Cautionary tales in the interpretation of systematic reviews of therapy trials

I. Scott,1* P. Greenberg,2* P. Poole4* and D. Campbell3*

1Department of Internal Medicine and Clinical Epidemiology, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, 2Department of General Medicine and Clinical Epidemiology and Health Care Evaluation Unit, Royal Melbourne Hospital, University of Melbourne, 3Monash Institute of Health Services Research, Monash University, Melbourne, Australia and 4Department of Medicine, University of Auckland, Auckland, New Zealand

Key words
clinical interpretation, systematic review, caution.

Correspondence
Ian Scott, Department of Internal Medicine and Clinical Epidemiology, Level 5A, Princess Alexandra Hospital, Ipswich Road, Brisbane, Qld 4102, Australia.
Email: ian_scott@health.qld.gov.au

Received 12 December 2005; accepted 16 March 2006.
doi:10.1111/j.1445-5994.2006.01140.x

Abstract
This is the second in a series of articles emphasizing the cautions in the interpretation of health-care studies. Systematic reviews are presented as comprehensive, unbiased summaries of evidence and are often referred to by clinicians, guideline developers and health policy-makers. Their strengths and limitations, and how their results can be subject to bias and misinterpretation, are discussed.

Introduction
Systematic reviews of therapy trials use explicit methods to identify, select and critically appraise trials around a specific clinical question and to collect, analyse and synthesize data from these trials (Table 1, Fig. 1).1 Reviews may be qualitative, using an impressionistic approach or quantitative (and termed a meta-analysis) when numeric outcome data from multiple studies are statistically combined to produce an overall (or summary) measure of effect.2 Compared with traditional narrative reviews, systematic reviews offer several advantages as listed in Table 2.3

As summaries, systematic reviews are becoming increasingly popular with clinicians,4 guideline developers,5 health policy-makers and consumers.6 The Cochrane Collaboration7 and better methods for retrieving reviews from published work8 have accelerated their use, with 1634 reviews indexed in Medline in 2004 compared with 250 in 1990.

By summarizing existing data and stressing uncertainty, systematic reviews can assist in decisions about whether additional trials are necessary (and ethical) and how they should be designed.9 From August 2005, the Lancet requires authors to place their new findings in context, by undertaking a systematic review themselves or by referring to a published review.10 Other journals are likely to follow suit.

However, systematic reviews are not without problems. Some yield results that are discordant with those of large randomized controlled trials (RCT),11 or with the results of other reviews ostensibly addressing the same question.12 Whereas there are guides for critically appraising the validity and applicability of systematic reviews,13 and improving their reporting in journals14, reviews can still be subject to bias and misinterpretation, as this article attempts to show.

Poorly formulated or irrelevant clinical questions
Questions forming the basis for a review should be clinically relevant and properly structured using the
Table 1  Anatomy of a systematic review

Reviews should establish a prospective protocol to assist in making what is essentially a retrospective exercise in literature searching and data review as prospective an endeavour as possible.

1. State the specific objectives of the review (based on a specific clinical question in which patient, intervention, comparator, outcome (PICO), and admissible study design are defined)
2. Decide selection criteria for studies to be included (based on trial design and methodological quality)
3. Undertake inclusive searches (using appropriately crafted search strategies)
4. Tabulate characteristics and assess methodological quality of each trial identified
5. Apply selection criteria and maintain a log of excluded studies with reasons for exclusion (carried out independently by at least two investigators, and reported as a flow diagram)
6. Extract from included studies the most complete dataset feasible (carried out independently by at least two investigators)
7. Analyse data using statistical synthesis (meta-analysis) where possible (Apply tests and investigate sources of heterogeneity; use fixed-effects and/or random-effects statistical models; carry out sensitivity analyses)
8. Present summary estimates of effect of treatment (Use graphical form such as forest plots (Fig. 1), in which, for each trial listed in order of year of publication, event rates (or means) and sample size in both treatment and control arms, point estimates of effect and 95% confidence intervals (CI) (or standard deviations) and weights are shown, followed at the bottom by the summary point estimate and its 95% CI (or standard deviation)
9. Discuss sources of heterogeneity of trial results, results of subgroup and sensitivity analyses, strengths and weaknesses of review, and results of other reviews if appropriate
10. State conclusions and recommendations

Special types of meta-analysis

- **Cumulative meta-analysis**: As individual trials are added in some specified order (such as year of publication or level of methodological quality), the summary estimate is recalculated and, at any time, is the cumulative estimate generated from all preceding trials.
- **Individual patient data meta-analysis**: Synthesis of data is undertaken using datasets of every individual patient in every trial, rather than just the summary data for treatment and control groups provided in published reports. Such meta-analyses require more time, effort and expense in acquiring large datasets from authors of primary trials.

