Critical appraisal for emergency medicine: 6 Systematic reviews

S Goodacre

Correspondence to:
Professor S Goodacre, Medical Care Research Unit, University of Sheffield, Regent Court, 30 Regent Street, Sheffield S1 4DA, UK; s.goodacre@sheffield.ac.uk

Accepted 6 January 2008

Systematic reviews are increasingly being seen as the optimal source of knowledge for evidence-based practice. A good systematic review will provide an unbiased summary of existing evidence and, provided it is applicable to local patients, should guide clinical practice. Being able to appraise systematic reviews is therefore a crucial skill for emergency physicians.

The use of complex statistical techniques in meta-analysis often distracts the clinician attempting to appraise a systematic review. As previously suggested in this series, complex statistical issues are best left to a statistician. Instead, we should focus upon the many important insights that clinical experience can bring to appraisal.

WHAT IS A SYSTEMATIC REVIEW?

A systematic review is a scientific study. It follows the introduction, methods, results and discussion approach. The conclusion should represent an unbiased synthesis of available data relating to a specific question. It may not be very entertaining to read but, if undertaken properly, will provide an objective answer based upon the best scientific evidence.

A narrative review is not a scientific study. The authors present their opinions of a particular topic with reference to primary studies they have selected. A good narrative review should be interesting, entertaining or provocative, but it should not be considered to provide scientific evidence. The differences between a systematic and a narrative review are summarised in table 1.

STAGES OF A SYSTEMATIC REVIEW

The process of identifying, selecting and assessing studies for inclusion in a systematic review should be open, explicit and objective. Data collection for a systematic review typically involves three stages: (1) literature searching and retrieval; (2) the selection of appropriate papers; (3) quality assessment of selected papers.

These three steps should be based upon explicit criteria and should ideally be carried out by two independent assessors who are blind to each other's decisions. The review should report the total number of articles identified by the search, the number selected after scanning titles/abstracts, the number selected after assessment of the full article and the number included in the review.

LITERATURE SEARCHING

An inadequate literature search may miss important articles leading to a biased conclusion. A literature search may include: electronic databases, such as Medline, Embase, Cinahl and the Cochrane Database; hand searching of key journals (ie, the reviewer searches the contents pages of all issues of a particular journal for potentially relevant articles); the grey literature: reports (government or academic), conference proceedings, the internet, libraries and professional societies; research registers, such as the national research register, clinictrials.gov and the health technology assessment database; searching the bibliographies of retrieved articles for relevant citations; contact with researchers or "experts"; contact with the pharmaceutical industry or equipment manufacturers.

Searching research registers can be particularly useful for a systematic review of a therapy. There is an increasing move towards ensuring that all funded studies are registered before they commence. Registers can therefore be used to identify unpublished studies. They can also identify whether a study is in progress that is likely to influence the findings of a systematic review when it is completed.

PUBLICATION BIAS

Publication bias occurs when the results of a study influence the likelihood that it will be written up, submitted for publication or published, and thus the likelihood that it will be included in a systematic review. Positive studies (ie, trials reporting a significant effect or diagnostic studies reporting high sensitivity/specificity) are more likely to be written up, submitted and published, so are more likely to be included in a systematic review. This may lead to an overestimate of treatment effect or diagnostic accuracy.

Publication bias can be minimised by undertaking a comprehensive search, but the possibility of publication bias can never be completely eliminated. Techniques such as the funnel plot can be used to search for publication bias (as shown in example 1 and fig 1), but these are often insensitive. Prospective registration of trials offers the best solution to publication bias in the future.

SELECTION OF RETRIEVED ARTICLES FOR ANALYSIS

Literature searches will retrieve large numbers of articles, most of which are irrelevant. A systematic review must therefore define the method by which retrieved articles are selected for inclusion. This should be directly related to the research question. Often the inclusion criteria will relate to the "PICO" of the research question—the defined patients or population, the intervention, the comparison and the outcome of interest.

The following criteria are sometimes used to exclude studies: (1) small studies; (2) English

Table 1 Differences between systematic and narrative reviews

Systematic review	Narrative review
Focussed question	Broad question
Methodology described	No methodology described
Systematic and comprehensive literature search	Based on authors' collected papers
Primary studies selected according to defined criteria	Primary studies selected at authors' discretion
Quality of primary data assessed objectively according to predefined criteria	Quality of primary data assessed subjectively according to authors' opinion
Synthesis of primary data may be attempted using statistical techniques	No formal statistical synthesis of primary data
Potential bias in selection of primary data may be assessed	Potential bias not considered
Conclusions result from a scientific study of the available data	Conclusions represent the authors' opinions

language only; (3) mainstream journals only; (4) insufficient data presented; (5) data presented in a form incompatible with planned analysis; (6) year of publication.

