Grading of Recommendations Assessment, Development, and Evaluation

Farid Najafi
MD, PhD
School of Public Health
Kermanshah University of Medical Sciences
2023

What is GRADE?

- An approach to rating quality of evidence and grading strength of recommendations
- An approach for rating quality of evidence in **systematic reviews** and **guidelines** and **grading strength of recommendations in guidelines**
- It offers a transparent and structured process for developing and presenting evidence summaries for systematic reviews and guidelines
- GRADE specifies an approach to framing questions, choosing outcomes of interest and rating their importance, evaluating the evidence, and incorporating evidence with considerations of values and preferences of patients and society to arrive at recommendations
- GRADE suggests somewhat **different approaches** for rating the quality of evidence for systematic reviews and for guidelines
- **Clear distinction between quality of evidence and strength of recommendation**

The GRADE process—defining the question and collecting evidence

Common to systematic review and guidelines

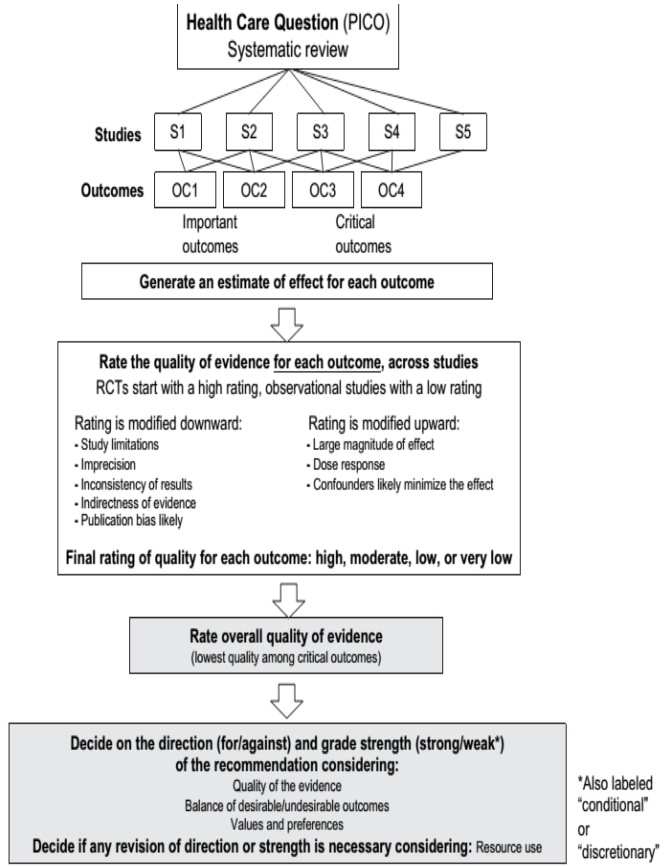


Fig. 1. Schematic view of GRADE's process for developing recommendations. *Abbreviation*: RCT, randomized controlled trials.

Specific for guideline development

PCI compared to CABG for Non-ST elevation ACS with multivessel or left main
[1,a]

Certainty assessment							Summary of findings					
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	№ of patients		Effect		Certainty	Importance
							PCI	CABG	Relative (95% CI)	Absolute (95% CI)		
MACE at 30 days follow-up (assessed with: Farid Najafi-Fatemeh Rajati)												
5	observational studies	serious [b]	serious [b,c]	not serious	serious [c]	publication bias strongly suspected all plausible residual confounding would suggest spurious effect, while no effect was observed [d]	0 cases 0 controls		**RR 0.96**	-	⨁◯◯◯ Very low	NOT IMPORTANT
							-	0.0%	(0.38 to 2.39)	**0 fewer per 1,000** (from 0 fewer to 0 fewer)		
All cause mortality after 30 days (assessed with: Farid Najafi-Fatemeh Rajati)												
5	observational studies	serious [e]	not serious	not serious	serious [c]	publication bias strongly suspected all plausible residual confounding would suggest spurious effect, while no effect was observed [d]	0 cases 0 controls		**RR 0.95**	-	⨁◯◯◯ Very low	
							-	0.0%	(0.54 to 1.76)	**0 fewer per 1,000** (from 0 fewer to 0 fewer)		
MI after 30 days (PCI vs. CABG (assessed with: Farid Najafi-Fatemeh Rajati)												
4	observational studies	serious [e]	serious [b]	not serious	serious [c]	publication bias strongly suspected	0 cases 0 controls		**RR 1.72**	-	⨁◯◯◯ Very low	

Quality of evidence

"I figure there's a 40% chance of showers, and a 10% chance we know what we're talking about."