![Figure 1](image-url)  
**Figure 1** Standard meta-analysis (left) and cumulative meta-analysis (right). The point estimates for the risk ratio (or relative risk) of each study and the pooled point estimates are shown by the points, and the horizontal lines show the confidence intervals (CI), typically 95%. N is the number of patients in the study. The studies are ordered according to year of publication. In the cumulative meta-analysis, N is the cumulative number of patients in the current and preceding trials combined. The points and lines are the point estimates and 95% CI of the pooled results after inclusion of each additional study, respectively. The CI narrow as more trials are added unless substantial heterogeneity exists. (Reproduced from reference 3 with permission American College of Physicians. The results displayed are derived from Ioannidis JP, Cappelleri JC, Lau J et al. Early or deferred zidovudine therapy in HIV-infected patients without an AIDS-defining illness. Ann Intern Med 1995; 122: 856–66).
population, intervention, comparison and outcomes (PICO) format. Questions must be formulated so as not to be too specific to be answerable or too broad to be clinically useful.

Example: The question: ‘Are anticoagulants useful in patients after stroke?’ does not specify patient type (e.g. ischaemic or haemorrhagic stroke), type of anticoagulant or treatment duration, comparisons (e.g. with aspirin) and outcomes of interest (e.g. death or disability).

Inadequate search strategies

Review authors must endeavour to retrieve all potentially relevant trials and reviews. The strategy should comprise

(i) multiple search text words and medical subject headings, (ii) no language or journal restrictions and (iii) as a minimum (a) search of at least two relevant electronic bibliographic databases (Table 3) over a specified time period, (b) review of all references cited by articles retrieved from (a) and (c) consultation with content experts. Medline searching alone misses, on average, 50% of relevant studies. Manual searching of journals, conference abstracts and unpublished technical reports (‘grey’ published work) may significantly increase the yield for uncommon, under-researched or poorly indexed topics.

Example: In reviewing trials of nutritional supplementation in patients after hip fracture, investigators identified 15 trials: 7 from two electronic databases (Medline being one) and 8 from other sources: four other databases, reference lists, contacting trial investigators and by manual searching of nutrition journals. The inclusion of ‘grey’ published work generates more realistic estimates of effectiveness for emerging therapies, but only one-third of reviews include such data.

Inappropriate or ambiguous selection criteria

Reviews will be seriously flawed if investigators exclude certain studies a priori because their results conflict with ‘first impressions’ existing at the time the review was initiated. Clearly defined and prospective selection criteria minimize such bias.

Example: A meta-analysis of cholesterol-lowering trials following myocardial infarction (MI) published up to 1990 included only RCT with at least 100 participants per group, 3 years follow up and no use of hormone treatment. The pooled mortality estimate from seven trials was favourable (odds ratio (OR) = 0.91; 95% confidence interval (CI) 0.82–1.02). However, one trial meeting selection criteria and not included was unfavourable (OR = 1.60; 95% CI 0.95–2.70). When sample size and follow-up criteria were relaxed, another 11 trials became analysable which, when pooled, suggested harm (OR = 1.14; 95% CI 1.03–1.26).

Table 2 Potential advantages of systematic reviews

- Explicit methods limit bias in identifying and rejecting studies
- Increased statistical power allows detection of treatment-related changes in rates of low-frequency events and greater precision of estimated treatment effects
- Large amounts of information from multiple trials can be assimilated quickly by researchers, clinicians and policy-makers
- Results of different studies may be formally compared to establish generalizability of findings and consistency (homogeneity) of results
- Outcomes can be analysed that were not objectives of the original trials
- Reasons for heterogeneity (inconsistency in results across studies) can be studied and new hypotheses generated about particular patient subgroups
- Publication bias may be investigated and adjusted for
- Reviews constitute useful summaries of knowledge around therapeutic interventions for purposes of guideline development, medical education and quality improvement

