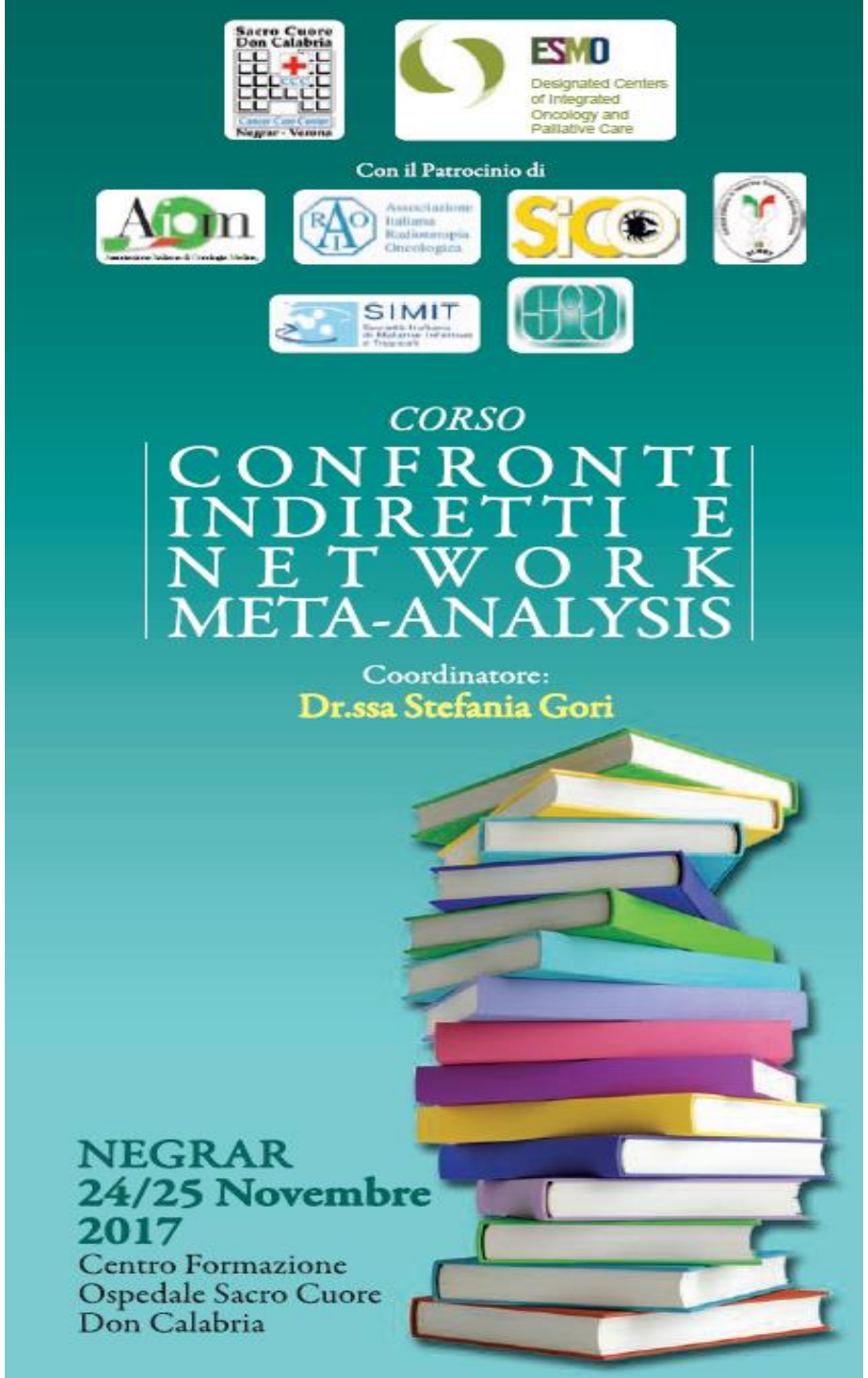


GRADE: breve introduzione

Negrar, 25 Novembre 2017

Michela Cinquini

Cosa è il GRADE ?

The “Grades of Recommendation, Assessment, Development, and Evaluation” (GRADE) approach provides guidance for rating quality of evidence and grading strength of recommendations in health care. It has important implications for those summarizing evidence for systematic reviews, health technology assessment, and clinical practice guidelines. GRADE provides a systematic and transparent framework for clarifying questions, determining the outcomes of interest, summarizing the evidence that addresses a question, and moving from the evidence to a recommendation or decision.

- Journal of Clinical Epidemiology 64 (2011)

GRADE Working Group

The Grades of Recommendation, Assessment, Development, and Evaluation (GRADE) Working Group is a group of health professionals, researchers, and guideline developers worldwide who, in 2000, began to work together to develop an optimal system of rating quality of evidence and determining strength of recommendations for clinical practice guidelines. The group now includes more than 200 members and continues, after a decade of work, to meet to refine and extend its methods.

BACKGROUND

- Le linee guida utilizzano vari e diversi metodi e criteri per valutare la qualità (livello) di evidenza derivante dalla letteratura e modi diversi per indicare la forza delle raccomandazioni

→ confusione per chi legge

- Il metodo seguito per muoversi dalle evidenze alle raccomandazioni cliniche è spesso poco trasparente , non esplicitato, lascia molto spazio alla valutazione soggettiva

→ Incertezza per chi legge (quanto mi posso fidare delle raccomandazioni? Su quali basi/elementi/considerazioni sono state formulate?)

