NERDCAT-RCT

A clinician's guide to appraising randomized controlled trials

Table of Contents

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: _____

TRIAL'S CLINICAL QUESTION

P Average patient (SCRAPP mnemonic):

- Sex:
- Comorbidities:
- Race:
- Age:
- Pathology, stage/severity of disease:
- Previous interventions:

Exclusion criteria:

I *Drug, dose, duration*

Co-interventions:

O *Clinical outcomes measured*

T *Follow-up duration*

GENERALIZABILITY – DO THESE RESULTS (NOT) APPLY TO MY PATIENTS?

11. Does my practice setting differ significantly from that in the trials? Some questions to consider: ▯ Era: Same diagnostic criteria used for disease/outcome being studied? ▯ Setting: 1°, 2°, or 3° care? ▯ Single centre or multicentre trial? ▯ Country: Were there Canadians in the trials?	
12. Were there important clinical differences between study participants and my patient (i.e. SCRAPP characteristics)? *E.g. HYVET, a RCT that assessed indapamide ± perindopril to target BP <150/80 in patients ≥80 y, included an elderly population that was overall less frail (e.g. diabetes in 7%, previous MI in 3%) than the average octogenarian*	
13. Are the interventions evaluated in the trial similar to those available in my practice?	☐ Same drug (or evidence supporting a class effect) ☐ Similar route, formulation and release mechanism ☐ Same dose & frequency/regimen ☐ Same monitoring plan is feasible, and all parameters are readily measurable
14. Are the outcomes evaluated in the trial readily measurable?	

RISK OF BIAS – ARE THE RESULTS RELIABLE?

😐 Unclear or inadequate sequence generation exaggerate relative benefits of an intervention by ~11%
😐 **Adequate randomization**: Computer-generated, random-number table, coin toss, drawing cards.
😐 **Inadequate randomization**: Quasi-randomized (alternation by case number or date of birth)

😐 Unclear or inadequate allocation concealment exaggerate relative benefits of an intervention by ~7%
✓ **Adequate allocation concealment**: Central randomization (including pharmacy-controlled), coded identical drug boxes, sequentially-numbered, sealed opaque envelopes (preferably lined with cardboard or foil), on-site locked computer system
💣 **Inadequate allocation concealment**: Allocation scheme posted on a bulleting board, non-opaque envelopes

To confirm that:
- Randomization was successful (no obvious biased imbalances in measured prognostic factors reassures us that there are no differences in unmeasured prognostic factors, i.e. confounding)

E.g. If 15% of patients in the intervention group were diabetic vs 25% in the control group, this difference could partly explain a slightly lower risk of cardiovascular events in the intervention group

😐 Lack of or unclear double-blinding is associated with a ~13% exaggeration of the relative benefits of an intervention for dichotomous outcomes, and a 68% exaggeration of relative benefits for subjective continuous outcomes
✓ **Adequate blinding of participants & personnel**: used identical placebo/control product without indication that treatments were distinguishable
💣 **Inadequate blinding of participants & personnel**: PROBE (prospective randomized open blinded endpoint), open-label
- *E.g. In an FDA re-analysis of the RECORD trial, it was discovered that MI event records in the rosiglitazone were selectively withheld from the study's blinded adjudication committee by non-blinded study personnel. When the primary analysis was re-calculated with these events added, the HR changed from 1.14 (95% CI 0.80-1.63) to 1.38 (95% CI 0.99-1.93), in line with results from a meta-analysis.*

✓ **Adequate blinding of outcome assessors**: Independent central adjudication committee adjudicated all outcomes
✓ **Situations difficult to blind**: The intervention has an effect on readily-measurable biomarker
- *E.g. HPS was a RCT evaluating the effect of simvastatin 40 mg/d vs placebo on mortality and CV events. By the 5th year of follow up, 32% of patients in the placebo group were receiving a non-study statin, likely due to higher LDL levels, therefore attenuating the difference seen in CV outcomes.*
  - Note: One group receiving the treatment specified for the other group is called contamination
- Some situations initially thought to be impossible to blind can be successfully blinded with some ingenuity
  - *E.g. In ROCKET-AF, INR was measured centrally and clinicians taking care of patients on rivaroxaban were given dummy INR values for which to adjust the warfarin-placebo dose*