Table 3 Commonly used electronic bibliographic databases

<table>
<thead>
<tr>
<th>General medical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medline</td>
</tr>
<tr>
<td>Embase (Excerpta Medica Database)</td>
</tr>
<tr>
<td>Cochrane Controlled Trials Register</td>
</tr>
<tr>
<td>Current Contents</td>
</tr>
<tr>
<td>BIDS Science Citation Index</td>
</tr>
<tr>
<td>Science/Social Sciences Citation Index</td>
</tr>
<tr>
<td>Biomed Central</td>
</tr>
<tr>
<td>PubMed Central</td>
</tr>
<tr>
<td>SumSearch</td>
</tr>
<tr>
<td>Nursing/allied health</td>
</tr>
<tr>
<td>CINAHL</td>
</tr>
<tr>
<td>BIOSIS</td>
</tr>
<tr>
<td>CABNAR</td>
</tr>
<tr>
<td>Johanna Briggs Institute</td>
</tr>
<tr>
<td>PEDro</td>
</tr>
<tr>
<td>Psychology/psychiatry</td>
</tr>
<tr>
<td>PsychLit</td>
</tr>
<tr>
<td>Oncology</td>
</tr>
<tr>
<td>CANCERLIT</td>
</tr>
<tr>
<td>Complementary medicine</td>
</tr>
<tr>
<td>Centralized Information Service for Complementary Medicine</td>
</tr>
<tr>
<td>Organizational/managerial/programmatic interventions</td>
</tr>
<tr>
<td>SIGLE</td>
</tr>
<tr>
<td>ERIC</td>
</tr>
<tr>
<td>HEALTHStar</td>
</tr>
</tbody>
</table>

© 2006 The Authors
Journal compilation © 2006 Royal Australasian College of Physicians

589
Absence of independent study selection and data extraction

In minimizing bias or confusion, at least two investigators should independently (i) apply selection criteria to retrieved abstracts and (ii) extract specified data from full-text copies of selected studies. The method for resolving disagreement between investigators, the number of studies retrieved, reasons for exclusion and whether attempts were made to contact trial authors for missing data should all be stated.

Example: Two meta-analyses of the effects on mortality of branched chain amino acids for hepatic encephalopathy came to differing conclusions – one suggesting a benefit27 the other not.28 On further analysis, the disagreement was entirely attributable to how the analysts had extracted data from the original studies.

Misinterpretation of measures of outcome and effect

Clinical outcomes may be (i) binary (or dichotomous), based on yes/no categorization such as death/survival, (ii) continuous, such as blood pressure or quality of life or (iii) time-to-event measures, such as time to death or readmission. Binary outcomes are aggregated as counts or proportions, continuous outcomes as means or medians. Differences in outcomes between treatment and control groups of each trial (called the treatment effect) may be statistically expressed in several ways (Table 4).2,29 Results for each trial are weighted (using methods described here) and then summed across all trials to yield a summary measure (or estimate) of effect (Fig. 1).

For binary outcomes, effect measures can be relative: OR or relative risk (RR) or absolute: absolute risk reduction (ARR), sometimes called risk difference (RD). For infrequent events (<15% of all patients), OR and RR are equivalent, but diverge at higher event frequencies (Fig. 2). Most meta-analyses report statistically robust RR or OR, but clinicians prefer absolute measures such as ARR, RD and number needed to treat (NNT) or harm.34

Example: A meta-analysis of lipid-lowering drugs in coronary artery disease reported similar summary mortality OR for primary and secondary prevention studies: 0.85 and 0.84, respectively.31 By contrast, ARR were quite different: 0.1 and 3.2%, corresponding to NNT of 1000 and 31, respectively.

Where meta-analyses do not report absolute measures, OR or RR can be converted to NNT using formulae,32 conversion tables32 or nomograms,33 which base the calculated NNT on the average underlying (or baseline) risk for all trial patients. There are methods for recalculating NNT for individual patients with different risk,34 although they assume constant relative treatment effects across the range of risk. Fortunately, for most treatments, this is a reasonable assumption.35

Failure to assess and explain heterogeneity

When studies are deemed to be relatively homogenous (‘all apples’), a relatively constant (or fixed) treatment effect is assumed.36 However, studies may be quite heterogenous (‘apples and oranges’) in clinical factors (patients, interventions, outcomes) or methodological aspects (Table 5). Such between-study differences may alter the size or direction of treatment effects. The challenge is quantifying and explaining if there is heterogeneity.

