Using GRADE to determine the quality of evidence and strength of recommendations

Gordon Guyatt

Plan

- Questions for you
- GRADE background
- Certainty in estimates
- Evidence profiles
- Strength of recommendation
- Exercise in applying GRADE

Who are you?

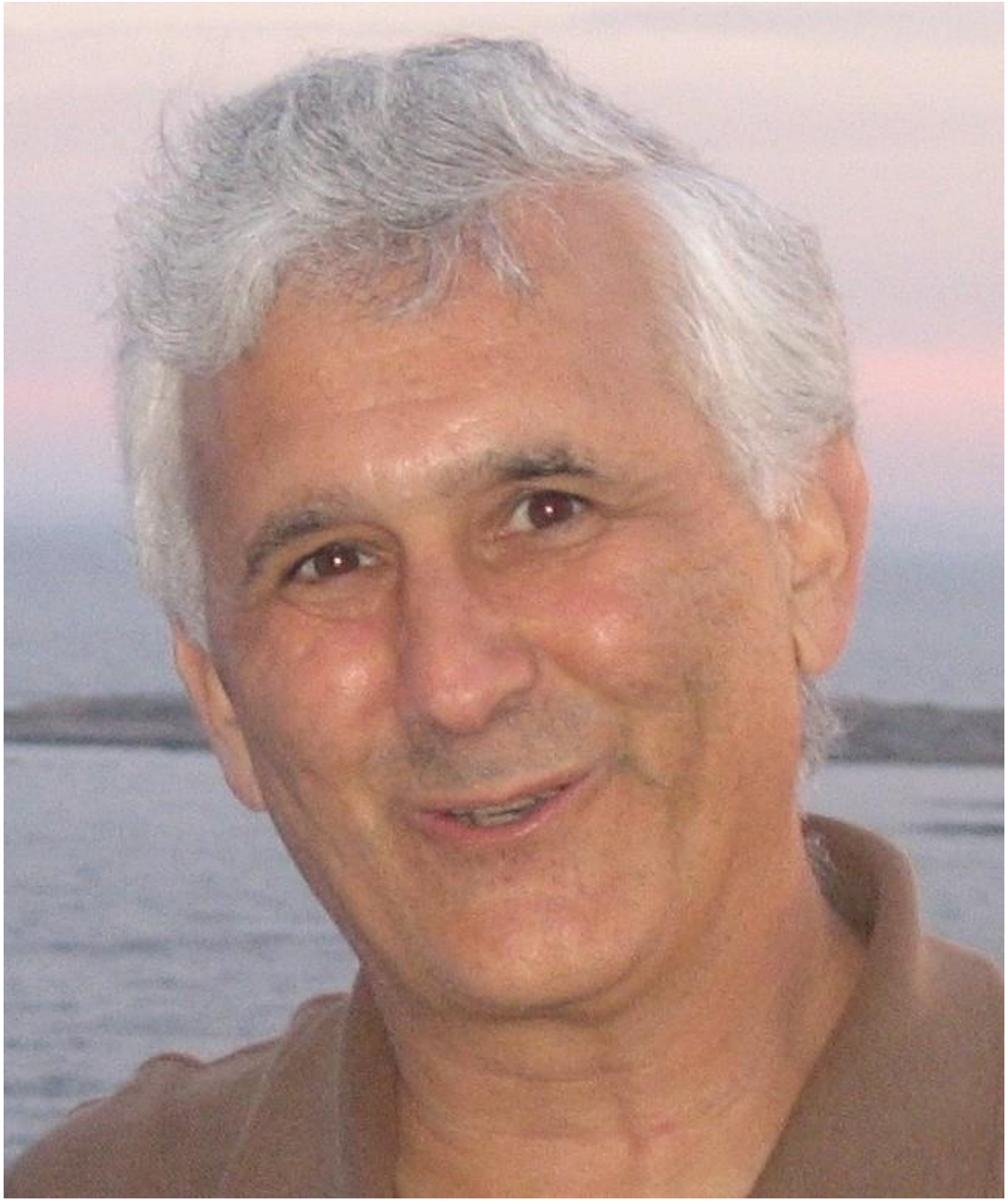
- Clinicians
- Experience systematic review guideline panels
- Use of grading systems, experience
- How will you be using GRADE
  - As a user of systematic reviews, guidelines
  - As a systematic review author
  - As a guideline panelist

Grading good idea, but which grading system to use?

- Many available
  - Australian National and MRC
  - Oxford Center for Evidence-based Medicine
  - Scottish Intercollegiate Guidelines (SIGN)
  - US Preventative Services Task Force
  - American professional organizations
    - AHA/ACC, ACCP, AAP, Endocrine society, etc….

Result

Confusion



Common international grading system?

- GRADE (*G*rades of *r*ecommendation, *a*ssessment, *d*evelopment and *e*valuation)
- International group
  - Australian NMRC, SIGN, USPSTF, WHO, NICE, Oxford CEBM, CDC, CC
- ~ 40 meetings over last 19 years
  - (~10 – 120 attendants)

GRADE GUIDANCE

- 2004 BMJ, first description
- 2008 BMJ six part series
  - for guideline users
- 2010-19, 26 part series
  - For systematic review authors, HTA practitioners, guideline developers

Where does GRADE certainty apply?

- Evidence regarding therapeutic interventions
- Evidence regarding screening interventions
- Evidence regarding diagnostic impact
- Evidence regarding diagnostic accuracy
- Evidence regarding prognosis

- Rating for a single study?
- Rating for a body of evidence
- Both single study and body of evidence

>110 organizations have adopted GRADE

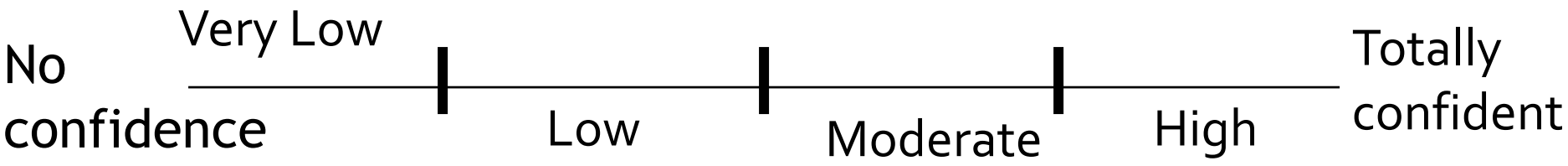
Figure 1. *Belief and confidence: a two-dimensional weather report. (Reprinted by permission from the Wall Street Journal).*

Likelihood of and confidence in an outcome

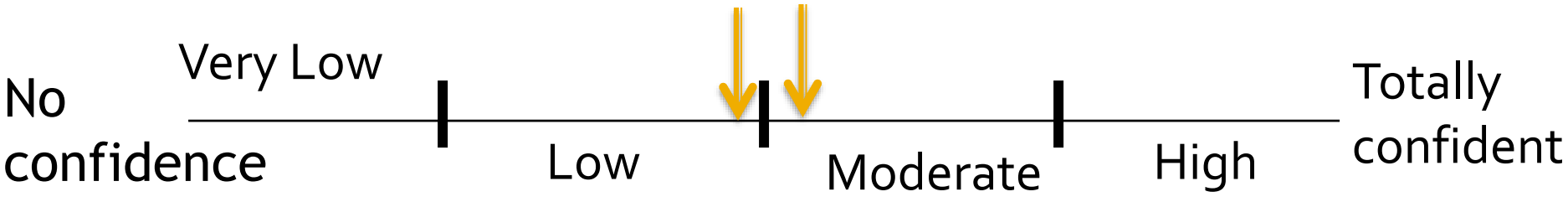
What are we grading?

- Two components
- Certainty/confidence in estimate of effect adequate to support decision (quality of body of evidence)
  - high, moderate, low, very low
- Strength of recommendation
  - strong and weak

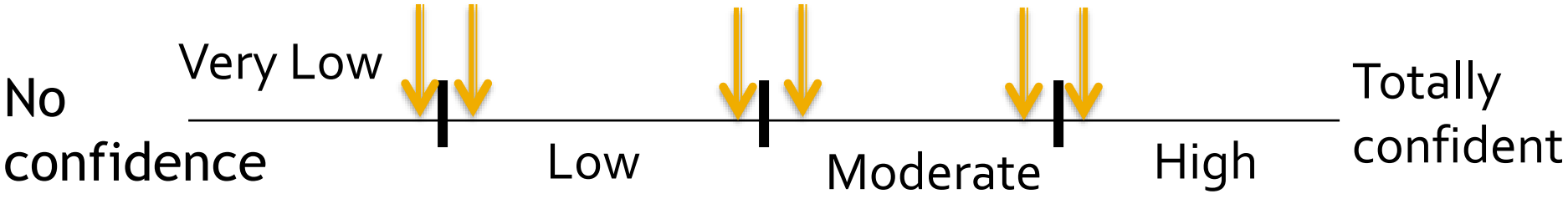
Confidence/Certainty in evidence (quality of evidence)



Apparent disagreement, true agreement



Apparent disagreement, true agreement



Structured question

- Patients:
  - Women considering breast cancer screening
  - Age 40-9; 50 to 74; ≥ 75
  - No risk genetic mutation, chest radiation
- Intervention
  - film mammography
- Alternative
  - no screening

Need to define all patient-important outcomes and evaluate their importance

- Desirable consequences
  - *Reduction in breast cancer mortality*
- Undesirable consequences
  - False positive screening results - anxiety
  - Invasive procedures from positive results
  - Complications of invasive procedures
  - Unnecessary diagnosis and treatment

Determinants of confidence

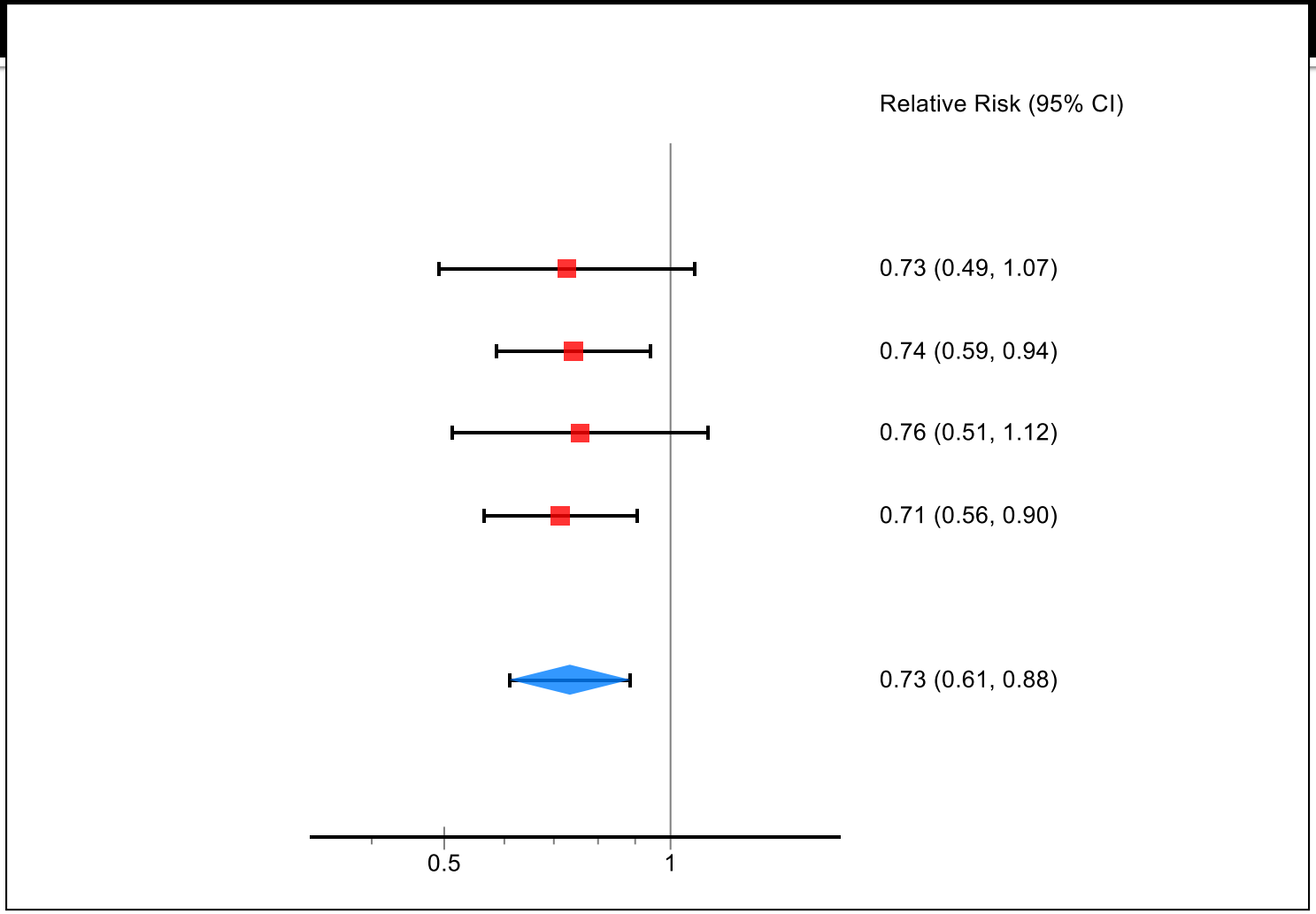
- RCTs start high
- Observational studies start low
- What can lower confidence?