Expert clinicians and organisations offering recommendations to the clinical community have often erred as a result of not taking sufficient account of the quality of evidence.[2] For a decade, organisations recommended that clinicians encourage postmenopausal women to use hormone replacement therapy.[3] Many primary care physicians dutifully applied this advice in their practices.

A belief that such therapy substantially decreased women's cardiovascular risk drove this recommendation. Had a rigorous system of rating the quality of evidence been applied at the time, it would have shown that because the data came from observational studies with inconsistent results, the evidence for a reduction in cardiovascular risk was of very low quality.[4] Recognition of the limitations of the evidence would have tempered the recommendations. Ultimately, randomised controlled trials have shown that hormone replacement therapy fails to reduce cardiovascular risk and may even increase it.[5 6]

The GRADE process—rating evidence quality

The quality of evidence reflects the extent to which our confidence in an estimate of the effect is adequate to support a particular recommendation

Study Design	Quality of Evidence	Lower if	Higher if
Randomized trial →	High	Risk of bias -1 Serious -2 Very serious Inconsistency -1 Serious -2 Very serious Indirectness -1 Serious -2 Very serious Imprecision -1 Serious -2 Very serious Publication bias -1 Likely -2 Very likely	Large effect +1 Large +2 Very large Dose response +1 Evidence of a gradient All plausible confounding +1 Would reduce a demonstrated effect or +1 Would suggest a spurious effect when results show no effect
	Moderate		
Observational study →	Low		
	Very low		

Important point: GRADE is “outcome centric”: rating is made for each outcome, and quality may differ

Certainty of Evidence

Quality of Evidence Grades	
Definition	**Grade**
We are very confident that the true effect lies close to that of the estimate of the effect.	**High**
We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different	**Moderate**
Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.	**Low**
We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect	**Very Low**

Quality of evidence

Table 5.2: Factors that can reduce the quality of the evidence	
Factor	Consequence
Limitations in study design or execution (risk of bias)	↓ 1 or 2 levels
Inconsistency of results	↓ 1 or 2 levels
Indirectness of evidence	↓ 1 or 2 levels
Imprecision	↓ 1 or 2 levels
Publication bias	↓ 1 or 2 levels

Table 5.3: Factors that can increase the quality of the evidence	
Factor	Consequence
Large magnitude of effect	↑ 1 or 2 levels
All plausible confounding would reduce the demonstrated effect or increase the effect if no effect was observed	↑ 1 level
Dose-response gradient	↑ 1 level

Study design

- Randomized trials generally provide stronger evidence for management strategies than observational studies, and rigorous observational studies are stronger than uncontrolled case series.
- In the GRADE approach, randomized trials without limitations provide high-quality evidence, while observational studies without special strengths or limitations provide low-quality evidence.
- Non-randomized experimental trials without limitations also provide high-quality evidence but may be downgraded for design limitations.
- Case series and case reports are observational studies that only investigate patients exposed to the intervention and usually warrant downgrading to very low-quality evidence.
- Expert opinion is not a category of evidence but represents an interpretation of evidence based on the expert's experience and knowledge. It is essential to specify the type of evidence used as the basis for interpretation.

Risk of bias in RCT

Lack of allocation concealment	Those enrolling patients are aware of the group (or period in a crossover trial) to which the next enrolled patient will be allocated (a major problem in "pseudo" or "quasi" randomized trials with allocation by day of week, birth date, chart number, etc.).
Lack of blinding	Patient, caregivers, those recording outcomes, those adjudicating outcomes, or data analysts are aware of the arm to which patients are allocated (or the medication currently being received in a crossover trial).
Incomplete accounting of patients and outcome events	Loss to follow-up and failure to adhere to the intention-to-treat principle in superiority trials; or in noninferiority trials, loss to follow-up, and failure to conduct both analyses considering only those who adhered to treatment, and all patients for whom outcome data are available. The significance of particular rates of loss to follow-up, however, varies widely and is dependent on the relation between loss to follow-up and number of events. The higher the proportion lost to follow-up in relation to intervention and control group event rates, and differences between intervention and control groups, the greater the threat of bias.
Selective outcome reporting	Incomplete or absent reporting of some outcomes and not others on the basis of the results.
Other limitations	• Stopping trial early for benefit. Substantial overestimates are likely in trials with fewer than 500 events and that large overestimates are likely in trials with fewer than 200 events. Empirical evidence suggests that formal stopping rules do not reduce this bias. • Use of unvalidated outcome measures (e.g. patient-reported outcomes) • Carryover effects in crossover trial • Recruitment bias in cluster-randomized trials