Assessing heterogeneity

Simply eyeballing forest plots (Fig. 1) may disclose wide separation of effect point estimates and non-overlapping CI for a significant proportion of trials. However, visual inspection is insufficient and statistical tests for heterogeneity (or inconsistency) are required,36,37 the most common being variants of the $\chi^2$ test, which determine whether variation exceeds that expected due to chance. Because of its low sensitivity, a more stringent level of significance ($P < 0.1$) is recommended.38 The variance test ($I^2$) measures the percentage of between-study variation not attributable to chance; >30% suggests significant heterogeneity.37

Responding to heterogeneity

If heterogeneity is significant, pooled analyses may be aborted or, alternatively, carried out assuming a variable (or random) treatment effect (discussed shortly) amid attempts to explain why there is heterogeneity.36 Failure to test for heterogeneity and assuming a constant (or fixed) treatment effect can lead to misleading results.

Example: A meta-analysis of 19 trials of prophylactic endoscopic variceal sclerotherapy in cirrhotic patients disclosed inconsistent OR for the outcome of death.39 Four trials showed benefit, 14 showed null effects, one showed harm, with the pooled result of 24% reduction in overall mortality (OR = 0.76). As expected, test of heterogeneity was highly significant ($\chi^2 = 43; P < 0.001$), challenging the generalizability of the calculated benefit. (A useful rule of thumb: heterogeneity is non-significant if $\chi^2 \leq n-1$ where $n$ is the number of trials.) The trials were shown to differ greatly in patient selection, baseline illness severity, endoscopic technique, management of variceal bleeding and duration of follow up.
Table 4  Statistical measures used to express estimates of effect

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Usage and interpretation</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary (or dichotomous) data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Odds ratio (OR)</td>
<td>Calculated by dividing odds of an event occurring in treatment group by the odds in the control group (where odds = probability of event / (1 − probability)).</td>
<td>Favourable treatment effect (i.e. reduction in unwanted events) represented by OR &lt; 1.0. Conversely results with OR &gt; 1.0 favour control group, suggesting treatment confers harm. Exception is where a treatment is used to promote a favourable outcome (e.g. increase in number of disease remissions) wherein benefit is expressed as OR &gt; 1.0.</td>
<td>Relatively easy to calculate and popular among statisticians.</td>
<td>Not easily interpreted in clinical settings as ‘odds’ are not intuitive for most clinicians. OR is not an absolute measure of benefit and may, in certain situations, be misinterpreted as relative risk.</td>
</tr>
<tr>
<td>Relative risk (RR)</td>
<td>Expresses the relative probability that an event will occur in the treatment group versus the control group. Also known as risk ratio or event rate ratio. Calculated as TER/CER</td>
<td>OR and RR are comparable when event rates across both treatment and control groups are less than 15%. Compared to RR, OR becomes smaller (suggesting greater benefit) or larger (suggesting greater harm) as event rates exceed 20% (Figure 2). RRR equals 1 − RR.</td>
<td>Popular among clinicians and provides a measure of relative effect that is presumed constant (but is not always so) across the range of baseline risk.</td>
<td>Provides no information about the event rate in the control group or the absolute difference in proportion of events between treatment and control groups.</td>
</tr>
<tr>
<td>Absolute risk reduction (ARR)</td>
<td>Defined as the absolute difference between event rate in treatment and control groups. Calculated as CER-TER</td>
<td>Indicates the absolute measure of effect which is the most useful measure to clinicians.</td>
<td>Relatively easy to calculate and interpret, with the reciprocal of ARR (1/ARR) being number needed to treat to prevent one event (NNT).</td>
<td>As for relative measures, it cannot always be assumed that ARR and NNT are constant across a range of variable risk, and therefore applicable to every patient.</td>
</tr>
<tr>
<td><strong>Continuous data</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted mean difference (WMD)</td>
<td>Expresses the treatment effect as the weighted difference in mean values between treatment and control groups. Examples: patients receiving treatment compared with controls walked, on average, an extra x metres on a walking test, or had pain scores that were 2 units less.</td>
<td>Results easy to interpret, as they are expressed in the same units of measurement used by the primary trials.</td>
<td>Requires all trials being combined to have used the same measurement tools and units of measurement.</td>
<td></td>
</tr>
<tr>
<td>Standardised mean difference (also called effect size; SMD)</td>
<td>Expresses treatment effect as the absolute difference in mean values between treatment and control groups divided by the standard deviation (SD) of the control group (or a pooled SD of all trials combined).</td>
<td>Trials assessing the same outcome often use different measurement tools or scales. To make use of the data, they must first be converted into a unit-less standardized measure, expressed as SD units (z scales). An SMD &gt; 0.60 shows a substantial clinical effect; 0.40–0.60 moderate effect; 0.20–0.39 small effect; &lt;0.20 minor, probably clinically insignificant effect.</td>
<td>Calculation is relatively easy and provides a scale-free estimate of treatment effect, allowing combination of trials using different outcome measures.</td>
<td>Interpreting results expressed as z scores is not intuitive, and some reviews do not provide an interpretation of the clinical significance of reported SMD.</td>
</tr>
</tbody>
</table>