- Rules of thumb (e.g. ≥20%) are misleading; loss-to-follow-up is significant when it is similar to or greater than the occurrence of the outcome of interest
- If there is differential loss to follow-up, do your own rudimentary "worst-case scenario" analysis: would the results remain similar if all participants lost-to-follow-up in one treatment group had suffered the bad outcome whilst all those lost-to-follow-up in the other group had had a good outcome?
- ITT analysis (see below) cannot correct the bias introduced by differences in loss-to-follow-up between groups
- *E.g. In a trial assessing quetiapine vs placebo for adjunctive treatment of depression, discontinuation due to adverse events in the placebo, quetiapine 150 mg, and quetiapine 300 mg groups were 1%, 11%, and 18%, respectively*

There are numerous methods to carry out an intention-to-treat analysis (e.g. last observation carried forward (LOCF), mixed model for repeated measurements (MMRM), sensitivity analyses). All of them rely on assumptions and no single method works in every situation
- E*.g. LOCF is the most common approach to ITT used in dementia trials evaluating the efficacy of cholinesterase inhibitors, despite violating the necessary LOCF assumption that, if left untreated, disease severity will remain stable. Patients given cholinesterase inhibitors tend to discontinue earlier in the trial (earlier in the decline) due to intolerable side-effects, giving the appearance that the patient's cognition has ceased to decline.*

Hierarchy of outcomes:
a) Death
b) Serious adverse events (SAE)/ quality of life (QoL)
c) Clinically important morbidity, adverse effects
d) Withdrawals
e) Surrogate markers

Does size matter?

😐 **A review found that, compared to large trials, small trials only show inaccurate treatment effects when they are not adequately randomized, allocation concealed or blinded**

RISK OF BIAS – ARE THE RESULTS RELIABLE?

Allocation bias	**1. Sequence generation** – Were patients appropriately randomized?	
	2. Allocation concealment – Was randomization concealed?	
	3. Baseline characteristics a) Were there any clinically important differences with respect to known prognostic factors at the start of the trial? b) Are the differences large enough to explain a difference in outcomes between groups?	
Performance (from patients & clinicians) **& detection bias** (from outcome assessors)	**4. Blinding of:** a) Participants & personnel b) Outcome assessors Note: Blinding is only possible if allocation is concealed.	
Attrition bias	**5. Loss to follow-up** – Was follow-up complete (i.e. were all patients accounted for at the end of the trial)? Did groups differ in: a) How many were lost to follow-up? b) Why patients were lost to follow-up? c) When patients were lost to follow-up?	
	6. Intention to treat (ITT) a) Were patients analyzed in the groups to which they were randomized (ITT), or were only patients who were adherent to their study treatment (per protocol) or completed the full trial duration (completer analysis) counted? b) Are the ITT methods used to account for loss-to-follow-up appropriate?	
Reporting bias	**7. Were all the important outcomes considered?**	

THE RESULTS

- Look at both relative and absolute differences
- Relative differences are typically assumed to be reasonably constant across populations; absolute differences depend on baseline risk
  - *E.g. Statins reduce the relative risk of all-cause mortality by 10-15% in both primary and secondary CV prevention. In patients WITH prior CHD, this translates to an absolute risk reduction in mortality of 2.2% over 5 years; in patients WITHOUT CHD and low risk of cardiovascular disease, the absolute reduction in mortality is 0.42%*

- Defined by the absolute - rather than a relative - risk reduction
  - *E.g. In CAPRIE, a RCT comparing clopidogrel vs aspirin in over 18,000 patients at high risk of CV events, the absolute risk reduction in the composite primary outcome (ischemic stroke, MI, vascular death) was 0.5% per year NNT = 200*
- Contrast with absolute risk reductions or NNTs achieved with other interventions used in a similar patient population
  - *E.g. In HPS, NNT = 58 for all-cause mortality over 5 y with simvastatin vs placebo in a high-risk population. In HOPE, NNT = 54 for ramipril vs placebo in a similar patient population over 5 y*

- Look at the width of the CI
  - *E.g. Narrow CI: RR 0.90 (0.85 to 0.95)*
  - *E.g. Wide CI: OR 1.25 (0.2 to 5)*
- Are the results clinically important at both bounds of the CI?

E.g. In CAPRIE, the lower end of the RRR CI ("worst-case") was 0.3% and the upper end ("best-case") was 16.5%, corresponding to a NNT of 5555 and 105 per year, respectively

SECONDARY OUTCOMES – Can conclusions be made from outcomes other than the primary one?

