NERDCAT-SRMA

A clinician's guide to appraising systematic reviews and meta-analyses

Table of Contents

FUNDAMENTALS

ADVANCED

Each section of this tool has 2 sides: Odd-numbered pages provide the rationale and supporting empiric evidence for assessing each study domain. Even-numbered pages include key questions to pose while appraising a study. When first learning to critically appraise, be sure to **read and consider every question** in the even-numbered pages, and refer to the rationale sections during interpretation.

To print only the key questions for critical appraisal, select the option to print only even-numbered pages in the print menu of your PDF software.

The latest version of this tool and other NERDCATs can be found at nerdcat.org

Full references to supporting literature and articles used as examples in this document can be found at: http://nerdcat.org/useful-references/

Article title: ____

CLINICAL QUESTION AND ELIGIBILITY CRITERIA

P *Age, condition, setting, etc.*

I *Drug, dose, frequency, duration, etc.*

C

O *Clinical outcomes assessed, etc.*

Trial Inclusion/Exclusion

GENERALIZABILITY (i.e. how do I apply the results to my patient)?

1. **Does my practice setting differ significantly from that in the trials?**

   Some questions to consider:

   - Timeframe: Have there been any large shifts in practice between now and when this trial was undertaken?
   - Setting: 1°, 2°, or 3° care?
   - Were the trials single center or multicenter?
   - Were there Canadians included in the trials?
   - Am I able to offer the same extent and quality of monitoring as that provided in the studies?

2. **Were there important clinical differences (PICO) between study participants and my patient?**

   Consider: Age, sex, race, diagnostic criteria, comorbidities, pathology, previous interventions, concomitant interventions

3. **Are the interventions evaluated in the trial similar to those available in my practice?**

   a. Same drug?
   b. Same dosage form?
   c. Same dose?
   d. Same release mechanism?
   e. Is there evidence supporting or opposing a class effect?
   f. Are the likely treatment benefits worth the potential harm and costs, considering the relative importance of these outcomes?

SEARCH

In one study, MEDLINE, EMBASE, & CENTRAL each identified only 69%, 65% and 78.5% of relevant trials. Combining these 3 databases together only misses 2.4% trials found searching an additional 26 databases.

Aims to minimize the effects of ***publication bias***, which typically leads to over-estimation (typically in favor) of the effect of an intervention.

Why is publication bias so concerning?

- 😐 Studies with statistically significant results (“positive” studies) are twice as likely to get published, and will typically get published faster (by median 1.3 years in one study) compared to trials with statistically non-significant results (“negative” studies).
- 😐 Published trials produce a 15% larger estimate of effect compared to unpublished trials.
- 😐 Although more common with industry-funded trials, government-funded studies are still prone to publication bias (32% vs 18% unpublished 5 years after completion)!
- 😐 In one study, 90-98% of meta-analyses with very large effects observed in early trials became substantially smaller once subsequent studies became available (e.g. median odds ratio decreased from ~11 to ~4 after more trials were added to the first trial).
- 😐 In one study of 42 meta-analyses, addition of unpublished FDA outcome data changed the efficacy summary estimate (either increased or decreased) compared to the meta-analysis based purely on published outcome data.

<u>**Bottom line:**</u> **Meta-analyses of only published trials will overestimate the effects of drugs and other interventions, especially when they are performed earlier on (i.e. before the “negative” trials get published). Thus, it is suggested that the risk of publication bias is greater in meta-analyses based on a few small studies.**

SEARCH

4. Databases of published literature – Was every relevant database searched?	□ Search timeframe: □ MEDLINE □ EMBASE □ CENTRAL □ Other relevant topic-specific databases:
5. Unpublished studies – Was a sufficient effort made to find unpublished studies (a.k.a. grey literature)?	Clinical trial registries: □ WHO International Clinical Trials Registry (includes clinicaltrials.gov & many more) □ ClinicalTrials.gov □ International Federation of Pharmaceutical Manufacturers and Associations [IFPMA] Clinical Trials Portal (pharmaceutical industry-sponsored trials) □ Others: Regulatory body websites □ drugs@FDA □ EMA □ Others: □ Poster presentations, conference proceedings, or abstracts:
6. Additional measures for comprehensiveness – Were sources of additional published/unpublished data sought out?	□ Handsearch of reference list of included studies □ Contacted study authors for unpublished data □ Not limited by language (e.g. English-only) □ Others:

RESULTS OF THE SYSTEMATIC REVIEW (i.e. which trials were included?)

