NERDCAT

NERDCAT

A Clinician's Guide to Appraising Randomized Controlled Trials, Systematic Reviews and Meta-Analyses

Ricky Turgeon and Blair MacDonald

NERDCAT by Ricky Turgeon is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](), except where otherwise noted.

Contents

Introduction 1

Part I. Generalizability

1. Do the results (not) apply to my patients? 5

Checklist Questions 6

Does my practice setting differ from that in the trials? 6

How do my patients differ from those included in the trial? 7

How do the trial interventions differ from those available in my practice? 10

Are the trial outcomes clinically important? 10

Does the trial reflect my patients risk of adverse events? 12

[Randomized Controlled Trials Only] Did the study design have a pre-randomization run-in period? 12

[Systematic Reviews/Meta-Analyses Only] Was each PICO element sufficiently reported to assess generalizability? 14

Do the differences above impede the generalizability of the study findings to my practice? 15

Part II. Randomized Controlled Trials

2. Risk of bias: Are the results internally valid? 19

Checklist Questions 20

Allocation Bias: Were patients appropriately randomized with allocation concealment? 21

Blinding: Were participants, treating clinicians, outcome assessors, or investigators aware of treatment assignment during the trial? 22

Crossover bias: Did participants from the comparator group receive the intervention from the intervention group (or vice versa)? 26

Missing data and loss to follow-up (LTFU): Was follow-up complete (i.e. were all patients accounted for at the end of the trial)? 27

Intention-to-Treat (ITT): Were patients analyzed in the groups to which they were randomized? 30

Reporting Bias: Are any important outcomes noted in the study protocol absent on publication? 31

3. Interpreting the results 33

Checklist Questions 34

Point estimate: What was the magnitude of effect for efficacy and harms? 34

Confidence interval: How precise were the estimates of treatment effect? 35

Is the difference clinically important? 37

Are these results consistent with other evidence? 38

4. Neutral trials: If the difference between interventions is not statistically significant, is there truly no difference? 41

Checklist Question 41

Does the confidence interval exclude a clinically important difference? 41

5. Composite outcome: Was the primary outcome a combination of outcomes? 43

Checklist Questions 43

Clinical importance: Are the components of the composite outcome all of similar importance to patients? 44

Statistical contribution: Did the component outcomes occur with similar frequencies? 45

Consistency in effect of therapy: Are the point estimates of treatment effect between each component consistent? Do the 95% CIs overlap? Are they sufficiently narrow? 46

Biologic rationale: Do the components of the composite outcome share a similar underlying biological mechanism? 48

6. Secondary outcomes: Can conclusions be made from outcomes other than the primary one? 51

Checklist Questions 52

Data-mining: Are we trying to find a difference in a secondary outcome when there was no statistically significant difference between groups for the primary outcome? 52

Minimizing multiplicity: Was the secondary outcome one of a small number of secondary endpoints defined in the original protocol? If there was a positive finding, were there appropriate adjustments made? 53

Consistency: Does the secondary endpoint result make sense in the context of the primary (and other secondary) outcome findings? 55

Was there an unexpected positive finding for a rare outcome? 56

7. Truncated studies: Was the trial stopped early for “overwhelming” evidence of benefit or futility? 57

Checklist Questions 59

Was there a predefined interim analysis plan with a stopping rule? 59

Did the stopping rule involve few interim looks and a stringent p-value (e.g. <0.001)? 60

Did enough endpoint events occur? 61

8. Subgroups analysis: Were additional comparisons made on segments of the study population? 63

Checklist Questions 64

Are statistically significant results in a subgroup being emphasized in the context of a neutral or negative trial? 64

Was the subgroup analysis pre-defined? 65

Was the direction of the subgroup effect correctly pre-defined? 65

Was the subgroup analysis one of a small number of hypotheses tested? 65

Is the subgroup variable a characteristic measured at baseline or after randomization? 66

Could treatment effect differences between subgroups be attributable to baseline imbalances? 66

Is the subgroup effect statistically significant? 67

Is the subgroup effect consistent within and across trials? 68

[Systematic Reviews/Meta-Analyses Only] Is the effect suggested by comparisons within rather than between studies? 68

9. Non-inferiority trials: Was the intervention compared to see if it is “no worse” than an established therapy? 71

Checklist Questions 74

Is the non-inferiority design justified by some other advantage of the intervention versus the comparator? 74

Did the trial use a non-inferiority margin based on a relative or an absolute risk difference? 76

Is the non-inferiority margin well justified based on statistical reasoning and clinical judgment? 77

Is the non-inferiority margin strict enough according to your own judgment? 78

Was non-inferiority demonstrated in both intention-to-treat (ITT) and per protocol analyses? 79

Was the comparator appropriate? 81

Part III. Systematic Reviews and Meta-Analyses

10. Search 85

Checklist Questions 85

Databases of published literature: Were a reasonable number of relevant databases searched? 86

Timeframe: When was the search conducted? Is it likely there have been subsequent publications that may alter the results? 87

Grey literature: Was a sufficient effort made to find unpublished studies (or unreported results of published studies)? 88

11. Results of the systematic review 93

Checklist Questions 93

Risk of bias within trials (internal validity): Did reviewers adequately assess for (& report) risk of bias? 94

Methodological & clinical heterogeneity: Is it appropriate to perform a meta-analysis? 94

12. Results of the meta-analysis 97

Checklist Questions 98

Statistical heterogeneity: What was the statistical heterogeneity? 98

Statistical models: Fixed-effects or random-effects? Is the model used appropriate? 101

Effect measure and precision 102

What proportion of included studies report on this outcome? 103

If performed, what GRADE rating was assigned to each outcome? 105

Appendix: Fundamental Statistics 109

P-Value Interpretation 109

Confidence Interval (CI) Interpretation 111

Sample Size Interpretation 112

Absolute Risk Differences and Relative Measures of Effect 113

Number Needed to Treat or Harm 115

Relative Risk, Odds Ratios, and Hazard Ratios 116

Kaplan Meier Curves 120

Forest Plots 123

Standardized Mean Difference Interpretation 124

Statistical Significance Is Not Everything 126

Glossary 129

References 141

Introduction

Welcome to NERDCAT: A Clinician's Guide to Appraising **Randomized Controlled Trials** and **Systematic Reviews/Meta-Analyses**. Led by Dr. Ricky Turgeon, NERDCAT was designed to help clinicians make sense of clinical research and has two core components: (1) The NERDCAT appraisal checklists, which facilitate the systematic appraisal of clinical studies; and (2) detailed guidance on how to address the NERDCAT appraisal checklist questions, along with rationales, supporting empiric evidence where available, and examples. While tools like CONSORT and PRISMA are aimed at researchers to facilitate adequate reporting of key details of their **randomized controlled trials** and **systematic reviews**, NERDCAT appraisal checklists are written "for clinicians, by clinicians" explicitly for the purpose of appraising clinical evidence and applying it to practice.

NERDCAT is organized into 3 chapters (Generalizability, Randomized Controlled Trials (RCTs), and Systematic Reviews and Meta-Analyses), an appendix of core statistical concepts (Appendix: Fundamental Statistics), and a comprehensive glossary of key terms. The **Generalizability** chapter is applicable to any clinical study type, including **RCTs** and **systematic reviews/meta-analyses**. NERDCAT can be read front-to-back or perused as a reference guide when appraising a study. Since the appendix describes the foundational statistical concepts necessary to understand the rest of the book, it is likely the best starting point for those unsure where to begin.

NERDCAT is structured around a core framework for appraising clinical studies adapted from the Users' Guides to the Medical Literature (Guyatt G et al.), which centers around 3 key questions:

1. **Generalizability:** Who was studied and how do these results apply to my patients?
2. **Internal validity:** How serious is the risk of **bias** and how might it impact the results?
3. **What are the results?:** What are the estimates of benefits & harms, how precise are those estimates, and are observed differences clinically important?

The NERDCAT checklists are available at the following Google Doc links:

- NERDCAT RCT (for **randomized controlled trials**)
- NERDCAT SR/MA (for **systematic reviews/meta-analyses**)

I

Generalizability

Generalizability, or **external validity**, refers to the extent to which the trial results are applicable beyond the patients included in the study. Clinicians typically understand **generalizability** in terms of how a study might apply to patients in their own practice. Even a perfectly-conducted trial may not be practically useful if there are important differences between your practice and the characteristics of the trial. The importance of such considerations is corroborated by a review (Kennedy-Martin T et al.) of 37 **RCTs** which found that roughly 70% of identified trials included participants that were not representative of patients in practice.

However, it should be emphasized that the trial population does not need to perfectly represent one's own practice. Use clinical judgement to determine to what extent differences between your practice and the trial characteristics impact the applicability of the results to your patients.

1.

Do the results (not) apply to my patients?

Generalizability is often understood in terms of **PICO**, which is an acronym for “patient, intervention, comparator, and outcome”. These are the four basic elements of a study. For instance, a study may examine an elderly population (P) to understand the effects of statin therapy (I) compared to placebo (C) in terms of cardiovascular events (O). The following questions are intended to comprehensively address each of these elements.

Most considerations of **generalizability** are independent of study type. So, unless explicitly noted otherwise, the following questions are applicable to both **randomized controlled trials** and **systematic reviews/meta-analyses**.

Checklist Questions

How does my practice setting differ from that in the trials?
How do my patients differ from those included in the trial?
How do the trial interventions differ from those available in my practice?
Are the trial outcomes clinically important?
Does the trial reflect my patients’ risk of adverse events? What differences exist?
[**Randomized controlled trials** only] Did the study design have a pre-randomization **run-in period**?
[**Systematic reviews**/**meta-analyses** only] Was each element of **PICO** (i.e. patient, intervention, comparator, and outcome) sufficiently reported to assess **generalizability**?
Do the differences above impede the **generalizability** of the study findings to my practice?

Does my practice setting differ from that in the trials?

Setting considerations:

- Country and type of healthcare system
- **Primary**, **secondary**, or **tertiary** care
- Outpatient vs. inpatient
- Inpatient unit type

How do my patients differ from those included in the trial?

Patient selection considerations:

- Diagnostic methods
- Inclusion / Exclusion criteria
- **Enrichment strategies**
- Proportion of patients not enrolled because of exclusion criteria
- Proportion of patients declining to participate

Patient characteristic considerations:

- Age
- Sex/Gender
- Race/ethnicity
- Stage/severity of disease
- Similar underlying pathologies (e.g. patients with a history of hemorrhagic stroke vs. patients with a history of ischemic stroke)
- Comorbidities
- Past interventions (e.g. proportion of patients previously having tried at least 3 antidepressants)
- Interventions at baseline (e.g. the proportion of patients taking aspirin at baseline in a trial of a SGLT2 inhibitor vs. placebo)
- Baseline clinical characteristics (e.g. blood

pressure, weight)

- Event rate in the control group

*E.g. #1 PARACHUTE (Yeh RW et al.) was a parody **RCT** examining whether the use of parachutes, compared to empty backpacks, prevented death and major trauma when jumping from an aircraft.The study did not find a difference in outcomes between the two groups. However, a major limitation was thatall participants jumped from a motionless (mean velocity 0 km/h), grounded (mean altitude 0.1 m) plane. Non-participants (declined or were ineligible) were on average moving much faster (800 km/h) and were at a much greater altitude (9146 m). Consequently, the results of this trial do not apply to the setting where a parachute may be used in practice (jumping out of an airborneplane).*

*E.g. #2 PARADIGM-HF was a **RCT** assessing the effects of sacubitril-valsartan vs. enalapril in patients with heart failure with reduced ejection fraction (McMurray JJV et al.). For the **primary outcome** of cardiovascular death or heart failure (HF) hospitalization the **HR** was 0.80 (95% CI 0.73-0.87) in favor of sacubitril-valsartan. To be included, patients were required to have elevated natriuretic peptides, such as a NT-proBNP ≥600 pg/mL (or ≥400 pg/mL if hospitalized within the last year). This was incorporated as an **enrichment criterion** (and not as a therapeutic target), as a higher serum natriuretic peptides concentration is associated with greater risk of HF-related events (Oremus M et al.), thus increasing trial event rates and reducing* *the required sample size to detect a difference between groups. However, elevated BNP is not the only prognostic factor in HF, as patients with "low" BNP can still be at high risk of HF hospitalization and death. Consider the following three patients with similar predicted risk (~35%) for HF hospitalization or death at 5 years:*

Table 1. Comparison of three patients with similar projected risk of heart failure hospitalization or death.
Estimates calculated using BCN-Bio-HF calculator on hfmedchoice.com

Characteristic	***Patient #1***	***Patient #2***	***Patient #3***
Age	*65*	*65*	*65*
Sex	*Male*	*Male*	*Male*
Ejection Fraction	*35%*	*35%*	*35%*
Type 2 Diabetes	*No*	***Yes***	*No*
NT-proBNP (pg/mL)	*1000*	***100***	***100***
New York Heart Association Class	*2*	*2*	***3***

*Since the **RRR** for this outcome with sacubitril-valsartan compared with an ACE inhibitor is the same regardless of NT-proBNP level, all 3 patients would be expected to have the same **absolute benefit** from sacubitril-valsartan despite* *patients #2 and #3 having NT-proBNP levels below trial inclusion criteria.*

How do the trial interventions differ from those available in my practice?

Intervention considerations:

- Intervention used (e.g. drug, dose, formulation (if relevant), duration)
- Timing of intervention
- Monitoring frequency
- Appropriate comparator
- Co-interventions – either pharmacological or non-pharmacological (e.g. both the intervention and comparator groups receiving lifestyle counselling in a trial evaluating the effects of a medication on weight loss)
- Changes in therapeutics / diagnostics since trial publication

Are the trial outcomes clinically important?

Outcome considerations:

- Clinical relevance of **surrogate outcomes**

- Clinical utility of measurement scales
- Consideration of all patient-centered outcomes
- Adequate follow-up duration
- Outcome assessor (i.e. patient or clinician)

When assessing the relative importance of outcomes and whether all important outcomes were evaluated it can be useful to construct a hierarchy of outcomes. These are specific to the clinical circumstance and patient preference, but the following is one example of a hierarchical ranking of outcomes:
1) Death or quality of life, depending on the goals of therapy
2) **Serious adverse events**
3) Clinically-important morbidity (e.g. heart failure hospitalizations, major bleed, symptom scores), withdrawals due to adverse events
4) Total adverse events, specific adverse events
5) **Surrogate markers** (e.g. change in a biomarker, **progression-free survival** in oncology trials)

E.g. A ***systematic review*** *and quantitative analysis (Kovic B et al.) examined the value of* ***progression-free survival (PFS)*** *as a surrogate endpoint for predicting health-related quality of life (HR-QoL) in cancer treatment trials. The slope of association between* ***PFS*** *and global HR-QoL was 0.1 (95% CI, −0.3 to 0.5), a non-statistically significant result suggesting that PFS is a poor surrogate for HR-QoL. In addition to concerns that* ***PFS*** *is also an unreliable predictor of overall survival, this casts doubt on the use of* ***PFS*** *as a predictor of patient important outcomes. Despite this,* ***PFS*** *remains a key endpoint of many oncology trials, and many oncology drugs are* *approved based on their impact on PFS without data on HR-QoL or overall survival.*

Does the trial reflect my patients risk of adverse events?

Adverse event considerations:

- Reporting of all clinically important adverse events
- Treatment discontinuations
- Trial site / clinician skill with treatment
- Exclusion of patients at elevated risk of adverse events
- Whether the duration of trial was adequate to detect adverse events of interest

[Randomized Controlled Trials Only] Did the study design have a pre-randomization run-in period?

Presence of a **run-in** period will require examination of the proportion of patients excluded during this phase, along with reasons for their exclusion.

Placebo **run-in periods** are usually used to:

- Obtain a pre-treatment baseline for clinical status (e.g. number of migraines/month in a trial of migraine prophylaxis)
- Ensure that the participants are sufficiently adherent to the assigned regimen

Active **run-in periods** are usually used to:

- Ensure short-term tolerability
- Ensure that the participants are sufficiently adherent to the assigned regimen

E.g. PARADIGM-HF (McMurray JJV et al.) was a ***RCT*** *assessing the effects of sacubitril/valsartan vs. enalapril in patients with heart failure with reduced rejection fraction with respect the* ***primary outcome*** *of cardiovascular death or heart failure hospitalization. This trial featured a single-blind* ***run-in*** *with enalapril followed by a single-blind* ***run-in*** *with sacubitril-valsartan. Approximately 11% of participants were excluded during the* ***run-ins*** *due to adverse events. After randomization, symptomatic hypotension occurred in 14% of patients receiving sacubitril-valsartan versus 9% of patients receiving enalapril. However, these rates are among patients who were able to tolerate both enalapril and sacubitril-valsartan during the* ***run-in periods****, and are therefore likely an underestimate of the true rate of this adverse event among reduced ejection fraction patients newly starting either medication.*

[Systematic Reviews/Meta-Analyses Only] Was each PICO element sufficiently reported to assess generalizability?

If the **PICO** characteristics are not reported sufficiently, or the review inclusion criteria too broad, it may not be possible to evaluate whether the results apply to a given patient or practice. As such, if the **PICO** elements are poorly described or excessively broad, consider looking for another **systematic review** with better reporting and scope.

E.g. A ***meta-analysis*** *by Ortiz-Orendain J et al. compared antipsychotic polypharmacy vs. antipsychotic monotherapy for the treatment of schizophrenia. The trial inclusion was not restricted based on particular patient characteristics (except being limited to those ≥18 years old), illness characteristics (e.g. severity or duration), treatment setting, nor drug characteristics (e.g. drug, dose, or formulation). Furthermore the results only reported average patient age and treatment setting, with no description of other demographic features nor illness characteristics. As a result, although comprehensive in its breadth, the study included a broad set of disparate studies with heterogeneous comparisons, rendering it difficult to apply the results to practice, or to determine if these patient-specific or treatment characteristics impacted outcomes.*

Do the differences above impede the generalizability of the study findings to my practice?