Risk of bias in observational studies

Table 5.5: Study limitations in observational studies

	Explanation
Failure to develop and apply appropriate eligibility criteria (inclusion of control population)	• Under- or over-matching in case-control studies • Selection of exposed and unexposed in cohort studies from different populations
Flawed measurement of both exposure and outcome	• Differences in measurement of exposure (e.g. recall bias in case-control studies) • Differential surveillance for outcome in exposed and unexposed in cohort studies
Failure to adequately control confounding	• Failure of accurate measurement of all known prognostic factors • Failure to match for prognostic factors and/or adjustment in statistical analysis
Incomplete or inadequately short follow-up	Especially within prospective cohort studies, both groups should be followed for the same amount of time.

Some consideration for overall assessment in SRs and guidelines

- When evaluating the quality of evidence, it is not appropriate to simply average across studies. Rather, each study should be evaluated individually, with a focus on high-quality studies.
- The contribution of each study to the overall estimate of effect should be considered, with larger studies carrying more weight.
- When rating down for risk of bias, reviewers should be conservative and confident in their assessment.
- Consider bias in relation to other limitations. When faced with a close call between two quality issues, such as bias and precision, rate down at least one of them.
- In close-call situations, reviewers should make their reasoning clear and explicit.

Table 5.6: Guidance to assess study limitations (risk of bias) in Cochrane Reviews and corresponding GRADE assessment of quality of evidence

Risk of bias	Across studies	Interpretation	Considerations	GRADE assessment of study limitations
Low	Most information is from studies at low risk of bias.	Plausible bias unlikely to seriously alter the results.	No apparent limitations.	No serious limitations, do not downgrade
Unclear	Most information is from studies at low or unclear risk of bias.	Plausible bias that raises some doubt about the results.	Potential limitations are unlikely to lower confidence in the estimate of effect.	No serious limitations, do not downgrade
			Potential limitations are likely to lower confidence in the estimate of effect.	Serious limitations, downgrade one level.
High	The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of results.	Plausible bias that seriously weakens confidence in the results.	Crucial limitation for one criterion, or some limitations for multiple criteria, sufficient to lower confidence in the estimate of effect.	Serious limitations, downgrade one level
			Crucial limitation for one or more criteria sufficient to substantially lower confidence in the estimate of effect.	Very serious limitations, downgrade two levels

Inconsistency

- There are some criteria for deciding about the inconsistency:

1. Wide variance of point estimates across studies (note: direction of effect is not a criterion for inconsistency)
2. Minimal or no overlap of confidence intervals (CI), which suggests variation is more than what one would expect by chance alone
3. Statistical criteria, including tests of heterogeneity which test the null hypothesis that all studies have the same underlying magnitude of effect, have a low p-value ($p < 0.05$), indicating to reject the null hypothesis

Inconsistency

- The I^2 statistic measures the variation in point estimates due to differences among studies.
- A large I^2 value indicates significant differences among studies.
- The interpretation of what constitutes a large I^2 value is subjective, but a rule-of-thumb suggests that values below 40% are low, 30-60% are moderate, 50-90% are substantial, and 75-100% are considerable.