Secondary analyses may be highlighted when the primary endpoint fails to cross the threshold of statistical significance

- *E.g. The FIELD trial, which evaluated fenofibrate vs placebo in patients with type 2 diabetes, failed to show a statistically significant reduction in coronary events at 5 years. In their conclusions, authors highlighted marginally statistically significant reduction in two secondary efficacy outcomes: total cardiovascular events and non-fatal MI*

😐 **More comparisons = greater risk of type 1 error (finding a difference when there is none)**

Outcomes with similar pathophysiology (e.g. MI and ischemic stroke with antihypertensives) should move in the same direction (both increased or both decreased), whereas outcomes with opposing pathophysiology (e.g. MI and bleeding with antiplatelets) should move in opposite directions

- *E.g. In FIELD, though the secondary outcome of non-fatal MI was statistically significantly improved with fenofibrate, all-cause mortality, coronary death, DVT, and PE occurred more frequently in the fenofibrate group*

- Sample size calculations are based on predicted rates of the primary outcome, which is frequently chosen on the basis of a higher event rate (i.e. will happen to more patients) than secondary outcomes
- One should be skeptical whenever an unexpected statistically significant reduction is found in a rare secondary outcome, particularly when there is no difference in the more common primary outcome
  - *E.g. The ELITE trial comparing losartan to captopril in 722 elderly heart failure patients failed to find a significant difference in the incidence of the primary outcome, increase in serum creatinine (10.5% in both groups). There was, however, an unexpected reduction in all-cause mortality with losartan vs captopril (4.8% vs 8.7%, p=0.035). The follow-up ELITE II trial with its larger sample of 3152 patients and a primary outcome of mortality found no reduction in – and in fact numerically higher – mortality with losartan vs captopril (18% vs 16%, p=0.16)*

"NEGATIVE" TRIALS – If the difference between interventions is not statistically significant, are they truly no different?

If the CI is wide enough to include a clinically important difference, it is still possible that the interventions differ

- *E.g. Authors of a trial evaluating the effect of adding N-acetylcysteine to prednisone in 180 patients with acute alcoholic hepatitis concluded that mortality was not reduced with the combination vs steroid alone. At 6 months, 27% of patients died in the combination group vs 38% of patients taking prednisone (ARR 11%; 95% CI ARI of 5 to ARR 22%). The uncertainty of the estimated reduction represented by the CI means that the trial could not exclude the possibility of an absolute reduction in mortality as high as 22%*

THE RESULTS	
8. Point estimate – How large were the treatment effects for benefits and harms?	
9. Are the results clinically important?	
10. Confidence interval – How precise were the estimates of treatment effect?	

SECONDARY OUTCOMES – Can conclusions be made from outcomes other than the primary one?	
Are we data-mining? – Was the primary endpoint found to be statistically significantly different?	
Multiplicity – Was the secondary endpoint one of a small number of secondary endpoints defined in the original protocol?	
Consistency – Does the secondary endpoint result make sense in the context of the primary – and other secondary - outcome findings?	
Power – Was the trial powered to find a difference in this secondary outcome?	

"NEGATIVE" TRIALS – If the difference between interventions is not statistically significant, are they truly no different?	
Does the confidence interval (CI) exclude a clinically important difference?	☐ Yes, so we can be reassured that the findings are truly "negative"

COMPOSITE OUTCOME – Was the primary outcome a combination of outcomes?

E.g. The primary outcome of CONDOR, a trial comparing celecoxib vs NSAID+PPI, was a composite of GI bleed, obstruction or perforation or clinically significant anemia (Hb drop ≥ 20 g/L or Hct drop ≥10%)

E.g. In CONDOR, component of the primary outcome and their rates for celecoxib vs NSAID+PPI:

- *GI bleed (0.17% vs 0.17%)*
- *GI obstruction (0% for both groups)*
- *GI perforation (0% for both groups)*
- *Clinically significant anemia (0.67% vs 3.4%) – the greatest contributor of events and least clinically important*

E.g. RRRs in CONDOR for celecoxib vs NSAID+PPI

- *Composite RRR = 75%*
- *GI bleed RRR = 0%*
- *Clinically significant anemia RRR = 80%*

E.g. In the UKPDS blood pressure target trial, the primary outcome was a composite of 21 outcomes including those resulting from vascular damage (e.g. stroke, renal failure), malignancy, and extremes in plasma glucose. Only the vascular events have a biological rationale for being reduced by improved blood pressure control

TRUNCATED STUDIES – Was the trial stopped early for "overwhelming" evidence of benefit or futility?

E.g. In JUPITER, a RCT of rosuvastatin vs placebo in a highly-selected primary CV prevention population, the pre-planned stopping rule was mentioned, though poorly described, in an early report: "Frequency of interim efficacy analyses and rules for early trial termination have been prespecified and approved by all members of this board."