There will almost always exist some differences between one's practice and the **PICO** of the trial. Use clinical judgement to evaluate whether these differences render the study results inapplicable to your practice or to an individual patient. If there are sufficient differences, then an attempt should be made to predict the effect of these differences (i.e. greater or less efficacy/harm).

*E.g. LoDoCo2 (Nidorf SM et al.) was a **RCT** of colchicine 0.5mg vs. placebo in patients with chronic coronary artery disease. Colchicine reduced the **primary** cardiovascular composite endpoint compared with placebo (**HR** 0.7 (95% CI 0.6 to 0.8), with an **absolute difference** of 1.5% at approximately 2 years. It is uncertain if these results could translate to cardiovascular benefit in patients without coronary artery disease. Even if colchicine was efficacious in patients without coronary artery disease, the **absolute difference** would be anticipated to be lower due to a lower event rate, and the benefit:harm trade-off may in turn also be quite different.*

II

Randomized Controlled Trials

RCTs are experimental studies that attempt to isolate the cause-and-effect relationship of an intervention on select outcomes. **RCTs** recruit participants that meet specified inclusion and exclusion criteria, randomly allocate these participants to two or more intervention groups, and then follow them over time to monitor for outcomes of interest. High-quality **RCTs** enable inferences regarding the effects of a treatment. However, deficiencies in the trial may introduce **bias**, obscuring the effect of the intervention.

This chapter will provide guidance in assessing **RCTs** for risk of bias and its potential impact on the results, as well as the clinical relevance of the findings.

2.

Risk of bias: Are the results internally valid?

Randomization is the core of the **RCT** and ensures that the play of chance dictates whether any given participant is assigned the intervention or comparator(s). Because of this, baseline characteristics tend to be similar between randomized groups, though imbalances can still occur by chance. So, at the start of the trial, each group should tend to have a similar probability of experiencing any outcome. If this similarity is properly maintained (i.e. neither **bias** nor confounding introduces differences between groups), then any differences in outcomes will either be due to either treatment allocation or to chance.

Towards this objective of maintaining similar groups, other strategies (such as blinding of participants, clinicians, and investigators) are often implemented to minimize differences in care between groups over the course of the trial. If these strategies are not successful, then any differences in outcomes between groups could also be attributable to these differences in care, thus introducing **bias**.

The **internal validity** sections of this chapter will focus on describing key sources of **bias** in **RCTs**, how to identify them, and how to evaluate their impact on observed study results.

Checklist Questions

Was the **sequence generation** random?
Was the allocation sequence concealed until participants were enrolled and assigned to interventions?
Were participants, clinicians, outcome assessors, and investigators blinded?
Were there deviations from the intended intervention that arose because of the above?
Could measurement or ascertainment of the outcome have differed between intervention groups?
Could assessment of the outcome have been influenced by knowledge of intervention received?
Did participants from the comparator group receive the intervention from the intervention group (or vice versa)?
Were data for key outcomes available for all, or nearly all, participants randomized?
Were patients analyzed in the groups to which they were randomized (**ITT**), or did researchers only count participants who were adherent to their study treatment (**per protocol**) or completed the full trial duration (completer analysis)?
Are the **ITT** methods appropriate?
Are any important outcomes included in the study protocol but absent from the publication? Is this justified?

Allocation Bias: Were patients appropriately randomized with allocation concealment?

Sequence generation (i.e. randomization)

Unclear or inadequate **sequence generation** exaggerates **relative benefits** of an intervention on average by ~11% (Savović J et al.).

Table 2. Adequate and inadequate randomization methods.

Adequate Randomization	**Inadequate Randomization**
Computer-generated random sequence generation **(preferred)**	Quasi-randomization (e.g. alternation by case number or date of birth)
Random numbers table	Treatment assignment left to the discretion of the clinician
Coin toss	
Drawing cards	

Allocation concealment

Unclear or inadequate **allocation concealment** exaggerate **relative benefits** of an intervention on average by ~7% (Savović J et al.).

Table 3. Adequate and inadequate allocation concealment methods.

Adequate allocation concealment	**Inadequate allocation concealment**
Central randomization (look for “interactive web-response system” or “interactive voice-response system” within a study manuscript) **(preferred)**	Allocation scheme posted on a bulletin board
Coded identical drug boxes/vials	Non-opaque, non-tamper proof envelopes
Sequentially-numbered, tamper-proof, sealed opaque envelopes (preferably lined with cardboard or foil)	
On-site locked computer system	

Blinding: Were participants, treating clinicians, outcome assessors, or investigators aware of treatment assignment during the trial?

Lack of or unclear blinding is associated with an average ~13% exaggeration of the **relative benefits** of an intervention for dichotomous outcomes (Savović J et al.), and a 68% exaggeration of **relative benefits** for subjective continuous outcomes (Hróbjartsson A et al.).

Note that **double-blinding** does not have a standardized definition and, consequently, further examinations are needed to ascertain exactly who was blinded (Lang TA et al).

Blinding of participants and clinicians

Table 4. Adequate and inadequate blinding methods for participants and clinicians.

Adequate blinding of participants and clinicians	**Inadequate blinding of participants and clinicians**
Used an identical placebo/ control product without indication that treatments were distinguishable	PROBE: Prospective randomized open-label, blinded endpoint trial (open-label refers to trial that has non-blinding as part of the design, and does not refer to cases where blinding is simply inadequate)

Blinding of outcome assessors

Awareness of treatment allocation by participants and clinicians may introduce performance bias, whereas awareness of allocation by outcome assessors may introduce detection bias. This is compounded when participants or their clinicians are also the outcome assessors (e.g. patient aware of treatment allocation asked to rate their pain or fill out a quality-of-life questionnaire). Lack of blinding is a particularly important source of **bias** with the use of subjective outcomes – one review (Wood L et al.) found that lack of blinding exaggerated the **OR** of subjective outcomes by ~30%. Conversely, the same review found no statistically significant **bias** was introduced by lack of blinding for objective outcomes. This review provided evidence that all-cause mortality is a particularly resistant to detection **bias** even when trials are not blinded.

Table 5. Adequate blinding methods for outcome accessors and difficult situations to blind.

Adequate blinding of outcome assessors	**Difficult situations to blind**
Independent central adjudication committee adjudicated outcomes	The intervention has an effect on a readily-measurable biomarker or the drug has an easily observable adverse effect profile (e.g. iron causing darkened stools)

*E.g. #1 Among several concerns raised by the FDA regarding the PLATO trial (DiNicolantonio JJ et al.), a **RCT** comparing ticagrelor vs. clopidogrel for patients with acute coronary syndrome, it was noted that blinding was not sufficiently protected. This is because the "dummy capsules" (identical in appearance to the ticagrelor containing capsule) could be opened, revealing a clopidogrel tablet cut in half. This could unblind both patients and sponsor site monitors (who were given unused capsules). There were also concerns that too many groups involved in the trial had access to treatment assignments (and could subsequently become unblinded). These concerns cast doubts on the **internal validity** of both the efficacy and safety outcomes, especially when combined with additional concerns by the FDA regarding the inaccuracy of reported events.*

Some situations initially thought to be impossible to blind can be successfully blinded with some ingenuity.

E.g. #2 In ROCKET-AF (Patel MR et al.), 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.

Were there differences between groups in the receipt of co-interventions?

Co-interventions may introduce **bias** if they affect the outcomes of interest and are distributed differently between groups.

E.g. #3 CONTACT (Roddy E et al.), an unblinded non-inferiority trial comparing naproxen and colchicine for acute gout attacks, found no difference in pain control between groups. However, co-intervention analgesic use (e.g. acetaminophen, ibuprofen) was 42% in the colchicine group and only 25% in the naproxen group. This raises the possibility that pain control would have been inferior in the colchicine group had it not been for the additional analgesic use.

Was outcome monitoring assessed consistently between groups? If no, then was this likely to bias the results?

Outcome measures ought to be consistent between groups with regards to:

- Which outcomes were examined
- How they were examined
- How frequently they were examined

E.g. #4 RATE-AF (Kotecha D et al.) was a ***RCT*** *comparing the impact of digoxin vs. bisoprolol on quality of life in participants with heart failure with preserved ejection fraction and atrial fibrillation. Participants were prompted to report adverse effects using adverse effects listed in the medication product monograph. It is unclear if an aggregate list was used for all participants or if a drug-specific list was used for each group. Given the extensive list of adverse effects listed on the bisoprolol monograph and lay perceptions of beta-blocker-related adverse effects, differential lists would be expected to* ***bias*** *the adverse effect outcomes in favor of digoxin. This would be an example of differential outcome monitoring between study groups.*

Crossover bias: Did participants from the comparator group receive the intervention from the intervention group (or vice versa)?

Crossover bias attenuates differences in outcomes between groups as a group accrues more participants that are taking the treatment intended for the other arm (e.g. patients in the placebo group receiving active treatment). This makes **superiority** harder to demonstrate and makes non-inferiority easier to demonstrate.

The extent of **bias** introduced will depend on the extent of crossover/contamination between groups.

E.g. HPS (Heart Protection Study Collaborative Group) was a ***RCT*** *evaluating the effect of simvastatin 40 mg daily vs. placebo on mortality and cardiovascular events. By the 5th year of follow up, 32% of patients in the placebo group were receiving a statin other than simvastatin, likely initiated due to higher LDL levels. If we assume these other statins were effective in reducing mortality and cardiovascular events, they may have attenuated the difference seen between the simvastatin and placebo group for these outcomes.*

Missing data and loss to follow-up (LTFU): Was follow-up complete (i.e. were all patients accounted for at the end of the trial)?

Rules of thumb (e.g. **LTFU** is only a problem if ≥20%) are misleading; **LTFU** is important when it is similar to or greater than the occurrence of the outcome of interest, or when differences in the frequency or timing of **LTFU** differ between groups. An **ITT analysis** (see below) cannot correct the bias introduced by differences in **LTFU** between groups. In addition to **LTFU**, there may also be missing data due to factors such as participants missing scheduled visits, variables not being measured during a visit, or data entry errors.

If there is **LTFU**, consider doing your own rudimentary “worst-case scenario” analysis: Would the results remain similar if all participants lost in one treatment group had suffered the bad outcome whilst all those lost in the other group had had a good outcome, and vice versa?

E.g. #1 In a trial (El-Khalili N et al.) comparing 2 doses of quetiapine vs. placebo for adjunctive treatment of depression, completion of trial follow-up to week 6 was 77% with quetiapine 150 mg/day, 70% with quetiapine 300 mg/day, and 85% with placebo. Differences between groups were driven by a dose-dependent increase in the risk of discontinuation due to adverse events with quetiapine (1% with placebo, 11% with quetiapine 150 mg/day, and 18% with quetiapine 300 mg/day,).

E.g. #2 In PARAMEDIC2, a ***RCT*** *comparing epinephrine vs. placebo for out-of-hospital cardiac arrest, the* ***primary outcome*** *was survival at 30 days.*

Table 6. Epinephrine vs. placebo in patients experiencing out-of-hospital cardiac arrest on the outcome of survival at 30 days.

	Epinephrine	**Placebo**	**OR (95% CI)**
Actual Analysis	130/4012 (3.2%)	94/3995 (2.4%)	1.39 (1.06 to 1.82)
LTFU	3 (<0.1%)	4 (<0.1%)	
Worst-Case Analysis	130/4015 (3.2%)	97/3999 (2.4%)	1.35 (1.03 to 1.76)

This worst-case scenario does not change the statistical or clinical significance of the result, so the ***LTFU*** *is not a concern for this outcome.*

E.g. #3 In PARAMEDIC2 a ***secondary outcome*** *was a favorable neurological outcome at three months.*

Table 7. Epinephrine vs. placebo in patients experiencing out-of-hospital cardiac arrest on favorable neurological outcome at three months.

	Epinephrine	**Placebo**	**OR (95% CI)**
Actual Analysis	82/3986 (2.1%)	63/3979 (1.6%)	1.31 (0.94 to 1.82)
LTFU	29	20	
Worst-Case Analysis	82/4015 (2.0%)	83/3999 (2.1%)	0.98 (0.72 to 1.34)

While the results are not statistically significant in both actual and worst-case analyses, the worst case analysis shifts the CI to be notably more pessimistic regarding the effects of epinephrine on this outcome. The ***absolute difference*** *between the actual and worse-case analysis is only 0.6%. However, in the context of the trial, where* ***absolute*** *survival was only 0.8% greater with epinephrine, this relatively small difference is nonetheless still important when considering the net benefit of epinephrine.*

Intention-to-Treat (ITT): Were patients analyzed in the groups to which they were randomized?

There are numerous methods to carry out an **ITT analysis** (e.g. **last observation carried forward (LOCF)**, mixed model for repeated measurements, sensitivity analyses). All of them rely on assumptions and no single method works in every situation.

E.g. In dementia trials evaluating the efficacy of cholinesterase inhibitors ***LOCF*** *is the most common approach to* ***ITT****. This occurs despite violating the* ***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 (Molnar FJ et al. 2008 and 2009).*

Reporting Bias: Are any important outcomes noted in the study protocol absent on publication?

If a trial does not report a clinically important outcome despite it being in the protocol, this warrants suspicion that the intervention did not provide benefit (or was possibly harmful) with respect to that outcome.

E.g. EPHESUS (Pitt B et al.) was a ***RCT*** *comparing eplerenone vs. placebo in patients with left ventricular dysfunction after myocardial infarction. None of the published reports of EPHESUS have reported on quality of life despite this being a pre-specified outcome of the trial (Spertus JA et al.). As such, it is not possible to determine the impact (beneficial, harmful or neutral) of eplerenone on quality of life in these patients.*

See “What proportion of the included studies report on this outcome?” here for a further discussion on outcome reporting bias.

3.

Interpreting the results

When examining the results of a trial, it is necessary to consider more than statistical significance. Not all statistically significant results are clinically important. Similarly, failure to find a statistically significant difference does not necessarily rule out that there is a clinically important difference. In addition, when making judgements of clinical relevance, it is necessary to examine not only the **relative effect** of treatment, but also the **absolute effect** (see here for further discussion).

In terms of effectively communicating results to patients, a review (Zipkin DA et al.) found that:
– Any type of difference (**absolute** or **relative**) is understood more accurately when baseline risk is provided;
– **Absolute differences** are understood more accurately than **relative differences**;
– Numbers needed to treat (NNTs) are often misunderstood and are inferior to reporting absolute differences;
– Addition of visual displays to numerical information increase understanding.

Checklist Questions

What was the magnitude of effect for efficacy and harms?
How precise were the estimates of treatment effect?
Is the difference clinically important?
Are these results consistent with other evidence?

Point estimate:What was the magnitude of effect for efficacy and harms?

Look at both the **absolute effect** and the **relative effect**. **Relative effects** are typically assumed to be reasonably consistent across populations, whereas **absolute effects** depend on baseline risk.

E.g. The ***relative risk reduction*** *of statins on all-cause mortality is similar in primary prevention (i.e. prevention in patients without cardiovascular disease) and secondary prevention (patients with cardiovascular disease), but the* ***absolute effect*** *is greater in secondary prevention (Wilt TJ et al., Tonelli M et al.):*

Table 8. Comparison of absolute risk reductions in primary and secondary prevention patients treated with statins.

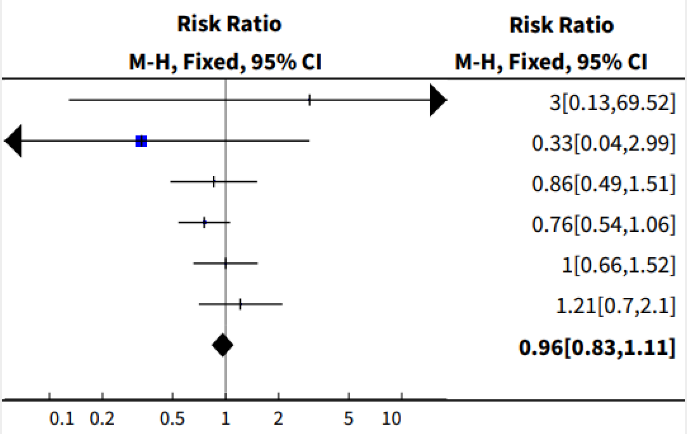
Population	***Relative risk reduction***	***Absolute Risk Reduction Over 5 Years***
Primary prevention patients – No coronary artery disease – A predicted <20% risk of a cardiovascular event in the next 10 years	*10-15%*	*0.4%*
Secondary prevention patients – Coronary artery disease	*10-15%*	*2%*

Confidence interval: How precise were the estimates of treatment effect?

Confidence intervals (CIs) provide information regarding the uncertainty of the results. The wider the CI, the greater the uncertainty. The width is based on the difference between the two ends of the CI. Wide and narrow do not have exact definitions.

E.g. #1 Narrow 95% CI: ***RR*** *0.90 (0.85 to 0.95)*
E.g. #2 Wide 95% CI: ***OR*** *1.25 (0.2 to 5)*

E.g. #3 A meta-analysis by Ortiz-Orendain J et al. illustrates visually how CI have varying widths:



Plot 1. Forest plot of any antipsychotic plus atypical antipsychotic vs. atypical antipsychotic in patients with schizophrenia on the outcome of no clinically important response.

The relevance of this uncertainty depends on whether the CI includes clinically important differences (see the following section). This involves examining both ends of the CI, and judging whether there is a meaningful difference between the two.

E.g. #4 In CAPRIE (CAPRIE Steering Committee), the lower end of the ***relative risk reduction*** *95% 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.*

Is the difference clinically important?

Clinical importance is determined by looking at the **absolute risk difference**, rather than a **relative risk reduction**. Note that clinical importance is dependent on an individual's preferences and values, therefore opinions will differ based on clinician and patient preferences, patient situation, intervention characteristics (e.g. adverse events, cost, convenience), and other factors.

*E.g. #1 A****meta-analysis*** *(ATT Collaboration) comparing ASA vs. placebo in secondary prevention patients found a statistically significant 1.5% per year* ***absolute risk reduction*** *in serious vascular events(myocardial infarction, stroke, or vascular death). The same* ***meta-analysis*** *also examined ASA vs. placebo in primary prevention patients found a much smaller, but still statistically significant, 0.07% per year* ***absolute reduction*** *for the same outcome. Many patients and clinicians would consider the benefit of ASA in secondary prevention to be clinically meaningful, whereas far fewer would be willing to take ASA for primary prevention knowing these* *numbers.*

Consider comparing the results with **absolute risk reductions** or NNTs achieved with other interventions used in a similar patient population.