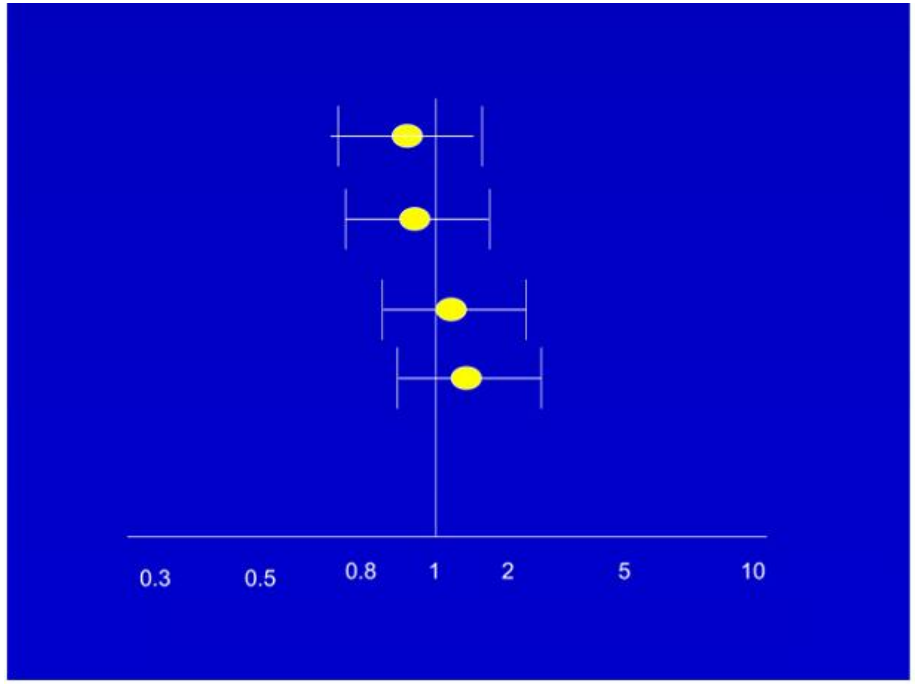


Fig. 1. Differences in direction, but minimal heterogeneity.

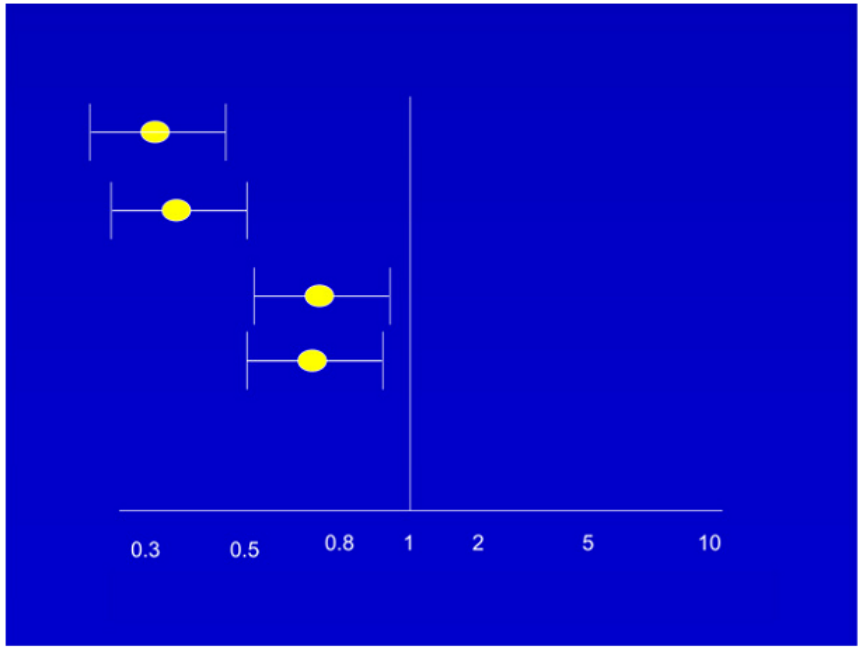


Fig. 2. Substantial heterogeneity, but of questionable importance.

If the effect size differs across studies, explanations for inconsistency may be due to differences in:
populations (e.g. drugs may have larger relative effects in sicker populations)
interventions (e.g. larger effects with higher drug doses)
outcomes (e.g. duration of follow-up)
study methods (e.g. RCTs with higher and lower risk of bias).