(Continued)
Explaining heterogeneity

Perusal of tables detailing the characteristics of included studies may disclose between-study differences in year of publication, baseline patient characteristics (age, disease severity), study setting, intensity and mode of administration of therapy, use of co-interventions, duration of follow up and choice of outcome measures.

A useful graphical method for exploring heterogeneity is the L’Abbe plot in which, for each trial, the outcome measure (event rate or mean) for the treatment group is plotted against that for the control group. A regression line (weighted for study size) is then applied, and plots lying well away from this line indicate ‘outlier’ trials which invite closer scrutiny (Fig. 3).

Analysing data by subgroups (or stratified groups) or applying meta-regression techniques are further means for examining heterogeneity. Both aim to quantify the extent to which variation in specific study factors (listed in Table 5) influence the magnitude and direction of treatment effect, as explained in more detail elsewhere.

Caution on methods for explaining heterogeneity

Subgroup analyses and meta-regression are observational studies across trials, not within-study comparisons. Consequently, any observed association between outcome and study characteristic is not necessarily causal. Such analyses should be pre-specified, avoided if few trials with

Table 4 Continued

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Usage and interpretation</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-to-event data</td>
<td>Hazard ratio (HR) Indicates the overall relative chance of an event occurring in the treatment group compared to controls during a specified period of time, taking into account censoring of patients once events have occurred.</td>
<td>Similar to OR but where there is adjustment of event rates over time (after applying proportional hazards models). HR &lt; 1.0 and HR &gt; 1.0 indicates lower (benefit) or higher (harm) chance of unfavourable event in the treatment group compared to control group at a specified point in time.</td>
<td>Can be derived from summary data using several methods or estimated from Kaplan-Meier time-to-event curves by splitting them into discrete time intervals, calculating an odds ratio for each period, and combining these over time.</td>
<td>In the absence of time-to-event curves, HR as a summary estimate gives no indication as to the exact timing (sooner or later) of events within a given time period.</td>
</tr>
</tbody>
</table>

Figure 2 Correlation between odds ratio and relative risk according to overall event rate.

1 Adapted from Moher et al. CER, control group event rate; TER, treatment group event rate; WMD, weighted mean difference.

Table 5 Examples of sources of heterogeneity

Clinical heterogeneity (clinical diversity)
- Patients
  - Comorbid clinical conditions
  - Eligibility criteria for entry into trials
  - Geographical variation
  - Other: age, sex etc.
- Treatments
  - Intensity/dose/duration of therapy
  - Mode of administration
  - Expertise of practitioners
  - Nature of control (placebo/none/standard care)
- Outcome
  - Definition of event
  - Follow-up duration
  - How outcome is measured and reported
  - Cut-off points on scales

Methodological heterogeneity (methodological diversity)
- Design
  - Randomized versus non-randomized trials
  - Cross-over versus parallel group versus cluster randomized
- Conduct
  - Allocation concealment
  - Blinding
  - Analytic method (e.g. intention to treat)
  - Imputation methods for missing data

© 2006 The Authors
Journal compilation © 2006 Royal Australasian College of Physicians
small sample size, restricted to investigating patient and trial design characteristics that vary across, but not within trials, and their results viewed as hypothesis-generating.