E.g. #2 In HPS (Heart Protection Study Collaborative Group), the ***absolute reduction*** *in risk of death over 5 years with simvastatin vs. placebo was 1.8% in a high-risk population. For comparison, ramipril reduced the 5-year risk of death by 1.7% versus placebo in a similar patient population within the HOPE trial (Yusuf S et al.).*

Are these results consistent with other evidence?

Differences between groups in any given study may occur by chance. Replication of the consistent results in multiple studies increases the confidence that the difference represents a true effect of the study intervention. Searching for a **systematic review** on the topic can efficiently provide insight into the context of surrounding literature and consistency between studies.

E.g. ASPEN (Knopp RH et al.), a 2006 ***RCT****, did not demonstrate a statistically significant difference in cardiovascular events between atorvastatin versus placebo in patients with diabetes (****HR*** *0.9; 95% CI, 0.7 to 1.1). In contrast, the 2004 CARDS trial (Colhoun HM et al.), also comparing atorvastatin vs. placebo in patients with diabetes, had previously shown benefit in a similar population (****HR*** *0.6; 95% CI 0.5 to 0.8) for a similar* ***primary endpoint****. The neutral findings of ASPEN should be understood in the context of the CARDS (and also the dozens of other trials that demonstrated benefits of statins vs. placebo for the prevention of cardiovascular events (Cholesterol Treatment Trialists' (CTT) Collaborators)).*

4.

Neutral trials: If the difference between interventions is not statistically significant, is there truly no difference?

Neutral trials (i.e. trials without a statistically significant difference in the **primary outcome**) should not all be interpreted equally, as they will differ in the degree in which they rule out a difference between interventions. It is important to be able to recognize that "no statistically significant difference" does not mean "no difference".

Neutral trials are also sometimes referred to as "negative trials".

Checklist Question

Does the confidence interval (CI) exclude a clinically important difference?

Does the confidence interval exclude a clinically important difference?

If the 95% CI is wide enough to include a clinically important difference, it remains possible that a future trial with greater precision, or a meta-analysis of multiple trials, may find a clinically meaningful difference.

E.g. Authors of a trial (Nguyen-Khac E et al.) 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. However, these were the results at 6 months:

Table 9. N-acetylcysteine plus prednisone vs. prednisone alone for patients with acute alcoholic hepatitis on the outcome of mortality.

Outcome	***Combination***	***Steroid Alone***	***Absolute risk difference (95% CI)***
Mortality	*27%*	*38%*	*-11% (-22% to +5%)*

The uncertainty of the estimated reduction represented by the 95% CI means that the trial could not exclude the possibility of an ***absolute reduction*** *in mortality with combination therapy as high as 22%.*

5.

Composite outcome: Was the primary outcome a combination of outcomes?

By combining several individual outcomes into a **composite outcome** a trial can increase its ability to detect a difference between groups. However, **composite outcomes** require careful interpretation of the individual components to avoid making erroneous conclusions.

Checklist Questions

Are the components of the **composite outcome** all of similar importance to patients?
Did the components occur with similar frequencies?
Are the point estimates of treatment effect (**HR**, **OR**, **RR**) similar between each component? Do the 95% CIs overlap? Are they sufficiently narrow?
Do the components share a similar underlying biological mechanism?

Clinical importance: Are the components of the composite outcome all of similar importance to patients?

E.g. #1 The ***primary outcome*** *of DAPA-HF (McMurray JJV et al.), a trial comparing dapagliflozin vs. placebo in patients with heart failure with reduced ejection fraction (HFrEF), was a* ***composite*** *of:*

- *Hospitalization for heart failure (HF) resulting in intravenous therapy*
- *Urgent HF visit resulting in intravenous therapy*
- *Death from cardiovascular (CV)causes*

These are all outcomes of significant importance to patients.

E.g. #2 The ***primary outcome*** *of CONDOR (Chan FKL et al.), a trial comparing celecoxib vs NSAID+PPI, was a* ***composite*** *of:*

- *Gastrointestinal (GI) bleed*
- *GI obstruction*

- *GI perforation*
- *Clinically significant anemia (decrease in hemoglobin ≥ 20 g/L or decrease in hematocrit ≥10%)*

The latter of which was notably less important than the other components.

Statistical contribution: Did the component outcomes occur with similar frequencies?

E.g. #1 Components of the ***primary outcome*** *in DAPA-HF (McMurray JJV et al.) and their rates for dapagliflozin vs. placebo:*

- *Hospitalization or urgent visit for HF (10% vs.14%)*
- *Death from CV causes (10% vs. 12%)*

The most important endpoint (death from CV causes) occurred only slightly less frequently than the other component.

E.g. #2 Components of the ***primary outcome*** *in CONDOR (Chan FKL et al.) and their rates for celecoxib vs NSAID+PPI:*

- *GI bleed (0.2% vs. 0.2%)*
- *GI obstruction (0% for both groups)*
- *GI perforation (0% for both groups)*
- *Clinically significant anemia (0.7% vs. 3%)*

The greatest contributor of events (clinically significant anemia) drove the difference between groups was also the least clinically important.

Consistency in effect of therapy: Are the point estimates of treatment effect between each component consistent? Do the 95% CIs overlap? Are they sufficiently narrow?

E.g. #1 ***HRs*** *in DAPA-HF (McMurray JJV et al.) for dapagliflozin vs. placebo:*

- ***Composite HR*** *= 0.74 (95% CI 0.65-0.85)*

- *Hospitalization or urgent visit for heart failure **HR** = 0.70 (95% CI 0.59-0.83)*
- *Cardiovascular death **HR** = 0.82 (95% CI 0.69-0.98)*

*Since all the CIs overlap and are sufficiently narrow, there can be greater confidence that the **composite outcome** is not misleading.*

*E.g. #2 **Relative risk reductions (RRRs)** in CONDOR (Chan FKL et al.) for celecoxib vs NSAID+PPI:*

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

*Since there is a very large difference in the point estimates, it is better to consider the individual endpoints rather than the **composite endpoint**.*

Biologic rationale: Do the components of the composite outcome share a similar underlying biological mechanism?

E.g. #1 In DAPA-HF (McMurray JJV et al.) for dapagliflozin vs. placebo all outcomes had a similar underlying mechanism, consisting of (i) hospitalization or urgent visit for heart failure and (ii) cardiovascular death.

E.g. #2 In CONDOR (Chan FKL et al.) for celecoxib vs NSAID+PPI all outcomes also had a similar underlying mechanism, consisting of (i) GI bleed, (ii) GI obstruction, (iii) GI perforation, and (iv) clinically significant anemia. Despite this coherence in underlying mechanism the ***composite outcome*** *was inadequately chosen, for reasons discussed above.*

E.g. #3 In the UKPDS blood pressure target trial (UK Prospective Diabetes Study Group), 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. As such, it is* *advisable to assess each of these outcomes individually rather than as a* ***composite****.*

6.

Secondary outcomes: Can conclusions be made from outcomes other than the primary one?

Most trials designate one outcome as a **"primary" outcome** (or 2-3 “co-primary outcomes”) and all other outcomes as **"secondary" outcomes**. Designation of an outcome as “primary” is done to determine and justify sample size calculations prior to conducting a study. In other words, the **primary outcome** is not necessarily the most clinically important (it often isn’t), and should not be the sole consideration as to whether an intervention is “better” than a comparator.

The interpretation of **secondary outcomes** requires additional considerations. The probability of finding a difference simply due to chance increases as the number of outcomes increases.

Checklist Questions

Are we trying to find a difference in a **secondary outcome** when there was no statistically significant difference between groups for the **primary outcome**?
Was the **secondary outcome** one of a small number of secondary endpoints defined in the original protocol? If there was a positive finding, were there appropriate statistical adjustments made?
Does the **secondary endpoint** result make sense in the context of the **primary** (and other secondary) outcome findings?
Was there an unexpected positive finding for a rare outcome?

Data-mining: Are we trying to find a difference in a secondary outcome when there was no statistically significant difference between groups for the primary outcome?

Authors may emphasize statistically significant differences in **secondary outcomes** when they fail to find a statistically significant difference in the **primary outcome**.

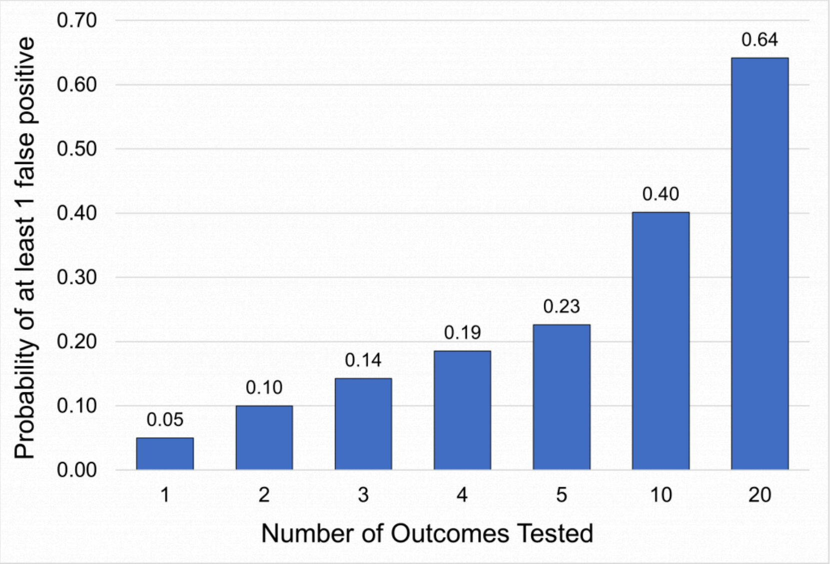
For example, a review (Khan MS et al.) of 93 cardiovascular **RCTs** found that spin (i.e. reporting strategies highlighting benefits despite a non-statistically significant **primary outcome**) was present in 57% of abstracts and 67% of main texts.

E.g. In the FIELD trial (Keech A et al.), the difference in

> *the **primary outcome** (coronary events up to 5 years) was not statistically significant between fenofibrate and placebo in patients with type 2 diabetes. In their conclusions, authors highlighted marginally statistically significant reduction in 3 of 9 **secondary** efficacy outcomes (total cardiovascular events, non-fatal myocardial infarction, and revascularization).*

Minimizing multiplicity: Was the secondary outcome one of a small number of secondary endpoints defined in the original protocol? If there was a positive finding, were there appropriate adjustments made?

More comparisons increase the risk of finding a difference when there is none, as depicted:



Graph 1. Probability of at least one false positive result by number of outcomes tested assuming no difference and threshold for statistical significance <0.05.

Depending on the context, it may be justified to adjust for multiplicity when considering multiple outcomes. Adjusting for multiplicity is a statistical method of requiring lower p-values to account for multiple comparisons. There are multiple methods for calculating the stricter margin (Bonferroni test, Holm test, etc.). There is no consensus on when to adjust for multiplicity, but the following can act as general guidance:

Table 10. Circumstances where adjusting for multiplicity may be necessary.

Circumstance	**Whether Adjustment is Necessary**
At least one outcome is positive and the outcome is intended to inform future research rather than be incorporated directly into clinical practice (i.e. exploratory)	Adjustments in the analysis are not warranted as such findings are used only to generate hypotheses
At least one outcome is positive and the outcome is intended to directly inform clinical practice (i.e. confirmatory)	Adjustments may be necessary if: – More than one dose is compared – More than one **primary outcome** is used – The **primary outcome** was assessed in multiple different population

Consistency:Does the secondary endpoint result make sense in the context of the primary (and other secondary) outcome findings?

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

E.g. In FIELD (Keech A et al.), the ***secondary outcome*** *of non-fatal myocardial infarction was statistically significantly* ***less*** *with the fenofibrate group compared to placebo. However, all-cause mortality, coronary death, deep vein thrombosis, and pulmonary embolism occurred* ***more*** *frequently in the fenofibrate group.*

Was there an unexpected positive finding for a rare outcome?

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 **primary outcome**.

E.g. The ELITE trial (Pitt B, Segal R, et al.) 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 (11% in both groups). There was, however, an unexpected reduction in all-cause mortality with losartan vs. captopril (5% vs 9%, p=0.04). The follow-up ELITE II trial (Pitt B, Poole-Wilson PA, et al.) 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).*

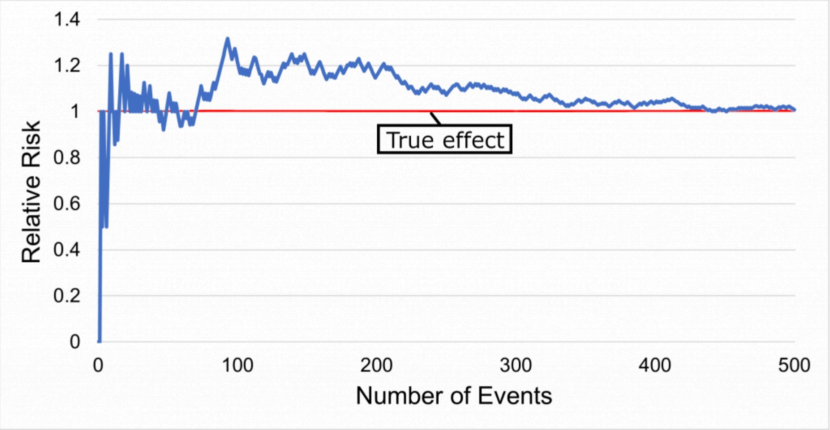
7.

Truncated studies: Was the trial stopped early for “overwhelming” evidence of benefit or futility?

Studies may be stopped early for efficacy as part of an ethical obligation to not expose participants to less effective treatment (or placebo) any longer than necessary. In other words, once it is sufficiently clear that an intervention is efficacious, there is reason to end the trial.

However stopping early runs the risk of overestimating the effect size of the intervention. The estimate of effect will randomly vary around the true effect over time (with more fluctuation with fewer events early in the trial), so interim looks may lead to premature stop due an exaggerated estimate of the true effect size.

Consider the following simulated trial where there is no true difference between the groups (i.e. **RR** = 1.0):



Graph 2. Relative risk vs. number of events in a simulated trial. Created via Microsoft Excel using the RAND function to generate randomized event-data for two groups.

As depicted in Graph 2 above, there is random deviation from the true effect as events accumulate. If the trial had interim analyses for benefit every 100 events, and the threshold for statistical significance was kept at the standard $p<0.05$ without accounting for interim looks, then the trial may have stopped at 100 events when the **RR** was 1.3, which we know to be an exaggeration of the true effect (**RR** = 1.0, i.e. no effect).

As a simplified example, imagine studying a chess player and trying to assess if they are an above-average player (and by what margin) by judging their win percentage. One approach is to wait 50 matches, then assess their win percentage and judge accordingly. However, this could waste time as it might be unnecessary to wait that long if they are quite skilled (e.g. winning 90% of their first 10 games). So instead there could be an assessment of skill every 5 matches (up to a maximum of 50 matches). If they seem sufficiently impressive at one of these midpoint assessments, then the observation could be stopped. While this might save time, it also has a risk: if by pure chance the player goes on a win streak, then the observation is likely to end early. Even if our player is truly above-average in skill, an early stop is most likely to occur when they are on such a hot streak, consequently introducing **bias** into our assessment (e.g. assessing their win probability to be 80% due to the win streak, when in fact it is only 60%).

This is the major concern with stopping rules: there is a systematic tendency for an early stop to be an overestimation. While such precautions cannot prevent **bias** towards overestimation, they can help reduce the extent of this **bias**, as discussed below.

Checklist Questions

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

Was there a predefined interim analysis plan with a stopping rule?

If there is no pre-planned stopping rule then there is no assurance that sufficient safeguards were in place to minimize **bias** from early stops.

E.g. In JUPITER (Ridker PM et al.), a ***RCT*** *of rosuvastatin vs. placebo in a highly-selected primary cardiovascular 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."*

Did the stopping rule involve few interim looks and a stringent p-value (e.g. <0.001)?

As the number of interim looks increases, then the probability of finding a false positive or overestimation also increases. This can be mitigated by (1) minimizing the number of interim looks and (2) having a stricter threshold for statistical significance that accounts for these multiple interim analyses.

Some common interim analysis strategies used (Schulz KF et al.) are:
Pocock: To keep the overall trial p-value threshold (alpha) = 0.05, the number of interim analyses are pre-defined & all have the same adjusted statistical significance threshold (i.e. $p<0.029$ for 2 planned analyses, $p<0.016$ for 5 planned analyses, and so forth).
Peto: Assign the final analysis p-value threshold = 0.05 (like in a conventional trial), but have a more stringent threshold (i.e. $p<0.001$) for the interim analyses.
O'Brien-Fleming: Begin with stringent interim analyses that start conservatively and then successively ease as they approach the final analysis (e.g. for 3 interim analyses & a final analysis,

sequence of p-value thresholds 0.0001, 0.004, 0.019, 0.043)
Lan-DeMets: An adaptable approach where the significance level changes and analysis timing changes in accordance to previously observed information.

E.g. JUPITER (Ridker PM et al.) 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.*

Did enough endpoint events occur?

Trials stopped early for benefit exaggerate the **relative effect** of an intervention by an average 29% compared with trials that conclude as planned (Bassler D et al.). As events accumulate, the fluctuations in effect size measures will become smaller and there will be less risk of **bias** (see graph above). Optimally ≥500 events (Bassler D et al.) should occur before stopping, after which the exaggeration decreases to an average of 12%.

For these reasons, skepticism is warranted for any **relative risk reduction (RRR)** ≥50% generated in truncated trials with <100 events (Pocock SJ et al., Montori VM et al.). The larger the number of events and the more plausible the **RRR** (e.g. ~20-30% is typical for the impact of cardiovascular pharmacotherapy on cardiovascular events), the more believable the results.