- Risk of bias
- Inconsistency
- Indirectness
- Imprecision
- Publication bias

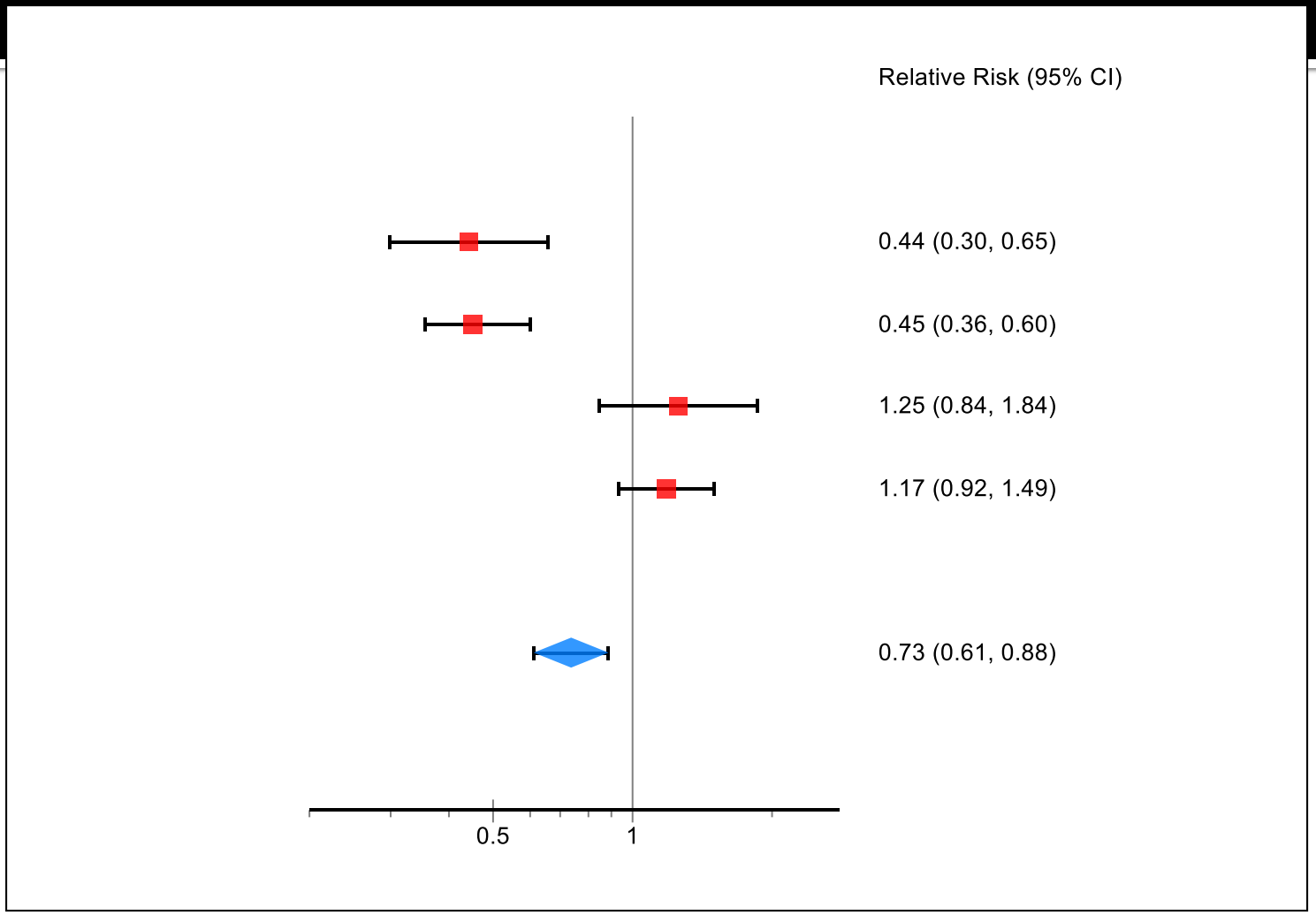
Risk of Bias - RCTs

- Well established
  - Concealment
  - Intention to treat principle observed
  - Blinding
  - Completeness of follow-up
- More recent
  - Selective outcome reporting bias
  - Stopping early for benefit

Inconsistency – happy with these results?

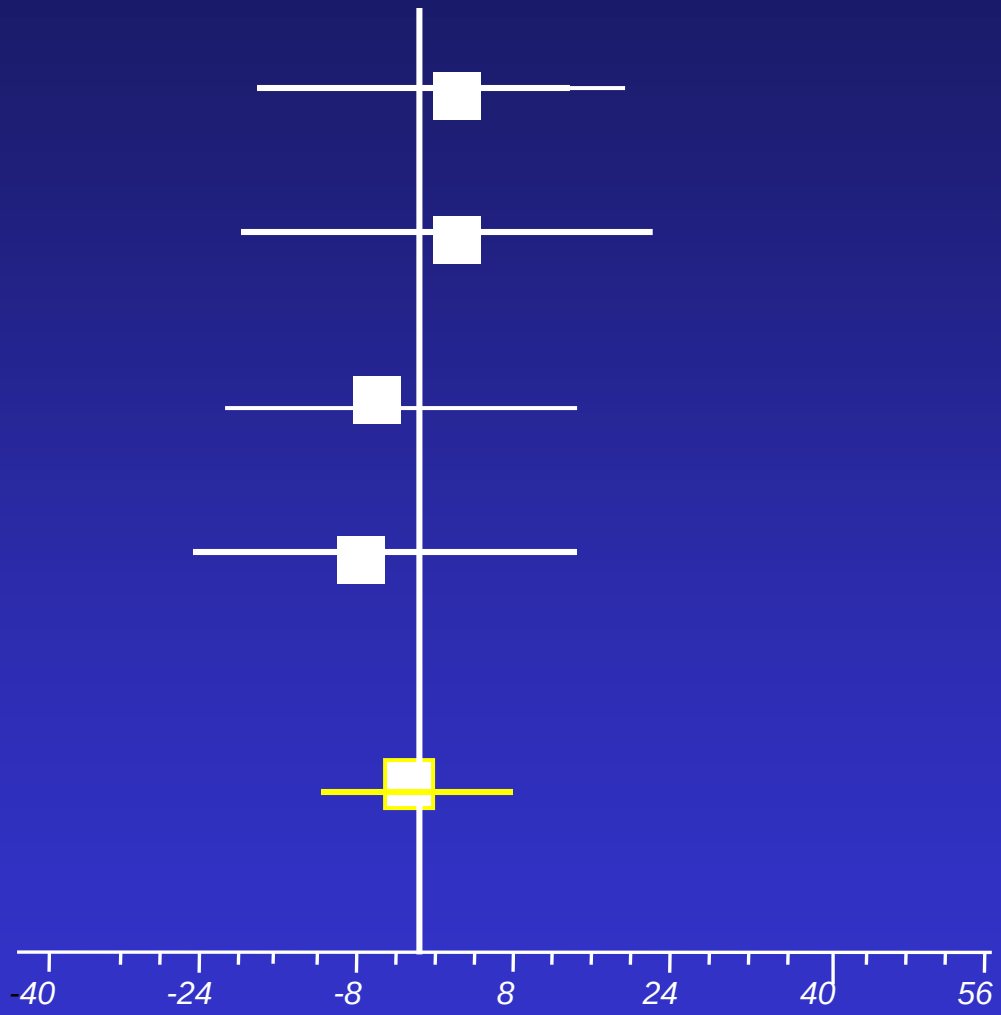


What about these?



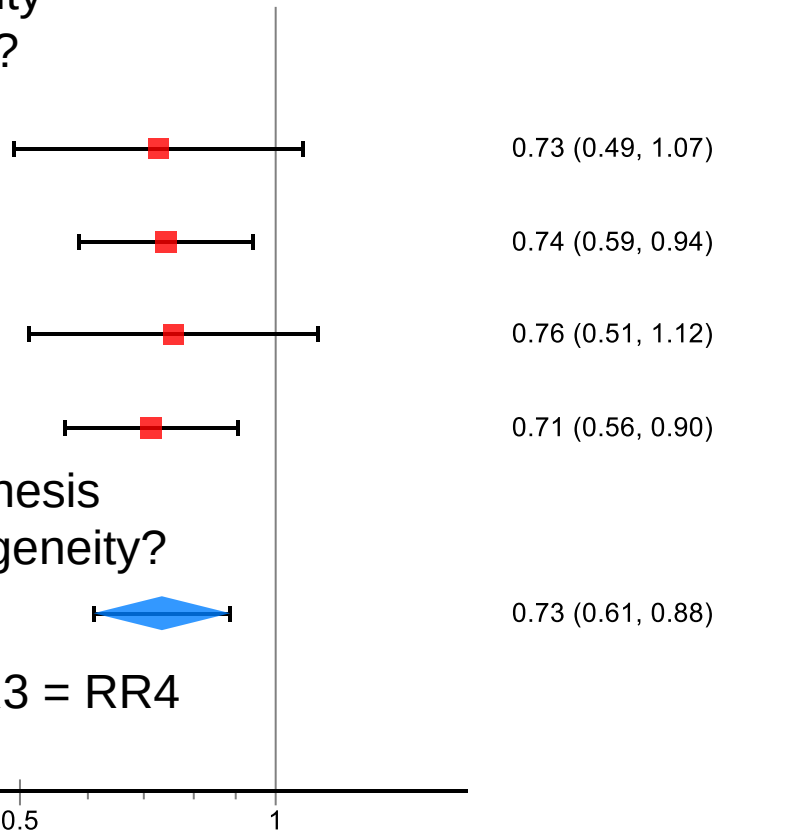
What criteria were you using?