Formulate question

P
I
C
O

Select outcomes

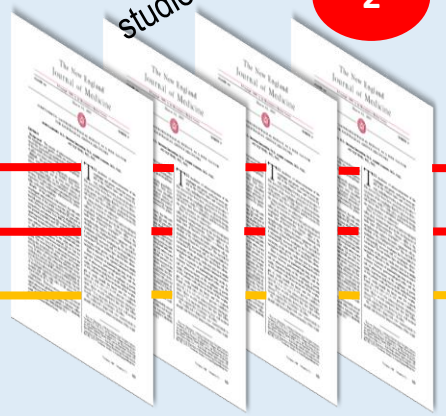
Rate importance

Outcome — Critical
Outcome — Critical
Outcome — Important
Outcome — Not important

1

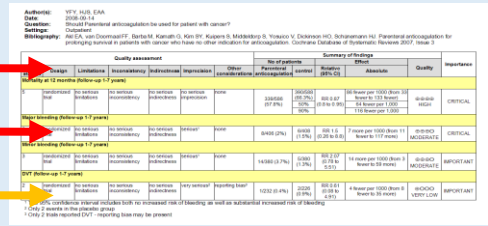
Outcomes across studies

2



Create evidence profile with GDT

3



Summary of findings & estimate of effect for each outcome

Rate quality of evidence for each outcome

High
Moderate
Low
Very low

Randomization raises initial quality
RCTs: high
Observational: low

Grade down

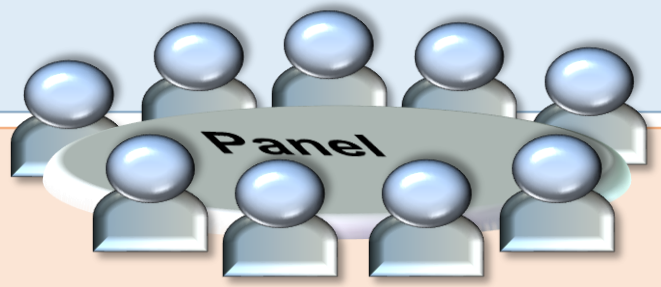
1. Risk of bias
2. Inconsistency
3. Indirectness
4. Imprecision
5. Publication bias

Grade up

1. Large effect
2. Dose response
3. Opposing bias & Confounders

4

Evidence synthesis (SR, HTA)



Recommendation/Decision

Grade overall quality of evidence
across outcomes based on lowest quality of ***critical*** outcomes

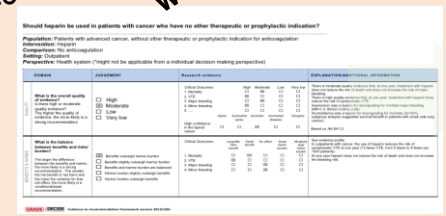
6

Grade recommendations (Evidence to Decision)
- For or against (direction) ↓↑
- Strong or conditional/weak (strength)

By considering balance of consequences (evidence to recommendations):
- ❑ Quality of evidence
- ❑ Balance benefits/harms
- ❑ Values and preferences
- ❑ Feasibility, equity and acceptability
- ❑ Resource use (if applicable)

5

EtD framework with GRADEpro



Guideline/Decision

Formulate Recommendations/Decision
"The panel recommends that ….should..."
"The panel suggests that ….should..."
"The panel suggests to **not** ..."
"The panel recommends to **not**..."

Transparency, clear, actionable
Research?

What are we grading?

two components

- **quality of body of evidence**

  extent to which confidence in estimate of effect adequate to support decision

  - high, moderate, low, very low

- **strength of recommendation**

  - strong and weak

Cos'è la Summary Of Findings

- **Summary of findings**: tabular presentation of key information about relevant outcomes of alternative health care interventions. It presents information about the body of evidence, key numerical results, and **summary judgment about the certainty of underlying evidence** for each outcome. SoF table has been chosen by the Cochrane Collaboration to present main findings of a systematic review.

Evidence profile: summary of evidence for a given question; it presentes relevant information about the body of evidence, key numerical results, and a **detailed quality assessment** and a explicit judgment of each factor that determines the quality. Used by guideline producers

Summary of findings

Summary of finding: antibiotics for acute otitis media in children

Antibiotics compared with placebo for acute otitis media in children

Patient or population: Children with acute otitis media
Setting: High- and middle-income countries
Intervention: Antibiotics
Comparison: Placebo

	Estimated risks (95% CI)				
	Control risk[a]	Intervention risk			
Outcomes	Placebo	Antibiotics	Relative effect (95% CI)	No. of Participants (studies)	Quality of the evidence (GRADE)
Pain at 24h	367 per 1,000	330 per 1,000 (286–382)	RR 0.9 (0.78–1.04)	1229 (5)	⊕⊕⊕⊕ High
Pain at 2–7 d	257 per 1,000	185 per 1,000 (159–213)	RR 0.72 (0.62–0.83)	2791 (10)	⊕⊕⊕⊕ High
Hearing, inferred from the surrogate outcome abnormal tympanometry—1 mo	350 per 1,000	311 per 1,000 (262–375)	RR 0.89 (0.75–1.07)	927 (4)	⊕⊕⊕○ Moderate[b]
Hearing, inferred from the surrogate outcome abnormal tympanometry—3 mo	234 per 1,000	227 per 1,000 (178–290)	RR 0.97 (0.76–1.24)	808 (3)	⊕⊕⊕○ Moderate[b]
Vomiting, diarrhea, or rash	113 per 1,000	156 per 1,000 (123–199)	RR 1.38 (1.09–1.76)	1,401 (5)	⊕⊕⊕○ Moderate[c]

Cos'è la Summary Of Findings

- **Summary of findings**: tabular presentation of key information about relevant outcomes of alternative health care interventions. It presents information about the body of evidence, key numerical results, and **summary judgment about the certainty of underlying evidence** for each outcome. SoF table has been chosen by the Cochrane Collaboration to present main findings of a systematic review.