E.g. In JUPITER (Ridker PM et al.), 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%.*

8.

Subgroups analysis: Were additional comparisons made on segments of the study population?

Most **RCT** publications report on additional analyses of subgroups from the overall trial population (e.g. examining only participants with diabetes, or only those age >70 years). In theory, such analyses could uncover more individualized treatment effects; however, in practice they are much more likely to be spurious and irreproducible, and therefore misleading. Subgroup analyses are often performed (and emphasized in publications) when studies do not find a statistically significant difference in the overall study population. Therefore, subgroup analysis is often a form of data-mining.

Checklist Questions

Are statistically significant results in a subgroup being emphasized in the context of a neutral or negative trial?
Was the subgroup analysis pre-defined?
Was the direction of the subgroup effect pre-defined?
Was the subgroup analysis one of a small number of hypotheses tested?
Is the subgroup variable a characteristic measured at baseline or after randomization?
Could treatment effect differences between subgroups be attributable to baseline imbalances?
Is the subgroup effect statistically significant?
Is the subgroup effect consistent within and across trials?
[**Systematic reviews/meta-analyses** only] Is the effect suggested by comparisons within rather than between studies?

Are statistically significant results in a subgroup being emphasized in the context of a neutral or negative trial?

Subgroup analyses can be used for data-mining when overall results are not statistically significantly different and may be highlighted when the **primary endpoint** fails to cross the threshold of statistical significance.

Was the subgroup analysis pre-defined?

Subgroup analyses that were not pre-defined in the protocol may be a form of data-mining, and are vulnerable to finding a difference by chance. Avoid making clinical decisions based on unanticipated significant subgroup differences (i.e. discovered post hoc) until they have been replicated in other studies.

This is particular concerning for continuous variables, such as age or cholesterol level, that are dichotomized via non-prespecified cutoffs (Schandelmaier S et al.). For example, LDL cholesterol could be dichotomized via numerous arbitrary cutoffs (e.g. >3.5, >4.0…). If not pre-specified, such cutoffs could be selected by whichever value showed the most impressive or statistically significant result. Such data-mining efforts are unlikely to uncover true subgroup differences.

Was the direction of the subgroup effect correctly pre-defined?

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

Was the subgroup analysis one of a small number of hypotheses tested?

More comparisons increase the likelihood of finding a difference by chance. See the multiplicity discussion here for more information.

Is the subgroup variable a characteristic measured at baseline or after randomization?

Subgroup analyses of variables measured after randomization may be affected by the interventions, thereby introducing confounding.

Examples of variables measured at baseline:

- *Age*
- *Sex*
- *Pre-treatment LDL cholesterol.*

Examples of variables measured after randomization:

- *LDL cholesterol achieved after 12 weeks of study intervention in fixed-dose statin trial*
- *Success of revascularization in a trial comparing coronary artery bypass grafting surgery to percutaneous coronary intervention in coronary artery disease*

Could treatment effect differences between subgroups be attributable to baseline imbalances?

Randomization ensures that confounders have equal probability of being distributed across intervention groups, but does not guarantee balance between subgroups. Subgroups are prone to imbalances of potential confounders, especially when these subgroups contain a small number of participants.

The exception is when **randomization is stratified** for the variable that defines the subgroups (e.g. stratified randomization by history of diabetes). In the case of a stratified subgroup, there is a reduced risk of confounder imbalance.

Is the subgroup effect statistically significant?

A review of 117 subgroup claims in 64 **RCTs** found that less than 40% of subgroup claims reported in the abstract were statistically significant (Wallach JD et al.).

Statistical significance is determined by examining the p-value for the test for interaction (which tests whether treatment effect differs across subgroups), not the p-value or 95% CI within a subgroup (Brookes ST et al.). “Positive” subgroup analyses that do not report the test for interaction p-value should be ignored.

E.g. In HPS (Heart Protection Study Collaborative Group), subgroup analysis based on sex (1 of 17 subgroup analyses reported) did not show a statistically significant test for interaction (p=0.18), meaning that the overall trial results applied to both males and females.

Is the subgroup effect consistent within and across trials?

Within a trial, consistent subgroup effect across multiple related outcomes (e.g. myocardial infarction, ischemic stroke and cardiovascular death) increases the credibility of there being a true subgroup effect.

E.g. Myocardial infarction, ischemic stroke, and cardiovascular death all being similarly reduced by an intervention on a subgroup of patients with diabetes.

A true subgroup effect is also more likely if additional studies replicate the effect; however, this rarely occurs. One review found that only approximately 10% of positive subgroup analyses were replicated in a subsequent trial designed to confirm the effect within the subgroup (Wallach JD et al.).

[Systematic Reviews/Meta-Analyses Only] Is the effect suggested by comparisons within rather than between studies?

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 (Steering Committee of the Physicians' Health Study Research Group), a study of men without previous cardiovascular disease, found that low-dose ASA statistically-significantly reduced the risk of myocardial infarction but not stroke. Many years later, the Women's Health Study demonstrated a statistically significant reduction in stroke but not myocardial infarction with ASA in women without previous cardiovascular disease. It would be inappropriate to conclude based on an indirect comparison of these two ***RCTs*** *that ASA has different benefits in men compared with women.*

9.

Non-inferiority trials: Was the intervention compared to see if it is “no worse” than an established therapy?

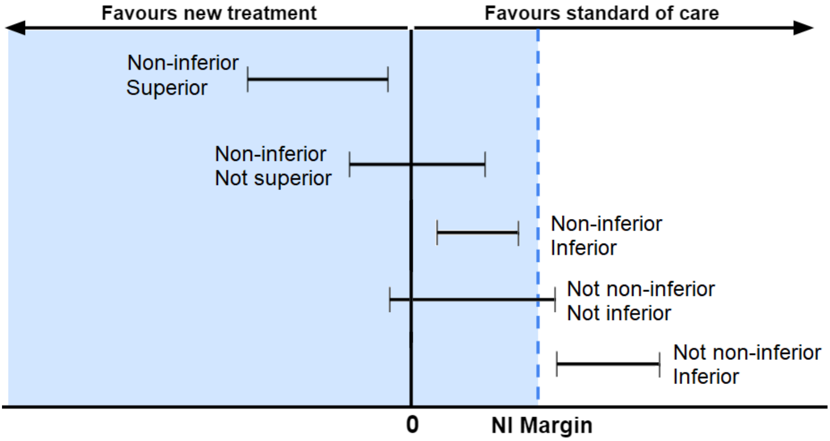
Most commonly trials test for **superiority** i.e. determining whether an intervention is superior to some comparator with respect to the **primary outcome**. Conversely, the objective of a non-inferiority trial is to test whether an intervention is “not much worse” than a comparator (usually the current standard of care) with regard to the **primary outcome**. The rationale for a non-inferiority design is that the new treatment offers some benefit other than increased efficacy, such as being safer, more affordable, or more convenient. While the fundamentals of non-inferior trials are similar to that of superiority trials, there are some unique concepts necessary when critically appraising them.

Non-Inferiority Margins

The non-inferiority margin is closely related to the **minimally important difference**, which is the smallest difference in the effect on an outcome that would be meaningful to a representative group of patients. The non-inferiority margin is the yardstick by which non-inferiority is defined, and is selected during the design of a non-inferiority trial. If the CI of the difference between the intervention and comparator crosses the non-inferiority margin, the intervention is deemed to *not be* non-inferior to the comparator. For example, consider a non-inferiority margin is a **RR** of 1.2 for stroke, and the actual **RR** is 0.9 with 95% CI 0.5 to 1.3. Since the observed upper end of the CI (1.3) is greater than the non-inferiority margin (>1.2), the conclusion is that the treatment is not non-inferior. If the upper end of the CI had been 1.1, the conclusion would be that the treatment is non-inferior given that 1.1 < 1.2.

Intuitively this should be equivalent to the **minimally important difference**, and ideally this is the case; however, researchers may choose a more "generous" non-inferiority margin (i.e. one that allows a difference greater than the **minimally important difference** to be considered "not much worse").

See the graphical depiction of concept below:


Favours new treatment
Favours standard of care
Non-inferior
Superior
Non-inferior
Not superior
Non-inferior
Inferior
Not non-inferior
Not inferior
Not non-inferior
Inferior
0
NI Margin

Plot 6. Graphical depiction of non-inferiority and related concepts.

Superiority and inferiority (consider the line of no difference):

- The treatment is considered superior when the upper end of the CI is below the line of no difference (0 in this case).

- The treatment is considered inferior when the lower end of the CI is above the line of no difference.

Non-inferiority and not non-inferiority (consider the non-inferiority margin):

- The treatment is considered non-inferior when the upper end of the CI falls to the left of the non-inferiority margin.
- The treatment is considered *not* non-inferior when the CI crosses to the right of the non-inferiority margin.

See Mulla SM et al. for more information on the questions asked below. See Hong J et al. for information concerning deficits in non-inferiority trial reporting.

Checklist Questions

Is a non-inferiority design justified by some other advantage of the intervention versus the comparator?
Did the trial use a non-inferiority margin based on a **relative** or an **absolute risk difference**?
Is the non-inferiority margin well justified based on statistical reasoning and clinical judgment?
Is the non-inferiority margin strict enough according to your own judgment?
Was non-inferiority demonstrated in both **intention-to-treat (ITT)** and **per protocol** analyses?
Was the comparator appropriate?
Has the active comparator demonstrated unequivocal **superiority** over placebo in previous trials?
Was the effect of the comparator in this trial consistent with that of previous trials?

Is the non-inferiority design justified by some other advantage of the intervention versus the comparator?

If the intervention is non-inferior but not **superior**, it should have another meaningful advantage that justifies considering it for your patients. Consider and quantify:

- Fewer, less frequent, or less-severe adverse effects
- Fewer drug interactions

- Easier to take
- Less intensive or less invasive monitoring required
- Lower cost

Note: The advantage of the non-inferior intervention should not be included in the **primary outcome** being tested for non-inferiority. This **biases** the results in favor of the new intervention.

E.g. In PRAGUE-17 (Osmancik P et al.), a ***RCT*** *comparing percutaneous left atrial appendage occlusion (LAAO) with direct-acting oral anticoagulants (DOACs) in patients with atrial fibrillation and a history of bleeding, the* ***primary outcome*** *was a composite of:*

- *Ischemic or hemorrhagic stroke*
- *Transient ischemic attack*
- *Systemic embolism*
- *Cardiovascular death*
- *Procedure-/device-related complications*
- *Major or non-major clinically relevant bleeding*

However, since the justification to see if LAAO was non-inferior to DOACs was that LAAO may offer a lower risk of bleeding, it was inappropriate to include bleeding in the ***primary outcome*** *being tested for non-inferiority. Indeed, bleeding events accounted for nearly half of all* ***primary outcome*** *events, and excluding these would not allow for*

> *the conclusion of non-inferiority (LAAO would be "not non-inferior" to DOAC).*

Did the trial use a non-inferiority margin based on a relative or an absolute risk difference?

- Non-inferiority margins based on **absolute risk** scales can falsely conclude an intervention to be non-inferior if event rates are lower than expected, which commonly occurs
- **Relative risk** non-inferiority margins are more conservative – and therefore preferable – as they scale to the incidence of outcomes in the trial

> *E.g. In SPORTIF V (Albers GW et al.), the intervention was non-inferior according to the **absolute risk difference** non-inferiority margin of 2%, but it would not have been non-inferior if a **relative risk** non-inferiority margin of 1.67 – based on the same previous study data – had been used. The discrepancy was caused by a lower-than-expected event rate of 1.2% in the warfarin group (vs. expected 3.1%).*

Is the non-inferiority margin well justified based on statistical reasoning and clinical judgment?

A trial's non-inferiority margin should be justified on the principle that the intervention being studied is (1) "not much worse" than (non-inferior to) the comparator, and (2) still better than nothing/placebo. Rules for an appropriate non-inferiority margin:

- Defined prior to undertaking the trial
- Justified relative to the **minimal important difference** (previously termed the minimal clinically important difference), which should be defined based on prior evidence
- Preserve the effect of the standard treatment over placebo

E.g. #1 In RE-LY (Connolly SJ et al.), a ***RCT*** *comparing dabigatran to warfarin for the prevention of stroke and systemic embolism in non-valvular atrial fibrillation, the pre-specified non-inferiority margin was a* ***relative risk*** *of 1.46. This was based on half the "worst case" end of the CI for benefit with warfarin vs. placebo. In other words, if RE-LY proved non-inferiority of dabigatran, it would, at its very worst, be ~2/3 (1÷1.46) as good as warfarin for this outcome.*

E.g. #2 In RESET (Kim B-K et al.), a ***RCT*** *comparing 3 months vs. 12 months of clopidogrel (added to aspirin) following drug-eluting stent placement, the non-inferiority margin 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 "worst case"* ***relative risk reduction*** *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 non-inferiority margin allowed for the shorter course of clopidogrel to be similar to or worse than placebo.*

Is the non-inferiority margin strict enough according to your own judgment?

Ultimately, you as the reader need to decide for yourself if the non-inferiority margin is reasonable and acceptable.

E.g. #1 In ROCKET-AF (Patel MR et al.), a ***RCT*** *comparing rivaroxaban to warfarin in patients with atrial fibrillation with the* ***primary outcome*** *ofstroke or systemic embolism, the non-inferiority margin was 1.46. Given the actual rate of occurrence of this outcome in the warfarin group(2.2 events per 100 patient-year), a 1.46 margin would have amounted to an increase of*

~1 event per 100 patient years. As such, this is a reasonable non-inferiority margin.

E.g. #2 In PRAGUE-17 (Osmancik P et al.), the non-inferiority margin was such that it allowed for 5% ***absolute risk increase*** *in the* ***primary outcome*** *(stroke, transient ischemic attack, systemic embolism, cardiovascular death, major or nonmajor clinically relevant bleeding, or procedure-/device-related complications) with LAAO versus DOAC. Many clinicians and patients would consider a 5%* ***absolute increase*** *in this* ***composite*** *(which includes the purported advantage of less bleeding with LAAO versus anticoagulation) to be clinically important and therefore reject non-inferiority of LAAO based on this margin.*

Note that the non-inferiority margin refers to an acceptable boundary for the “worst case” end of the CI, not the **point estimate** itself.

Was non-inferiority demonstrated in both intention-to-treat (ITT) and per protocol analyses?

- As is the case with superiority trials, **ITT** analysis is preferred as the primary analysis as it preserves the advantages of randomization and minimizes attrition bias. However, **ITT** may attenuate outcome differences between groups and make it easier to demonstrate non-inferiority.

- **Per-protocol analysis** aims to isolate the effect of the intervention by excluding patients who did not receive study treatment “per-protocol”, such as patients who dropped out or received the intervention intended for the other treatment group (“crossover”). In many cases, dropouts and crossovers are due to intervention inefficacy/intolerance and/or associated with patient prognosis, which introduces **bias**. Some falsely believe that this makes the **per-protocol analysis** the more conservative analysis for non-inferiority trials; however, that is only the case if the **bias** that is introduced favors the comparator. In other words, using the **per-protocol analysis** where protocol violations or crossovers occur more frequently in the comparator group will bias the results in favor of concluding that the intervention is non-inferior.
- In most cases, discrepancies between **ITT** and **per-protocol** analyses suggest that **bias** has been introduced into the trial. As a general rule, non-inferiority should only be accepted/concluded if it is demonstrated in both the **ITT** and **per-protocol** analysis.

In a systematic review of 231 non-inferiority **RCTs** published in five high-impact journals from 2005 to 2014, only 45% of non-inferiority **RCTs** reported both **ITT** and **per-protocol** analyses. When both were reported, discrepancies between analyses (in terms of demonstrating non-inferiority) occurred in 6% of comparisons. Neither analysis was consistently more conservative, with the **ITT** being more conservative in 50% of discrepancies (Turgeon RD, Reid EK, et al.).

Was the comparator appropriate?

The comparator intervention should:

- Be consistent with the current **standard of care.** This can be assessed by scanning local institution policy and/or national guidelines
- Be **more effective than nothing/placebo.** This can be assessed by scanning tertiary references such as DynaMed and UpToDate for high-quality evidence demonstrating clinically important benefits of the comparator
- Have an effect that is consistent with that of previous trials

E.g. In RE-LY (Connolly SJ et al.),the yearly incidence of stroke in the warfarin group was 1.6%. In a meta-analysis of older trials, the yearly incidence of stroke was 2.2%. This indicates there may be differences in the warfarin administration, monitoring, or population studied compared to previous trials.

III

Systematic Reviews and Meta-Analyses

Systematic reviews and **meta-analyses** are methods of aggregating research. By synthetizing multiple studies, the intent is to provide an answer that is more comprehensive and precise (in the case of a **meta-analysis**) than can be provided by a single trial.

Systematic reviews begin with a transparent, systematic search for all potentially relevant studies addressing a well-defined research question. Only studies that meet the specified inclusion and exclusion criteria are included in the review. These studies are then evaluated for factors such as the risk of **bias** of individual studies and **heterogeneity** between studies. If the data is also combined quantitatively, it is also referred to as a **meta-analysis**.

While this can potentially lead to high-quality evidence, the review itself must be conducted properly. A **systematic review** can produce invalid results if the search systematically missed studies or if the research question was not sufficiently focused. To address such concerns, this chapter will provide guidance on thoroughly appraising a **systematic review** with or without **meta-analysis**.

This chapter is focused on **systematic reviews** and **meta-**

analyses of **RCTs**, however these methods can also be applied to observational trials. Caution should be exercised with regards to such **meta-analyses**. This is because pooling observational trials will not reduce the potential **biases** and confounders found in the original studies. See Stroup DF et al. for a further discussion.

10.

Search

A comprehensive search is at the core of all **systematic reviews**, and is essential to ensure that all relevant trials were included. A search that is not sufficiently thorough will be more vulnerable to **publication bias**. In the case of **publication bias** "an ounce of prevention is worth a pound of cure" since the tools available to identify and adjust for **publication bias** are insensitive and cannot discriminate between **publication bias** and alternative causes for **small-study effects**. Other considerations (such as when the search was last completed) are also necessary to ensure the search is sufficiently complete.

