Southern African Journal of Anaesthesia and Analgesia. 2021;27(6 Suppl 1):S94-97
https://doi.org/10.36303/SAJAA.2021.27.6.S1.2739

South Afr J Anaesth Analg
ISSN 2220-1181 EISSN 2220-1173
© 2021 The Author(s)
FCA REFRESHER COURSE

Systematic reviews

BM Biccard

Department of Anaesthesia and Perioperative Medicine, Groote Schuur Hospital and University of Cape Town, South Africa
Corresponding author, email: *bruce.biccard@uct.ac.za*

This article is based predominantly on the works of Guyatt and colleagues.[1] This is an excellent text for understanding the principles of evidence-based medicine.

Keywords: research, systematic reviews, meta-analysis

Introduction

Definitions

- A systematic review is a summary of research that addresses a focused clinical question in a systematic and reproducible manner.
- A meta-analysis is a statistical pooling of results from different studies to provide a single best estimate of effect.

Why should we conduct systematic reviews?

1. Single studies may be unrepresentative of the total body of evidence.
2. An accompanying meta-analysis will provide the best estimate of effect, and increase the precision of that estimate of effect. These data aid clinical decision making.
3. A systematic review provides information to inform our confidence in the current evidence.

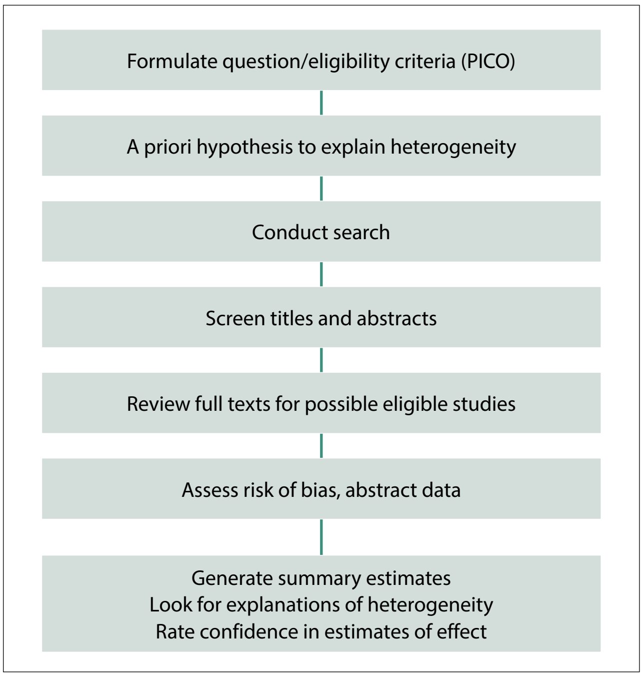
Formulate question/eligibility criteria (PICO)
A priori hypothesis to explain heterogeneity
Conduct search
Screen titles and abstracts
Review full texts for possible eligible studies
Assess risk of bias, abstract data
Generate summary estimates
Look for explanations of heterogeneity
Rate confidence in estimates of effect

Figure 1: The process of a systematic review
PICO – patient, intervention, comparison, outcome

The process of a systematic review

The process of a systematic review is shown in Figure 1.

The credibility of the effect estimates

The two fundamental problems which may adversely affect the credibility of the effect estimates in a meta-analysis:

1. the credibility of the review i.e. to what extent did the design and conduct of the review protect against misleading results, i.e. what are the methodological standards of the review process; and
2. the individual studies may include studies with a high risk of bias which will decrease confidence in the estimates.

The rest of this text will address the strategies adopted to either minimise and/or understand whether either of these points has a significant effect on our interpretation of the effect estimates from a meta-analysis.

The credibility of the systematic review process

A systematic review has eight strategies to increase the credibility of the review. These are summarised in the following questions:

1. Was the review prospectively registered?
2. Did the review explicitly address a sensible clinical question?
3. Was the search of relevant studies exhaustive?
4. Was the risk of bias of the primary studies assessed?
5. Did the review address possible explanations of between-study differences in the results?
6. Did the review present results that are ready for clinical application?
7. Were the selection and assessment of studies reproducible?
8. Did the review address confidence in effect estimates?

Was the review prospectively registered?

To ensure credibility and remove the ability of the authors to bias the results, registration of the systematic review prospectively with PROSPERO (https://www.crd.york.ac.uk/prospero/) is recommended. It must be ensured that there is no reporting bias, i.e. where the reviewers report the experimental intervention associated most strongly with the favourable outcome.

Did the review explicitly address a sensible clinical question?

One needs to consider if it is appropriate to aggregate the various studies together in the systematic review. It is important to consider if the underlying biology suggests that across the range of interventions aggregated, one would expect a similar treatment effect. Appropriate eligibility criteria for study inclusion in the systematic review is important:

a. Are the results likely to be similar across the range of included patients?
b. Are the results likely to be similar across the range of studied interventions?
c. Are the results likely to be similar across the range of ways in which the outcome was measured, e.g. duration of follow up?

Explicit eligibility criteria will ensure that the authors' own biases are less likely to influence which studies are included.

Was the search of relevant studies exhaustive?