- Similarity of point estimates
  - less similar, less happy
- Overlap of confidence intervals
  - less overlap, less happy


-40
-24
-8
8
24
40
56
RRR (95% CI)

Homogenous

test for heterogeneity
what is the p-value?



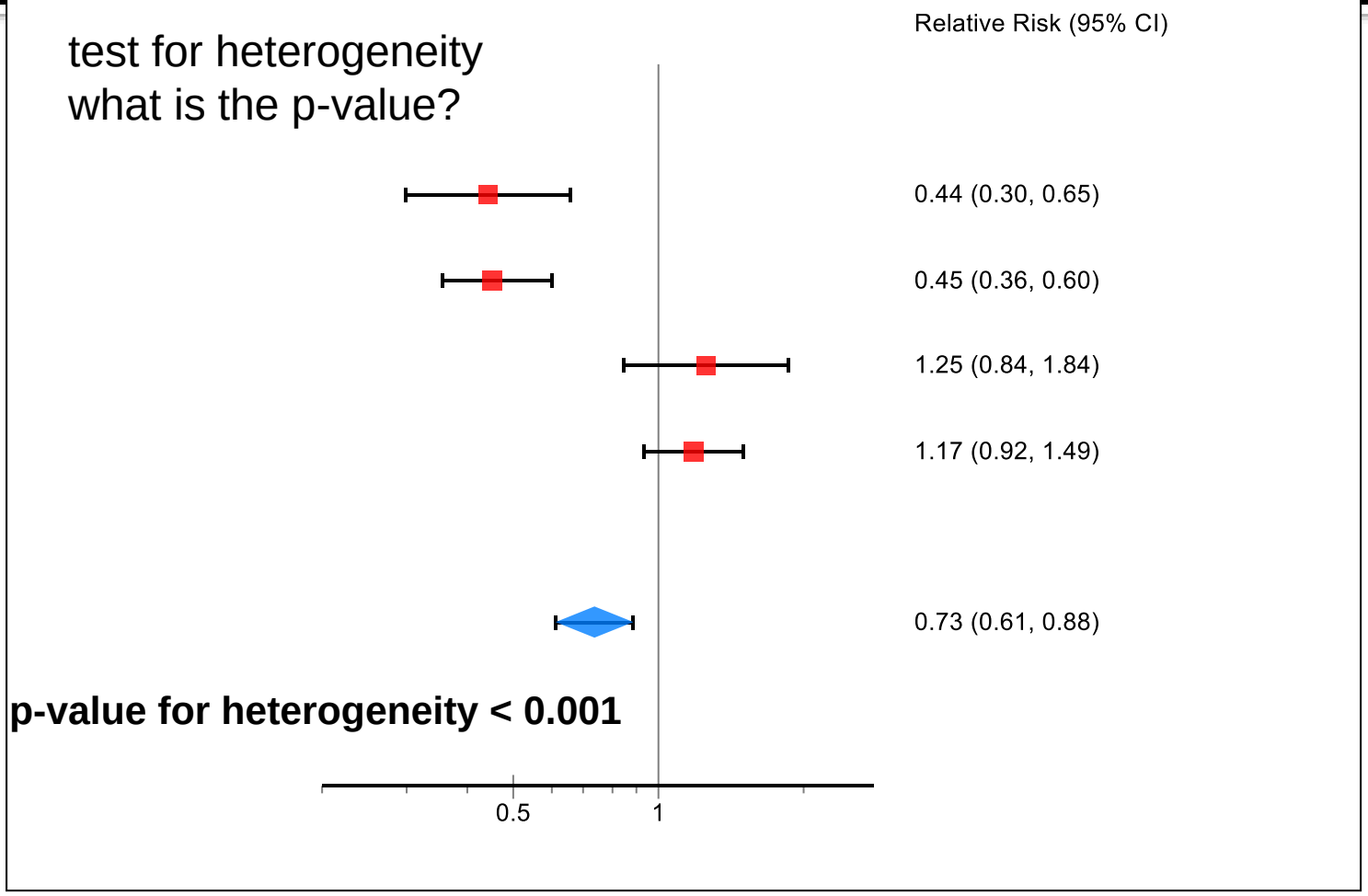
what is the null hypothesis
for the test for heterogeneity?

Ho: RR1 = RR2 = RR3 = RR4

p=0.99 for heterogeneity

Heterogeneous

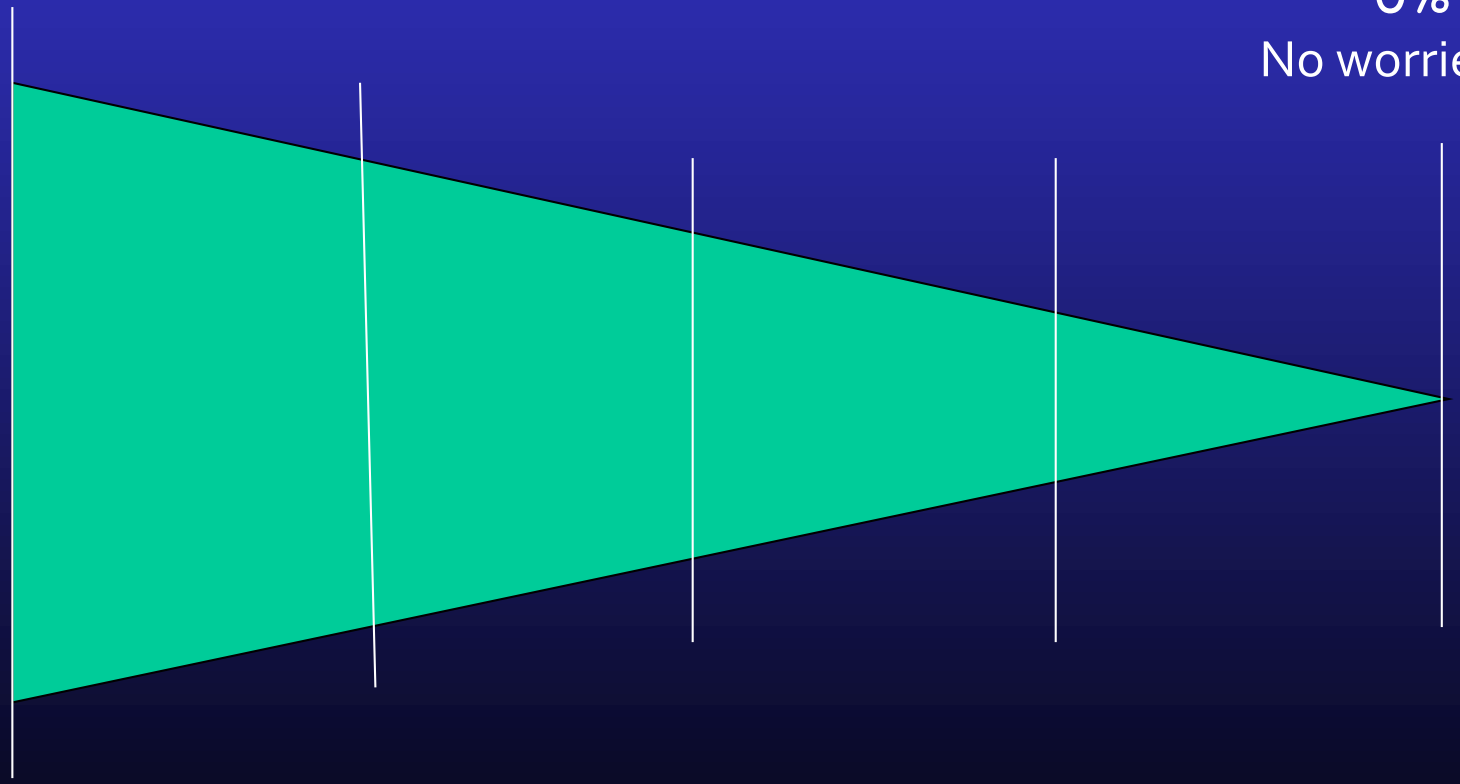
test for heterogeneity
what is the p-value?



I^2 Interpretation

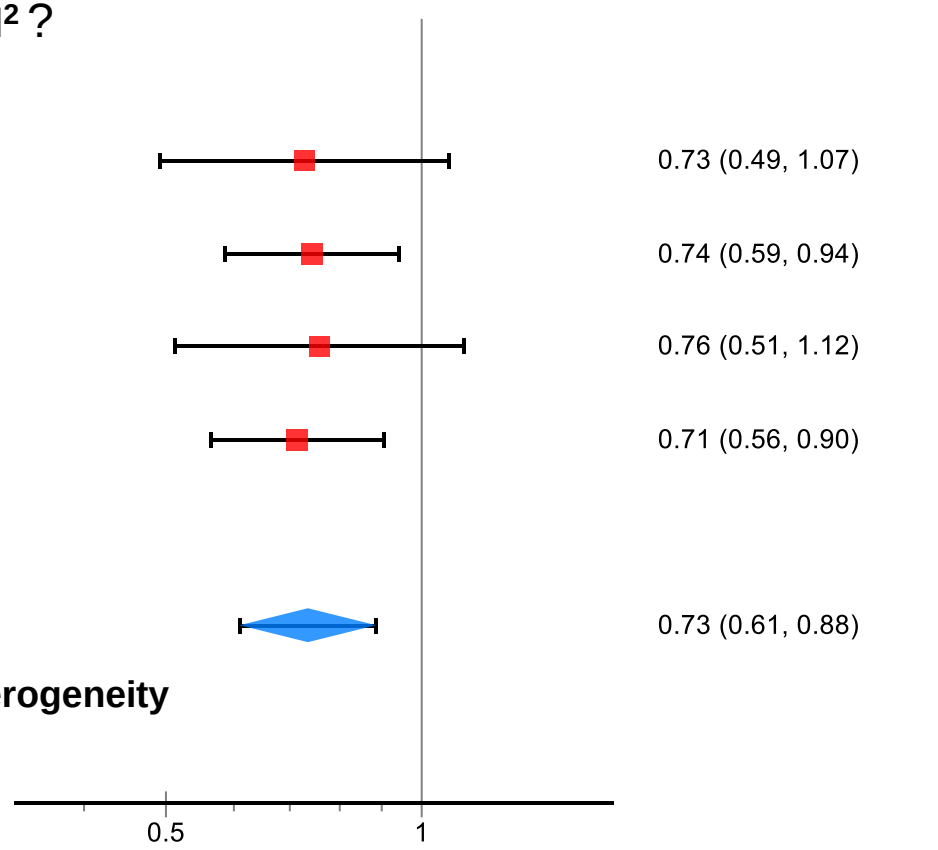
100%
Why are we pooling?

0%
No worries



Homogenous

What is the I^2?