See ref 12 for most commonly used interim analysis statistical "stopping boundaries"

E.g. JUPITER was stopped after the first of two interim analyses using "O'Brien-Fleming stopping boundaries determined by means of the Lan-DeMets approach," (which requires a p-value <0.005). The actual p-value for the primary endpoint was <0.00001

😐 Studies stopped early for benefit exaggerate the relative effect of an intervention by an average 29%

- As events accumulate, the likelihood that chance is inflating the true effect decreases
- Optimal: ≥500 events
- You should not believe RRRs ≥50% generated in truncated trials with <100 events
- The larger the number of events and the more plausible the RRRs (~20-30%), the more you can believe the results

E.g. In JUPITER, 393 primary (composite) endpoint events occurred between the two groups by the interim analysis. The RRR for the primary endpoint was 44%, and the RRRs for individual components ranged from 18-54%.

COMPOSITE OUTCOME – Was the primary outcome a combination of outcomes?

- ✓ If "yes" to all 4 questions below: Feel comfortable using the effect on the composite outcome as the basis for decision-making
- 💣 If no to any: Look at the effect on each of the individual components of the endpoint for decision-making

Importance – Are the component endpoints of the composite endpoint all of similar importance to patients?	
Statistical contribution – Did the more and less important endpoints occur with similar frequencies?	
Consistency in effect of therapy a) Are the point estimates of treatment effect (HR, OR, RR) similar between each component? b) Do the CIs overlap? Are they sufficiently narrow?	
Biologic rationale – Do the components share a similar underlying biological mechanism?	

TRUNCATED STUDIES – Was the trial stopped early for "overwhelming" evidence of benefit or futility?

Was there a pre-planned stopping rule?	
Did the stopping rule involve few interim looks and a stringent p-value (e.g. <0.001)?	
Did enough endpoint events occur?	

SUBGROUPS – Were additional comparisons made on segments of the study population?

- Subgroup analyses that were not predefined in the protocol may be a form of data-mining, and are vulnerable to finding a difference by chance
- Do not believe unanticipated significant subgroup differences (i.e. discovered post *hoc*) until they have been replicated in other studies

Subgroup effects that are significant but go in the direction opposite to what was expected are less credible than correct predictions

More comparisons = more likely to find a difference by chance

- Subgroup analyses of variables measured after randomization may be affected by the interventions, and thus can only demonstrate an association rather than causation
- *E.g. Measured at baseline: age, gender*
- *E.g. Measured after randomization: achieved LDL in fixed-dose statin trial, achieved BP in trial comparing 2 fixed-dose antihypertensives*

- Randomization ensures that confounders are spread evenly between groups in the overall study population, but not within subgroups, especially when these subgroups contain a small number of subjects
- Randomization stratified for the subgroup results in separate randomization within each subgroup, which minimizes baseline differences in confounders

- Subgroup effects identified between studies, such as in two trials in a systematic review, may be due to methodological or clinical differences between trials rather than true associations with the different subgroups
- *E.g. The Physicians' Health Study, a study of men without previous CV disease, found that low-dose ASA reduced the risk of MI but not stroke to a statistically significant degree. Many years later, the Women's Health Study demonstrated a statistically significant reduction in stroke but not MI with ASA in women without previous CV disease. It would be inappropriate to conclude based on an indirect comparison of these two RCTs that ASA has different benefits in men and women*

Subgroup analyses can be used for data-mining when overall results are "negative"

😐 A review of 117 subgroup claims in 64 RCTs found that <40% of subgroup claims reported in the abstract were statistically significant

- Determined by looking for the test for interaction (i.e. treatment effect differs across subgroups, similar to test for heterogeneity conducted in meta-analysis)
- "Positive" subgroup analyses that do not report the test for interaction p-value should be ignored
- *E.g. In HPS, subgroup analysis based on gender, one of the 17 subgroup analyses reported, did not show a statistically significant test for interaction (p=0.18), meaning women were not less likely to benefit from statin therapy than men*

E.g. MI and ischemic stroke; hallucinations and agitation; weak urine flow and straining

😐 A review found that attempts with a subsequent RCT or meta-analysis were made in only ~10% of cases to corroborate the 117 subgroup claims in 64 RCTs, of which none replicated the subgroup effect

SUBGROUPS – Were additional comparisons made on segments of the study population?

Design	**A. Was the subgroup analysis pre-defined (i.e. defined *a priori*)?**	
	B. Was the direction of the subgroup effect pre-defined?	
	C. Was the subgroup analysis one of a small number of hypotheses tested?	
	D. Is the subgroup variable a characteristic measured at baseline or after randomization?	
	E. Could treatment effect differences between subgroups be attributable to baseline imbalances?	
	F. Is the effect suggested by comparisons within rather than between studies?	
Results	**G. Are the results in the overall study population statistically significant?**	
	H. Is the magnitude of the difference clinically important?	
	I. Is the subgroup effect statistically significant?	
	J. If test for interaction is significant, is the difference in the subgroup statistically significant?	
Reproducibility	**K. Is the direction of effect consistent across closely related outcomes within the study?**	
	L. Do other RCTs demonstrate this subgroup difference?	