✓ The Cochrane risk of bias tool evaluates the risk of individual trial biases & offers the most transparent assessment of trial internal validity (see NERDCAT-RCT for more)

💣 “Quality scores” such as the well-known Jadad score are more closely related to reporting quality than methodological issues, & lead to wide variability in conclusions on “quality” based on the score used

See NERDCAT-RCT to learn more on how to appraise validity of subgroup effects

RESULTS OF THE SYSTEMATIC REVIEW (i.e. which trials were included?)

7. Eligible trials – a. Do all inclusions & exclusions of trials make sense? b. Are all the trials that you would expect to see included in the systematic review?	□ Inclusions/exclusions adequately described (e.g. PRISMA flowchart) □ Irrational eligibility criteria: □ Missing trial(s):
8. Risk of bias within trials (trial internal validity) – a. Did reviewers adequately assess individual trials for risk of bias? b. Was each component reported separately, or summarized with a composite quality score?	□ "Quality control": Do you agree with the risk of bias reported for the largest weighted trial?
9. Clinical and methodological heterogeneity – a. Are there any differences in clinical characteristics between the individual trials (i.e. any component of PICO) that preclude pooling the trials together in meta-analysis? b. Are there methodological differences (i.e. risk of bias) between studies? c. Is the impact of any of these characteristics tested in subgroup analysis? Examples of sources of clinical heterogeneity based on PICO format: P: Year/era, geographic location, age, sex, diagnostic criteria, duration of illness, comorbidities, past interventions, co-interventions I/C: Dose, route, formulation, duration, adherence, monitoring O: Outcome definition, diagnostic or measurement criteria, follow-up duration	

RESULTS OF THE META-ANALYSIS (i.e. what do the pooled results of the trials show?)

Assessing for the presence & extent of ***outcome reporting bias***

Why is outcome reporting bias so concerning?

- 😐 In one study, identified RCTs only reported on 50% of efficacy and 65% of harm outcomes. Additionally, ~2/3 of the trials had their primary outcome changed in the final published reported compared to the original protocol.
- 😐 A study by the same author found outcome reporting bias present in government-funded studies. In the identified study, negative studies were most likely to have reporting issues (i.e. reported as "not statistically significantly different" without reporting absolute values)
- 😐 In one study of 42 meta-analyses, addition of unpublished FDA outcome data changed the efficacy summary estimate (either increased or decreased) compared to the meta-analysis based purely on published outcome data

<u>Bottom line:</u> As with whole trials, negative outcome results are less likely to be published than positive results. Since most systematic reviews rely heavily on published outcome data, outcome reporting bias poses a serious risk to the accuracy of intervention effect estimates (i.e. overestimation of benefits and underestimation of harms, distorting the true benefit/harm balance).

- ✓ Statistical heterogeneity is assessed using either Cochran's Q test, or I^2 statistic (preferred)
  - Cochran's Q is a yes/no test that shows statistical evidence of heterogeneity if $p < 0.10$ (analogous to the test for interaction used in subgroup analyses)
  - I^2 ranges from 0-100% & represents the amount of variability in the point estimate across trials. Rule-of-thumb (one of many): I^2 <25% = minimal variability; I^2 >50% = substantial variability (may not be appropriate to meta-analyze trials)
- ✓ If present, requires either/both:
  - Different statistical approach to pool the results (i.e. random-effects model, see below)
  - Evaluation of clinical & methodological sources of heterogeneity

Note: Trials with very different point estimates but wide confidence intervals may falsely show little or no heterogeneity with statistical tests. The opposite is true for trials with very small confidence intervals. Thus, heterogeneity tests should always be considered with visual evaluation of differences in individual-trial point estimates & confidence intervals

Either the *fixed-effects model* or *random-effects model* may be applied

- The fixed-effects model assumes that all trials estimate the same underlying "true" effect, and thus that any differences between trials are due to chance
- The random-effects model does not assume that all trials estimate the exact same underlying effect (i.e. that different populations may vary in their response to intervention), & thus incorporates heterogeneity
- In many cases, both models produce very similar meta-analytic results.
  - With significant statistical heterogeneity that does not preclude pooling results, the random-effects model will be more conservative (i.e. have a wider confidence interval to illustrate the increased uncertainty)
  - HOWEVER, the random-effects model gives greater weight to trials with less statistical information (i.e. smaller trials or trials with fewer outcome events). In rare cases (where often pooling the results was not appropriate in the first place), the random-effects model can "pull" the summary estimate towards the smaller trials (which may be more prone to within-study bias & publication bias). Statistics cannot fix poor data.