Summary treatment effects based only on fixed-effects statistical models

Meta-analysis calculates a summary effect estimate as a weighted average of the treatment effect estimates of all individual trials, where weights for each trial can be derived using a fixed-effects or a random-effects model.44

Fixed-effects model

This model assumes a single ‘true’ effect for each study, which, if very large or repeated many times, would yield a single result; thus any between-study variation is simply due to chance. Trial weights are calculated $w = 1/d^2$ (where $d$ (variance) measures within-study random error and is inversely proportional to numbers of events (or sample size)). The precision of the summary effect estimate (as indicated by the width of the CI) is thus largely determined by the largest trials. For binary outcomes, the Peto or Mantel–Haensel fixed-effect methods are most often quoted.44

Random-effects model

This model assumes treatment effect actually varies between studies, which are but a random sample of the total population of possible trials and that these different treatment effects assume a normal distribution around a mean value. Accordingly, study weights will vary inversely with the sum of within-study variation ($d$) and between-study variation ($\tau$): $w = 1/(d^2 + \tau^2)$. These weights will be more evenly valued, being no longer determined by variance alone. Hence, the summary effect estimate, compared to that of a fixed-effects model, is more conservative (averaged across more evenly weighted trials), and bounded by a wider CI (greater dispersion of effect sizes independent of trial size). The DerSimonian–Laird method is the random-effects model used for binary outcomes.44

When to use random-effects models

Whenever there is significant trial heterogeneity, the random-effects model is preferred, as the fixed-effects model will likely overestimate both effect size and its precision. Ideally, results should be calculated using both methods, the only exceptions being highly homogenous trials or where one large trial accounts for most (>80%) events within a small number of trials.

Example: In a meta-analysis of nine RCT of alendronate in prevention of fractures in sites not usually associated with osteoporosis, point estimates for three studies showed benefit, one showed no effect, five studies suggested harm. The fixed-effects estimate was a trend toward benefit with a narrow CI (RR = 0.80; 95%CI 0.74–1.09), whereas the random-effects estimate trended in the opposite direction with wider CI (RR = 1.05; 95%CI 0.72–1.53), reflecting more accurately the uncertainty about the treatment effect.

Failure to assess quality of primary trials

Systematic reviews are flawed if included studies are of poor methodological quality: garbage in is equal to garbage out.46

Important determinants of trial quality

Overestimation of treatment benefits is more likely when investigators fail to (i) conceal allocation to groups at randomization, (ii) blind patients and clinicians (double-blinding) or outcome assessors, (iii) account for patients who dropped out, (iv) analyse outcomes using intention-to-treat method and (v) allow studies to run to completion.47,48

How to assess and report trial quality

Reviewers should pre-specify how trial quality will be evaluated and assess whether effect estimates change if
lower quality trials are omitted from analysis. Cochrane, compared with non-Cochrane reviews, assess quality more often: 94% versus 60% of included studies.49

Examples: In a review of perioperative use of β-blockers in non-cardiac surgery, pooling of low-quality trials suggested substantial benefit in reducing cardiovascular events (RR = 0.13; 95%CI 0.03–0.54) compared with no effect seen in high-quality trials (RR = 0.82; 95%CI 0.49–1.36).50

Component versus composite analyses of trial quality

Quality of each trial may be evaluated according to either specific methodological attributes (or components) or by assigning an overall (or composite) quality score (the Chalmers et al. score51 and the Jadad et al. score52 being most commonly used). Component analysis is preferred, as different composite scales can rate the same trials within a meta-analysis as being of high or low quality.53 The methodological flaws in trials most likely to invalidate a particular review should be identified, ideally a priori, and, for each trial, assessed individually.54

Omission of relevant sensitivity analyses

The preceding discussion on quality is one example of sensitivity (or influence) analysis – examining the extent to which the summary estimate of treatment effect is ‘sensitive’ to various ‘what if?’ questions.55 The estimate is recalculated on subsets of the original trials after excluding, for example, trials that: are small or remote; have incomplete follow-up or short duration; use irregular outcome definitions or statistical methods; or differ in design features considered important by content experts.