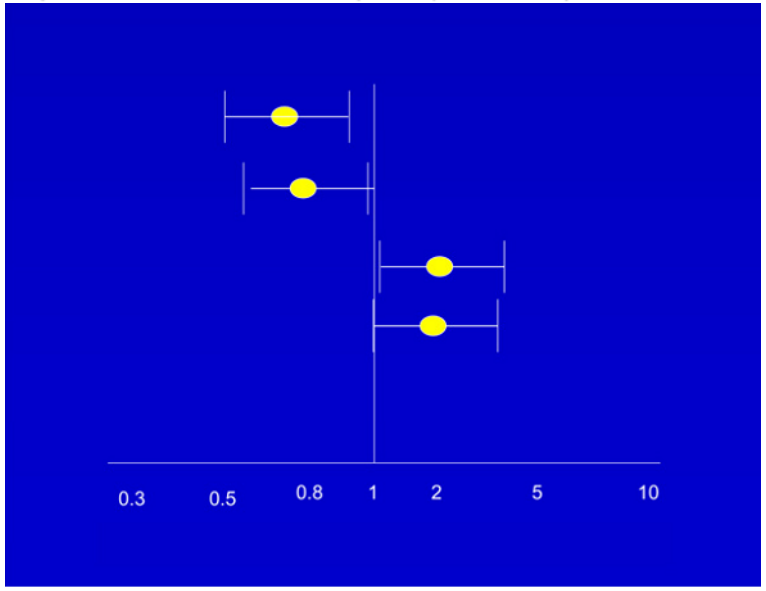


Fig. 3. Substantial heterogeneity, of unequivocal importance.

Indirectness

- Direct evidence, which involves comparing interventions and measuring patient outcomes of interest, increases our confidence in results. Systematic review authors and guideline panels should consider their uncertainty about evidence applicability and potentially downgrade quality ratings.
- There are four sources of indirectness:
    1. Differences in population
    2. Differences in intervention
    3. Differences in outcomes measures
    4. Indirect comparison

Imprecision in guidelines

- Guideline panels rate evidence quality based on the threshold for management decisions and weighing the pros and cons of outcomes.
- Assessing the CI boundaries in comparison to the guideline threshold and evaluating if the optimal information size criteria are fulfilled, can aid in determining if downgrading for imprecision is necessary.

Imprecision

not serious

serious

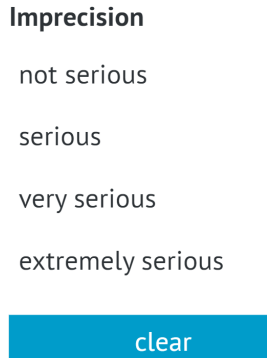
very serious

extremely serious

clear

Imprecision
Criteria for downgrading

- First consider whether the boundaries of the CI are on the same side of their decision-making threshold. **Does the CI cross the clinical decision threshold between recommending and not recommending treatment**? If the answer is **yes** (i.e. the CI crosses the threshold), **rate down** for imprecision irrespective of where the point estimate and CI lie.
- If the threshold is **not crossed**, are criteria for an **optimal information size** met? (*see note on OIS and Example 3*)
- **Or,**
- Is the event rate very low and the sample size very large (at least 2000, and perhaps 4000 patients)? (*see Exception note*)
- If **neither criterion is met**, **rate down** for imprecision.