Evidence profile: summary of evidence for a given question; it presentes relevant information about the body of evidence, key numerical results, and a **detailed quality assessment** and a explicit judgment of each factor that determines the quality. Used by guideline producers

Evidence profile: use of antibiotics (penicillin) versus no use of antibiotics in children with sickle cell disease. Source: Hirst et al. 4

Quality assessment							№ of patients		Effect		Quality	Importance
№ of studies	Study design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Penicillin prophylaxis	Standard care	Relative (95% CI)	Absolute (95% CI)		
Incidence of pneumococcal infection, for initiation of treatment - Initiation of penicillin												
2	Randomized trials	Not serious [1]	serious [2]	not serious [3]	serious [4]	none [5]	9/248 (3.6%)	19/209 (9.1%)	**OR 0.37** (0.16 to 0.86)	55 fewer per 1000 (from 12 fewer to 75 fewer)	⊕⊕◯◯ LOW	CRITICAL
Deaths, for initiation												
1	rand trial						/110 .6%)		**OR 0.11** (0.01 to 2.11)	32 fewer per 1000 (from 36 fewer to 37 more)	⊕⊕⊕◯ MODERATE	CRITICAL
Adverse drug effects - Nausea and vomiting												
1	randomized trials	not serious [6]	not serious	not serious	serious [4]	none [5]	2/201 (1.0%)	1/199 (0.5%)	**OR 1.99** (0.18 to 22.12)	5 more per 1000 (from 4 fewer to 95 more)	⊕⊕⊕◯ MODERATE	CRITICAL

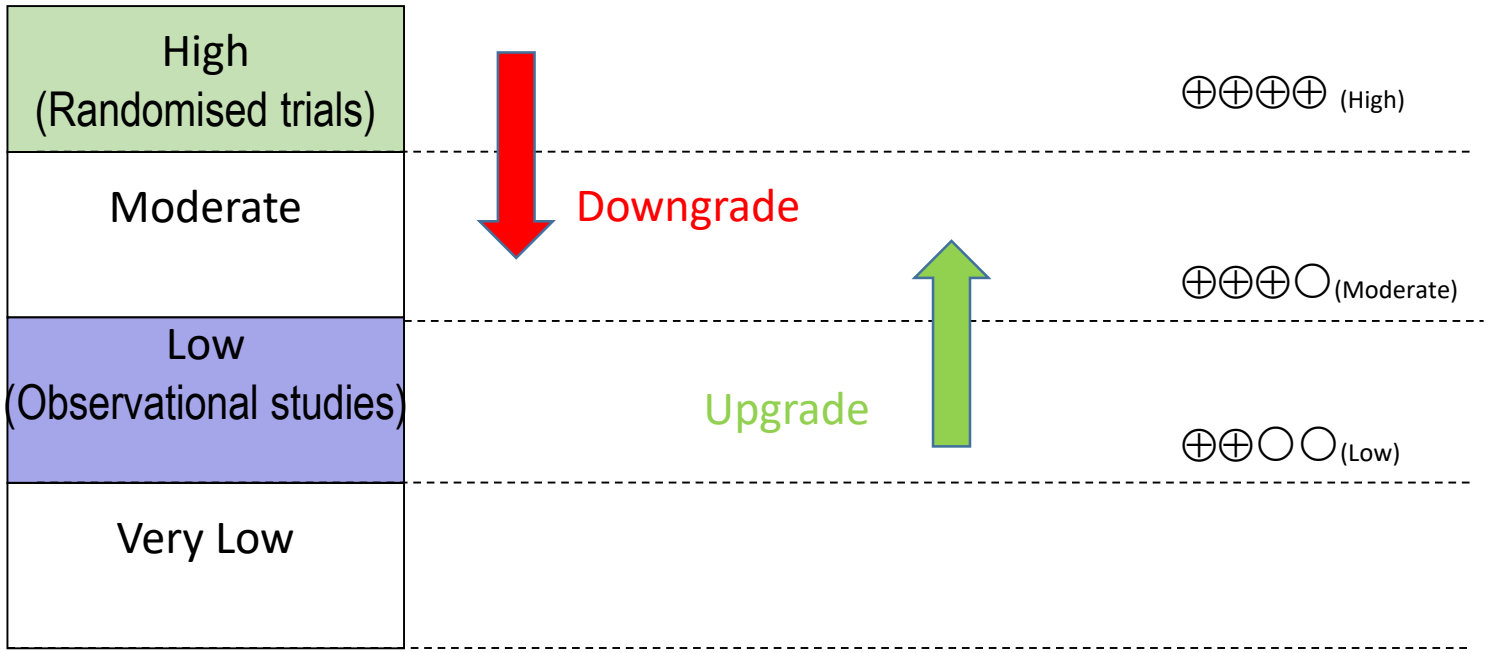
1.blinding and concealment were not clear for one of the two studies
2.heterogeneity exists; p-value for testing heterogeneity is 0.07 and I2=69%
3.the question addressed is the same for the evidence regarding the population, intervention, comparator and outcome
4.total sample size is small and the total number of events is <300 (a threshold rule-of-thumb value)
5.insufficient number of studies to assess publication bias
6.unclear allocation concealment

Quality of evidence

- GRADE is ‘‘**outcome centric**’’: rating is made for each outcome, and quality may differ -indeed, is likely to differ - from one outcome to another **within a single study** and **across a body of evidence**
- E.g: subjective outcomes are prone to performance and detection bias, while objective outcomes are not
- E.g. one outcome within a review could have imprecision in the pooled estimate of the effect, while another could have not
- E.g. one outcome could have high attrition bias (use of substance) while another could have not (drop out)

Rating quality of evidence

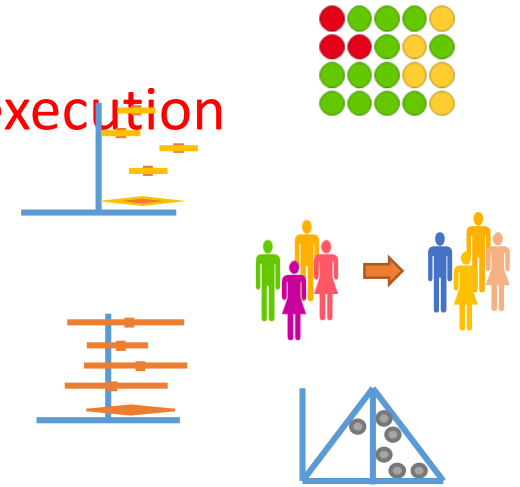
GRADE's approach begins with the study design. Randomized controlled trials (RCTs) start as high-quality evidence and observational studies as low-quality ~~evidence supporting estimates of intervention effects~~