NON-INFERIORITY TRIALS – Was the intervention compared to see if it is "no worse" than an established therapy?

- ✓ *E.g. In RE-LY, a trial comparing dabigatran to warfarin for the prevention of stroke and systemic embolism in non-valvular atrial fibrillation, the pre-specified MCID was a relative risk of 1.46. This was based on half the "worst case" end of the confidence interval (CI) for benefit with warfarin vs placebo. In other words, if RE-LY proved non-inferiority of dabigatran, it would, at its very worse, ~2/3 (1÷1.46) as good as warfarin for this outcome*
- 💣 *E.g. In RESET, a trial comparing a 3-month vs a 12-months duration of clopidogrel (added to aspirin) following drug-eluting stent placement, the MCID was set as an absolute risk difference of 4% without rationale. At the expected control-group event rate of 11%, this would allow for a "worse case" relative risk reduction CI of 43%. For comparison: in CREDO, the addition of clopidogrel to aspirin vs aspirin alone reduced the primary outcome by only an absolute 3% (relative risk reduction of 27%) in a similar population. In other words, the chosen MCID allows for the shorter course of clopidogrel to be as good or worse than placebo, which is clearly irrational*

- Absolute risk difference MCIDs can bias results towards non-inferiority if event rates are lower than expected
- Relative risk MCIDs are more conservative – and therefore preferable - as they scale to the event rates
- *E.g. In SPORTIF V, the intervention was non-inferior according to the absolute risk difference MCID of 2%, but it would not have been non-inferior if a relative risk MCID of 1.67 - based on the same previous study data - would have been used. The cause of this was an event rate that was lower than expected (1.2% in warfarin group vs expected 3.1%)*

- *E.g. In EINSTEIN-PE, a RCT comparing rivaroxaban to standard enoxaparin and warfarin for the treatment of acute pulmonary embolism (PE), the margin of non-inferiority for recurrent VTE was a relative risk of 2. If rivaroxaban indeed did have twice the risk of recurrent VTE compared to standard therapy following PE, would you feel comfortable offering it to your patient?*
- Note that the margin of non-inferiority refers to an acceptable boundary for the "worst case" end of the CI, not the point estimate itself

- Optimally, ≥90% power to find a difference rather than the typical 80% used in superiority trials
- Should use a 1-sided alpha of 0.025 for assessment of non-inferiority

- ITT is preferred as the primary analysis as it preserves the advantages of randomization
- Per protocol analysis is more likely to find a difference between groups, and is therefore more conservative in non-inferiority trials
- Non-inferiority should be accepted only if is demonstrated in both these analyses

Scan local institution policy or national guidelines

Scan DynaMed, UpToDate or similar references for high-quality evidence demonstrating clinically important benefits of the control treatment

e.g. In RE-LY,[Error! Bookmark not defined.] *the yearly incidence of stroke in the warfarin group was 1.57%. In a meta-analysis of older trials, the yearly incidence of stroke was 2.2%.*

Consider and quantify:

- Fewer or less-severe adverse effects
- Fewer drug interactions
- Easier to take
- Less intensive monitoring required
- Lower cost

NON-INFERIORITY TRIALS – Was the intervention compared to see if it is "no worse" than an established therapy?

Non-inferiority margin	**Was the minimally clinically important difference (MCID; a.k.a. non-inferiority margin) defined prior to undertaking the trial on the basis of statistical reasoning and clinical judgment?**	
	Did the trial use a MCID based on a relative or an absolute risk difference?	
	Is the non-inferiority margin strict enough according to your own judgment?	
Statistical Analysis	**Sample size calculation** – Was the non-inferiority assessment adequately powered to minimize statistical uncertainty?	
	Was non-inferiority demonstrated in both intention-to-treat (ITT) and per protocol analyses?	
Comparator	**Comparator is the standard of care** – Is the active control (drug, dose, interval, formulation, mode of administration, etc) comparable to the standard of care for the population of interest in your practice?	
	Comparator is better than nothing – Has the active control demonstrated unequivocal superiority over placebo in previous trials?	
	Was the active control effect in this trial consistent with that of previous trials?	
Other benefits	**If the intervention is non-inferior but not superior, what other benefits make it worth considering for your patients?**	