Checklist Questions

Were a reasonable number of relevant databases searched?
When was the search conducted? Is it likely there have been subsequent publications that may alter the results?
Was a sufficient effort made to find unpublished studies (or unreported results of published studies)?
Were sources of additional published/unpublished data sought out?

Databases of published literature:Were a reasonable number of relevant databases searched?

It is important to search multiple databases to maximize the identification of all relevant studies, as no single database includes all studies. One study (Royle P et al.) compared three major databases to a set of relevant studies established by searching twenty-six additional databases:

Table 11. Proportion of relevant trials identified by different databases.

Database	Proportion of relevant trials identified
MEDLINE	69%
EMBASE	65%
CENTRAL	79%
Combining all three	97%

The optimal selection of which (and how many) databases to search will depend on the discipline, topic area, and type of intervention. For example, studies evaluating nursing and physiotherapy interventions should at minimum include CINAHL and PEDro, respectively. A good rule-of-thumb is to search MEDLINE and at least 1-2 other topic-specific databases (e.g. EMBASE and CENTRAL for pharmacotherapy studies).

Timeframe: When was the search conducted? Is it likely there have been subsequent publications that may alter the results?

There are no strict rules as to how long is too long before a review becomes outdated, as this largely depends on the rate of evidence generation in a given field or topic area. It is important to consider the rate at which new publications are being added to the literature (i.e. considering if it is a "hot" topic) and whether the results would likely be sensitive to new publications (i.e. a **meta-analysis** with low-to-moderate certainty). If there are already several large high-quality trials showing consistent results it is less likely that any new literature would substantially change results.

"Hot" topic: A living ***meta-analysis*** *(i.e. a* ***meta-analysis*** *that is actively updated with new evidence) (Siemieniuk RA et al.) of drug treatments for COVID-19 illustrates an instance of rapidly changing evidence. The first version, published July 2020, included 32* ***RCTs*** *and evaluated 17 therapies. The fourth version, published in March 2021, included 196 trials and evaluated 27 therapies. Under such circumstances,* ***meta-analyses*** *become quickly outdated.*

"Cold" topic: The evidence surrounding the cardiovascular risk associated with rosiglitazone has

changed minimally for over a decade. ***Meta-analyses*** *from 2007 and 2010 (Nissen SE et al. 2007, 2010) demonstrated increases in the risk of myocardial infarction with rosiglitazone. Both reviews had large patient sample sizes (27,847 and 35, 531 respectively), a factor which weighed in favor of their persisting relevance. As such, the evidence on this topic has remained largely unchanged since those reviews.*

Grey literature:Was a sufficient effort made to find unpublished studies (or unreported results of published studies)?

A thorough search of unpublished literature aims to minimize the effects of **publication bias**.

E.g. In a ***meta-analysis*** *(Siu JT et al.) of N-acetylcysteine for non-acetaminophen-related acute liver failure the authors searched all of:*

- *The following databases (without language restrictions): Cochrane Hepato-Biliary Group Controlled Trials Register, the Cochrane Central Register of Controlled Trials, MEDLINE Ovid, Embase Ovid, LILACS,* *Science Citation Index Expanded, and Conference Proceedings Citation Index – Science*

- *The reference lists of all included studies and relevant papers*
- *The following online clinical trial registries: ClinicalTrial.gov, European Medicines Agency, World Health Organization International Clinical Trial Registry Platform, the Food and Drug Administration, and pharmaceutical company sources for ongoing or unpublished trials*

The authors of relevant papers were also contacted to inquire regarding any further published or unpublished work.

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 a median of 1.3 years in one study) compared to trials with statistically non-significant results ("neutral" studies) (Hopewell S et al., Ioannidis JP).

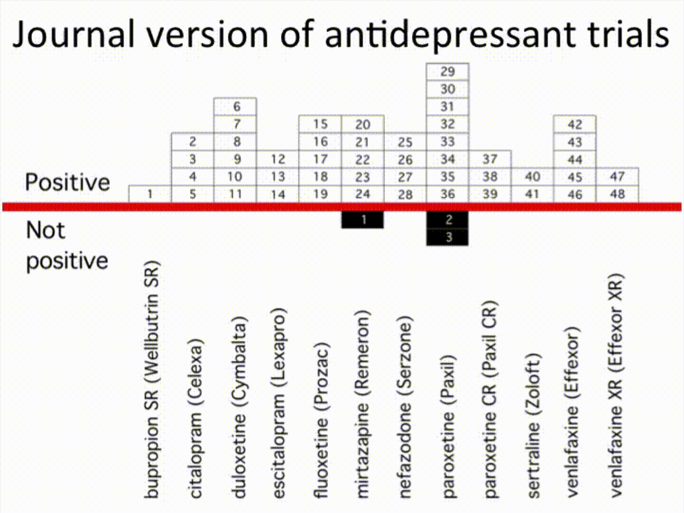
- Published trials have a 15% larger estimate of effect compared to unpublished trials (McAuley L et al.)
- Although more common with industry-funded trials, government-funded studies are still prone to publication bias (32% vs. 18% unpublished 5 years

after completion) (Jones CW et al.)

- 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) (Pereira TV et al.)
- In one study of 42 **meta-analyses**, in 93% of cases the 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 (Hart B et al.)

Bottom line: **Meta-analyses** of only published trials will overestimate the effects of drugs and other interventions, especially when **meta-analyses** are conducted "earlier on" (before the neutral trials get published). Consequently, there is likely a greater risk of **publication bias** in **meta-analyses** based on a few small studies.

E.g. A review (Turner EH et al.) of antidepressants found that 94% of published trials demonstrated a statistically significant difference with respect to the ***primary outcome****. However, when combined with unpublished FDA review data, only 51% of total trials demonstrated a statistically significant difference with respect to the* ***primary outcome****. Including only published studies increased the* ***relative effect*** *size by 32%.*



Gif 1. ***Publication bias*** *among antidepressant trials as reported by Turner EH et al. GIF created by Turner EH.*

11.

Results of the systematic review

The quality of the **systematic review** depends both on the quality of the individual studies and the aggregate characteristics of these studies. If the aggregate results are missing studies, contain predominantly poorly conducted studies, or are highly **heterogeneous** then this will likely warrant lower confidence in the results.

Checklist Questions

Do all inclusions & exclusions of trials make sense?
Are you aware of any relevant studies that were not identified/ included in this review?
Did reviewers adequately assess individual trials for risk of **bias**?
Was each component reported separately, or summarized with a composite quality score?
Are there any differences between studies that should preclude **meta-analysis**?

Risk of bias within trials (internal validity): Did reviewers adequately assess for (& report) risk of bias?

Risk of **bias** should be evaluated by using a tool that is specific to **RCTs**. The Cochrane risk of bias tool (version 1 (Higgins JPT et al. 2011) or 2 (Sterne JAC et al. 2019)) evaluates the risk of individual trial **biases** and offers the most transparent assessment of trial **internal validity** (see NERDCAT-RCT for more information regarding **internal validity**). ROBIS-I (Sterne JA et al.) is a similar tool available for appraising risk of **bias** in observational trials.

Quality Scores

“Quality scores” such as the Jadad score are more closely related to reporting quality than methodological issues, and lead to wide variability in conclusions on “quality” based on the score used. In particular, the Jadad score is considered obsolete and is a poor measure of risk of **bias**.

Methodological & clinical heterogeneity: Is it appropriate to perform a meta-analysis?

- Methodological **heterogeneity**: Are there methodological differences (e.g. risk of **bias**) between studies?
- Clinical **heterogeneity**: Are there any differences in clinical characteristics between the individual trials (i.e. any component of **PICO**) that preclude pooling

the trials together in a **meta-analysis**?

- Is the impact of any of these characteristics tested in a subgroup analysis or meta-regression?

Testing possible sources of **heterogeneity** may identify causes for statistical **heterogeneity** identified in the **meta-analysis** (e.g. the intervention may only appear beneficial in trials at high risk of **bias**, but not in those at low risk).

See NERDCAT-RCT to learn more on how to appraise validity of subgroupeffects.

12.

Results of the meta-analysis

As with **RCTs**, outcomes ought to be interpreted beyond just statistical significance to assess the magnitude of effect and clinical relevance. Interpretation also requires considerations beyond what is necessary when appraising **RCTs**. It is also important to consider how many trials reported on a particular outcome, and what the quality of those specific trials were. Additionally, even if the trials are otherwise clinically and methodologically similar, statistical **heterogeneity** identified by visual inspection and/or formal statistical testing may preclude confidently combining trial results.

Checklist Questions

I^2 Value – What was the statistical **heterogeneity**?
Appropriate to pool the results & interpret the summary statistics?
Fixed-effects or random-effects?
Is the model used appropriate?
Which effect measure was used? (e.g. **OR**, **RR**, **SMD**)
What is the baseline risk for your patient from the individual trial they would fit best?
What was the calculated absolute effect? (e.g. ARR, NNT)
What proportion of the included studies report on this outcomes?
If performed, what GRADE rating was assigned to each outcome?

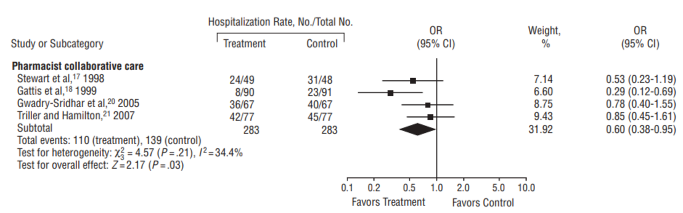
Statistical heterogeneity: What was the statistical heterogeneity?

For information regarding the interpretation of forest plots refer to Appendix: Fundamental Statistics.

Table 12. Different methods of assessing heterogeneity.

Methods to Assess Heterogeneity	**Description**
Visual assessment	An intuitive visual evaluation of **heterogeneity** (see examples below)
Cochran's Q	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	I^2 ranges from 0-100% and represents the amount of variability in the point estimate across trials. Rule-of-thumb (one of many): I^2 <25% = minimal heterogeneity; I^2 >50% = substantial **heterogeneity** (may not be appropriate to meta-analyze trials) **(preferred over Cochran's Q)**

E.g. #1 A forest plot from a review (Koshman SL et al.) evaluating the impact of pharmacist involvement in the care of patients with heart failure on all-cause hospitalization rate:



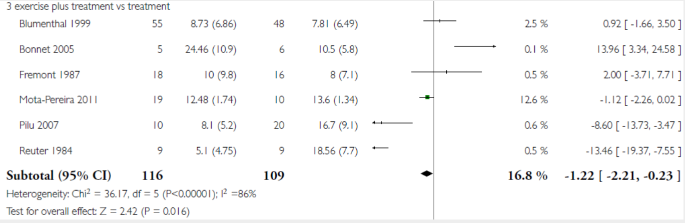
Study or Subcategory	Hospitalization Rate, No./Total No. Treatment	Hospitalization Rate, No./Total No. Control	Weight, %	OR (95% CI)
Pharmacist collaborative care				
Stewart et al,[17] 1998	24/49	31/48	7.14	0.53 (0.23-1.19)
Gattis et al,[18] 1999	8/90	23/91	6.60	0.29 (0.12-0.69)
Gwadry-Sridhar et al,[20] 2005	36/67	40/67	8.75	0.78 (0.40-1.55)
Triller and Hamilton,[21] 2007	42/77	45/77	9.43	0.85 (0.45-1.61)
Subtotal	283	283	31.92	0.60 (0.38-0.95)

Total events: 110 (treatment), 139 (control)
Test for heterogeneity: $\chi^2_3 = 4.57$ ($P = .21$), $I^2 = 34.4\%$
Test for overall effect: $Z = 2.17$ ($P = .03$)

Plot 2. Pharmacist collaborative care vs. usual for patients with heart failure on the outcome of all-cause hospitalization.

Visually it can be seen that the ***point estimates*** *are directionally consistent and all the CIs overlap. Consequently meta-analyzing the results for this outcome is appropriate. Notably, this is a case of appropriate* ***meta-analysis*** *despite there being "moderate" statistical heterogeneity as measured by* I^2 *(34.4%), as discussed in the note below.*

E.g. #2 A forest plot from a review of exercise for depression (Cooney GM et al.) evaluating the effects of exercise plus treatment vs. treatment alone:



3 exercise plus treatment vs treatment

Blumenthal 1999	55	8.73 (6.86)	48	7.81 (6.49)	2.5 %	0.92 [-1.66, 3.50]
Bonnet 2005	5	24.46 (10.9)	6	10.5 (5.8)	0.1 %	13.96 [3.34, 24.58]
Fremont 1987	18	10 (9.8)	16	8 (7.1)	0.5 %	2.00 [-3.71, 7.71]
Mota-Pereira 2011	19	12.48 (1.74)	10	13.6 (1.34)	12.6 %	-1.12 [-2.26, 0.02]
Pilu 2007	10	8.1 (5.2)	20	16.7 (9.1)	0.6 %	-8.60 [-13.73, -3.47]
Reuter 1984	9	5.1 (4.75)	9	18.56 (7.7)	0.5 %	-13.46 [-19.37, -7.55]
Subtotal (95% CI)	**116**		**109**		**16.8 %**	**-1.22 [-2.21, -0.23]**

Heterogeneity: $Chi^2 = 36.17$, df = 5 ($P<0.00001$); $I^2 = 86\%$

Test for overall effect: $Z = 2.42$ ($P = 0.016$)

Plot 3. Exercise plus treatment vs. treatment alone for patients with depression on the outcome of reduction in depression symptoms post-treatment.

Visually it can be seen that the ***point estimates*** *have unreasonable variation and the CIs have minimal overlap. Consequently* ***heterogeneity*** *is a concern and additional considerations are necessary, as discussed more below.*

If **heterogeneity** is judged to be too high, this requires either:

- Different statistical approach to pool the results (i.e. random-effects model, see below)
- Evaluation of clinical & methodological sources of **heterogeneity**
- A decision not to meta-analyze the results for the outcome in question

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

Statistical models: Fixed-effects or random-effects? Is the model used appropriate?

Either the fixed-effects model or random-effects model may be used to pool results. In many cases, both models produce very similar meta-analytic results. However, some differences can be noted:

Table 13. Differences between fixed-effects and random-effects models.

Fixed-Effect Model	**Random-Effects Model**
Assumes all trials measure same “true” underlying effect	Does not assume that all trials estimate the exact same underlying effect (e.g. different populations may vary in their response to intervention)
Less conservative if statistical **heterogeneity** present (uses narrower CIs)	More conservative if statistical **heterogeneity** present (uses wider CIs)
Statistical weight of a trial is proportional to the number of participants/events (i.e. larger trials given more weight).	Compared with a fixed-effect model, a random-effects model will give relatively more weight to smaller trials when studies are heterogeneous.

In cases where there is evidence of **small-study effect**, the random-effects model can “pull” the summary estimate towards the smaller trials (which are more prone to **publication bias**). In other words, statistical analysis cannot fix poor data.

Effect measure and precision

Refer to Randomized Controlled Trials: Interpreting the results for a discussion of how to assess **point estimates** and CIs for clinical importance.

Refer to Appendix: Fundamental Statistics for a discussion of different measures of effect. Depending on the studies included and the outcome types, some effect measures may be more appropriate than others (e.g. if multiple different symptom scales were used between studies, it would be most appropriate to use **standardized mean difference** and not raw mean difference scores)

What proportion of included studies report on this outcome?

Why is outcome reporting bias so concerning?

- In one study of 122 **RCTs**, 50% of efficacy outcomes and 65% of harm outcomes were incompletely reported. Additionally, 62% of the trials had their **primary outcome** changed in the final published reported compared to the original protocol (Chan A-W, Hróbjartsson A et al.).
- A study by the same lead author also found outcome reporting bias present in government-funded studies. Additionally, it found that neutral studies were most likely to have reporting issues (i.e. reporting results as “not statistically significantly different” without reporting absolute values) (Chan A-W, Krleza-Jerić K et al.).
- In one study of 42 **meta-analyses**, in 93% of cases the 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 (Hart B et al.).

Bottom line: As with individual trials, neutral 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 threat to the accuracy of intervention effect estimates (i.e. overestimation of benefits and underestimation of harms, distorting the true trade-off between benefits and harms).

Outcome reporting bias should be considered when data on a clinically important outcome is only available for a minority of included studies, which in turn should raise concerns regarding the certainty of evidence (see the discussion of GRADE ratings below).

> *E.g. A* ***meta-analysis*** *by Ortiz-Orendain J et al. compared antipsychotic polypharmacy vs. antipsychotic monotherapy for the treatment of schizophrenia. It found no statistically significant difference with regards to drowsiness between the groups (****RR*** *1.0; 95% CI 0.5-2.0). However only 12 of 62 trials reported on this outcome. There is consequently reason to suspect selective reporting, and this lowers the certainty of evidence with regards to this outcome.*

The evaluation of selectively reported outcomes is more nuanced when the outcome can be measured in many different ways (e.g. 10% of studies may report on depression score change as measured by the HAM-D scale, but 70% of studies may have reported on depression score change as measured by PHQ-9). In these cases it is necessary to consider the overarching outcome (e.g. depression score change by any scale) to evaluate whether there was selective reporting.

If performed, what GRADE rating was assigned to each outcome?

GRADE (Grading of Recommendations, Assessment, Development and Evaluations) is a method of transparently assessing the certainty of evidence for a particular outcome as either high, moderate, low, or very low.

Certainty is determined by two factors: the type of studies examined (**RCTs** or observational studies), and the characteristics of those studies. **RCTs** start at “high certainty” and observational trials at “low certainty”. Studies are then rated up or down – either by one or two levels per characteristic. For example, for a **meta-analysis** of **RCTs** the evidence would start at high certainty, but then may be downgraded to moderate certainty due to serious risk of bias, and then rated down again to low certainty due to inconsistency.

Certainty can be rated down for any of:

Table 14. Reasons to downgrade GRADE certainty.

