine questions. The following guidelines can be used to assist with deriving a summary score: If the "can't tell" option is used one or more times on the preceding questions, a review is likely to have minor flaws at best and it is difficult to rule out major flaws (i.e., a score of 4 or lower). If the "no" option is used on Questions 2, 4, 6 or 8, the review is likely to have major flaws (i.e., a score of 3 or less, depending on the number and degree of the flaws).

The guide has been applied to systematic reviews of pain topics [17], and the results are worth noting because they teach us much about the need for scepticism when reading systematic reviews. Seventy reports were included in the quality assessment. The earliest report was from 1980. Over two thirds appeared after 1990. Reviews considered between two and 196 primary studies (median 28). Sixty reviews reached positive conclusions, seven negative, twelve uncertain and one did not manage any conclusion. All were based on published data only (no individual patient data analysis), without validity checks with the study investigators.

The median agreed overall Oxman & Guyatt score for the systematic reviews was 4 (range 1 to 7). Systematic reviews of high quality were significantly less likely to produce positive results (Figure 3). Sixteen of 19 systematic reviews with negative or uncertain results had overall quality scores above the median, compared with only 20 of the 60 with positive results. Systematic reviews restricted to RCTs were significantly less likely to produce positive conclusions (19 of 31) than those which included other study architectures (41 of 49). All conclusions from systematic reviews of psychological interventions were positive. In only one of those reviews was quality scored above the median. All abstracts scored below the median, and 6 out of 8 abstracts received the minimum possible score.

Figure 3: The Oxman & Guyatt system applied to systematic reviews with pain as an outcome



Meta-analyses of high quality were less likely to produce positive results (chi-square 18.2, p = 0.006)

Jadad & McQuay J Clin Epidemiol 1996

Quality scoring in observational studies

This is an area where there are few clear guidelines on what constitutes a quality study, and where we have little evidence that higher quality produces different results from those of lower quality. A useful guide was used in a meta-analysis of studies examining the association between homocysteine and coronary heart disease risk [18].

Table 10 shows the scoring system as it was used in the study. It asks questions about study design, response rate, exclusion criteria, the type of controls and matching and adjustment for confounders. In the analysis, studies with higher quality scores (7/10 or more) were associated with lower risk than lower quality studies (6/10 or less), but not significantly so. The scoring system could be adjusted for use in other settings, especially by thinking about the types of controls and confounders that would be appropriate.

Reporting standards

We now have a series of guidelines about how randomised trials [19], systematic reviews [20], and epidemiological studies and their meta-analyses [21] should be reported. Though all aspects of the guidelines are unlikely to be followed in every instance, the application of the guidelines would make for far fewer poor quality reports, and even fewer reports in total if journals applied the guidelines. These are always likely to be updated, but cover many of the issues in this essay.

Table 10: A quality scoring system for observational studies

Criterion	Score
Study design	not given = 0
	Cross-sectional study, series, or angiographic = 1
	Case-control =2
	nested case-control =3
Response rate	less than 75% =1
	more than 75% = 2
Exclusion criteria	not given = 0
	criteria specified = 1
Type of controls	hospital or mixed = 1
	community = 2
Matching, adjustment for confounders	none = 0
	any confounder = 1
	age, smoking, hypertension, and cholesterol = 2

References:

1 KF Schulz, et al. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995 273:408-412.

2 D Carroll et al. Randomization is important in studies with pain outcomes: Systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. British Journal of Anaesthesia 1996 77:798-803.

3 M Bennett & R Heard. Treatment of multiple sclerosis with hyperbaric oxygen therapy. Undersea Hyperbaric Medicine 2001 28:117-122.

4 E Ernst & AR White. Acupuncture for back pain. Archives of Internal Medicine 1998 158:2235-2241.

5 M Tramèr et al. Impact of covert duplicate publication on meta-analysis: a case study. British Medical Journal 1997 315:635-639.

6 RA Moore et al. Quantitative systematic review of topically-applied non-steroidal anti-inflammatory drugs. British Medical Journal 1998 316:333-338.

7 KS Khan et al. The importance of quality of primary studies in producing unbiased systematic reviews. Archives of Internal Medicine 1996 156:661-666.

8 D Moher et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998 352:609-613.

9 AR Jadad et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Controlled Clinical Trials 1996 17:1-12.

10 JE Edwards & RA Moore. Finasteride in the treatment of clinical benign prostatic hyperplasia: a systematic review of randomised trials. BMC Urology 2002 2: 14. (http://www.biomedcentral.com/1471-2490/2/14)

11 HJ McQuay et al. Some patients don't need analgesics after surgery. Journal of the Royal Society of Medicine 1982 75:705-8.

12 LA Smith et al. Teasing apart quality and validity in systematic reviews: an example from acupuncture trials in chronic neck and back pain. Pain 2000; 86:119-132.

13 FA McAlister & DL Sackett. Active-control equivalence trials and antihypertensive agents. American Journal of Medicine 2001 111: 553-558.

14 D Moher et al. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994 272: 122-124.

15 FA McAlister et al. Users' guides to the medical literature XIX Applying clinical trial results B. Guidelines for determining whether a drug is exerting (more than) a class effect. JAMA 1999 282: 1371-1377.

16 AD Oxman & GH Guyatt. Validation of an index of the quality of review articles. Journal of Clinical Epidemiology 1991 44:1271-1278.

17 AR Jadad & HJ McQuay. Meta-analyses to evaluate analgesic interventions: a systematic qualitative review of their methodology. Journal of Clinical Epidemiology 1996 49:235-43.

18 ES Ford et al. Homocyst(e)ine and cardiovascular disease: a systematic review of the evidence with special emphasis on case-control studies and nested case-control studies. International Journal of Epidemiology 2002 31: 59-70.

19 Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. Clin Oral Investig 2003 7:2-7.

20 D Moher et al.Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet 1999 354: 1896-1900.

21 DF Stroup et al. Meta-analysis of observational studies in epidemiology. A proposal for reporting. JAMA 2000 283: 2008-2012.