Determinants of quality/certainty of a body of evidence

- **RCTs ⊕⊕⊕⊕**
- **observational studies ⊕⊕◯◯**

- **5 factors that can lower quality**
  1. limitations in detailed study design and execution *(risk of bias criteria)*
  2. Inconsistency *(or heterogeneity)*
  3. Indirectness *(PICO and applicability)*
  4. Imprecision
  5. Publication bias



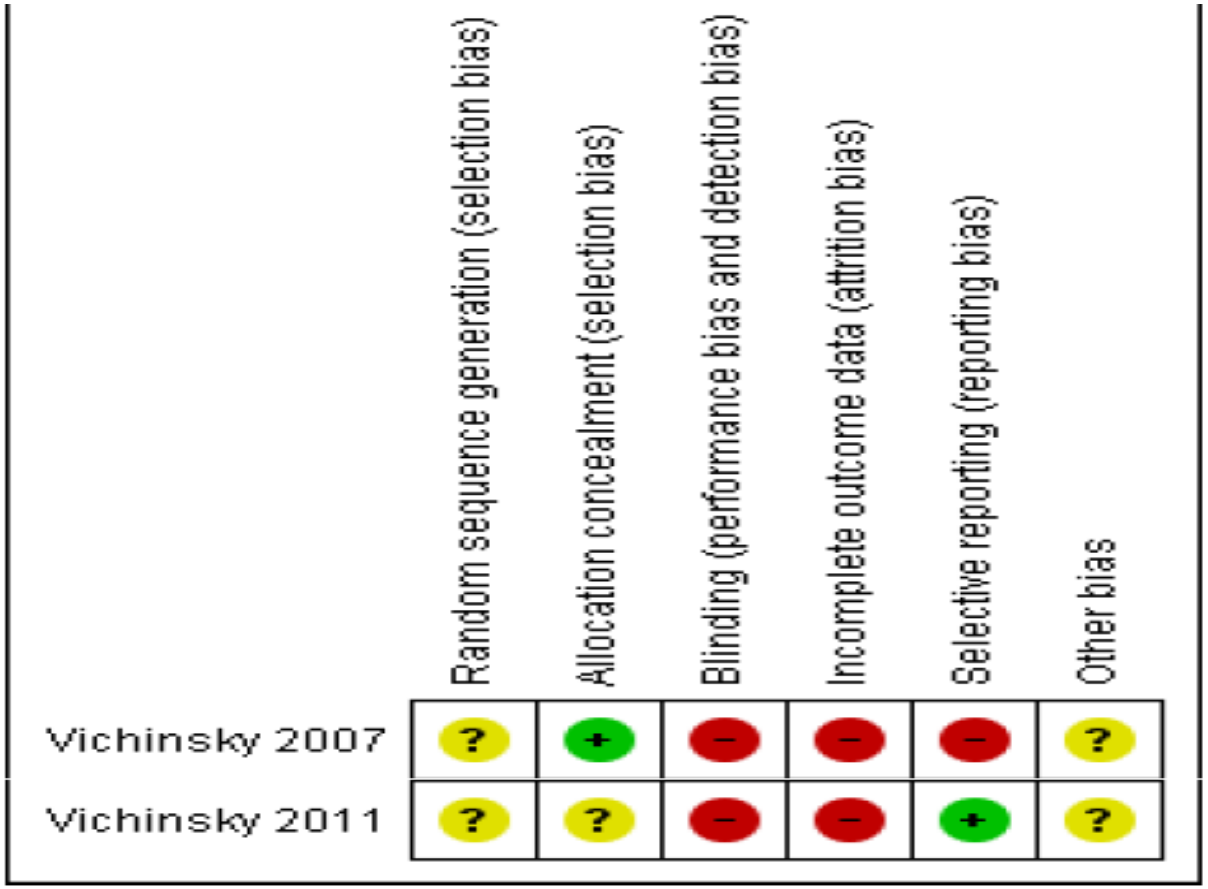
- **3 factors can increase quality**
  1. large magnitude of effect
  2. opposing plausible residual bias or confounding
  3. dose-response gradient

Risk of bias table for RCTs Cochrane Collaboration

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk of bias High risk of bias Unclear risk of bias	
Allocation concealment (selection bias)	Low risk of bias High risk of bias Unclear risk of bias	
Blinding of participants and personnel (performance bias)	Low risk of bias High risk of bias Unclear risk of bias	
Blinding of outcome assessment (detection bias)	Low risk of bias High risk of bias Unclear risk of bias	
Incomplete outcome data (attrition bias)	Low risk of bias High risk of bias Unclear risk of bias	
Selective reporting (reporting bias)	Low risk of bias High risk of bias **Unclear risk of bias**	

1. risk of bias

Deferasirox for managing transfusional iron overload in people with sickle cell disease (Review)



Risk of bias

- Outcome specific
- Do not average risk quality across the studies
- Evaluate the extent to which each trial contributes toward the estimate of magnitude of effect. This contribution will usually reflect study sample size and number of outcome events -larger trials with many events will contribute more, much larger trials with many more events will contribute much more (look at the weight of each study in the forest plot)