Risk of bias	Refers to **internal validity** limitations due to factors such as inadequate randomization, **allocation concealment**, blinding, or selective reporting. See the here section for more information on how to assess risk of bias.
Imprecision	Refers to a CI which spans clinically important differences. For instance, a **RR** with a 95% CI of 0.5 to 2.0 for mortality is imprecise as the CI includes both possibilities that the intervention halves or doubles deaths. In contrast, a **RR** with a 95% CI of 0.6 to 0.65 for schizophrenia symptom reduction is very narrow and would be considered precise. Imprecision can be assessed formally by comparing the achieved sample size to the calculated optimal information size as described by Guyatt et al.
Inconsistency	Refers to the presence of between-study **heterogeneity**. This can be assessed visually and statistically – see the Statistical Heterogeneity discussion above for more information.

Indirectness	Refers to results which are not directly applicable to one or more of the study **PICO** elements (i.e. in terms of patient characteristics, interventions, or treatment settings. For example, using studies of adults as indirect evidence of the effects of treatment in children. Indirectness can also apply to outcomes, such as when **surrogate outcomes** act as indirect evidence of clinically important outcomes.
Publication bias	Refers to a systematic tendency for results to be published based upon the direction or statistical significance of the results. Such tendency can lead to bias when aggregating evidence if the methods are more likely to include published literature than unpublished literature.

Certainty of evidence based on observational studies can be rated up for any of:

Table 15. Reasons to upgrade GRADE certainty.

Large magnitude of effect	**Confounding** alone is unlikely to explain large associations (e.g. risk ratio <0.50 or >2.0).
Dose-response gradient	Refers to an increasing effect size as the dose increases. If such a gradient is apparent then this increases the likelihood of a true effect.
All residual confounding would decrease magnitude of effect (in situations with an effect)	Residual confounding refers to unknown or unmeasurable confounding that could not be accounted for in an observational study. It is seldom possible to completely eliminate all residual confounding in observational studies as there is always the possibility of imbalance of yet-unknown prognostic variables. If all of such residual confounders were expected to decrease the effect size, then the effect estimate is a conservative measure. If this conservative analysis demonstrates a benefit, then this warrants greater confidence in the result.

It is important to emphasize again that these assessments are specific to each outcome. For instance, the evidence for the comparison of an intervention versus a comparator may be of high certainty for one outcome, but low certainty for another outcome. All of these judgements are made subjectively, ideally with rationales provided. The intention is not for this to be a mechanistic rating scheme, but rather to transparently communicate the thought process behind ratings.

Appendix: Fundamental Statistics

This appendix covers the fundamental statistical concepts necessary to critically appraise **randomized controlled trials (RCTs)** and **systematic reviews/meta-analyses**.

P-Value Interpretation

P-values are sometimes misinterpreted to mean “the probability that the results occurred by chance”. This is problematic on at least two counts: the probability of any particular result occurring by chance will be extremely low, and also “by chance” requires further definition to be meaningful. A more technical definition is that the p-value is the probability of finding a result at least as extreme as the observed result if the **null hypothesis** (usually “no difference”) is correct and all assumptions used to compute the p-value are met.

*E.g. #1 A **RCT** finds a mean difference in pain of 2.3 on a 10-point scale between treatment A and treatment B (p=0.04). This means that, if there truly were no difference between treatment A and B, the probability of finding a mean difference of ≥2.3 by chance alone is 4%.*

E.g. #2 A ***RCT*** *of hydralazine-nitrate vs. placebo demonstrated a* ***relative risk reduction (RRR)*** *of 10% for heart failure hospitalization with p=0.34. This means that, if there truly is no difference between hydralazine-nitrate vs. placebo, the probability of finding a RR* ***RRR*** *of 10% or greater in heart failure hospitalization by chance alone is 34%.*

Errors in p-value interpretation usually involve confusing the following two probabilities:

- The probability that the treatment is ineffective given the observed evidence (the misinterpretation of a p-value)
- The probability of the observed evidence if the treatment were ineffective (what the p-value provides)
    - For more information on this common inference mistake (known as the "The Prosecutor's Fallacy") see Westreich D et al.

By convention, a p-value ≤0.05 is considered statistically significant, though this is increasingly recognized as an oversimplification and ignores consideration of clinical importance.

The most important takeaway from this discussion is that the typical understanding of a p-value is incorrect and such misunderstanding can lead to erroneous conclusions. For a further discussion of p-value misinterpretation in medical literature see Price R et al. For a more advanced discussion on why research findings are often false despite statistically significant results see Ioannidis JPA.

Confidence Interval (CI) Interpretation

The technical definition of a 95% confidence interval (CI) is: If we were to repeat the study an infinite number of times, 95% of 95% CIs would contain the true effect, if all assumptions used to calculate the interval are correct. Consequently, a 95% CI does not entail "there is a 95% chance the true value is within this range" (a common misinterpretation). As with p-values, the true meaning is more nuanced. See here for visual CI simulations as illustrative examples.

The 95% CI provides all of the information of a p-value (and is derived using the same information), but also adds information on a plausible range of the effect size. When interpreting CIs, it is important to examine both ends of the CI and judge whether there is a clinically important difference between them. For example, a point estimate of 5% **absolute risk reduction** in stroke risk over 5 years with a 95% CI of 3% to 7% will include a narrow range that many clinicians would consider clinically important difference at both ends of the interval. This examination (along with considerations of **bias**) will help establish the degree of uncertainty in the result. See McCormack J et al. for a discussion of how considering only statistical significance without proper regard for CIs can cause confusion.

By convention, a 95% CI that does not include the null (e.g. a **relative effect** 95% CI that includes 1.0 or an **absolute risk difference** 95% CI that includes 0%) is considered statistically significant (i.e. consistent with $p<0.05$).

Sample Size Interpretation

It is sometimes believed that sample size (i.e. how many participants were included in the study) is a determinant of **internal validity**. However, a review (Kjaergard LL et al.) found that smaller trials (<1,000 participants) only exaggerated treatment effects compared with larger trials (≥1,000 participants) when they had inadequate randomization, **allocation concealment**, or blinding. As such, sample size itself is not indicative of **bias**. Furthermore, if there were too few participants enrolled to detect a difference between groups this will be reflected in the corresponding wide CI (see Confidence Intervals – How precise were the estimates of treatment effect? for a discussion of wide CIs and how they illustrate precision).

Advanced discussion (included for completeness, but rarely applicable to appraisal)

An exception may be true for some "very small" trials – as parametric statistical tests rely on the central limit theorem and require a minimum sample size (e.g. n≥30 is often suggested). See Fagerland MW for a more discussion. The details of this particular statistical concern are beyond the scope of this resource, but two simplified takeaways are that:

- "Very small" sample size may be a concern for the proper use of certain statistical methods when measuring continuous outcomes
- A minimum sample of 30 is an arbitrary rule-

of-thumb to prevent this – nonetheless 30 does provide an approximation of the sample sizes where this may be a concern (e.g. this will almost certainly not be a concern for a trial with several hundreds of participants)

Absolute Risk Differences and Relative Measures of Effect

Absolute Risk Difference

The **absolute risk difference** between groups refers to the risk of an event in one group minus the risk in another group. Consider the following example of a theoretical 2 year trial examining insomnia rates:

Table 16. Absolute risk difference example.

Outcome	**Intervention Group**	**Comparator Group**	**Absolute Difference**	**Duration**
Insomnia	15%	5%	+10%	2 years

In this case, the **absolute difference** was calculated by subtracting the intervention group event rate (15%) by the comparator group event rate (5%), which equals 10% (15% – 5%). This difference is “absolute” because the number (e.g.

+10% risk of insomnia over 2 years) is independently meaningful.

Absolute differences also need to be communicated in the context of time. For example, a 1% **absolute risk reduction** over 1 month is quite different from a 1% **absolute risk reduction** over 10 years. As such, **absolute differences** should be stated as a **__**% increase/decrease over [timeframe].

Relative Measures of Effect

This contrasts with **relative effect** measures. One example of a **relative effect** is **relative risk (RR)**, which is calculated by dividing the risk of event in the intervention group by that in the comparator group.

Table 17. Relative effect example.

Outcome	**Intervention Group**	**Comparator Group**	**Relative Risk (RR)**	**Duration**
Insomnia	15%	5%	3.0	3 months

In this case, the **RR** was calculated by dividing the intervention group rate (15%) by the comparator group event rate (5%), which equals 3.0 (15% ÷ 5%). This difference is "relative" because the number (e.g. a 3.0 **RR** of experiencing insomnia) is dependent on the risk in the comparator group to be meaningful. **RR** 3.0 means that the risk has tripled, but without knowing the baseline risk that is being tripled, then the number is not fully interpretable.

This dependence can be problematic if not properly considered.

Consider the following example, where the **RR** is identical in both cases:

Table 18. Relative effect dependency on baseline risk example.

Relative Risk (RR)	**Baseline Risk**	**Risk on Treatment (RR x Baseline Risk)**	**Absolute Risk Difference**	**Duration**
0.5	30%	15%	15%	10 years
0.5	2%	1%	1%	10 years

As demonstrated, **RR** considered in isolation lacks crucial information. The same concept is relevant when (responsibly) buying a product during a sale. Knowing that a particular product is 50% off is not sufficient for a rational choice, as there needs to also be information about the original price (e.g. $20 versus $20,000) before deciding if the purchase is desirable.

Note: **RR** is just one relative measure – see discussion below for information on **relative risks**, **odds ratios**, and **hazard ratios**.

Number Needed to Treat or Harm

Both number needed to treat (NNT) and number needed to harm (NNH) are measures of how many patients have to receive the treatment of interest for one additional person to experience the outcome of interest (NNT being for beneficial outcomes, and NNH for harmful outcomes).

It is calculated as: 100 ÷ **Absolute risk difference**, with the result always rounded up.

For example, if a treatment has a 7% **absolute risk increase** of causing urinary retention over 3 months then the NNH is 15 (100 ÷ 7 = 14.3, then round up to 15). This means that 15 patients will have to be treated for one of them to have urinary retention over the next 3 months (always including timeframe, as with **absolute risk differences**).

This is an alternative way to understand **absolute risk differences** that may be more intuitive to some (though it is more poorly understood by patients than other measures, as discussed here).

Relative Risk, Odds Ratios, and Hazard Ratios

Before delving into the details of each type of **relative effect** it should be noted that all of them have the following features:
Any relative measure = 1.0 means there was no difference between groups
Any relative measure > 1.0 means the outcome was more likely with the intervention than the comparator
Any relative measure < 1.0 means the outcome was less likely with the intervention than the comparator

To demonstrate the differences between these measures of **relative effect**, consider the following table:

Table 19. Example 2×2 chart of aspirin vs. placebo for stroke prevention.

	Stroke	**No stroke**
Aspirin	10 (A)	90 (B)
Placebo	20 (C)	80 (C)

Relative Risk (RR)

Calculating the **RR** consists of dividing the risk of event in the aspirin group by risk of event in the placebo group.

Using the above table:
The risk of event in the treatment group: A ÷ (A+B)
The risk of event in the comparator group: C ÷ (C+D)
With numbers imputed: 10 ÷ 100 = 0.1 (or 10%) in the aspirin group and 20 ÷ 100 = 0.2 (or 20%) in the placebo group.
The **RR** is then 0.1 ÷ 0.2 = 0.5.

Odds Ratio (OR)

Calculating the **OR** consists of dividing the odds of an event in the aspirin group by the odds of an event in the placebo group.

Using the above table:
The odds of event in intervention group: A ÷ B
The odds of event in the comparator group: C ÷ D
With numbers imputed: 10 ÷ 90 = 0.11 in the aspirin group and 20 ÷ 80 = 0.25 in the placebo group.
The **OR** is then 0.11 ÷ 0.25 = 0.44.

OR are similar to **RR** when events are rare (A ÷ (A+B) ≈ A ÷ B when A is very small) (Holcomb WL et al.). As events become more common, these measures diverge and **ORs** will overestimate **RRs** (such as in this example where RR=0.5 and OR=0.44). The ClinCalc tool can be used to convert **OR** to **RR**.

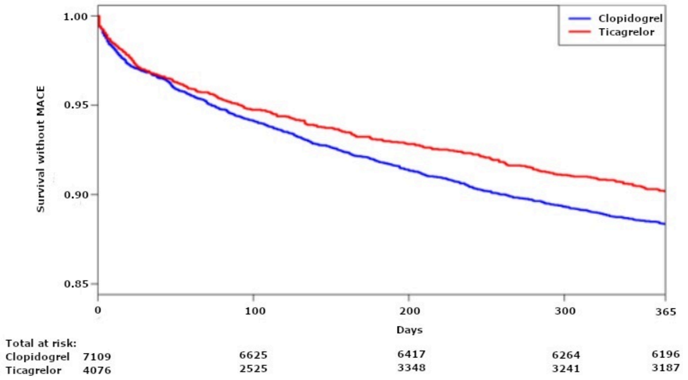
Hazard Ratio (HR)

Hazard ratios (HRs) represent the average of the instantaneous incidence rate at every point during a trial. Consider an example of a 5-year trial that has a **HR** of 0.70 for the outcome of death comparing an intervention against some comparator. This means that a participant assigned to intervention will be 30% less likely to die relative to the comparator at any point during the trial:

- Year 1: If 5% have died in the comparator group, then 3.5% are expected to have died in the intervention group (5% * 0.70 = 3.5%)
- Year 2: If 10% have died in the comparator group, then 7% are expected to have died in the intervention group (10% * 0.70 = 7%)
- Year 5: If 20% have died in the comparator group, 14% are expected to have died in the intervention group (20% * 0.70 = 14%)

The same is approximately true at any given timepoint during the trial follow-up. These are all approximations as the **HR** is an average, it (almost certainly) will not be exactly true at every time point. For instance, the final **HR** might be 0.70, but it could be 0.80 during the first half of the trial and 0.60 during the latter half.

For example, consider the unadjusted analysis from an observational study (Turgeon RD, Koshman SL, et al.) that compared the use of ticagrelor vs. clopidogrel in patients who had undergone percutaneous coronary intervention following acute coronary syndrome (ACS). For the outcome of survival without major adverse coronary events (MACE) the **HR** was 0.84 before adjustment for potential confounding variables, depicted visually below:



Graph 3. Kaplan Meier curve of survival without major adverse coronary events

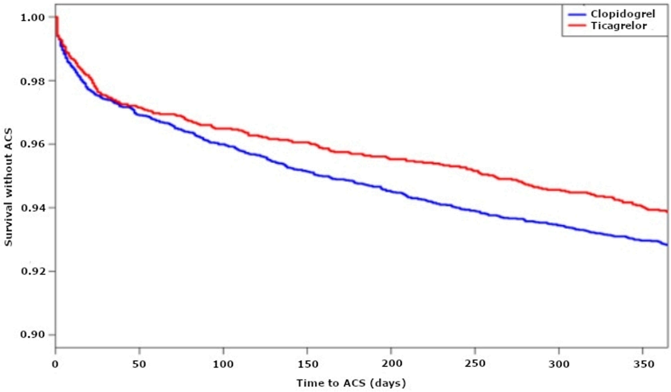
While not exactly true at every time point, past the first 100 days the cumulative proportion patients experiencing death or MACE in the ticagrelor group appears to be roughly 84% of the cumulative proportion in the clopidogrel group fairly consistently (see “Kaplan Meier Curves” for more information below on how to interpret these types of graphs). This coheres with the **HR** of 0.84 discussed above.

HRs are usually similar to **RRs** (Sutradhar R et al.). **HRs** examine multiple timepoints over trial follow-up, whereas **RRs** evaluate cumulative proportions at the end of the trial (or at another single timepoint). **HRs** can account for differential follow-up times, and contain more information than **RRs**/**ORs** since they include the added dimension of time (Guyatt G et al.). **HRs** are limited in their ability to convey fluctuations in effect over time, as a **HR** of 1.0 could mean that there was consistently no effect, or it could mean that there was beneficial effect during the first half and a proportional detrimental effect during the latter half (Hernán MA). However, some of these limitations can be overcome by combining a **HR** with the use of a Kaplan Meier curve, as discussed below.

Kaplan Meier Curves

Cumulative Hazards

Kaplan Meier curves are graphical representations comparing event accrual between two groups over time. Consider another example from the aforementioned study comparing ticagrelor vs. clopidogrel:



Graph 4. Kaplan Meier curve of survival without acute coronary syndrome (ACS)

Each curve displays the cumulative proportion of patients in that group who have experienced the outcome of interest. As time passes more participants experience the outcome and the curve progresses downwards. The **relative differences** in outcome accumulation can be expressed as a **HR**, as discussed above. Note that, while not displayed, each curve has a CI surrounding it at every time point.

Onset of Benefit

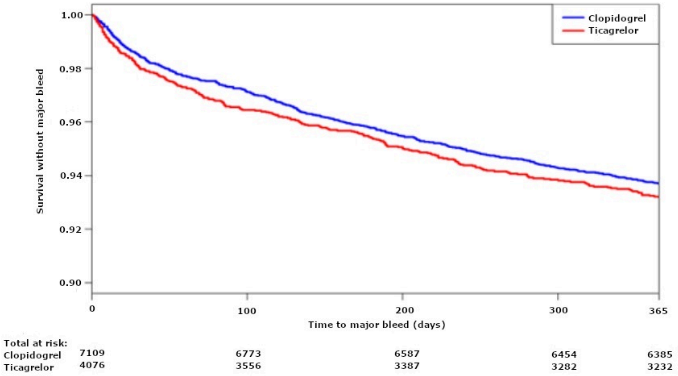
Consider the above Kaplan Meier curve. During the first 50 days, the curves for the two groups overlap. However, after this point the two curves begin to separate. This curve provides insight into the onset of benefit of the intervention and if the benefit is sustained over time. In this case, onset of benefit begins after approximately 50 days and is sustained as time elapses.

Course of Condition

The graph also gives insight into event rates over time. As can be seen above, the curve is steepest initially – indicating that the risk of death or ACS is highest immediately following the intervention. The slope of the curve then flattens and remains relatively stable – indicating the event rate after the initial period is relatively constant during the first year. This demonstrates how Kaplan Meier curves can be useful to understand the course of a condition over time.