- ✓ Effect measures for dichotomous data (yes/no): Relative risk (RR), odds ratio (OR), hazard ratio (HR), risk difference (RD; a.k.a. absolute risk reduction/increase)
- ✓ Common effect measures for continuous data: Weighted mean difference (WMD), standard mean difference (SMD)

<u>Confidence intervals do not always represent an accurate estimate of the precision of the summary estimate</u>, particularly when based on a small number of patients or outcome events
Learn more in the Advanced Topics section *"How do I know when there is enough evidence?"*

RESULTS OF THE META-ANALYSIS (i.e. what do the pooled results of the trials show?)

10. Clinical importance & completeness – a. Are all the important outcomes included in the review? b. Are all of the reported outcomes clinically important? c. What fraction of the included studies report on these outcomes? <u>Hierarchy of outcomes in order of importance:</u> i. Death ii. Serious adverse events (SAEs) or quality of life (QoL) iii. Morbidity, complication of condition, adverse effects iv. Changes in surrogate markers	
11. Statistical heterogeneity – a. Based on visual assessment of the forest plot, does it make sense to pool the results? b. Is there significant variability between the studies? ▪ If so, was it dealt with appropriately given the level of heterogeneity? ▪ If not, were individual trials precise enough to truly rule-out heterogeneity?	Assess the following for each outcome: □ I^2 = □ Appropriate to pool the results & interpret the summary statistics
12. Statistical models: Fixed- vs random-effects – Was the appropriate statistical method used to pool results?	Assess the following for each outcome: □ Fixed-effects or □ Random-effects; is it appropriate?
13. Effect measure and precision – a. Is the effect measure chosen appropriate for meta-analysis? b. Is the absolute effect clinically important? c. Are the results clinically important at both ends of the confidence interval (i.e. of the absolute effect)?	□ <u>Relative</u> effect measure (HR, OR, RR) used for meta-analysis □ Calculate the baseline risk for your patient from the individual trial they would fit best in □ Calculate <u>absolute</u> effect NNT NNH

ADVANCED TOPIC #1: TRIAL SEQUENTIAL ANALYSIS (TSA) (i.e. how do I know if there is enough evidence?)

Why perform trial sequential analysis (TSA)?