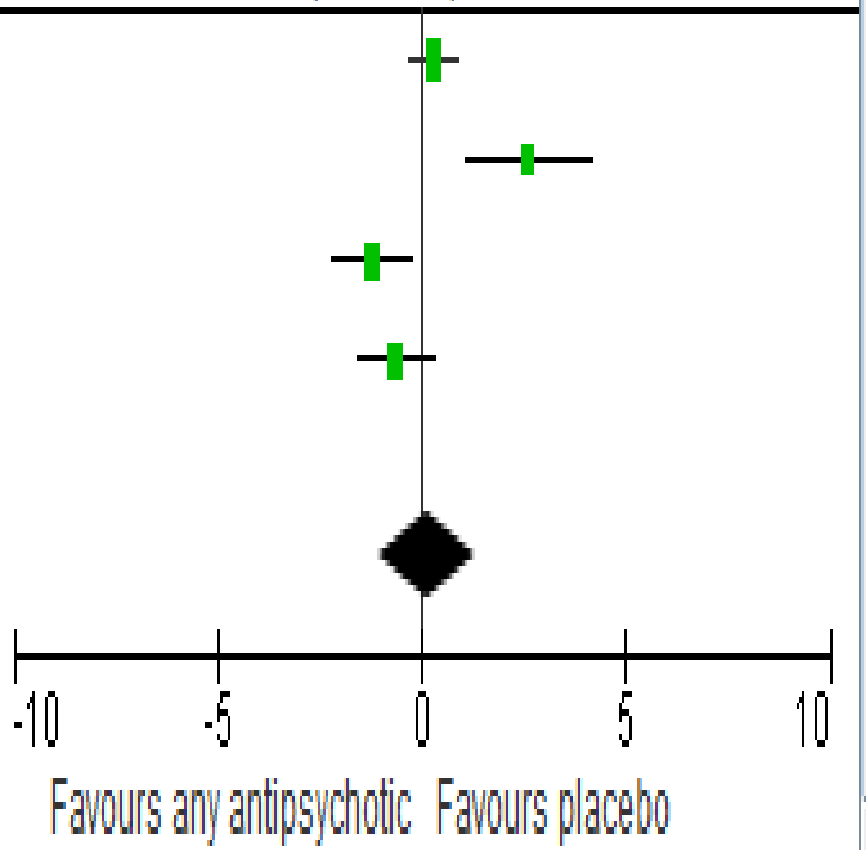
2.Inconsistency (heterogeneity) between studies results

- Variation in size of effect (**Point estimates** vary widely across studies)
- **Confidence intervals** (CIs) show minimal or no overlap
- The statistical test for heterogeneity which tests the null hypothesis that all studies in a meta-analysis have the same underlying magnitude of effect shows a low **P-value** (< 0.05)
- The $\mathbf{I^2}$ which quantifies the proportion of the variation in point estimates due to among-study differences (< 40% : low, 30 e 60% :moderate, 50% e 90% :substantial, 75 e 100% : considerable)
- All statistical approaches have limitations, and their results should be seen in the context of a subjective examination of the variability in point estimates and the overlap in CIs.

Study or Subgroup	Any antispsychotic Mean	SD	Total	Placebo Mean	SD	Total	Weight	Mean Difference IV, Random, 95% CI
Kampman 2003	1.3	0.8	15	1	0.9	15	28.3%	0.30 [-0.31, 0.91]
Reid 2005	6.1	2.1	16	3.5	2.3	15	20.3%	2.60 [1.05, 4.15]
Tapp 2015	0.4	1.26	29	1.63	2.39	31	25.6%	-1.23 [-2.19, -0.27]
Winhusen 2007	2.8	2.45	60	3.44	2.76	59	25.8%	-0.64 [-1.58, 0.30]
Total (95% CI)			120			120	100.0%	0.13 [-1.08, 1.35]

Heterogeneity: $Tau^2 = 1.26$; $Chi^2 = 19.78$, df = 3 ($P = 0.0002$); $I^2 = 85\%$

Test for overall effect: $Z = 0.22$ ($P = 0.83$)



3. Directness of Evidence generalizability, transferability, applicability

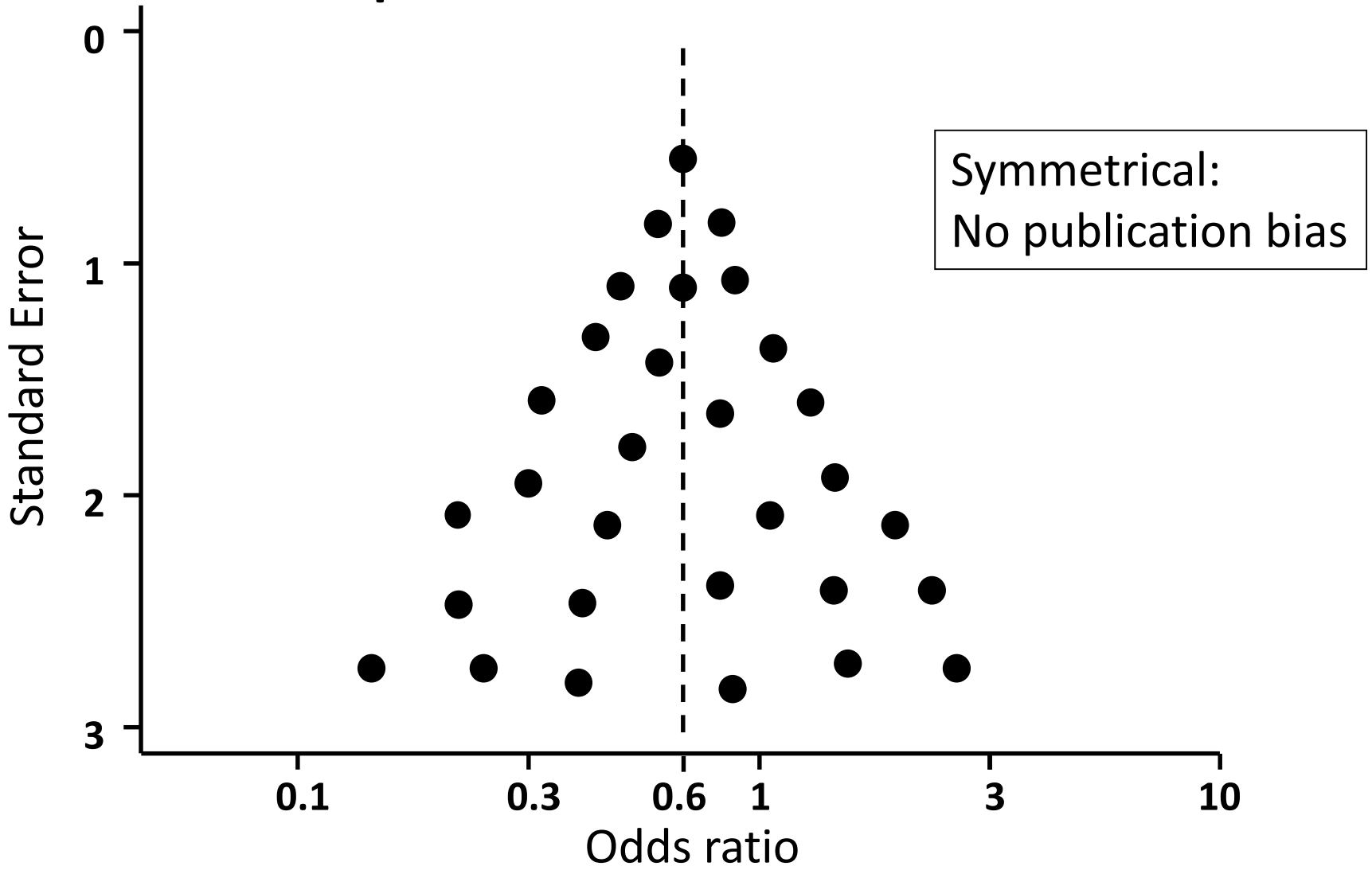
- differences between PICO and available evidence in
  - populations/patients (interested in children found adults population)
  - interventions (interested in high dosage, found low dosage, interested in long treatment, found short, etc)
  - outcomes (interested in important but we found surrogate; e.g hip fracture vs bone density; interested in long term but found short term results)
- indirect comparisons
  - interested in A versus B
  - found A versus C and B versus C