Total at Risk (or Number at Risk)

Consider another curve from the same study, this one examining survival without major bleeds:



Graph 5. Kaplan Meier curve of survival without major bleeding

As depicted, sometimes there is also a “Total at risk” (sometimes called “Number at risk”) table beneath the curve. This can give additional information regarding the participants as they progressed through the trial. All participants begin “at risk”, but as the study progresses the number decreases. Below are the following reasons the number may decrease, as well as possible implications if the decreases are not balanced between groups:

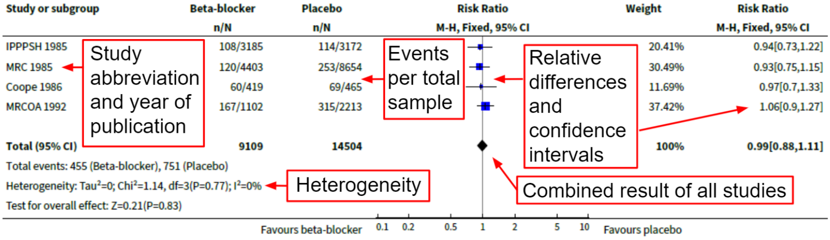
Table 20. Reasons and implications of total at risk decreases.

Reasons for total at risk decreasing	**Implications if imbalanced between groups**
Outcome of interest occurred	If there is a difference in effect between the intervention and comparator, then the total at risk may decrease more quickly in one group. This is evidence of effect, not **bias**.
Death	If there are differences in mortality rates then this should prompt consideration of the relative safety of the comparators, as well as consideration of death as a competing event within the analysis.
Loss to follow-up	This could result in systematic **bias** if the reasons for loss to follow-up are not random (for more see the discussion on **loss to follow-up** here).
The study ended before the participant had outcome data at that time point	If by chance there is a difference in how many patients were enrolled early in one group (and thus had more time to accrue events) this could **bias** a **RR** or **OR**. For example, by chance one group might have patients enrolled for an average of 4 years and another group had patients enrolled for an average of 5 years. However, since the **HR** incorporates the timing of events, this should not result in **bias**.

Thus the total at risk table can serve as a clue that further examination should be undertaken to see if there is **bias**.

Forest Plots

Forest plots are used in **meta-analyses** to graphically depict the effects of an intervention across multiple studies. Consider the labeled example below from the "Beta-blockers for hypertension" Cochrane Review (Wiysonge CS et al.):



Plot 4. Forest plot of beta-blockers vs. placebo in patients with hypertension for the outcome of mortality.

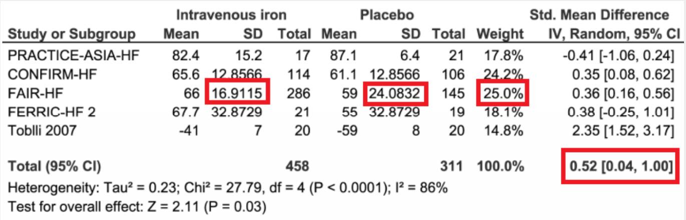
As showcased, forest plots show information about each individual study included for that outcome, and also the combined results. This information is displayed visually as well as numerically. Trials with more events or participants are generally given greater weight. **Heterogeneity** is typically measured via I^2, which is 0% in this case. For more information on **heterogeneity** see here.

Standardized Mean Difference Interpretation

The **standardized mean difference (SMD)** is a method of combining multiple continuous outcome scoring systems into one measurement. For example, when performing a meta-analysis on the effects of antidepressants on depression symptom reduction, trials may use different scales to rate depression symptoms (HAM-D, PHQ-9, etc.). **SMD** will allow the aggregation of the results of all these studies. Notably, using **SMD** assumes that differences between studies are due to differences in scales (and not in intervention/population characteristics). Arbitrary “rule-of-thumb” cutoffs (e.g. **SMD** of 0.2 = “small effect”) may not reflect the minimal important difference.

An alternative approach is to transform the **SMD** into a more familiar scale (Higgins JPT et al.). Multiply the **SMD** by the standard deviation (SD) of the largest trial to convert to its scale.

E.g. These are the results of a ***meta-analysis*** *which assesses the effects of IV iron on health-related quality of life at 6 months in patients with HFrEF (Turgeon RD, Barry AR, et al.):*



	Intravenous iron			Placebo				Std. Mean Difference
Study or Subgroup	**Mean**	**SD**	**Total**	**Mean**	**SD**	**Total**	**Weight**	**IV, Random, 95% CI**
PRACTICE-ASIA-HF	82.4	15.2	17	87.1	6.4	21	17.8%	-0.41 [-1.06, 0.24]
CONFIRM-HF	65.6	12.8566	114	61.1	12.8566	106	24.2%	0.35 [0.08, 0.62]
FAIR-HF	66	16.9115	286	59	24.0832	145	25.0%	0.36 [0.16, 0.56]
FERRIC-HF 2	67.7	32.8729	21	55	32.8729	19	18.1%	0.38 [-0.25, 1.01]
Toblli 2007	-41	7	20	-59	8	20	14.8%	2.35 [1.52, 3.17]
Total (95% CI)			**458**			**311**	**100.0%**	**0.52 [0.04, 1.00]**

Heterogeneity: $Tau^2 = 0.23$; $Chi^2 = 27.79$, $df = 4$ ($P < 0.0001$); $I^2 = 86\%$

Test for overall effect: $Z = 2.11$ ($P = 0.03$)

Plot 5. Forest plot of IV iron vs. placebo in patients with heart failure with reduced ejection fraction on health-related quality of life.

Step 1: *Identify the trial with the most weight, FAIR-HF in this case.*

Step 2: *Calculate average SD. Average SD in this case is ~20.50 (the average of 16.9115 and 24.0832).*

Step 3: *Multiply average SD by* ***SMD****. In this case this gives 10.66 (20.50 * 0.52).*

Step 4: *Contextualize this result in the scale used in the trial. In this case, that would be equal to a 10.66 out of 100 improvement at 6 months (per the scale used in FAIR-HF).*

Step 5: *Compare this value to the* ***minimally important difference (MID)*** *if known. In the case of the FAIR-HF* *scale the **MID** was 5. Therefore the mean effect was greater than the **MID**.*

Ideally the review will present this information along with a comparison of the proportion of participants within each group who experienced clinically important improvement or decline (the so-called “responder analysis”). This is useful because calculating the mean alone will not provide any information about the distribution. The responder analysis unfortunately cannot typically be calculated by readers if it is not already reported by the reviewers.

Statistical Significance Is Not Everything

While this section has focused on statistical fundamentals it is important to emphasize that statistics are only one aspect of critical appraisal. Even if a study shows statistical significance, there needs to be considerations of **bias**, clinical significance, and **generalizability**.

Bias: P-values and CIs are contingent on all the assumptions being used to calculate them being correct. In other words, they assume there is absolutely no **bias** present. As such, if the study is poorly conducted (e.g. a **RCT** without adequate **allocation concealment** and blinding) this will not be reflected in the statistical analysis of the results. This is why it is necessary to appraise the conduct of the trial to evaluate the credibility of the results.

Clinical significance: Even if a result is statistically significant, it may be too small of an effect to matter to a patient. For instance, with enough participants, a 1-point reduction in pain on a 100-point scale could be statistically significant, but very unlikely to be felt by an individual patient.

Generalizability: Even if the results are unbiased and clinically significant, they will only be useful if they can be applied to practice. If there are substantial differences between the features of the trial and your own practice, then the result may not be applicable.

Other sections of this resource will go into these concepts in more depth, but these are the fundamental reasons why a comprehensive approach to appraisal is necessary, and simply looking at statistical significance in the results section will not be sufficient to understand the clinical implications of a trial.

Glossary

Absolute risk difference (a.k.a. absolute risk increase or decrease)

Absolute risk difference is the risk in one group compared to (minus) the risk in another group over a specified period of time. For example, if the absolute risk of myocardial infarction over 5 years was 15% for the comparator and 10% for the intervention, then the absolute risk difference was 5% (15% - 10%) over 5 years. See here for further discussion.

Allocation concealment

Refers to the process that prevents patients, clinicians, and researchers from predicting which intervention group the patient will be assigned. This is different from blinding; allocation concealment refers to patients/clinicians/outcome assessors/etc. being unaware of group allocation prior to randomization, whereas blinding refers to remaining unaware of group allocation after randomization. Allocation concealment is a necessary condition for blinding. It is always feasible to implement.

Bias

Systematic deviation of an estimate from the truth (either an overestimation or underestimation) caused by a study design or conduct feature. See the Catalog of Bias for specific biases, explanations, and examples.

Composite outcome

An outcome which consists of multiple component endpoints. For example, a cardiovascular composite may include stroke, myocardial infarction, and death.

Confidence interval (CI)

See here for a discussion of confidence intervals.

Confounders

See here for discussion regarding confounders.

Crossover bias

Occurs when participants receive treatment intended for the other study group (a phenomenon known as contamination). For example, a participant assigned to the placebo group may end up taking active treatment. This bias results in underestimating the difference between groups.

Double-blinding

Double-blinding does not have a standardized definition and, consequently, further examinations are needed to ascertain exactly who was blinded (Lang TA et al).

Enrichment strategies

A trial strategy to identify populations where the intervention will show the greatest effect. There is no singular method. One method is to enroll subjects and put them all on active treatment, then randomize only those who responded to treatment to either continue active treatment or switch to placebo (withdrawal trial). Another method is to include risk factors for the outcome of interest in the study as inclusion criteria (enrichment criteria) (e.g. recent diabetes trials assessing cardiovascular outcomes have selectively enrolled patients with established atherosclerotic cardiovascular disease (ASCVD) or multiple additional ASCVD risk factors to be included).

External validity

Refers to the extent to which the trial results are applicable beyond the patients included in the study. Also known as generalizability.

Fixed-effects model

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.

Generalizability

Refers to the extent to which the trial results are applicable beyond the patients included in the study. Also known as external validity.

Hazard ratio (HR)

Hazard ratios are a relative measure of effect. Hazards refer to average instantaneous incidence rate at every point during the trial. This differentiates it from other measures, such as relative risk, which rely only on cumulative event rates. See here for a more detailed discussion.

Heterogeneity

Refers to variability between studies in a systematic review. It can refer to clinical differences, methodological differences, or variable results between studies. Heterogeneity occurs on a continuum and, in the case of heterogeneity amongst results, can be expressed numerically via measures of statistical heterogeneity. See here for a further discussion of statistical heterogeneity.

Intention-to-treat (ITT)

Participant outcomes are analyzed according to their assigned treatment group, irrespective of treatment received. A common "modified ITT" approach used in pharmacotherapy trials considers only participants who received at least one dose of the study drug (thereby excluding participants who were randomized but did not receive any study intervention).

Internal validity

The extent to which the study results are attributable to the intervention and not to bias. If internal validity is high, there is high confidence that the results are due to the effects of treatment (with low internal validity entailing low confidence).

Last observation carried forward (LOCF)

A method of evaluating patients who have dropped out partway through a trial when performing an intention-to-treat analysis. It treats the patients as if they were still in the trial and their outcome status remained the same as when they were last observed. For example, a patient who reported a pain score of 7/10 at day 3 and dropped out prior to the 1-week follow-up would be analyzed as having 7/10 pain at the end of 1 week (despite no outcome data being recorded past day 3).

Loss to follow-up (LTFU)

Loss to follow-up may occur when participants stop coming to study follow-up visits, do not answer follow-up phone calls, and cannot otherwise be assessed for study outcomes. This leads to missing data from the time they became "lost". Underlying reasons may include leaving the trial without informing investigators, moving to a new location, debilitation due to illness, or death.

Meta-analysis

A meta-analysis is a quantitative combination of the data obtained in a systematic review.

Minimally important difference

The minimum difference in a value that would be of importance to a patient. There are various methods of calculating a minimally important difference.

Null hypothesis

In superiority analyses, this is the hypothesis that there is no difference in the outcome of interest between the intervention group and the comparator group. In non-inferiority analyses, this is the hypothesis that there is a difference in the outcomes of interest between the treatment group and the control group.

Odds ratio (OR)

Odds ratios are the ratio of odds (events divided by non-events) in the intervention group to the odds in the comparator group. For example, if the odds of an event in the treatment group is 0.2 and the odds in the comparator group is 0.1, then the OR is 2 (0.2/0.1). See here for a more detailed discussion.

P-value

See here for a p-value discussion.

Per-protocol analysis

This type of analysis examines patients only if they sufficiently adhered to the treatment group in which they were assigned.

PICO

An acronym for "patient, intervention, comparator, and outcome". These are the four basic elements of a study. For instance, a study may examine an elderly population (P) to understand the effects of statin therapy (I) compared to placebo (C) in terms of cardiovascular events (O). Sometimes extended to PICO(T) to include the time at which outcomes were assessed, or (D)PICO to incorporate the study design.

Placebo

An inert intervention, such as a sugar pill, that does not have a physiological mechanism of influencing any of the outcomes of interest. Typically given to the comparator group in an effort to blind participants and clinicians.

Point estimate

A single value given as an estimate of the effect. For example, results may be listed as a relative risk of 0.5 (95% CI 0.4-0.6). In this case 0.5 is the point estimate, and 0.4-0.6 is the 95% confidence interval.

Primary care

This is the most accessible healthcare setting where generalist services are provided. For example, a family medicine clinic.

Primary outcome

A primary outcome is an outcome from which trial design choices are based (e.g. sample size calculations). Primary outcomes are not necessarily the most important outcomes.

Progression-free survival (PFS)

A measure of time to disease progression or death. This outcome is frequently used in cancer trials where disease progression is typically defined as an increase in radiographic tumor mass above a certain threshold.

Publication bias

Refers to a systematic tendency for results to be published based upon the direction or statistical significance of the results. This results in bias when aggregating evidence if methods are more likely to include published literature than unpublished literature.

Random-effects model

The random-effects model does not assume that all trials estimate the exact same underlying effect (e.g. different populations may vary in their response to intervention).

Randomized controlled trial (RCT)

Randomized controlled trials are those in which participants are randomly allocated to two or more groups which are given different treatments.

Relative effect

Calculates the effect of an intervention via a fractional comparison with the comparator group (i.e. intervention group measure ÷ comparator group measure). Used for binary outcomes. Relative risk, odds ratio, or hazards ratio are all expressions of relative effect. For example, if the risk of developing neuropathy was 1% in the treatment group and 2% in the comparator group, then the relative risk is 0.5 (1 ÷ 2). See the Absolute Risk Differences and Relative Measures of Effect discussion here for more information.

Relative risk or risk ratio (RR)

Relative risk (or risk ratio) is the risk in one group relative to (divided by) risk in another group. For example, if 10% in the treatment group and 20% in the placebo group have the outcome of interest, the relative risk in the treatment group is 0.5 (10% ÷ 20%; half) the risk in the placebo group. See here for a more detailed discussion.

Relative risk reduction (RRR)

The difference between two relative risks (RRs). If the intervention has a RR of 70% and the comparator a risk of 100%, then the relative risk reduction is 30% (100% - 70%).

Run-in period

A pre-randomization trial phase where all patients are assigned to active treatment, placebo, or no treatment (observation only). A run-in phase may be implemented for several reasons, including to restrict randomization only to patients who can adhere to study follow-up or treatment, or to exclude patients who cannot tolerate the intervention. Run-in periods by design select a certain subgroup of patients for enrolment, which introduces selection bias (i.e. potential issues with generalizability), which may be important in some cases. Note that this selection bias occurs prior to randomization, and therefore does not introduce differences between randomized groups (i.e. allocation bias).

Secondary care

Healthcare services provided via specialists in settings less advanced than tertiary care. For example, an outpatient cardiology clinic.

Secondary outcome

A secondary outcome is any outcome that is not a primary outcome (i.e. secondary outcomes are not the focal point of design choices like sample size). Secondary outcomes may be more clinically important than the primary outcome.

Sensitivity analyses

This type of analysis explores to what degree the results are dependent upon certain decisions and assumptions. It can be thought of as a "stress test" of study assumptions. For example, a meta-analysis including trials performed across a date range of 1960 to 2020 may perform a sensitivity analysis to explore if the estimated effect size differed across decades.

Sequence generation

The process by which allocation of participants to groups is conducted. Computer generation and coin tosses are examples of methods of random sequence generation.

Serious adverse events

Standardized definition encompassing any adverse event that:
(1) Results in death or is life-threatening;
(2) Requires or prolongs hospitalization;
(3) Results in persistent, significant, or permanent disability or incapacity;
(4) Causes congenital malformation;
(5) Per the clinician's judgement led to an important medical event.

Small-study effects

A tendency for smaller published studies to demonstrate a larger effect size than larger published studies. One possible cause is publication bias. However, other possible causes include systematic differences between smaller and larger studies (e.g. stricter enrolment criteria, adherence and/or follow-up in smaller studies, more pragmatic design in larger studies).

Standardized mean difference (SMD)

Transformation of continuous data that consists of dividing the difference in means between two groups by the standard deviation of the variable. In clinical research, this is often used to summarize and/or pool continuous outcomes that are measured in several ways. For example, a meta-analysis of antidepressants may need to use the SMD if trials used different scales (e.g. Beck Depression Inventory, Hamilton Depression Rating Scale) to report change in depression symptoms. See here for further discussion on SMD interpretation.

Stratified randomization

A multistage approach to randomization in which participants are initially allocated to strata based on certain defined commonalities (e.g. stratified according to LDL levels). After stratification these participants are then randomized within their respective stratum.

Superiority trial

A superiority trial tests for whether an intervention has a greater effect than a comparator with respect to the primary outcome. This is contrasts with non-inferiority trials.

Surrogate markers or outcomes

These markers or outcomes act as proxies for clinical outcomes under the assumption that the proxy is sufficiently predictive of the clinical outcome. For example, LDL cholesterol lowering may be used as a surrogate marker for lowering the risk of cardiovascular events. Surrogate markers are typically used because they are more convenient to measure.

Systematic review

A review that systematically identifies all potentially relevant studies on a research question. The aggregate of studies is then evaluated with respect to factors such as risk of bias of individual studies or heterogeneity among results. The qualitative combination of results is a systematic review.