These criteria are applied for reasons of convenience, rather than methodology. However, judgement is required to determine whether excluding these articles is a reasonable way of avoiding fruitless work or whether this may influence the overall findings of the analysis. Excluding articles published before a certain date, for example, is entirely appropriate for a systematic review of a technology that has only recently been developed. Many would also argue that studies that fail to present data in an interpretable manner are likely to be poor quality, and the analysis may suffer little from their exclusion.

ASSESSMENT OF STUDY QUALITY

Ideally, all studies selected for inclusion should be assessed for quality. This will allow the authors to determine the overall quality of the available data and to explore the impact of excluding poor quality studies.

Quality assessment should be objective and based upon criteria that are known to influence study quality. The only factors proved to impair quality in trials are lack of allocation concealment, lack of blinding, inadequate follow-up and failure to use intention-to-treat analysis. These factors are combined in a commonly used quality score, the Jadad score.

HETEROGENEITY

Studies of a similar intervention, using similar methodology in a similar environment, should give similar results. The only differences between results will be due to random error. "Heterogeneity" is the term used to describe the amount of variation in the results of trials included in a systematic review.

The usual assumption behind a systematic review is that included studies are measuring the same result. This is particularly important if there is to be any attempt to combine results (meta-analysis). If there is substantial heterogeneity between results then studies may not be measuring the same thing and any conclusions based on assumptions of a common effect will be suspect.

It is therefore important to assess results for heterogeneity of effect. This can be done in several ways: (1) The results of a systematic review are usually presented as a forest plot (see fig 2). Individual study results, with 95% confidence intervals, are plotted alongside each other. Simply observing the overlap of confidence intervals gives a crude estimate of heterogeneity. If there is little overlap between the confidence intervals then heterogeneity is present. (2) Various statistical methods can test the null hypothesis that all the studies come from the same population and are estimates of the same value. If the test is statistically significant this gives good evidence that studies are heterogeneous. However, a non-significant test does not rule out potentially important heterogeneity.

META-ANALYSIS

This is the synthesis of data from various sources to provide an estimate of common effect. Meta-analysis should not consist of simply adding results together or calculating a mean effect. This does not take into account the size or variance of each individual study. Although meta-analysis software is available free on the internet, the involvement of someone with statistical expertise is usually required.

Meta-analysis assumes that all the individual studies are estimates of the same value. Combining results provides a more precise estimate and reduces the chances of a type II (false negative) statistical error (ie, missing a potentially important treatment effect). This is the principal value of meta-analysis. It does not overcome bias in the original data. Combining biased data (such as the results of historically controlled trials) will just give a precise, but inaccurate, estimate.

Clearly meta-analysis is much more controversial if there is any evidence of heterogeneity of effect. Combining the results of fundamentally different studies simply does not make sense. Clinicians may feel intimidated by fancy statistical tests and the discussion of "fixed effects" and "random effects" models. However, clinicians are often well placed to comment on heterogeneity and the inappropriate combination of results.

Rather than trying to decipher the statistics, have a look at the studies that have been combined. What were the patient inclusion/exclusion criteria? What was the setting? What exactly was the intervention? What was the control?

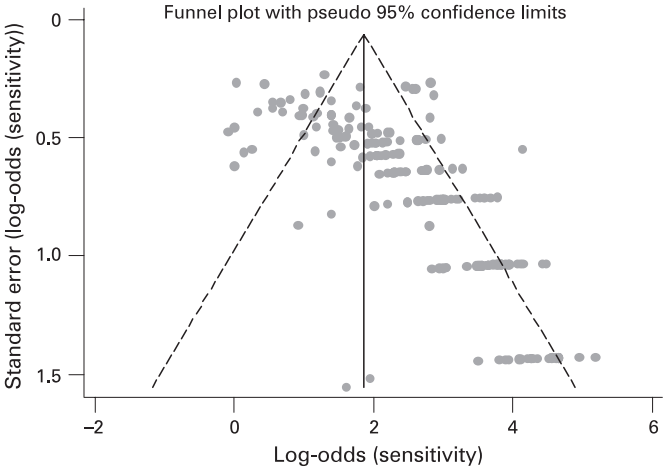

Figure 1 Funnel plot of studies evaluating the sensitivity of D-dimer for deep vein thrombosis.