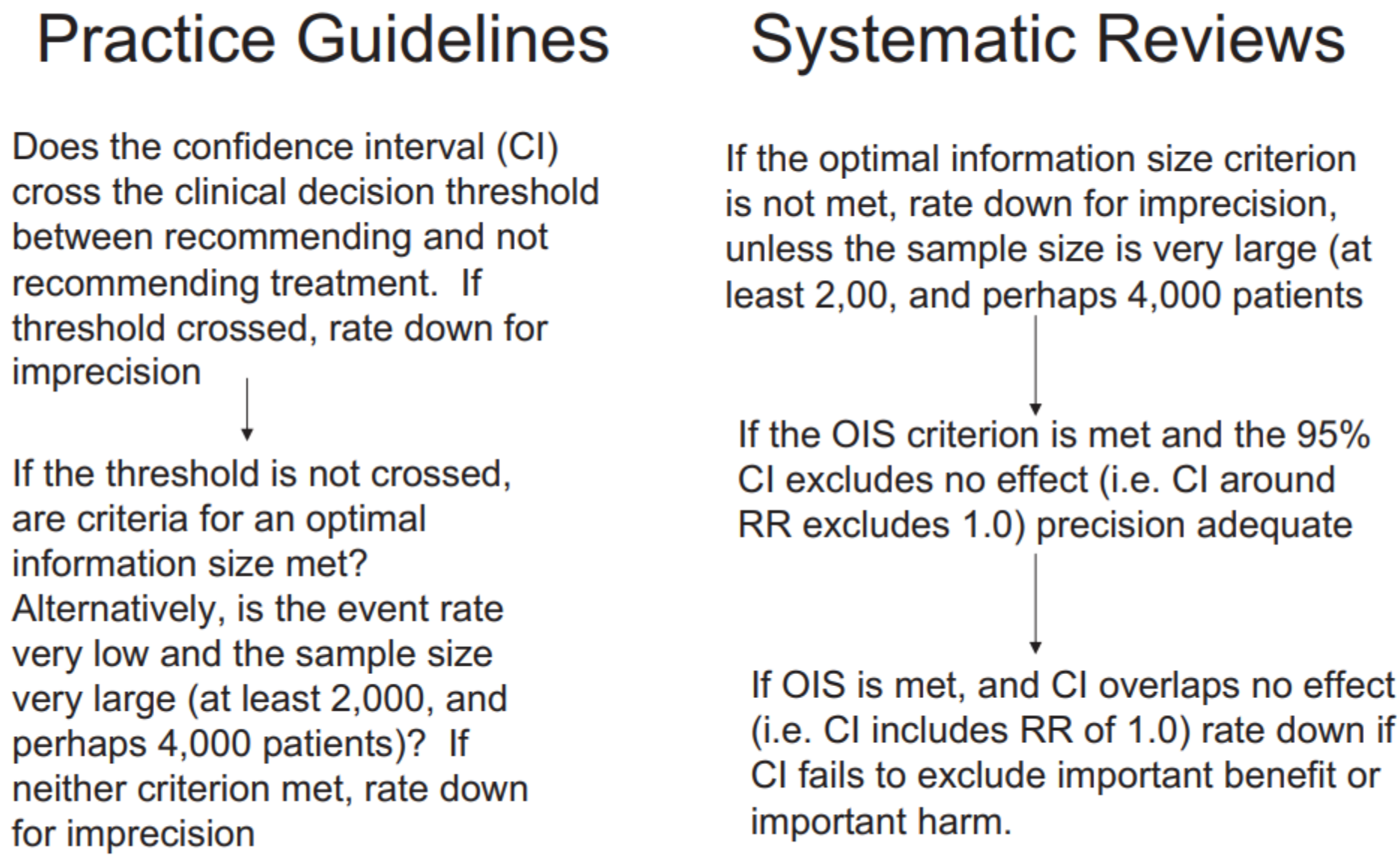


Fig. 3. Deciding whether to rate down for imprecision in guidelines and systematic reviews of binary variables.

Publication bias

- Publication bias is the selective publication of studies that leads to an under- or over-estimation of the true beneficial or harmful effect.

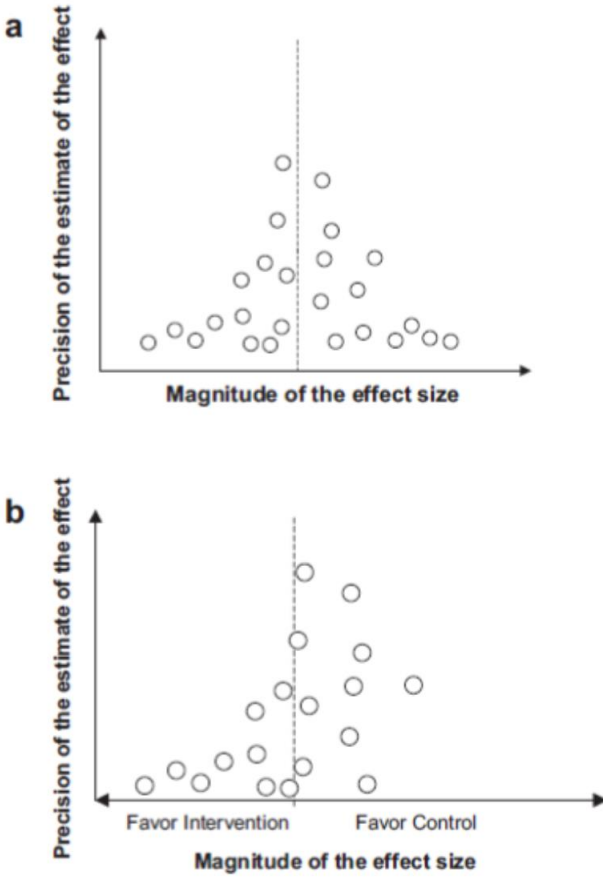
Table 5.8: Possible sources of publication bias throughout the publication process