Example: A meta-analysis of 17 trials of β-blockers for secondary prevention after MI suggested mortality benefit (RR = 0.80).56 Sensitivity analyses based on statistical modelling (fixed vs random effects), blinding (double-blinding vs other forms of blinding), trial size (<25 vs >99 deaths), follow up (<1 vs >2 years), and protocol compliance (trials stopped early vs completed) showed overestimation of benefit with small trial size (RR = 0.58 vs 0.80) and absence of blinding (RR = 0.75 vs 0.81).

Post-hoc subgroup analyses

Clinicians are interested in knowing whether treatment effects are larger or smaller in particular patient subgroups. Like sensitivity analyses, subgroup analyses should be pre-specified (not post-hoc); limited in number to minimize false-positive errors, and centred on clinical factors known, or strongly suspected, to influence treatment effects.57

Example: A meta-analysis of 25 trials of β-blockers after MI published in 1985 included a post-hoc subgroup analysis comparing mortality for agents with and without intrinsic sympathomimetic activity (ISA).58 Drugs without ISA were associated with significantly fewer deaths than those with ISA (OR = 0.68 vs OR = 0.88; p = 0.009 by test of interaction). However, no significant difference was seen in an updated 1999 analysis of 33 trials.57

Failure to assess possible publication bias

Systematic reviews rely on published trials that are over-represented by those yielding positive results, in contrast to unpublished trials, which tend to show null effects or harm.59 This bias will diminish now that all trials, to be published, have to be registered at inception in a publicly accessible register.

However, for reviews analysing trials commenced before September 2005, testing for publication bias should be included. Most cited tests are based on the ‘funnel plot’ where sample size (or standard error) of every trial is plotted against its treatment effect (usually OR or log OR).60 If publication bias is absent, a symmetrical ‘funnel’ of plots should form around a vertical line centred on the plot of the largest trial (Fig. 4).

Approximately 25% of meta-analyses show funnel plot asymmetry and, although suggestive of publication bias, may indicate other forms of reporting bias (e.g. bias towards

Figure 4 Funnel plot where sample size is plotted against risk ratio. Note the asymmetry with less small studies to the right of the line of effect centred on the largest trial showing little benefit or favouring control compared to the left of the line favouring treatment. (Reproduced from reference 3 with permission of American College of Physicians. Data displayed are derived from Lau J, Antman EM, Jimenez-Silva J et al. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med 1992; 327: 248–54).
English-speaking journals, selective outcome reporting, data irregularities, heterogeneity and chance variation.61

**Reporting of treatment benefits based on small numbers of events**

Pooling data from multiple studies allows detection of significant treatment effects, which individual trials have insufficient power to show definitively. However, as most modern therapies do not decrease relative risk by more than 10–15%, exaggerated benefits based on comparatively small numbers of events (‘small-study effect’)62 must be viewed cautiously.

**Example:** A meta-analysis of seven small trials evaluating effects of magnesium after MI in 1301 patients with 78 deaths showed this treatment to significantly decrease mortality by 55% (95%CI 29–72%; $P < 0.001$) compared to placebo.63 A moderate-sized trial (LIMIT-2) published 6 months later with 2316 patients and 208 deaths showed a lower risk reduction of 24% (95%CI 1–43%; $P = 0.04$).64 Finally, in the ISIS-4 mega-trial with 58 050 patients and 4319 deaths, magnesium had no effect, indeed showing a trend towards excess deaths.65

**Preventing small-study effects**

Several methods may circumvent small-study effects: (i) including only trials that enrol a minimum number of participants (e.g. 500 or 1000 for common conditions such as MI; 100 in less common disorders);62 (ii) adopting more stringent significance testing (99%CI ($P < 0.01$) rather than traditional 95%CI ($P < 0.05$)) or (iii) specifying, a priori, the ‘total sample size’ (or optimum information size) required to detect significant, clinically worthwhile treatment effects.65 Only when the cumulative effect size (generated from cumulative meta-analysis) crosses a pre-specified monitoring boundary of significance66 can a treatment effect be said to be conclusively proven.

**Example:** The previously cited review of perioperative use of β-blocker therapy50 calculated that 6124 patients were needed to prove efficacy in preventing cardiovascular events. The apparent marked benefit (RR = 0.44; 95%CI 0.20–0.97) rested on only 83 deaths in total and became insignificant at the 1% level of significance (99%CI 0.16–1.24). The authors concluded that the evidence for β-blockers, although encouraging, was not definitive and more trials were needed.