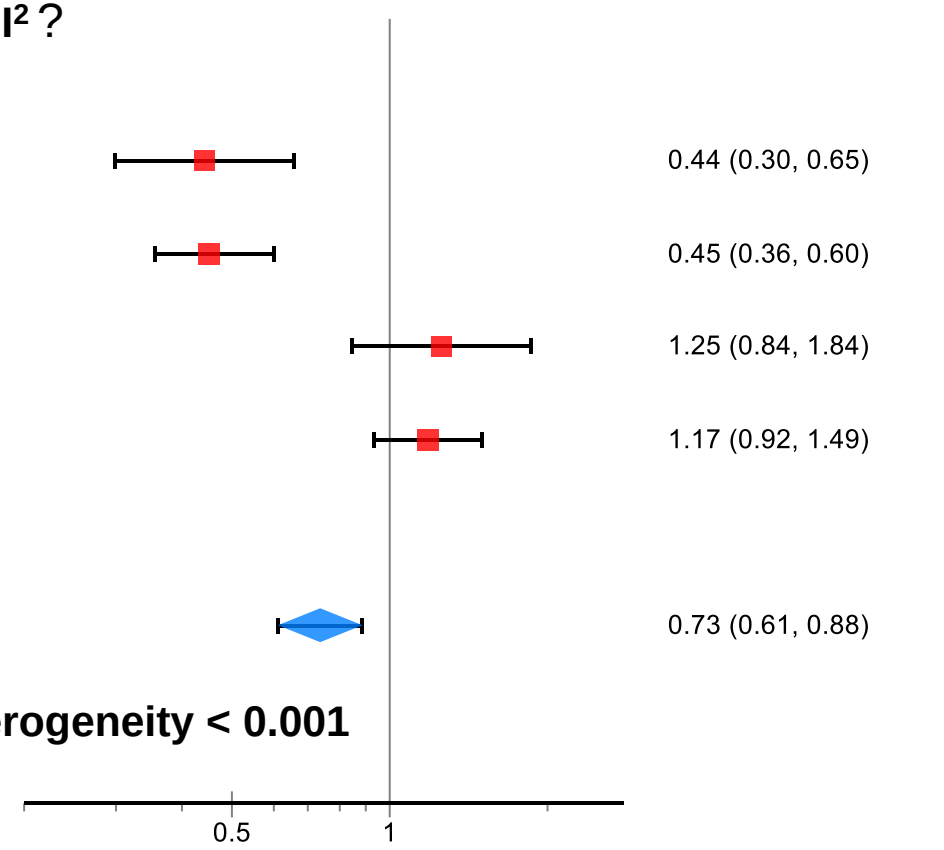


p=0.99 for heterogeneity

I^2=0%

Heterogeneous

What is the I^2?



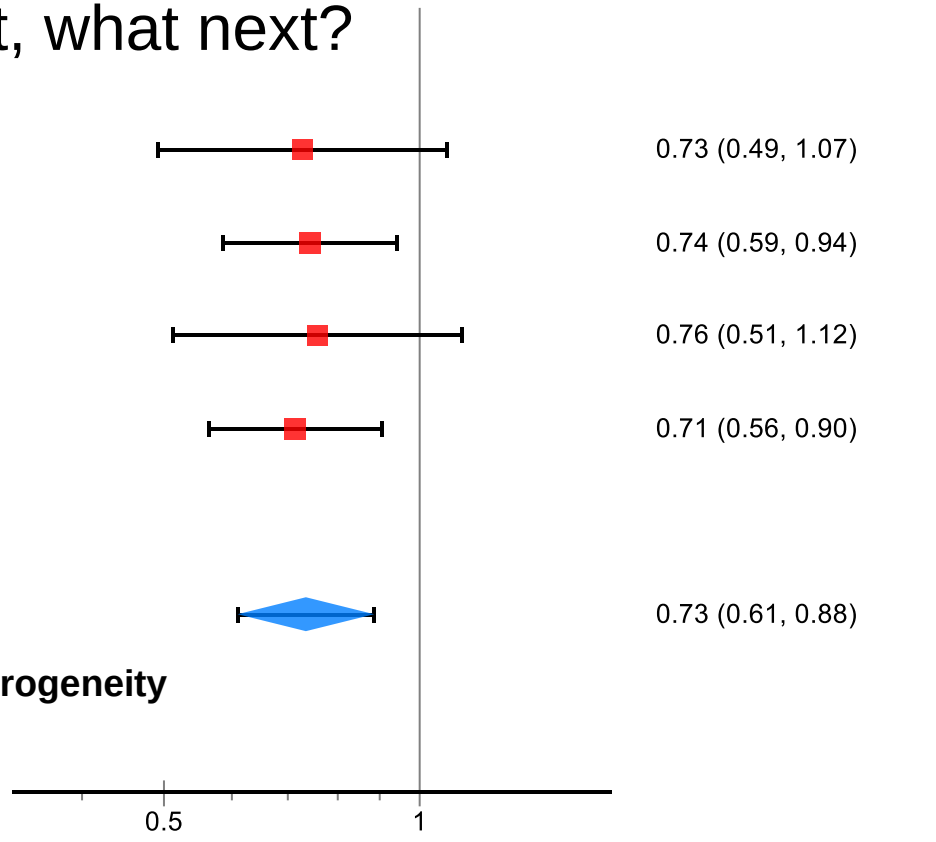
p-value for heterogeneity < 0.001
I^2=89%

Consistency of results

- Judgment of consistency
- Variation in size of effect
- Overlap in confidence intervals
- Statistical significance of heterogeneity
- I^2

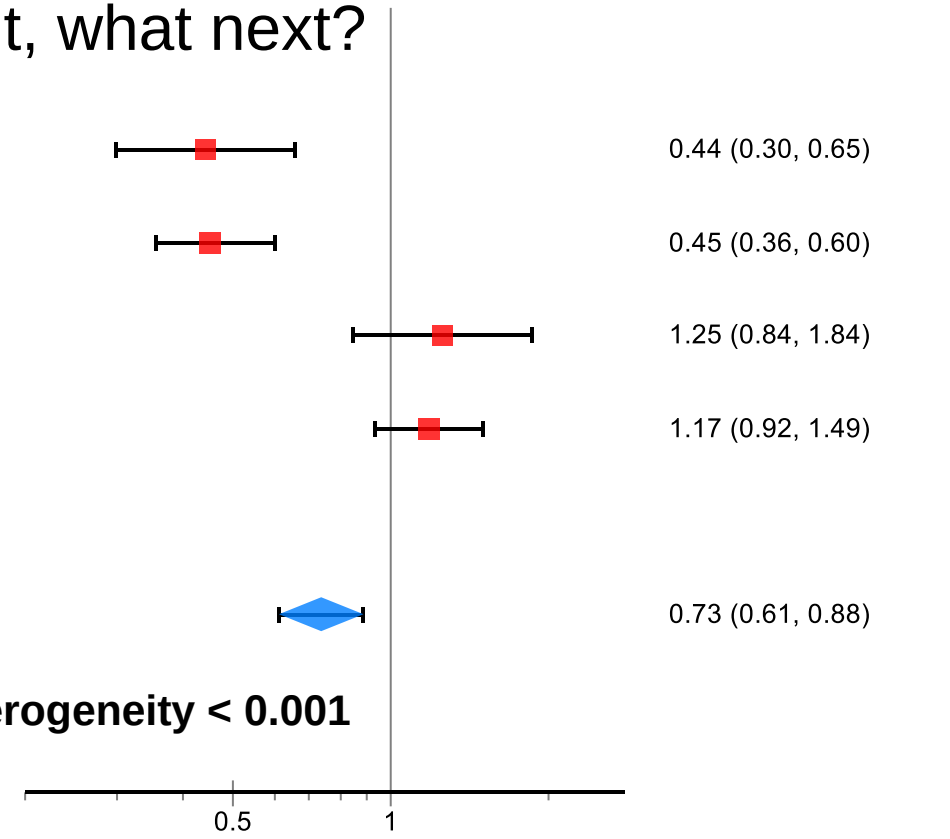
Homogenous

If this result, what next?



Heterogeneous

If this result, what next?



p-value for heterogeneity < 0.001
I^2=89%

Heterogeneity

- Look for explanation: Where?

- Patients
- Interventions
- Comparators
- Outcomes
- Risk of bias

- No good explanation? What to do?

- Decrease confidence in effect estimates

Relative Risk with 95% CI for Vitamin D Non-vertebral Fractures

Chapuy et al, (1994) 0.79 (0.69, 0.92)

Lips et al, (1996) 1.10 (0.87, 1.39)

Dawson-Hughes et al, (1997) 0.46 (0.24, 0.88)

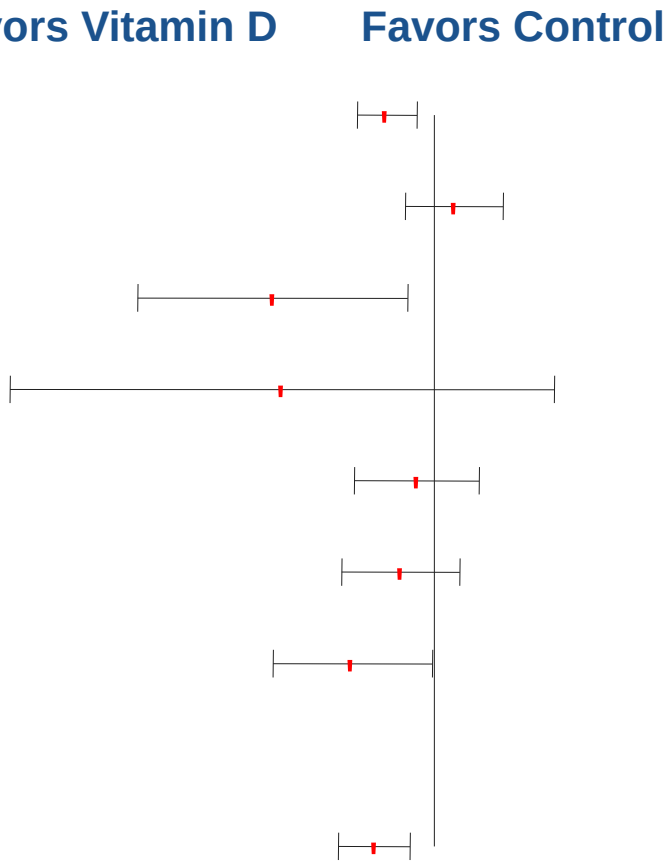
Pfeifer et al, (2000) 0.48 (0.13, 1.78)

Meyer et al, (2002) 0.92 (0.68, 1.24)

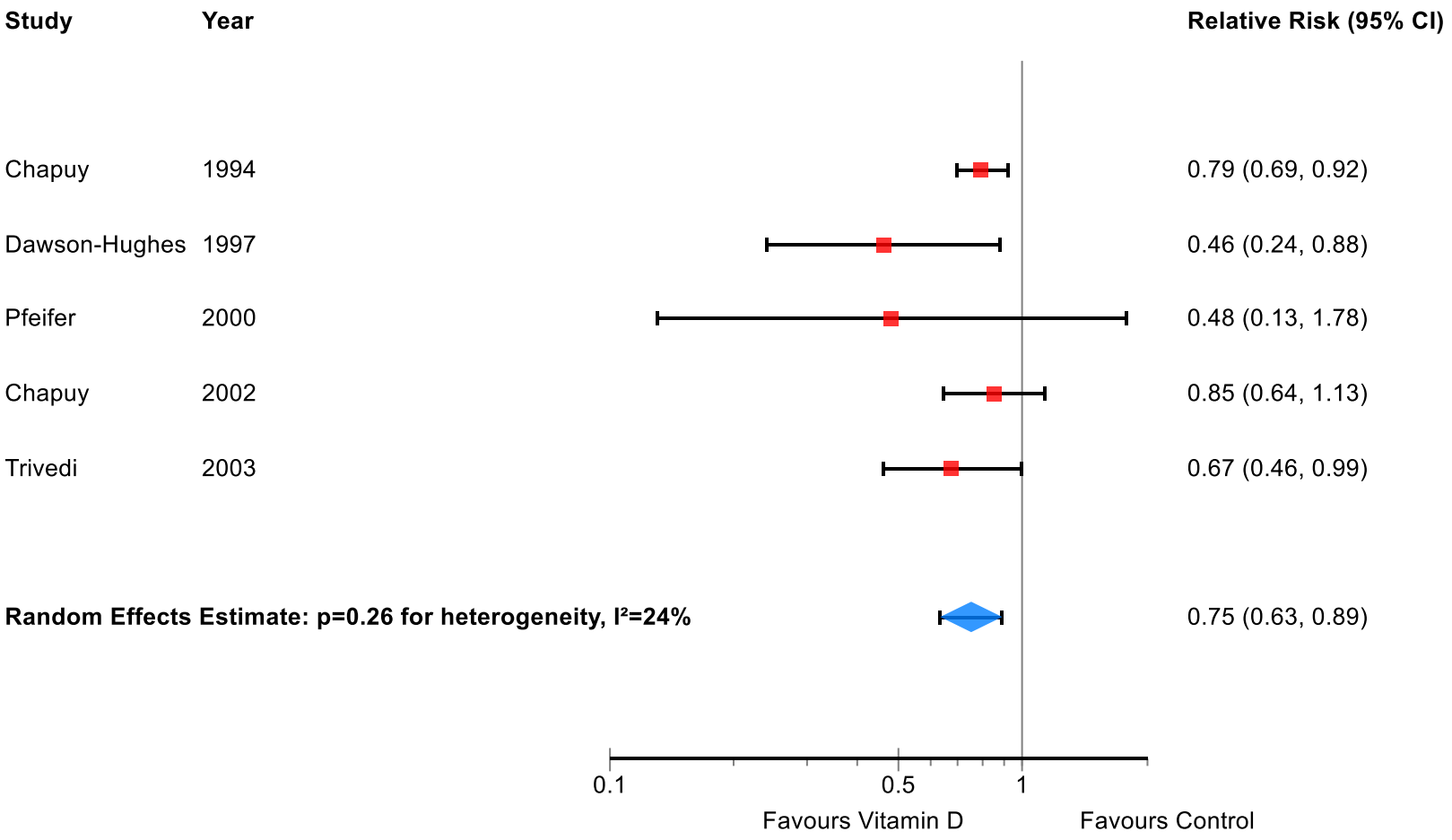
Chapuy et al, (2002) 0.85 (0.64, 1.13)