The literature search needs to be exhaustive, covering a number of biographic databases, e.g. MEDLINE, EMBASE, Cochrane Central Register of Controlled Trials, etc. The reference lists of included articles need to be scrutinised for other studies which may have been missed. Other strategies include the review of abstracts of scientific meetings, and databases of ongoing trials. The appendix of the systematic review should include the exact search strategy used, including search terms for each database.

To limit reporting bias, attempts to identify unpublished studies should also be made, through 'grey literature' searching. Ideally, the full reports of unpublished studies (as opposed to an abstract) should be included.

Was the risk of bias of the primary studies assessed?

Studies with less rigorous methodology are more likely to overestimate the effectiveness of the intervention. A classic example is trials stopped early for efficacy.[2]

The determinants of bias are dependent on the type of study: therapy, diagnosis, harm, or prognosis. Key factors associated with limited bias are listed below:

a. Therapy: randomisation, was complete follow-up complete?
b. Diagnosis: Is the patient sample representative; was the diagnosis verified by credible criteria?
c. Harm: Adjusted for known determinants of outcome; was follow-up sufficiently complete?
d. Prognosis: Was there a representative sample of patients; was follow-up sufficiently complete?

There are a number of tools to assess bias. Some include:

For randomised trials (RoB2, https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials), and for mixed methods studies (Mixed Methods Appraisal Tool [MMAT][3]). A list of tools can be found in this paper,[4] and the EQUATOR network has lists of reporting guidelines for each type of study (https://www.equator-network.org/).

Did the review address possible explanations of between-study differences in the results?

The explanations for expected differences in outcomes should be stated *a priori*, that is, in the systematic review protocol. Subgroup analyses where differences are expected may include age cohorts or specific comorbidities, amongst others.

Did the review present results that are ready for clinical application?

For binary outcomes, the results are presented as proportions (e.g. outcomes of death, myocardial infarction, etc.), and the 'relative' efficacy of an intervention should generally be consistent across the entire cohort. Therefore, the preference is to present the relative effects of the intervention, i.e. relative risk (RR), odds ratio (OR) or hazards ratio (HR). To understand a specific patient's risk, we would then need to estimate the patient's baseline risk, and then calculate the patient's absolute risk difference from the RR. Using this information, one could calculate the 'number-needed-to-treat'.[5]

For continuous outcomes (e.g. walking disease, forced expiratory volume, etc.), we usually present the weighted mean difference (WMD) and standardised mean difference (SMD). The SMD is the mean difference divided by the standard deviation. The SMD is used for continuous data where different measurement instruments have been used to assess a similar outcome between studies.

The effect size between SD units is important to understand clinical effect; 0.2 SD is small, 0.5 SD is moderate, and 0.8 SD is large. A difference of 0.5 is generally considered to be of clinical importance. To understand the impact of the intervention, it would be possible to calculate the number-needed-to-treat for the number of patients who achieve a 'clinically important' threshold.

Were the selection and assessment of studies reproducible?

Data extraction should be conducted in duplicate by two independent reviewers. The reason for this is that two reviewers extracting data prevents mistakes (i.e. random errors) and bias (i.e. systematic errors). Good agreement between reviewers (e.g. chance-corrected agreement, such as the κ statistic) should also be reported to establish the agreement between the independent reviewers. The appendix should document the search strategy, data extraction plan and assessment of data extraction.

Did the review address confidence in effect estimates?

Addressing bias can increase the confidence in the effect estimates. A meta-analysis would decrease the imprecision (by decreasing the width of the confidence interval [CI]), and document any inconsistencies through the heterogeneity between study results. Authors need to make an explicit assessment of the confidence in the estimates of effect.

The credibility of the individual studies

Interpretation and understanding the effect estimate of a meta-analysis includes assessing the credibility of the individual studies. An example is the 'risk of bias' tool (RoB2, https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials).

The forest plot

Different studies have different weightings, based on i) size of the study, and ii) number of events. Studies with a higher weighting have narrower CIs, and the point estimate is represented by a square, which is larger due to the increased weighting.

The pooled estimate of the individual studies is shown as a diamond, with the width showing the CI. There is a vertical line of no effect. If the CI crosses the line of no effect, it is uncertain whether there is a difference between interventions.

Presentation of outcomes

Dichotomous outcomes are reported as RR or OR, and continuous data as WMD or SMD, as discussed above.