Tertiary care

Care provided in a specialized institutional centre. For example, neurosurgery or severe burn treatment.

References

1. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015 Nov 3;16:495.
2. Kim B-K, Hong M-K, Shin D-H, Nam C-M, Kim J-S, Ko Y-G, et al. A new strategy for discontinuation of dual antiplatelet therapy: the RESET Trial (REal Safety and Efficacy of 3-month dual antiplatelet Therapy following Endeavor zotarolimus-eluting stent implantation). J Am Coll Cardiol. 2012 Oct 9;60(15):1340–8.
3. CAPRIE Steering Committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). CAPRIE Steering Committee. Lancet. 1996 Nov 16;348(9038):1329–39.
4. Ridker PM, Cook NR, Lee I-M, Gordon D, Gaziano JM, Manson JE, et al. A randomized trial of low-dose aspirin in the primary prevention of cardiovascular disease in women. N Engl J Med. 2005 Mar 31;352(13):1293–304.
5. Holcomb WL, Chaiworapongsa T, Luke DA, Burgdorf KD. An odd measure of risk: use and misuse of the odds ratio. Obstet Gynecol. 2001 Oct;98(4):685–8.
6. McMurray JJV, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR, et al. Angiotensin-neprilysin inhibition versus enalapril in heart failure. N Engl J Med. 2014 Sep 11;371(11):993–1004.
7. Ortiz-Orendain J, Castiello-de Obeso S, Colunga-Lozano LE, Hu Y, Maayan N, Adams CE. Antipsychotic combinations for schizophrenia. Cochrane Database Syst Rev. 2017 Jun 28;6:CD009005.

8. Antithrombotic Trialists' (ATT) Collaboration, Baigent C, Blackwell L, Collins R, Emberson J, Godwin J, et al. Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. Lancet. 2009 May 30;373(9678):1849–60.
9. Turgeon RD, Koshman SL, Youngson E, Har B, Wilton SB, James MT, et al. Association of Ticagrelor vs Clopidogrel With Major Adverse Coronary Events in Patients With Acute Coronary Syndrome Undergoing Percutaneous Coronary Intervention. JAMA Intern Med. 2020 Mar 1;180(3):420–8.
10. Wiysonge CS, Bradley HA, Volmink J, Mayosi BM, Opie LH. Beta-blockers for hypertension. Cochrane Database Syst Rev. 2017 Jan 20;1:CD002003.
11. Oremus M, Don-Wauchope A, McKelvie R, Santaguida PL, Hill S, Balion C, et al. BNP and NT-proBNP as prognostic markers in persons with chronic stable heart failure. Heart Fail Rev. 2014 Aug;19(4):471–505.
12. Chan FKL, Lanas A, Scheiman J, Berger MF, Nguyen H, Goldstein JL. Celecoxib versus omeprazole and diclofenac in patients with osteoarthritis and rheumatoid arthritis (CONDOR): a randomised trial. Lancet. 2010 Jul 17;376(9736):173–9.
13. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions [Internet]. 2020 [cited 2021 Aug 7]. Available from: https://doi.org/10.1002/9781119536604
14. Nidorf SM, Fiolet ATL, Mosterd A, Eikelboom JW, Schut A, Opstal TSJ, et al. Colchicine in Patients with Chronic Coronary Disease. N Engl J Med. 2020 Nov 5;383(19):1838–47.
15. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ. 2012 Mar 15;344:e1553.
16. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J,

Oldgren J, Parekh A, et al. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med. 2009 Sep 17;361(12):1139–51.
17. McMurray JJV, Solomon SD, Inzucchi SE, Køber L, Kosiborod MN, Martinez FA, et al. Dapagliflozin in Patients with Heart Failure and Reduced Ejection Fraction. N Engl J Med. 2019 Nov 21;381(21):1995–2008.
18. Turgeon RD, Reid EK, Rainkie DC. Design and Interpretation of Noninferiority Trials. J Gen Intern Med. 2018 Aug;33(8):1215.
19. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ. 2020 Aug 10;192(32):E901–6.
20. Molnar FJ, Hutton B, Fergusson D. Does analysis using “last observation carried forward” introduce bias in dementia research? CMAJ. 2008 Oct 7;179(8):751–3.
21. McAuley L, Pham B, Tugwell P, Moher D. Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses? Lancet. 2000 Oct 7;356(9237):1228–31.
22. Siemieniuk RA, Bartoszko JJ, Ge L, Zeraatkar D, Izcovich A, Kum E, et al. Drug treatments for covid-19: living systematic review and network meta-analysis. BMJ. 2020 Jul 30;370:m2980.
23. Steinhubl SR, Berger PB, Mann JT, Fry ETA, DeLago A, Wilmer C, et al. Early and sustained dual oral antiplatelet therapy following percutaneous coronary intervention: a randomized controlled trial. JAMA. 2002 Nov 20;288(19):2411–20.
24. Kotecha D, Bunting KV, Gill SK, Mehta S, Stanbury M, Jones JC, et al. Effect of Digoxin vs Bisoprolol for Heart Rate Control in Atrial Fibrillation on Patient-Reported Quality of

Life: The RATE-AF Randomized Clinical Trial. JAMA. 2020 Dec 22;324(24):2497–508.
25. Pitt B, Poole-Wilson PA, Segal R, Martinez FA, Dickstein K, Camm AJ, et al. Effect of losartan compared with captopril on mortality in patients with symptomatic heart failure: randomised trial–the Losartan Heart Failure Survival Study ELITE II. Lancet. 2000 May 6;355(9215):1582–7.
26. Hart B, Lundh A, Bero L. Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. BMJ. 2012 Jan 3;344:d7202.
27. Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. N Engl J Med. 2007 Jun 14;356(24):2457–71.
28. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA. 1998 Jan 28;279(4):281–6.
29. Wilt TJ, Bloomfield HE, MacDonald R, Nelson D, Rutks I, Ho M, et al. Effectiveness of statin therapy in adults with coronary heart disease. Arch Intern Med. 2004 Jul 12;164(13):1427–36.
30. Heart Outcomes Prevention Evaluation Study Investigators, Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, et al. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. N Engl J Med. 2000 Jan 20;342(3):145–53.
31. Keech A, Simes RJ, Barter P, Best J, Scott R, Taskinen MR, et al. Effects of long-term fenofibrate therapy on cardiovascular events in 9795 people with type 2 diabetes mellitus (the FIELD study): randomised controlled trial. Lancet. 2005 Nov 26;366(9500):1849–61.
32. Knopp RH, d’Emden M, Smilde JG, Pocock SJ. Efficacy and safety of atorvastatin in the prevention of cardiovascular end points in subjects with type 2 diabetes: the Atorvastatin Study for Prevention of Coronary Heart Disease Endpoints in

non-insulin-dependent diabetes mellitus (ASPEN). Diabetes Care. 2006 Jul;29(7):1478–85.
33. Cholesterol Treatment Trialists' (CTT) Collaborators, Kearney PM, Blackwell L, Collins R, Keech A, Simes J, et al. Efficacy of cholesterol-lowering therapy in 18,686 people with diabetes in 14 randomised trials of statins: a meta-analysis. Lancet. 2008 Jan 12;371(9607):117–25.
34. Tonelli M, Lloyd A, Clement F, Conly J, Husereau D, Hemmelgarn B, et al. Efficacy of statins for primary prevention in people at low cardiovascular risk: a meta-analysis. CMAJ. 2011 Nov 8;183(16):E1189-1202.
35. Pereira TV, Horwitz RI, Ioannidis JPA. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012 Oct 24;308(16):1676–84.
36. Chan A-W, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA. 2004 May 26;291(20):2457–65.
37. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008 Mar 15;336(7644):601–5.
38. Westreich D, Iliinsky N. Epidemiology visualized: the prosecutor's fallacy. Am J Epidemiol. 2014 May 1;179(9):1125–7.
39. Pitt B, Remme W, Zannad F, Neaton J, Martinez F, Roniker B, et al. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. N Engl J Med. 2003 Apr 3;348(14):1309–21.
40. Kovic B, Jin X, Kennedy SA, Hylands M, Pedziwiatr M, Kuriyama A, et al. Evaluating Progression-Free Survival as a Surrogate Outcome for Health-Related Quality of Life in Oncology: A Systematic Review and Quantitative Analysis.

JAMA Intern Med. 2018 Dec 1;178(12):1586–96.
41. Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JPA. Evaluation of Evidence of Statistical Support and Corroboration of Subgroup Claims in Randomized Clinical Trials. JAMA Intern Med. 2017 Apr 1;177(4):554–60.
42. Zipkin DA, Umscheid CA, Keating NL, Allen E, Aung K, Beyth R, et al. Evidence-based risk communication: a systematic review. Ann Intern Med. 2014 Aug 19;161(4):270–80.
43. Cooney GM, Dwan K, Greig CA, Lawlor DA, Rimer J, Waugh FR, et al. Exercise for depression. Cochrane Database Syst Rev. 2013 Sep 12;(9):CD004366.
44. Spertus JA, Tooley J, Jones P, Poston C, Mahoney E, Deedwania P, et al. Expanding the outcomes in clinical trials of heart failure: the quality of life and economic components of EPHESUS (EPlerenone's neuroHormonal Efficacy and SUrvival Study). Am Heart J. 2002 Apr;143(4):636–42.
45. El-Khalili N, Joyce M, Atkinson S, Buynak RJ, Datto C, Lindgren P, et al. Extended-release quetiapine fumarate (quetiapine XR) as adjunctive therapy in major depressive disorder (MDD) in patients with an inadequate response to ongoing antidepressant treatment: a multicentre, randomized, double-blind, placebo-controlled study. Int J Neuropsychopharmacol. 2010 Aug;13(7):917–32.
46. Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. N Engl J Med. 1989 Jul 20;321(3):129–35.
47. Nguyen-Khac E, Thevenot T, Piquet M-A, Benferhat S, Goria O, Chatelain D, et al. Glucocorticoids plus N-acetylcysteine in severe alcoholic hepatitis. N Engl J Med. 2011 Nov 10;365(19):1781–9.
48. Molnar FJ, Man-Son-Hing M, Hutton B, Fergusson DA.

Have last-observation-carried-forward analyses caused us to favour more toxic dementia therapies over less toxic alternatives? A systematic review. Open Med. 2009 Mar 24;3(2):e31-50.
49. McCormack J, Vandermeer B, Allan GM. How confidence intervals become confusion intervals. BMC Med Res Methodol. 2013 Oct 31;13:134.
50. Mulla SM, Scott IA, Jackevicius CA, You JJ, Guyatt GH. How to use a noninferiority trial: users' guides to the medical literature. JAMA. 2012 Dec 26;308(24):2605–11.
51. DiNicolantonio JJ, Tomek A. Inactivations, deletions, non-adjudications, and downgrades of clinical endpoints on ticagrelor: serious concerns over the reliability of the PLATO trial. Int J Cardiol. 2013 Oct 9;168(4):4076–80.
52. Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012 Sep 18;157(6):429–38.
53. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ. 2010 Mar 30;340:c117.
54. Osmancik P, Herman D, Neuzil P, Hala P, Taborsky M, Kala P, et al. Left Atrial Appendage Closure Versus Direct Oral Anticoagulants in High-Risk Patients With Atrial Fibrillation. J Am Coll Cardiol. 2020 Jun 30;75(25):3122–35.
55. Khan MS, Lateef N, Siddiqi TJ, Rehman KA, Alnaimat S, Khan SU, et al. Level and Prevalence of Spin in Published Cardiovascular Randomized Clinical Trial Reports With Statistically Nonsignificant Primary Outcomes: A Systematic Review. JAMA Netw Open. 2019 May 3;2(5):e192622.
56. Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. Eur Heart J. 2019 May 1;40(17):1378–83.
57. Royle P, Milne R. Literature searching for randomized

controlled trials used in Cochrane reviews: rapid versus exhaustive searches. Int J Technol Assess Health Care. 2003;19(4):591–603.
58. Lewis G, Marston L, Duffy L, Freemantle N, Gilbody S, Hunter R, et al. Maintenance or Discontinuation of Antidepressants in Primary Care. N Engl J Med. 2021 Sep 30;385(14):1257–67.
59. Stroup DF, Berlin JA, Morton SC, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA. 2000 Apr 19;283(15):2008–12.
60. Schulz KF, Grimes DA. Multiplicity in randomised trials II: subgroup and interim analyses. Lancet. 2005 May 7;365(9471):1657–61.
61. Siu JT, Nguyen T, Turgeon RD. N-acetylcysteine for non-paracetamol (acetaminophen)-related acute liver failure. Cochrane Database Syst Rev. 2020 Dec 9;12:CD012123.
62. Jones CW, Handler L, Crowell KE, Keil LG, Weaver MA, Platts-Mills TF. Non-publication of large randomized clinical trials: cross sectional analysis. BMJ. 2013 Oct 29;347:f6104.
63. Hong J, Tung A, Kinkade A, Tejani AM. Noninferiority drug trials fail to report adequate methodological detail: an assessment of noninferiority trials from 2010 to 2015. J Clin Epidemiol. 2019 Apr;108:144–6.
64. Hróbjartsson A, Thomsen ASS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ. 2013 Mar 5;185(4):E201-211.
65. Roddy E, Clarkson K, Blagojevic-Bucknall M, Mehta R, Oppong R, Avery A, et al. Open-label randomised pragmatic trial (CONTACT) comparing naproxen and low-dose colchicine for the treatment of gout flares in primary care. Ann Rheum Dis.

2020 Feb;79(2):276–84.
66. Chan A-W, Krleza-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. CMAJ. 2004 Sep 28;171(7):735–40.
67. Yeh RW, Valsdottir LR, Yeh MW, Shen C, Kramer DB, Strom JB, et al. Parachute use to prevent death and major trauma when jumping from aircraft: randomized controlled trial. BMJ. 2018 Dec 13;363:k5094.
68. Koshman SL, Charrois TL, Simpson SH, McAlister FA, Tsuyuki RT. Pharmacist care of patients with heart failure: a systematic review of randomized trials. Arch Intern Med. 2008 Apr 14;168(7):687–94.
69. Turgeon RD, Barry AR, Hawkins NM, Ellis UM. Pharmacotherapy for heart failure with reduced ejection fraction and health-related quality of life: a systematic review and meta-analysis. Eur J Heart Fail. 2021 Apr;23(4):578–89.
70. Colhoun HM, Betteridge DJ, Durrington PN, Hitman GA, Neil HAW, Livingstone SJ, et al. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicentre randomised placebo-controlled trial. Lancet. 2004 Aug 21;364(9435):685–96.
71. Price R, Bethune R, Massey L. Problem with p values: why p values do not tell you if your treatment is likely to work. Postgrad Med J. 2020 Jan;96(1131):1–3.
72. Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009 Jan 21;(1):MR000006.
73. Pitt B, Segal R, Martinez FA, Meurers G, Cowley AJ, Thomas I, et al. Randomised trial of losartan versus captopril in patients over 65 with heart failure (Evaluation of Losartan in the Elderly Study, ELITE). Lancet. 1997 Mar

15;349(9054):747–52.
74. Heart Protection Study Collaborative Group. Randomized trial of the effects of cholesterol-lowering with simvastatin on peripheral vascular and other major vascular outcomes in 20,536 people with peripheral arterial disease and other high-risk conditions. J Vasc Surg. 2007 Apr;45(4):645–54; discussion 653-654.
75. Montori VM, Devereaux PJ, Adhikari NKJ, Burns KEA, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA. 2005 Nov 2;294(17):2203–9.
76. Sutradhar R, Austin PC. Relative rates not relative risks: addressing a widespread misinterpretation of hazard ratios. Ann Epidemiol. 2018 Jan;28(1):54–7.
77. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med. 2001 Dec 4;135(11):982–9.
78. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med. 2011 Sep 8;365(10):883–91.
79. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019 Aug 28;366:l4898.
80. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016 Oct 12;355:i4919.
81. Home PD, Pocock SJ, Beck-Nielsen H, Curtis PS, Gomis R, Hanefeld M, et al. Rosiglitazone evaluated for cardiovascular outcomes in oral agent combination therapy for type 2 diabetes (RECORD): a multicentre, randomised, open-label trial. Lancet. 2009 Jun 20;373(9681):2125–35.
82. Nissen SE, Wolski K. Rosiglitazone revisited: an updated

meta-analysis of risk for myocardial infarction and cardiovascular mortality. Arch Intern Med. 2010 Jul 26;170(14):1191–201.
83. Ridker PM, Danielson E, Fonseca FAH, Genest J, Gotto AM, Kastelein JJP, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. N Engl J Med. 2008 Nov 20;359(21):2195–207.
84. Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med. 2008 Jan 17;358(3):252–60.
85. Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA. 2010 Mar 24;303(12):1180–7.
86. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess. 2001;5(33):1–56.
87. Fagerland MW. t-tests, non-parametric tests, and large studies–a paradox of statistical practice? BMC Med Res Methodol. 2012 Jun 14;12:78.
88. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. BMJ. 2011 Oct 18;343:d5928.
89. Hernán MA. The hazards of hazard ratios. Epidemiology. 2010 Jan;21(1):13–5.
90. Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38. UK Prospective Diabetes Study Group. BMJ. 1998 Sep 12;317(7160):703–13.
91. Guyatt G, Rennie D, Meade M, Cook D, American Medical Association, editors. Users’ guides to the medical literature. A

manual for evidence-based clinical practice. Third edition. New York: McGraw-Hill Education Medical; 2015. (JAMAevidence).
92. Pocock SJ. When (not) to stop a clinical trial for benefit. JAMA. 2005 Nov 2;294(17):2228–30.
93. Lang TA, Stroup DF. Who knew? The misleading specificity of “double-blind” and what to do about it. Trials. 2020 Aug 5;21(1):697.
94. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005 Aug;2(8):e124.
95. Albers GW, Diener H-C, Frison L, Grind M, Nevinson M, Partridge S, et al. Ximelagatran vs warfarin for stroke prevention in patients with nonvalvular atrial fibrillation: a randomized trial. JAMA. 2005 Feb 9;293(6):690–8.