Phases of research publication	Actions contributing to or resulting in bias.
Preliminary and pilot studies	Small studies more likely to be "negative" (e.g. those with discarded or failed hypotheses) remain unpublished; companies classify some as proprietary information.
Report completion	Authors decide that reporting a "negative" study is uninteresting; and do not invest the time and effort required for submission.
Journal selection	Authors decide to submit the "negative" report to a nonindexed, non-English, or limited-circulation journal.
Editorial consideration	Editor decides that the "negative" study does not warrant peer review and rejects manuscript.
Peer review	Peer reviewers conclude that the "negative" study does not contribute to the field and recommend rejecting the manuscript. Author gives up or moves to lower impact journal. Publication delayed.
Author revision and resubmission	Author of rejected manuscript decides to forgo the submission of the "negative" study or to submit it again at a later time to another journal (see "journal selection" above).
Report publication	Journal delays the publication of the "negative" study. Proprietary interests lead to report getting submitted to, and accepted by, different journals.

Detection of publication bias

- If there are at least 10 studies included in the meta-analysis (or even 5 studies), the following methods can be used:

1. Visual inspection
2. Test for asymmetry of funnel plot
3. Trim and fill method



Example 6: Publication bias detected

Rating up the quality of evidence

- The three primary reasons for rating up the quality of evidence are as follows:

1.When a large magnitude of effect exists,

2.When there is a dose–response gradient, and

3.When all plausible confounders or other biases increase our confidence in the estimated effect.

Large effect

- When observational studies provide strong evidence without any downgrades, we can have more confidence in the results.
- Even though observational studies tend to overestimate the true effect, the study design itself is unlikely to explain all of the apparent benefit or harm.
- When considering whether to rate up the quality of evidence based on large effects, we should not only look at the point estimate but also the precision (width of the confidence interval) around that effect. If the confidence interval overlaps with effects smaller than the chosen threshold of clinical importance, we should rarely and cautiously rate up the quality of evidence based on apparent large effects.

Definition of large effect

- It is suggested that confounding (from nonrandom allocation) alone is unlikely to explain associations with a relative risk (RR) greater than 2 (or less than 0.5), and very unlikely to explain associations with an RR greater than 5 (or less than 0.2)
- the GRADE group has previously suggested guidelines for rating quality of evidence up by one category (typically from low to moderate) for associations greater than 2, and up by two categorie for associations greater than 5

Some considerations for large effect

- All thresholds for large effect are about risk ratio and need to be adjusted for OR
- Rapidity of treatment response as well as previous underlying trajectory of the condition need to be considered.
- Indirect evidence usually provides further support for large treatment effects
  - Oral anticoagulation in mechanical heart valves has not been compared with placebo in an RCT, but evidence from observational studies suggests a large effect of oral anticoagulation in decreasing thromboembolic events
  - Impact of routine colonoscopy vs. no screening for colon cancer on the rate of perforation associated with colonoscopy

Dose-response gradient

- An important criterion for causation specifically in observational studies when the confirmation of causation is under question.
  - infant growth is slowest in infants fed exclusively with breast milk, accelerated to some extent in infants fed in part with breast milk and part formula, and further accelerated in infants fed exclusively with formula

Plausible confounding can increase confidence in estimated effect

- It is less likely that all possible confounding factors are measured and have been adjusted in our model
  - The problem of residual confounding or residual biases
- If all plausible unmeasured confounding factors in a observational study, would result in an underestimate of an observed intervention effect, this confirms that the true effect size (after adjustment of all plausible confounding factors) is even larger than what we have seen.
  - Higher death rates in private for-profit vs. private not-for-profit hospitals in a systematic review of observational studies

Example

- an unpublished systematic review addressed the effect of condom use on HIV infection among men who have sex with men. The pooled effect estimate of RR from the five eligible observational studies was 0.34 [0.21, 0.54] in favor of condom use compared with no condom use.
- Condom users were more likely to have more partners (but did not adjust for this confounding factor in their analyses).
- Considering the number of partners would, if anything, strengthen the effect estimate in favor of condom use.

Plausible confounding can increase confidence in estimated effect (cont)

- In some situations the plausible confounding would suggest spurious effect, even if there is no effect.
- In such situation we upgrade the quality of evidence
- An example comes from the studies that failed to confirm the association between vaccination and autism
  - parents of autistic children diagnosed after the publicity associated with the previous article would be more likely to remember their vaccine experience than parents of children diagnosed before the publicity and presumably, than parents of nonautistic children
  - The negative findings despite this form of recall bias suggest rating up the quality of evidence

موفق باشید