Trivedi et al, (2003) 0.67 (0.46, 0.99)

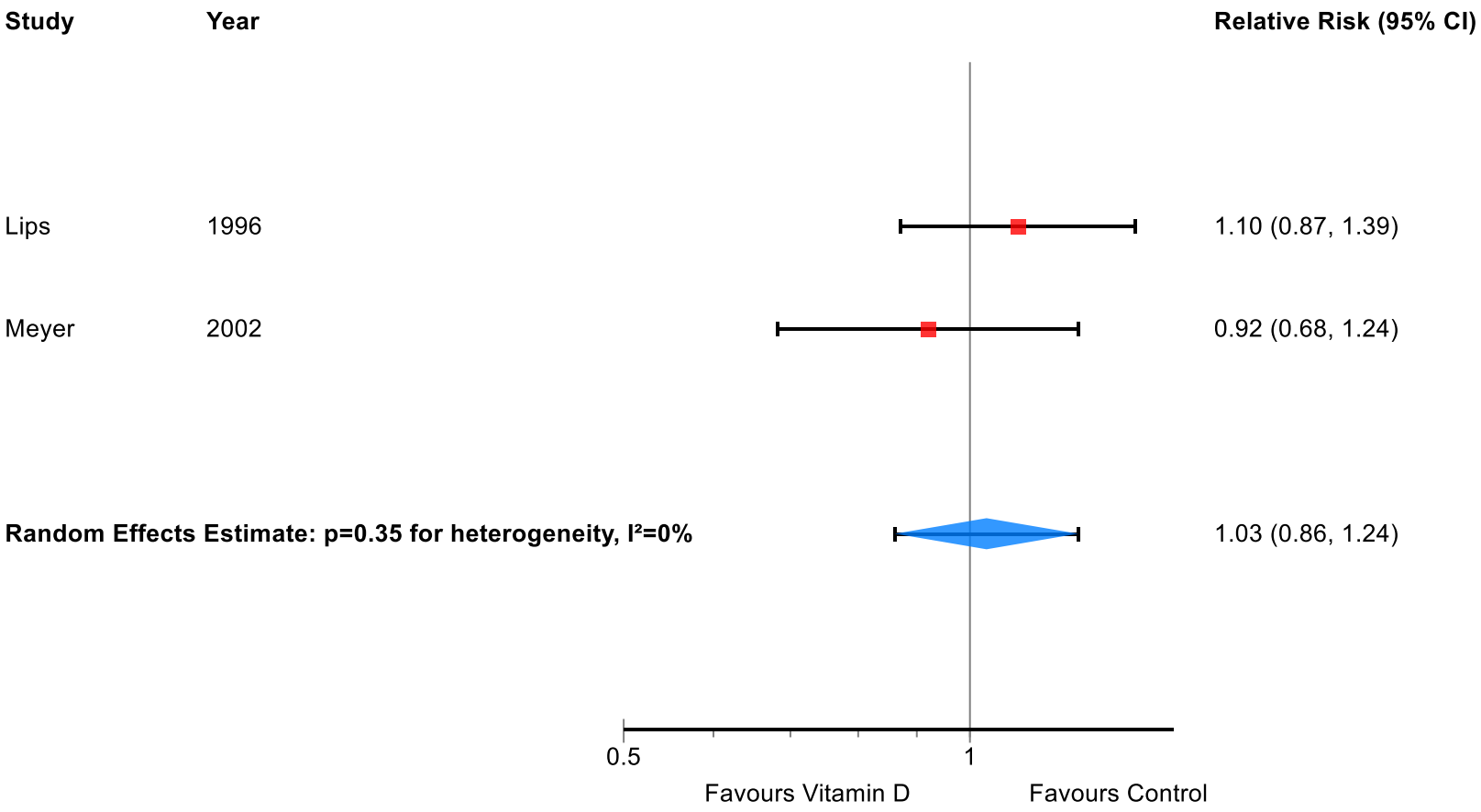
Pooled Random Effect Model
0.82 (0.69 to 0.98)
p= 0.05 for heterogeneity, I^2=53%



Relative Risk with 95% CI for Vitamin D (Non-Vertebral Fractures, Dose >400)



Relative Risk with 95% CI for Vitamin D (Non-Vertebral Fractures, Dose = 400)



Credibility of Subgroup Analsysis

- Within-study comparison?
- Unlikely chance
- A priori hypothesis, direction specified
- One of small number hypotheses
- Biologically compelling

Within and between study

Situation 1

- Study 1 includes only men
  - RR of outcome with treatment X: 0.5
- Study 2 includes only women
  - RR of outcome with treatment X: 1.0

Situation 2

- Study 1 includes both men and women
  - RR of outcome with treatment X in men: 0.5
  - RR of outcome with treatment X in women: 1.0

Within and between study

Situation 1

- Study 1 includes only men
  - RR of outcome with treatment X: 0.5
- Study 2 includes only women
  - RR of outcome with treatment X: 1.0

Possible explanations?

- Men were older, sicker, etc.
- Study 1 used different doses
- Study 1 failed to blind, high LFUP, etc.
- Chance
- Teatment x really does benefit men not women

Within and between study

Within study

- Study 1 treatment x benefits men
- Study 1 treatment x fails to benefit women

Possible explanations?

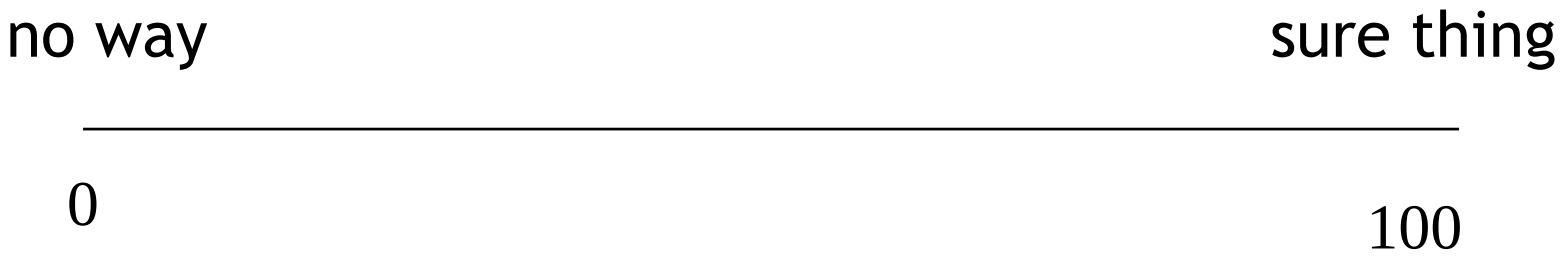
- Men were older, sicker – No
- Study 1 used different doses – No
- Study 1 failed to blind, high LFUP, etc. – No
- Chance
- Treatment x really does benefit men not women

Within-study much stronger than between

Believe sub-group analysis high vs low dose vitamin?

- Unlikely chance — *p = 0.006*
- Consistent across studies — *yes*
- Small # a priori direction right — *yes*
- Biologically compelling — *yes*
- Within-study comparison — *no*

Credibility of sub-group analysis

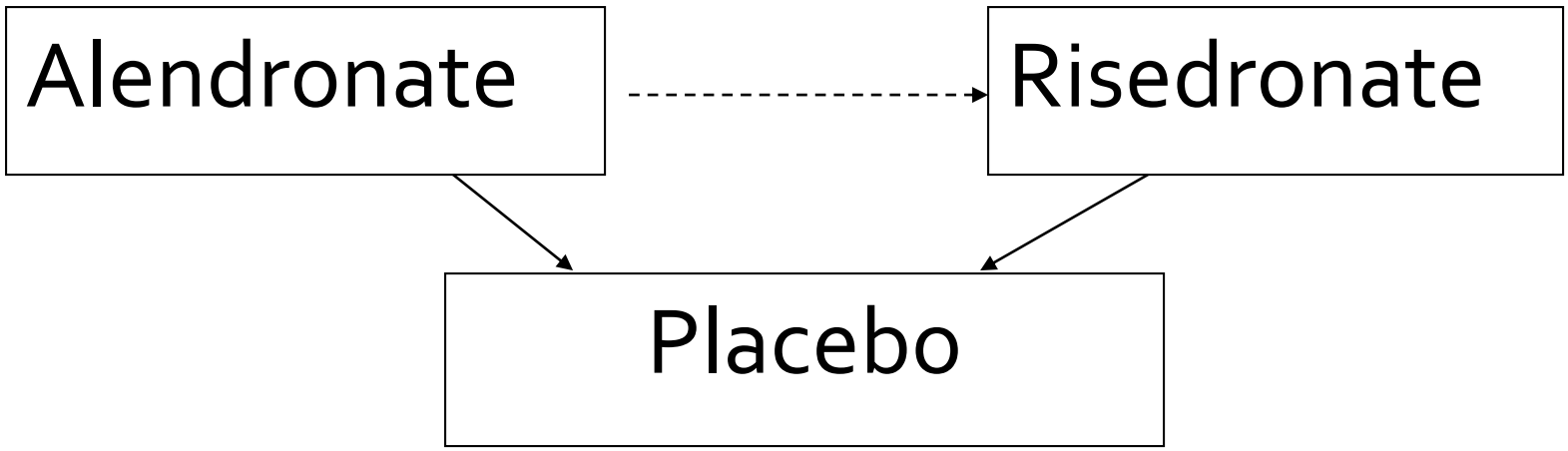


Confidence judgments: Directness

- Populations
  - Older, sicker or more co-morbidity
- Interventions
  - Warfarin in trials vs clinical practice
- Comparators
  - Standard care
- Outcomes
  - Important versus surrogate outcomes
  - Glucose control versus CV events