4. Publication Bias

Consider rating down if:

- You find **systematic reviews performed early**, when only few initial studies are available, that will overestimate effects when ‘‘negative’’ studies face delayed publication. Early positive studies, particularly if small in size, are suspect.
- You **find only small “positive” studies, mainly if sponsored by industry**
- **Funnel plot showing asimmetry** but
- Funnel plot should be seen as a generic means of displaying small-study effects – a tendency for the intervention effects estimated in smaller studies to differ from those estimated in larger studies (Sterne 2000). Small-study effects may be due to reasons other than publication bias (low methodological quality, chance, patients characteristics).
- **Funnel plot should be used only when there are at least 10 studies** included in the meta-analysis, because when there are fewer studies the power of the tests is too low to distinguish chance from real asymmetry

Funnel plot

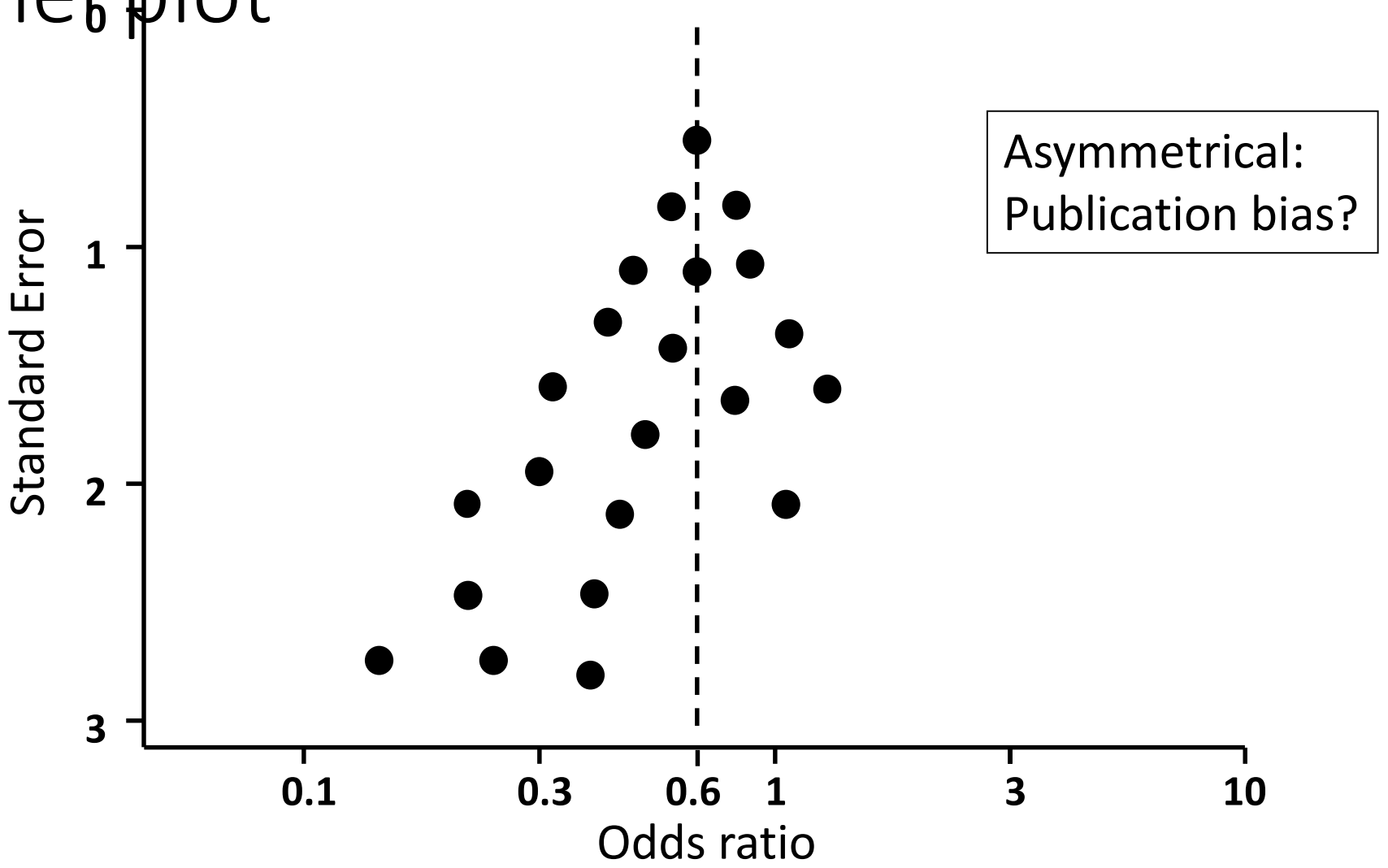


Egger M, Cochrane Colloquium Lyon 2001

Funnel plot

- On the **horizontal axis** : measure of treatment effect
- On the **vertical axis**: standard error (SE) of the intervention effect estimate: measure of **precision of the estimate** ; SE is determined by sample size, and by the number of participants experiencing the event for dichotomous outcomes, and the standard deviation of responses for continuous outcomes.
- Precision of the estimated intervention effect increases as the size of the study increases. Effect estimates from small studies will therefore scatter more widely at the bottom of the graph, with the spread narrowing among larger studies**. In the absence of bias the plot should approximately resemble a symmetrical funnel plot**.