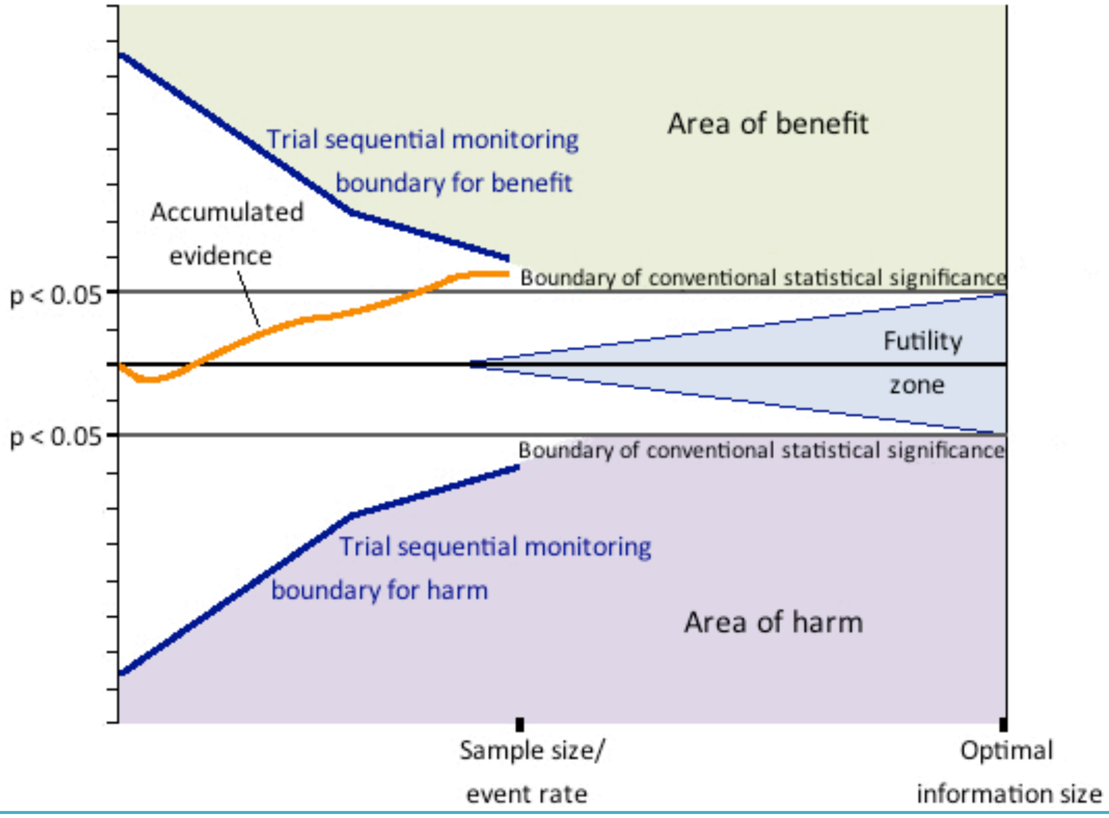
Systematic reviews are typically updated or repeated as new trials are published. For example, the Cochrane Collaboration requires that Cochrane reviews be updated at least every two years to remain up-to-date. These frequent updates lead to repeat significance testing. In RCTs, unadjusted interim analyses, wherein statistical tests with a threshold of p<0.05 are repeatedly performed as patients enter the trial and experience outcome events, have been shown to lead to frequent spurious false-positive findings and inappropriate early stoppage of trials. The same issue occurs in updated meta-analyses when a new trial is identified and added. Statistical simulations suggest that the false-positive rate of meta-analyses may be as high as 10-30%.

TSA has been proposed to minimize this problem. The first step is the calculation of a reasonable sample size (or number of events), which for meta-analyses is termed the optimal information size (OIS) that would be required to demonstrate a clinically important difference between the intervention and control. In addition to the usual alpha, beta (1-power), baseline event rate and effect size assumptions required for a RCT sample size calculation, the OIS also incorporates knowledge of the statistical heterogeneity present in the data.

The OIS is the point at which the conventional threshold for statistical significance (p<0.05) holds true. If there is a total sample size or event rate lower than the OIS, the threshold must be adjusted (i.e. made stricter) for the suboptimal accumulated information, which is called the trial sequential monitoring boundary. If the calculated p-value crosses the trial sequential monitoring boundary before the OIS has been reached, it is likely that the effect is true and believable.

Caution: Relying purely on TSA is the same as worshipping the p-value; it is foolish. TSA simply acts as a barrier to making strong conclusions prematurely in an intervention's evidence timeline. Always apply logical clinical reasoning when weighing and interpreting TSA, just as you do throughout your appraisal of the literature!

Example TSA graph:



ADVANCED TOPIC #1: TRIAL SEQUENTIAL ANALYSIS (TSA) (i.e. how do I know if there is enough evidence?)

The steps below must be repeated for each outcome with a TSA, as TSA is outcome-specific

A. Optimal information size (OIS) - a. Has an OIS been calculated? b. Is it calculated based on reasonable assumptions of treatment effect?	□ Alpha (usually 0.05) = □ Beta (usually 0.1 or 0.2) = □ Baseline event rate (i.e. control group event rate) = □ Effect size (e.g. HR, RR, OR) = □ Heterogeneity = □ OIS =
B. Accumulated evidence – Does the total sample size/event rate of the meta-analysis achieve the OIS?	
C. Conventional statistical significance – Are the results statistically significant based on the conventional threshold for statistical significance (Z ≥1.96 or p ≤0.05)?	
D. Trial sequential monitoring boundaries – Does the accumulated evidence cross any of the trial sequential monitoring boundaries (i.e. benefit, harm, or futility/equivalence)?	
E. For each clinically relevant outcome, is there sufficient evidence to conclude benefit, harm, or equivalence?	