Directness

Interested in A versus B
available data A vs C, B vs C

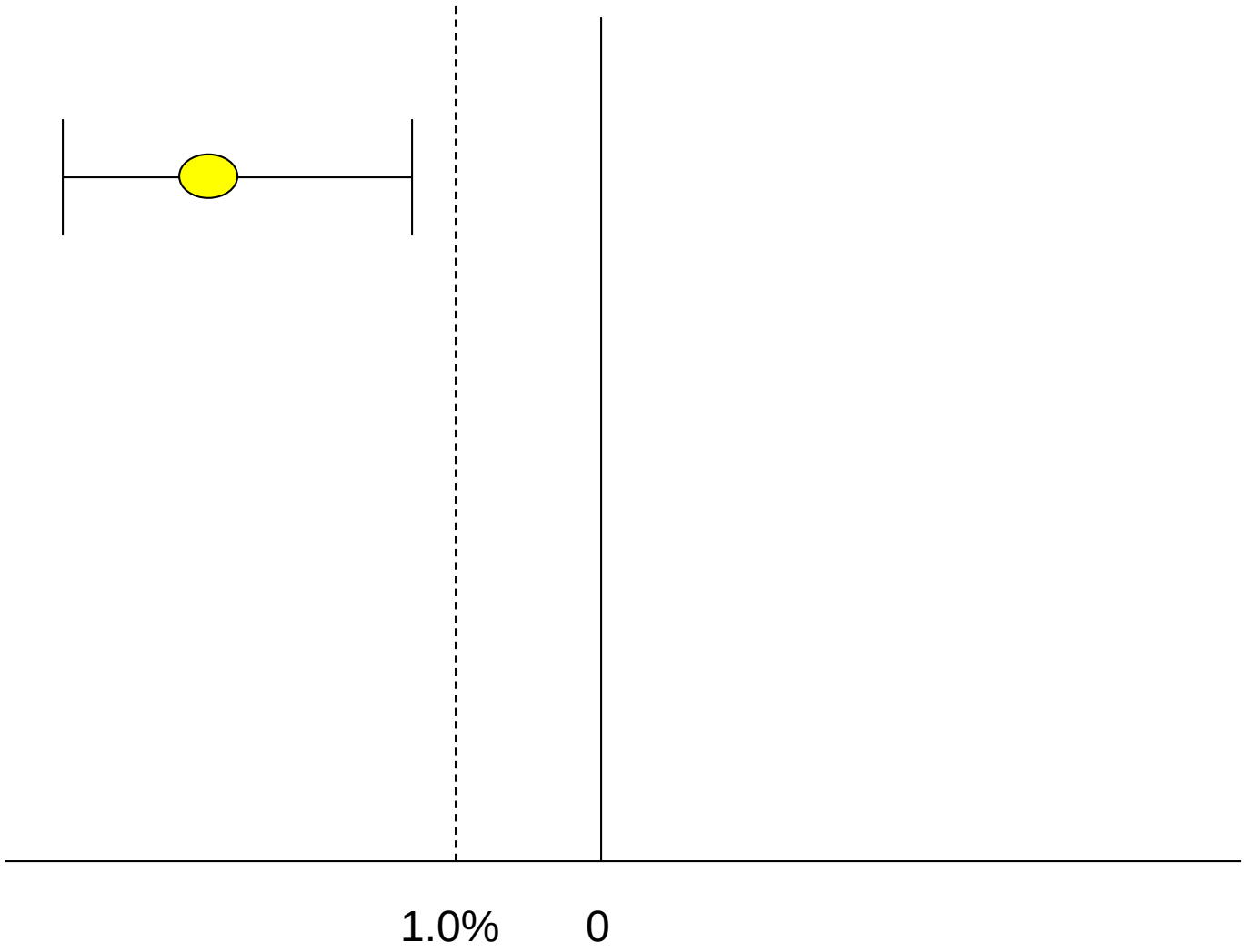


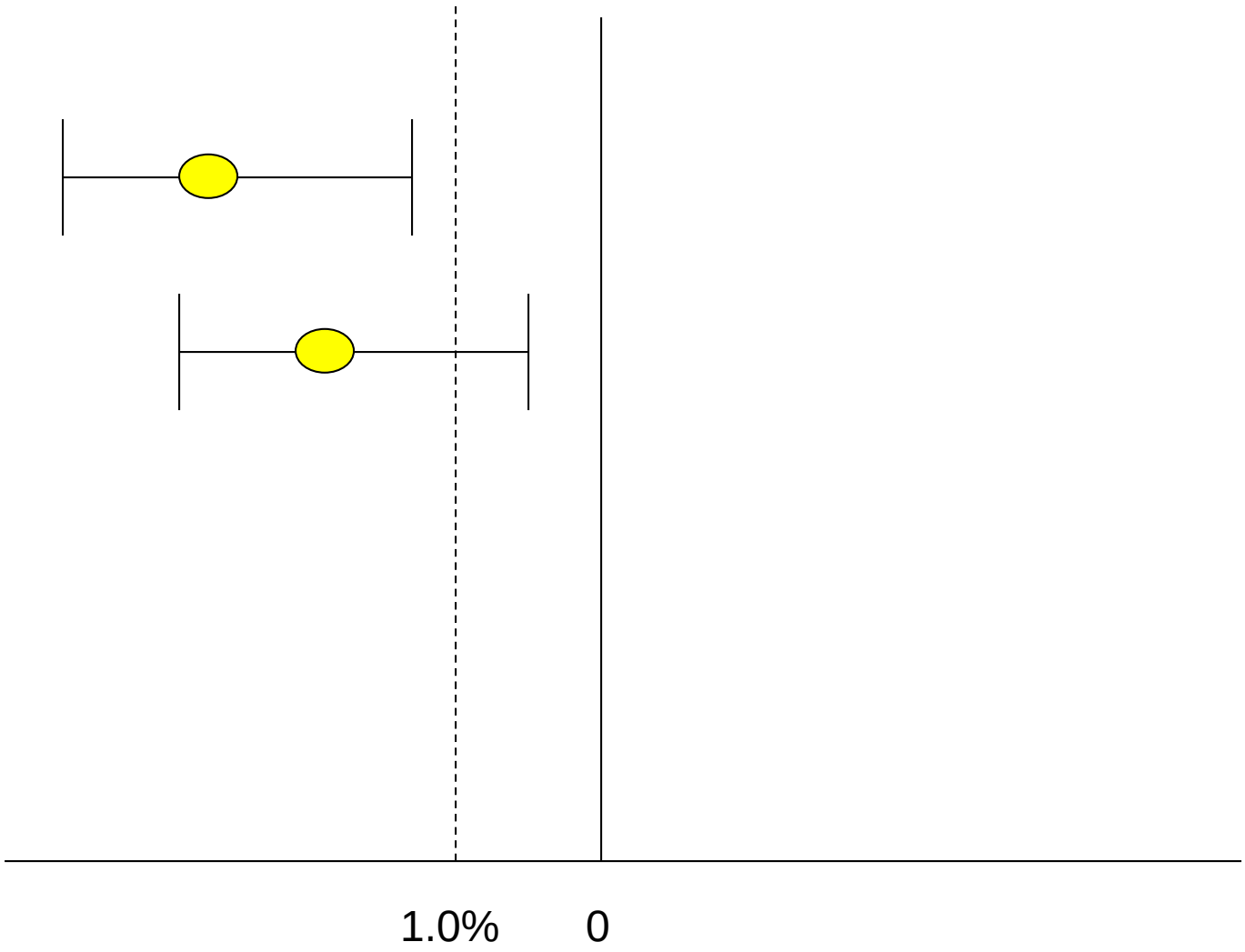
Imprecision

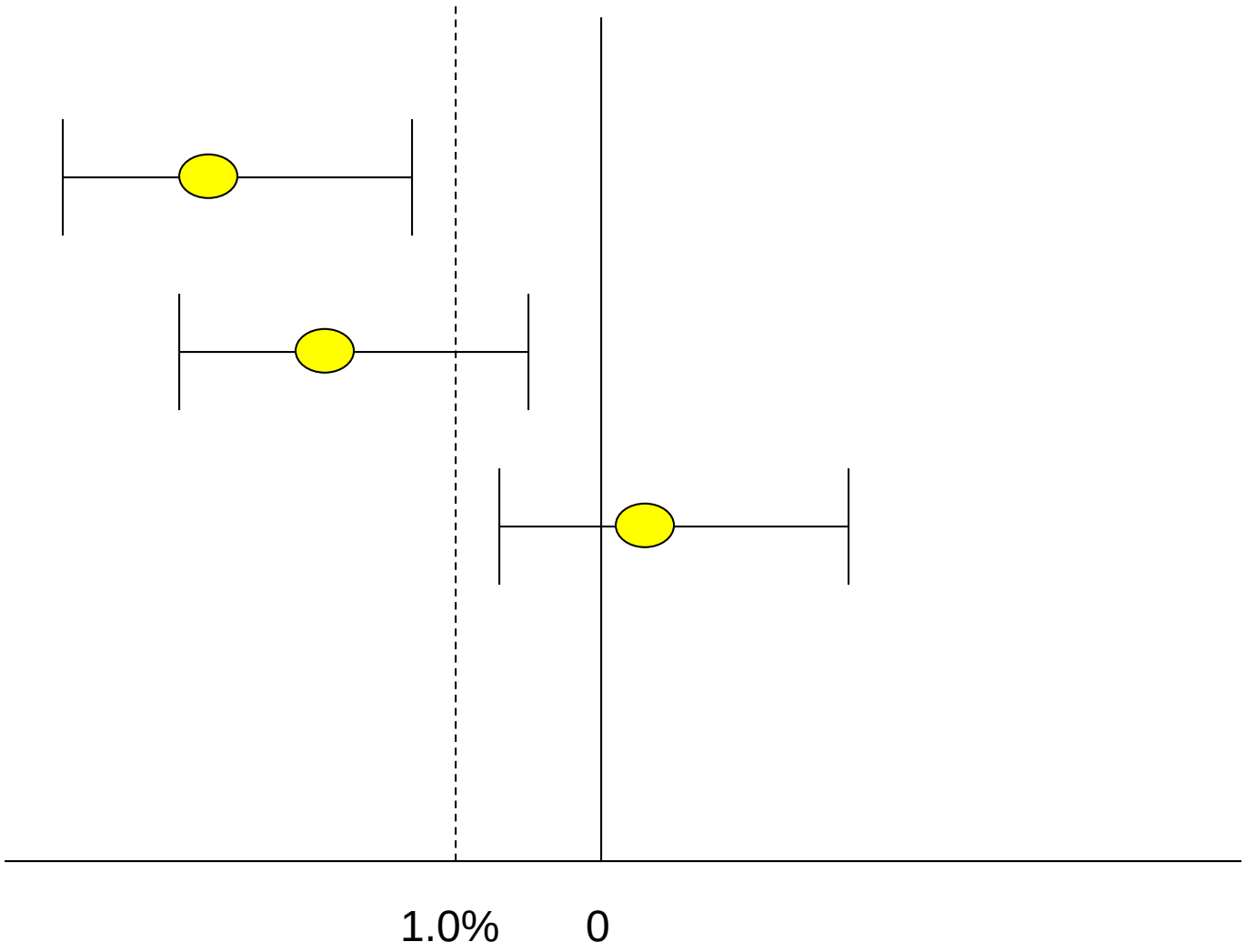
- Small sample size
  - Small number of events
- Wide confidence intervals
  - Uncertainty about magnitude of effect
- How do you decide what is too wide?