Funnel plot



Asymmetrical:
Publication bias?

Egger M, Cochrane Colloquium Lyon 2001

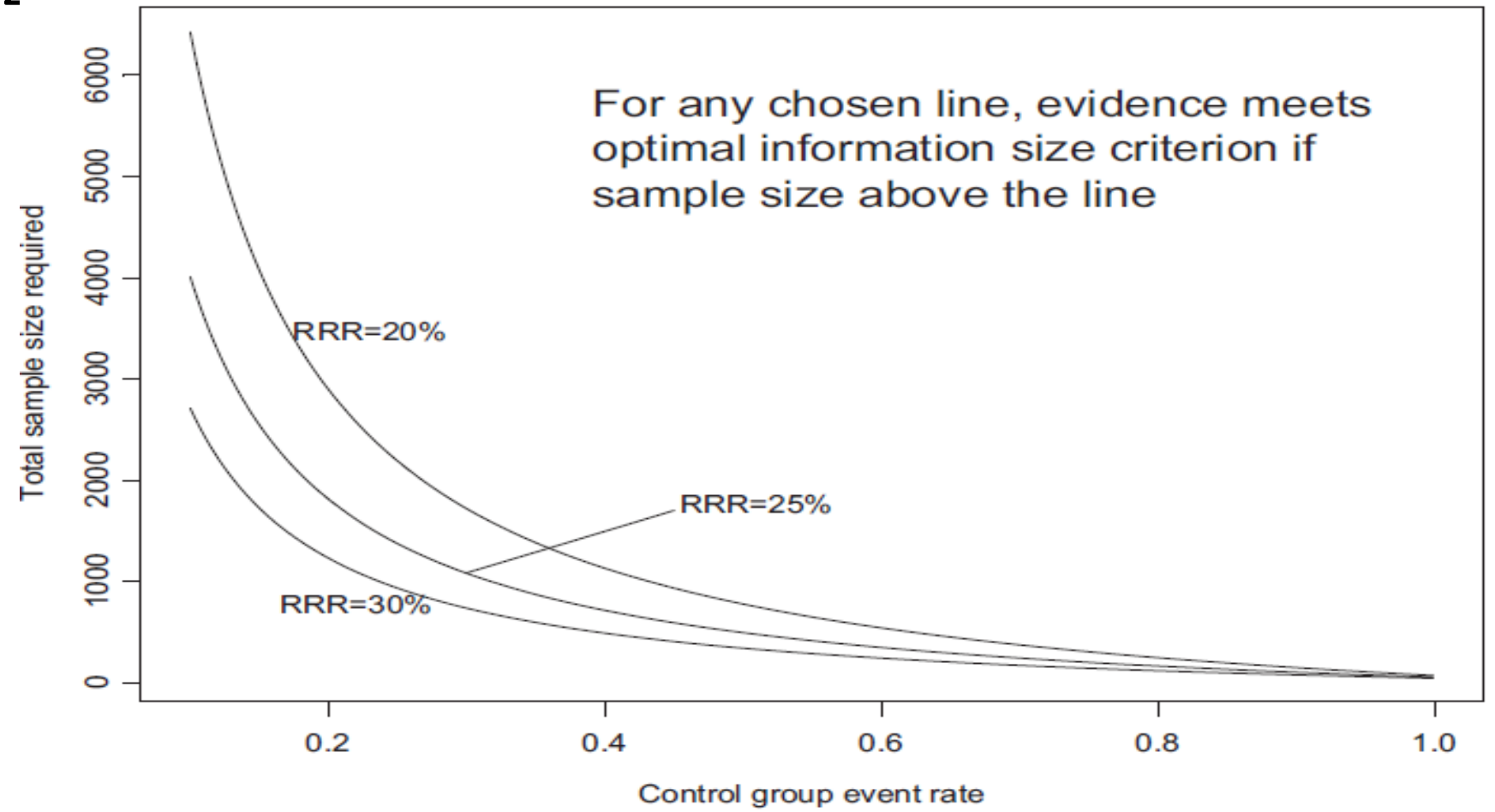
5. Imprecision of the overall estimate

- **Wide confidence intervals** (CIs inform the impact of random error on evidence quality; CI expresses the range in which the truth plausibly lies)
- **Small number of events**
- **Small sample size**
- recommendation or **clinical course of action would differ** if the **upper versus the lower boundary** of the CI represented the truth

Optimal information size

- We suggest the following: if the total number of patients included in a systematic review is less than the number of patients generated by a conventional sample size calculation for a single adequately powered trial, consider rating down for imprecision. Authors have referred to this threshold as the “optimal information size” (OIS)

Required sample size (assuming α of 0.05, and β of 0.2) for RRR of 20%, 25%, and 30% across varying control event rates. For example, if the best estimate of control event rate was 0.2 and one specifies an RRR of 25%, the OIS is approximately 2,000 patients (GRADE guideline n.6 Journal of Clinical Epidemiology 64 (2011) 1283e1293)



ptimal information size given α of 0.05 and β of 0.2 for varying control event rates and relative risks.