**Lack of reporting of therapy-related adverse effects**

Only one in four reviews of therapeutic trials provides information on therapy-related adverse events (AE),67 with very few primarily devoted to evaluating therapy-specific AE. Although such analysis is precluded if trials do not measure or report AE, many reviews do not aim to assess safety, or consider it in sufficient detail, despite data being available in primary trials.

**Example:** A review of drug therapy for neuropsychiatric symptoms of dementia concluded that atypical antipsychotics showed ‘minimal adverse effects at low dose’.68 A more recent meta-analysis of 15 trials (original 6 plus 9 unpublished studies), specifically aimed at assessing drug-related mortality, disclosed a 54% increase in the odds of death (1% absolute increase) compared to placebo.69

**Unjustified conclusions or recommendations**

Reviews may state conclusions or recommendations not justified by marginal or statistically insignificant effect sizes.70 One recent study of meta-analyses of antihypertensive drugs showed industry sponsorship, compared to non-commercial funding, increased the reporting of unsubstantiated conclusions favouring study drug by a factor of five.71 Cochrane reviews are usually more conservative in statements of support compared to non-Cochrane reviews.72

**Example:** A review of trials of corticosteroids in acute exacerbations of COPD concluded treatment ‘significantly reduces treatment failure and need for additional treatment’, and ‘increases the rate of improvement in lung function and dyspnoea’.72 However, mortality, readmission rates and length of stay were unaffected, changes in dyspnoea scores and spirometric results were clinically marginal and definition of ‘treatment failure’ included return visit to doctors or simply up-titration of other therapies.

**Reviews of similar questions that yield discordant results**

Where independent reviews of the same question report conflicting results, the most methodologically rigorous review is likely to be closest to the truth. Cochrane reviews rate better in this regard than non-Cochrane reviews.72 Fortunately, most discrepancies between meta-analyses are those of degree, rather than direction, of effect size, although some have reached opposite conclusions.74,75 In such circumstances, if reviews show comparable methodological quality, then differences in review design may explain discordant results. An approach for exploring such ‘meta-heterogeneity’ is outlined in Figure 5.76

**Example:** Four reviews of stress ulcer prophylaxis in critically ill patients77–80 yielded inconsistent results which, on reappraisal, were explained by differences in prophylactic regimens assessed, identification of relevant
studies (non-English language journals and non-randomized trials), definitions of bleeding, contact with authors of primary trials in obtaining additional data and statistical methods.12

How to avoid being misled

Methods for synthesizing data from multiple studies need to be explicit, prospective and rigorous. Systematic reviews are limited in being essentially retrospective hypothesis-testing exercises and are prone to special biases over and above those of their component trials: publication bias, heterogeneity and false-positive errors resulting from multiple comparisons.

Reviews reporting unexpected or counterintuitive results should be reappraised and discussed with experts,81 particularly as only a quarter of reviews, irrespective of subspecialty, show high-quality scores across all domains listed in Table 1.82,83 Physicians are likely to feel more confident about results of meta-analyses and favour treatment when individual trial results are statistically homogenous and the overall effect size is large.84

We recommend that readers of reviews apply the cautions discussed above and look for attributes predictive of greater validity, namely (i) well-defined protocols, (ii) inclusion of RCT only, (iii) explicit assessment of trial quality, (iv) inclusion of sensitivity analyses and (v) peer-reviewed publication. In the end, reliable meta-analysis is predicated on high-quality reporting of data from primary trials, a need which cannot be overstated.

References

71 Yank V, Rennie D, Bero L. Are authors' financial ties with pharmaceutical companies associated with positive results or conclusions in meta-analyses on antihypertensive medications [abstract]. Proceedings of the 5th International Congress on Peer Review and Biomedical Publication; 2005 Sept; Chicago.


76 Jadad AR, Cook DJ, Browman GP. A guide to interpreting discordant systematic reviews. CMAJ 1997; 156: 1411–16.


83 Bramwell VH, Williams CJ. Do authors of review articles use systematic methods to identify, assess and synthesize information? Ann Oncol 1997; 8: 1185–95.