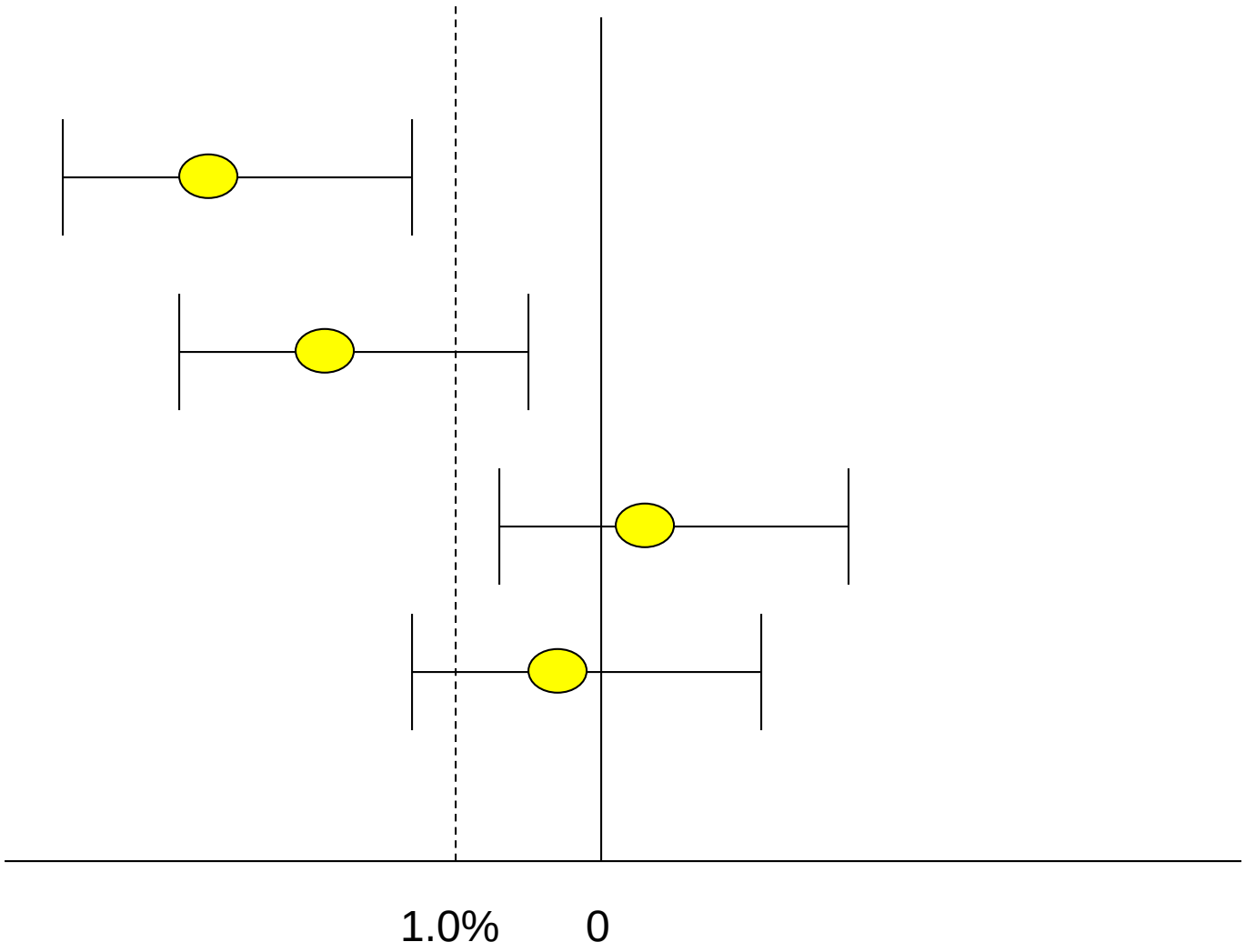
Precision

- Atrial fib at risk of stroke
- Anticoagulants increases serious gi bleeding
  - 3% per year
- 1,000 patients 1 less stroke
  - 30 more bleeds for each stroke prevented
- 1,000 patients 100 less strokes
  - 3 strokes prevented for each bleed
- Where is your threshold?
  - How many strokes in 100 with 3% bleeding?


1.0%
0


1.0%
0


1.0%
0


1.0%
0

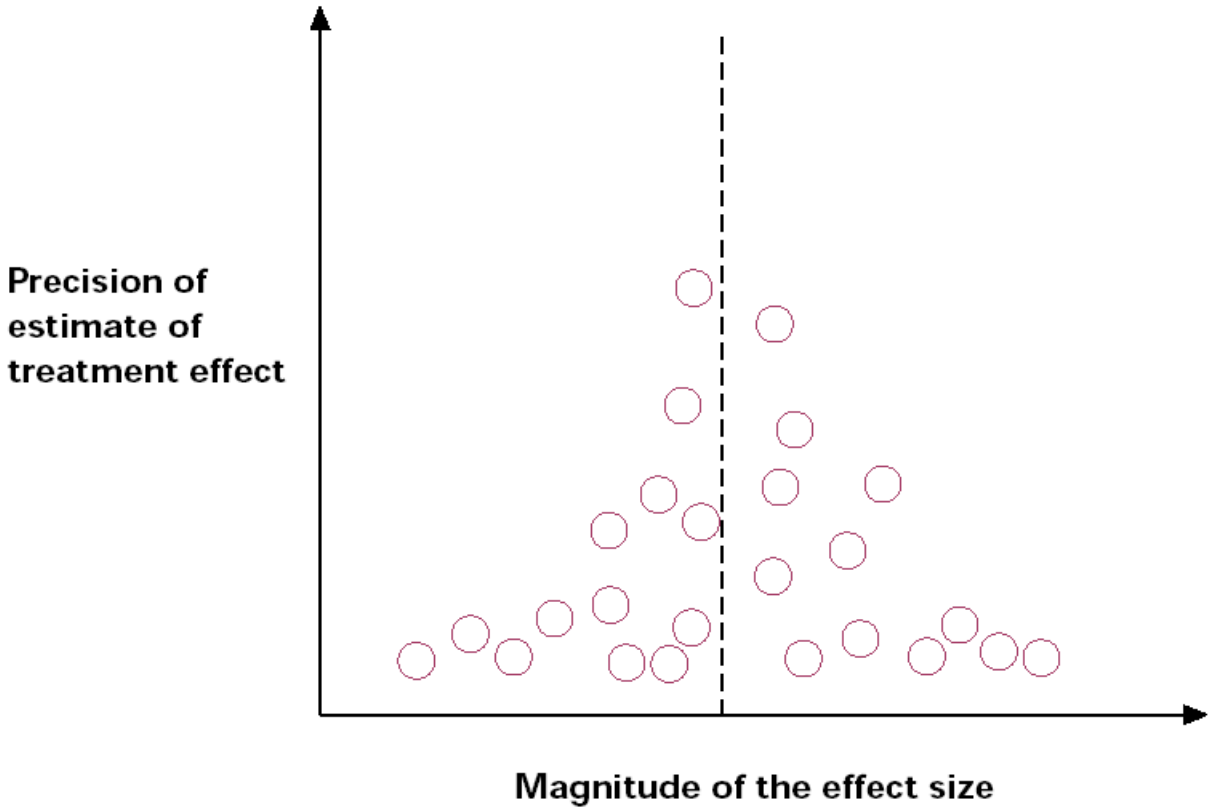
Imprecision – additional problem

- Small trials, large effect
  - Likely to be overestimate
- Analogy to stopping early
- Lack of prognostic balance
- Solution: optimal information size
  - # of pts from conventional sample size calculation
  - specify control group risk, α, β, Δ

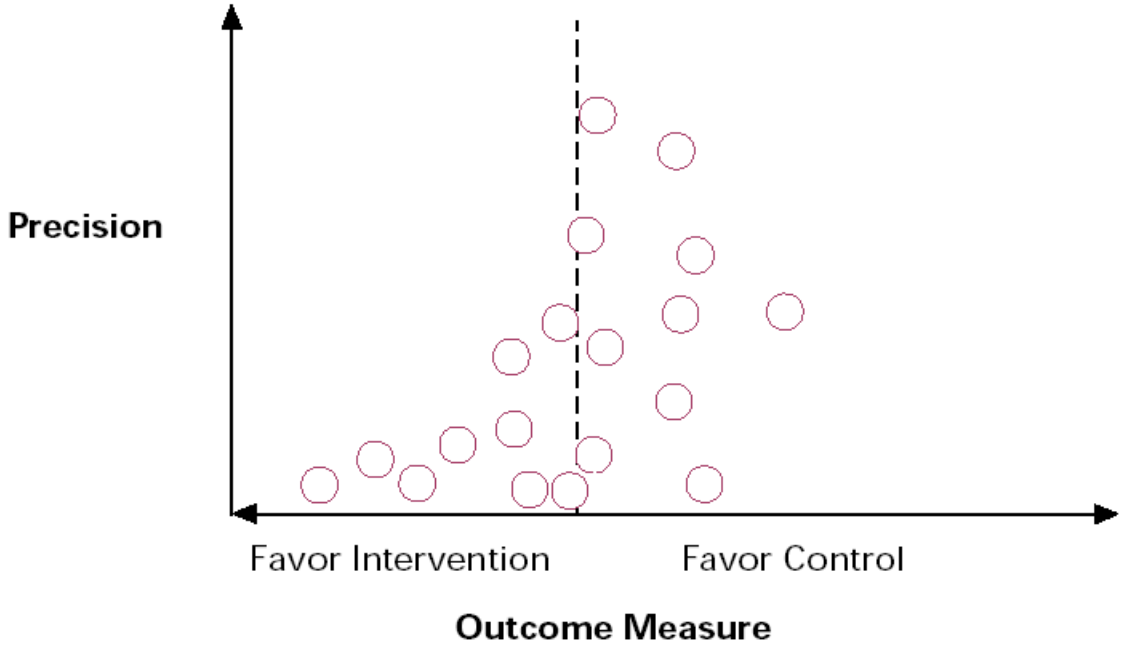
Publication bias

- High likelihood could lower quality
- When to suspect
  - Number of small studies
  - Industry sponsored