Power is more closely related to number of events than to sample size

(GRADE guideline n.6 Journal of Clinical Epidemiology 64 (2011) 1283e1293)

Calculating the OIS for **dichotomous outcome** requires specifying:

- probability of detecting a false effect – type I error (α; usually 0.05)
- probability of detecting a true effect – power (usually 80% [power = 1 – type II error; β; usually 0.20])
- realistic relative risk reduction (RRR; we suggest a default of 25%)
- control event rate (we suggest the median of the available trials, or the rate from a dominating trial, if it exists).

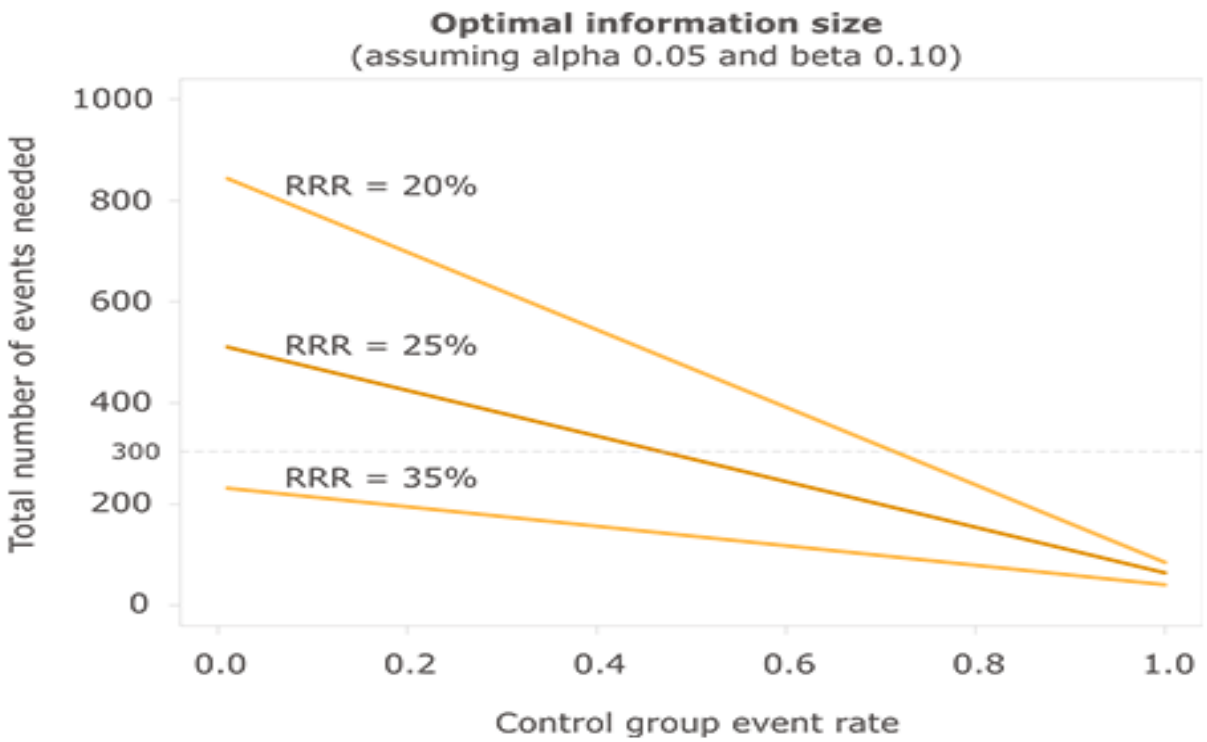


Table 1: Optimal information size implications from Figure above

Total Number of Events	Relative Risk Reduction	Implications for meeting OIS threshold
100 or less	≤ 30%	Will almost never meet threshold whatever control event rate
200	30%	Will meet threshold for control event rates for ~ 25% or greater
200	25%	Will meet threshold for control event rates for ~ 50% or greater
200	20%	Will meet threshold only for control event rates for ~ 80% or greater
300	≥ 30%	Will meet threshold
300	25%	Will meet threshold for control event rates ~ 25% or greater
300	20%	Will meet threshold for control event rates ~ 60% or greater
400 or more	≥ 25%	Will meet threshold for any control event rate
400 or more	20%	Will meet threshold for control event rates of ~ 40% or greater

OIS for continuous outcomes

- Authors can calculate the OIS for continuous variables in exactly the same way they can for binary variables by specifying the α and β errors (we have suggested 0.05 and 0.2) and the Δ (i.e. the difference one wishes to detect as clinically relevant), and choosing an appropriate standard deviation from one of the relevant studies.
- A particular challenge in calculating the OIS for continuous variables arises when studies have used different instruments to measure a construct, and the pooled estimate is calculated using a standardized mean difference.
- we suggest authors choose one of the available instruments (ideally, one in which an estimate of the minimally important difference is available) and calculate an OIS using that instrument

OIS for continuous outcomes

whenever there are sample sizes that are less than 400, review authors and guideline developers should certainly consider rating down for imprecision.

Downgrading and OIS

- if OIS not met downgrade for imprecision
- If OIS met and the 95% CI excludes a relative risk (RR) of 1.0 (statistically significant results), precision is adequate.
- if OIS met but the 95% CI includes a RR of 1 (null effect) , authors should consider whether CIs include appreciable benefit or harm (we suggest a RR of under 0.75 or over 1.25 as a rough guide) ; if yes downgrading for imprecision may be appropriate.

Grades of evidence and Interpretation

Symbol	Quality	Interpretation
⨁⨁⨁⨁	**High**	We are very confident that the true effect lies close to that of the estimate of the effect
⨁⨁⨁◯	**Moderate**	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
⨁⨁◯◯	**Low**	Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
⨁◯◯◯	**Very low**	We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Per approfondire

- Guyat G. et al. GRADE guidelines: 1. Introduction: GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology 64 (2011) 383e394
- Balshem H et al. GRADE guidelines: 3. Rating the quality of evidence. Journal of Clinical Epidemiology 64 (2011) 401e406
- Guyatt G. et al. for the GRADE Working Group. Rating quality of evidence and strength of recommendations: What is "quality of evidence" and why is it important to clinicians? BMJ. 2008 May 3;336(7651):995-8