Assessing the confidence in the estimates and GRADE recommendations

Based on the assessment of the quality of the individual studies, the consistency of study results, and the local applicability, the Grading of Recommendations, Assessment, Development and Evaluations (GRADE) provides a transparent framework for developing and presenting summaries of evidence and provides a systematic approach for making clinical practice recommendations.[6]

The GRADE confidence in estimates of effect has four categories: high, moderate, low and very low. The lower the confidence, the more likely it is that the observed effect estimate is substantially different from the true effect. Confidence in the effect estimate is determined by:

- *The study design.* Randomised controlled trials are assumed to have higher confidence, and observational studies are low. These initial estimates of confidence are further modified by risk of bias.
- *Risk of bias.* This is systematic rather than random error. It may be due to inappropriate or suboptimal i) randomisation sequence, ii) allocation concealment, iii) blinding of patients, caregivers or study personnel, and iv) lost to follow-up. Risk of bias can be assessed by the Cochrane Collaboration risk of bias assessment tool. Bias can lead to 'no change' assigned to confidence ratings, or to a 1 or 2 level downgrading.
- *Inconsistency.* The assumption is that the treatment effect applies to a broad range of patients. However, this may not be the case across different groups of patients, and may require an *a priori* defined subgroup analysis. Consistency is evaluated by:
  - i. The visual assessment of variability. Visual assessment would show point estimates (on the same side of the line of no effect) and the CIs of the various studies overlapping if the studies provided consistent findings. Causes for concern, or inconsistency between studies, would be associated with study point estimates which are far apart, and CIs which do not overlap.
  - ii. Yes or no statistical test of heterogeneity. The Cochran Q is chi-squared test which assumes the difference between studies is due to chance. A significant finding, therefore, suggests significant 'inconsistency' of results between studies. A word of caution for studies with large sample sizes, as larger studies may generate a statistically significant result, although there may be no clinically important heterogeneity.
  - iii.Magnitude of heterogeneity (variability). The I^2 statistic focuses on the magnitude of variability as opposed to the statistical significance. At about 25%, we would be getting concerned about the consistency of the findings between studies, and at 75%, we would consider the findings inconsistent between studies.

When the between-study variability is large, one needs to consider factors which may have contributed to this. These may include different population effects, e.g. ill versus less ill patients, differences in the intervention between studies, e.g. different doses, and differences between comparators, e.g. control receiving other treatment versus control receiving placebo. A test of interaction for these subgroups is necessary to determine whether this occurred by chance. A significant finding suggests that the differences in effect estimates cannot be attributed to chance alone, and these may be real differences between the subgroups. Remember, if these are *post hoc* analyses, then these data are only 'hypothesis generating'. Inconsistency would lead one to consider whether it was appropriate to include these studies in a meta-analysis in the first place. Any residual inconsistency would require downgrading of the confidence in the estimates in the GRADE recommendations.

- *Imprecision.* Precision is dependent on the width of the CI. If our clinical decision making remains consistent across the 95% CI, i.e. lower and upper boundary, then this would increase our confidence in the effect estimate. If our clinical decisions would change from the lower to upper boundary, then we would have less confidence in the effect estimate, and this is due to the imprecision of the findings. To test this, we would need to determine the absolute risk difference, and number-needed-to-treat at the lower and upper boundary. This would determine the clinical utility of the intervention.

- *Indirectness.* Directness means that the research applies to our population of interest, that the interventions are appropriate in our population, and the outcomes are important to our population. Indirectness would include studies where the populations differ from ours, interventions which are tested against a placebo and not our standard of care, and outcomes which are surrogates for the real outcomes of interest.
- *Publication bias.* This is most likely when negative studies are not published (reporting bias), when specific outcomes are reported (selective outcome reporting), and reporting in less prominent journals (dissemination bias). Reporting bias can be assessed by a funnel plot, where we would expect studies to be symmetrically arranged around the summary estimate, with the larger studies closer to the summary estimate, and all quadrants populated with studies. If this is not the case, reporting bias may be a concern. Selective outcome reporting can be identified by assessing the registration protocols of studies and the listed primary outcome. Studies registered late, or unregistered should raise concerns about reporting bias.
- *Effect size.* A larger effect size should increase confidence. However, remember if it is implausible, or due to studies of low quality, then this may be the reason for the large effect size.

Conclusion

Once a systematic review and meta-analysis has been conducted, ideally an evidence-based summary of the findings should be produced as an evidence profile. This allows for knowledge translation and communication with patients concerning informed choices about care. A good example is the 'Living WHO guideline on drugs for COVID-19' (https://www.bmj.com/content/370/bmj.m3379).

ORCID

BM Biccard https://orcid.org/0000-0001-8666-4104

References

1. Guyatt GH, Rennie D, Meade MO, Cooke DJ. Users' guides to the medical literature: A manual for evidence-based clinical practice. Third ed. New York: McGraw-Hill Education; 2015.
2. Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005;294(17):2203-9. https://doi.org/10.1001/jama.294.17.2203.
3. Pluye P, Gagnon MP, Griffiths F, Johnson-Lafleur J. A scoring system for appraising mixed methods research, and concomitantly appraising qualitative, quantitative and mixed methods primary studies in Mixed Studies Reviews. Int J Nurs Stud. 2009;46(4):529-46. https://doi.org/10.1016/j.ijnurstu.2009.01.009.
4. Ma L-L, Wang Y-Y, Yang Z-H, et al. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? Mil Med Res. 2020;7(1):7. https://doi.org/10.1186/s40779-020-00238-8.
5. Barratt A, Wyer PC, Hatala R, et al. Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. CMAJ. 2004;171(4):353-8. https://doi.org/10.1503/cmaj.1021197.
6. Siemieniuk R, Guyatt G. What is GRADE? London: BMJ Best Practice; 2021. Available from: https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/. Accessed 15 Sept 2021.