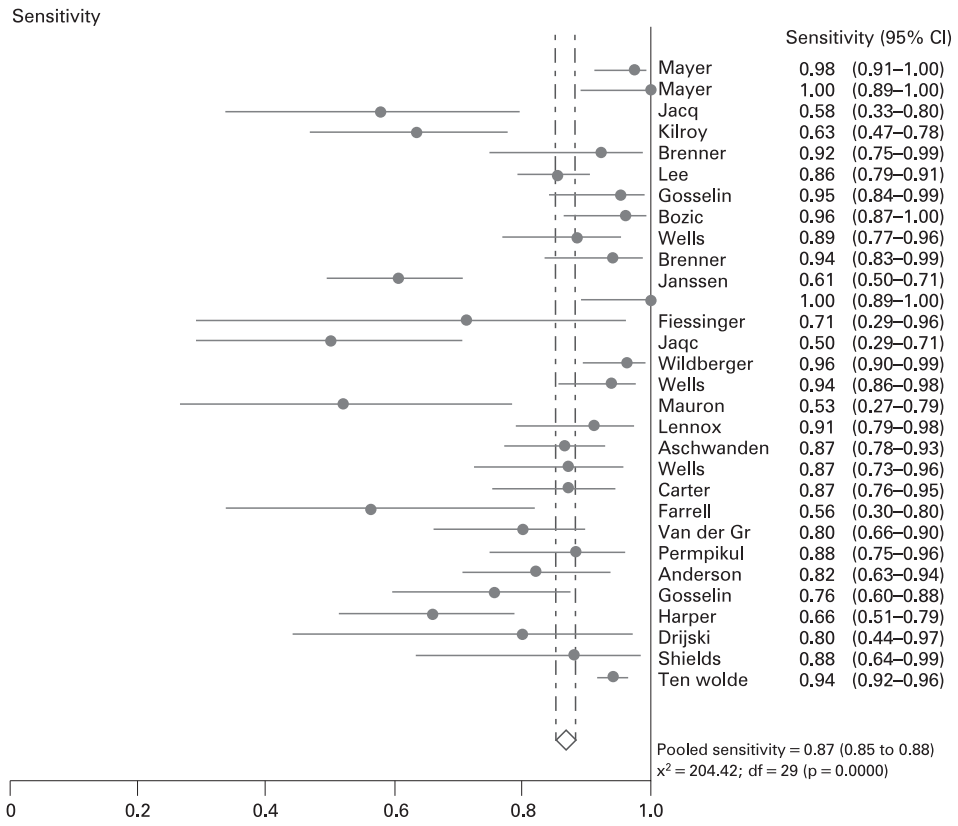

Figure 2 Forest plot of studies evaluating the sensitivity of D-dimer for deep vein thrombosis.

If there are important differences in these characteristics between the studies in the meta-analysis then it may be inappropriate to combine them. It may also be inappropriate to extrapolate conclusions from the meta-analysis to the various specific treatments or patient groups included in the analysis.

Meta-analysis is sometimes described as the statistical equivalent of combining apples and oranges. However, it may become apparent that the statisticians, for all their fancy tests, are not just trying to combine apples and oranges. They are trying to combine apples, oranges, potatoes and cabbages, with the odd sock thrown in as well.

SUMMARY

Systematic reviews are often undertaken according to well-established protocols that ensure high quality, whereas understanding meta-analysis requires a certain amount of expertise. These factors can make critical appraisal of systematic reviews seem to be a rather unrewarding experience. Nevertheless, the clinician can bring a lot to the appraisal of systematic reviews, particularly in assessing heterogeneity and deciding when the findings are applicable.

EXAMPLE 1

A meta-analysis combined sensitivity data from studies of D-dimer for deep vein thrombosis. A funnel plot was made to examine for evidence of publication bias (see fig 1). A measure of the precision of the study (the standard error of the log odds of the sensitivity) was plotted against the log odds of the sensitivity. We would expect the funnel plot to be symmetrical and shaped like an upside-down funnel. The more precise (ie, larger) studies would be close to the "true" value for log odds (sensitivity), whereas the less precise (ie, smaller) studies would be more scattered.

However, the funnel plot in this study was asymmetrical, with smaller studies tending to produce higher estimates of sensitivity. One possible explanation for this is publication bias. Smaller studies are more likely to be published if they produce high estimates for sensitivity (authors may not submit, and journals may not publish, small studies showing poor sensitivity), whereas larger studies are likely to be published regardless of their findings.

There are other possible causes for an asymmetrical funnel plot. Differences between small and large studies in methodology, patient selection, use of the intervention and outcome measurement may produce different estimates of the outcome of interest.[1]

EXAMPLE 2

The forest plot in fig 2 is taken from the meta-analysis of the diagnostic accuracy of D-dimer for deep vein thrombosis. The sensitivity reported in each study of SimpliRED D-dimer is plotted with a 95% confidence interval. There is clearly substantial heterogeneity between the studies: point estimates of sensitivity vary from 50% to 100%, and the confidence intervals of estimates from different studies do not overlap in many cases. This heterogeneity may be due to differences in the study populations, the way the test was used or the reference standard used for diagnosing deep vein thrombosis. We should be cautious about using the overall estimate of sensitivity for D-dimer from this analysis.

Competing interests: None.

REFERENCE

1. **Goodacre S,** Sampson FC, Sutton AJ, *et al*. Variation in the diagnostic performance of D-dimer for suspected deep vein thrombosis: systematic review, meta-analysis and meta-regression. *Q J Med* 2005;**98**:513–17.

Critical appraisal for emergency medicine: 6 Systematic reviews

S Goodacre

Emerg Med J 2009 26: 114-116
doi: 10.1136/emj.2007.057356