Funnel Plot


Precision of estimate of treatment effect
Magnitude of the effect size

Publication Bias



Funnel Plot

Fish oil on mortality

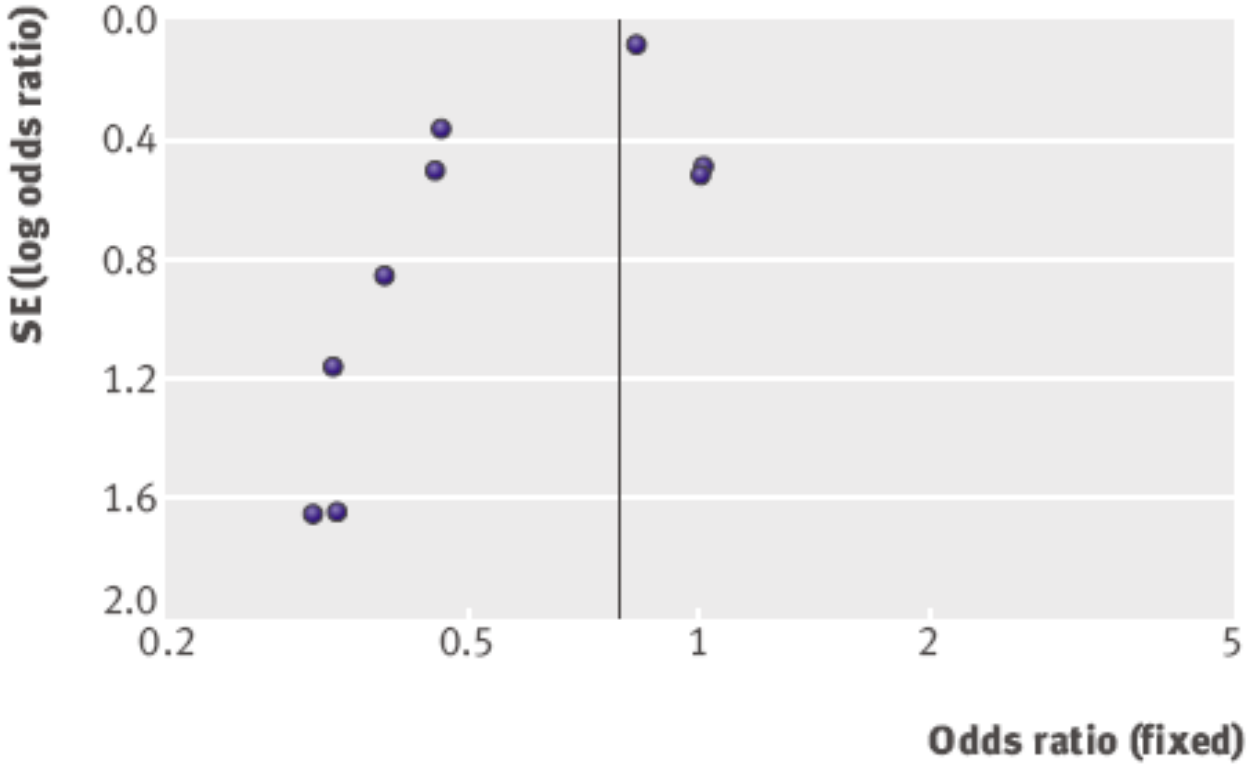


Fig 4 | Funnel plot for assessment of publication bias for death from cardiac causes in 11 included studies reporting data on this outcome

What can raise confidence?

- Clinicians: no RCTs, high certainty?
- Large magnitude can rate up one level
  - Very large two levels
- Common criteria
  - Everyone used to do badly
  - Almost everyone does well
  - Quick action
- Hip replacement for hip osteoarthritis

Dose-response gradient

- Childhood lymphoblastic leukemia
- Risk for CNS malignancies 15 years after cranial irradiation
- No radiation: 1% (95% CI 0% to 2.1%)
- 12 Gy: 1.6% (95% CI 0% to 3.4%)
- 18 Gy: 3.3% (95% CI 0.9% to 5.6%).

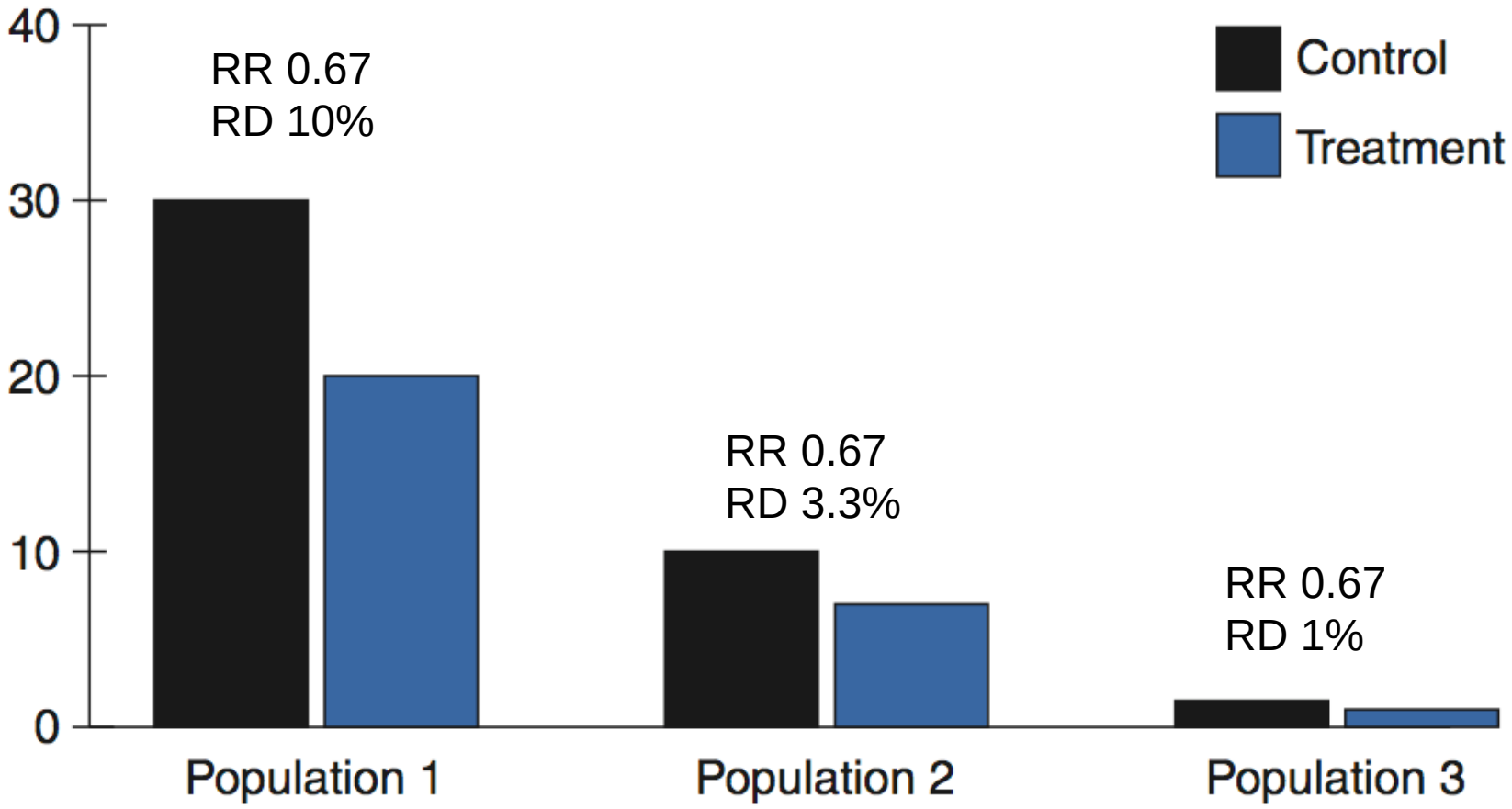
Certainty assessment criteria

Study Design	Confidence in estimates	Lower if	Higher if
Randomized trials	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 + 1 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 studies	Low		
	Very Low		

Trading off

- What do patients/clinicians need to know
  - Relative risk reduction?
  - Absolute risk difference?
- Why do meta-analyses always report relative?

Constant Relative Risk With Varying Risk Differences



Trading off

- What do patients/clinicians need to know
  - Relative risk reduction?
  - Absolute risk difference?
- Why do meta-analyses always report relative?
- Body of evidence
  - How do we get risk difference?

How to get absolute?

- Meta-analysis get pooled relative risk
- Obtain baseline risk and multiply
- BR 10%, RRR 50%, RD 5%

		Quality Assessment					Summary of Findings			
							Quality	Relative Risk (95% CI) p-value	Illustrative risks	
Outcome	No. of patients (studies)	Risk of Bias	Inconsistency	Indirectness	Imprecision	Publication Bias			control rate	vaccinated rate
Zoster episodes	38,546 (1)	No serious risk	only one study	Direct	Precise	Undetected	High	not reported	11.12 per 1,000 patient-years	5.42 (difference 5.7 per 1,000 pt-years (p< 0.001)
Post-herpetic neuralgia	38,546 (1)	No serious risk	only one study	Direct	Precise	Undetected	High	not reported	1.38 per 1,000 patient-years	0.46 (difference 0.92 per 1,000 pt-years (p< 0.001)
Serious adverse events	38,546 (1)	No serious risk	only one study	Direct	Precise	Undetected	High	Not reported	13 per 1,000	19 (difference 6 per 1,000)

Zoster vaccine

Beta blockers in non-cardiac surgery

Quality Assessment							Summary of Findings		
Outcome	Number of participants (studies)	Risk of Bias	Consistency	Directness	Precision	Publication Bias	Quality	Relative Effect (95% CI)	Absolute risk difference
Myocardial infarction	10,125 (9)	No serious limitations	No serious imitations	No serious limitations	No serious limitations	Not detected	High	0.71 (0.57 to 0.86)	1.5% fewer (0.7% fewer to 2.1% fewer)
Mortality	10,205 (7)	No serious limitations	No serious limiations	No serious limitations	Imprecise	Not detected	Moderate	1.23 (0.98 – 1.55)	0.5% more (0.1% fewer to 1.3% more)
Stroke	10,889 (5)	No serious limitaions	No serious limitations	No serious limitations	Serious limitations	Not detected	Moderte	1.67 (1.00 – 2.80)	0.3% more (0 more to 1.5% more)

Overall level of evidence

- Most systems just use evidence about primary benefit outcome
- But what about others (risk)?
- What to do?
- Options
  - Ignore all but primary
  - Lowest of any outcome
  - Some blended approach
  - Lowest of critical outcomes

Strength of Recommendation

- Strong recommendation
  - Benefits clearly outweigh risks/hassle/cost
  - Risk/hassle/cost clearly outweighs benefit

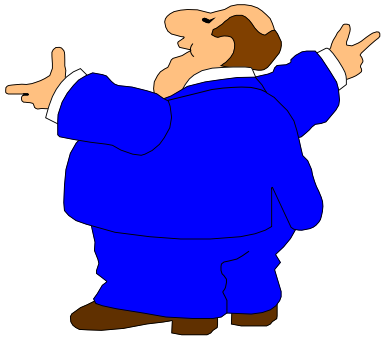
- What can downgrade strength?
- Low confidence in estimates
- Close balance between up and downsides

Risk/Benefit tradeoff

- Aspirin after myocardial infarction
  - 25% reduction in relative risk
  - side effects minimal, cost minimal
  - benefit obviously much greater than risk/cost
- Anticoagulants in low risk atrial fibrillation
  - anticoagulants reduce stroke vs ASA by 50%
  - but if risk only 1% per year, ARR 0.5%
  - increased bleeds by 1% per year

Strength of Recommendations

Aspirin after MI – do it



Anticoagulant rather than ASA in Afib

-- probably do it

-- probably don't do it

Determinants of strength of recommendation

Factor	Comment
Balance between desirable and undesirable effects	The larger the difference between the desirable and undesirable effects, the higher the likelihood that a strong recommendation is warranted. The narrower the gradient, the higher the likelihood that a weak recommendation is warranted
Quality of evidence	The higher the quality of evidence, the higher the likelihood that a strong recommendation is warranted
Values and preferences	The more values and preferences vary, or the greater the uncertainty in values and preferences, the higher the likelihood that a weak recommendation is warranted
Costs (resource allocation)	The higher the costs of an intervention—that is, the greater the resources consumed—the lower the likelihood that a strong recommendation is warranted

Additional criteria in evidence to decision frameworks;
Importance of the problem
Acceptability
Feasibility
Equity

Significance of strong vs weak

- Variability in patient preference
  - strong, almost all same choice (> 90%)
  - weak, choice varies appreciably
- Interaction with patient
  - strong, just inform patient
  - weak, ensure choice reflects values
- Use of decision aid
  - strong, don't bother; weak, use the aid
- Quality of care criterion
  - strong, consider; weak, don't consider

Flavanoids for Hemorrhoids

- Venotonic agents
- Popularity
  - 90 venotonics commercialized in France
  - None in Sweden and Norway
  - France 70% of world market
- Possibilities
  - French misguided
  - Rest of world missing out

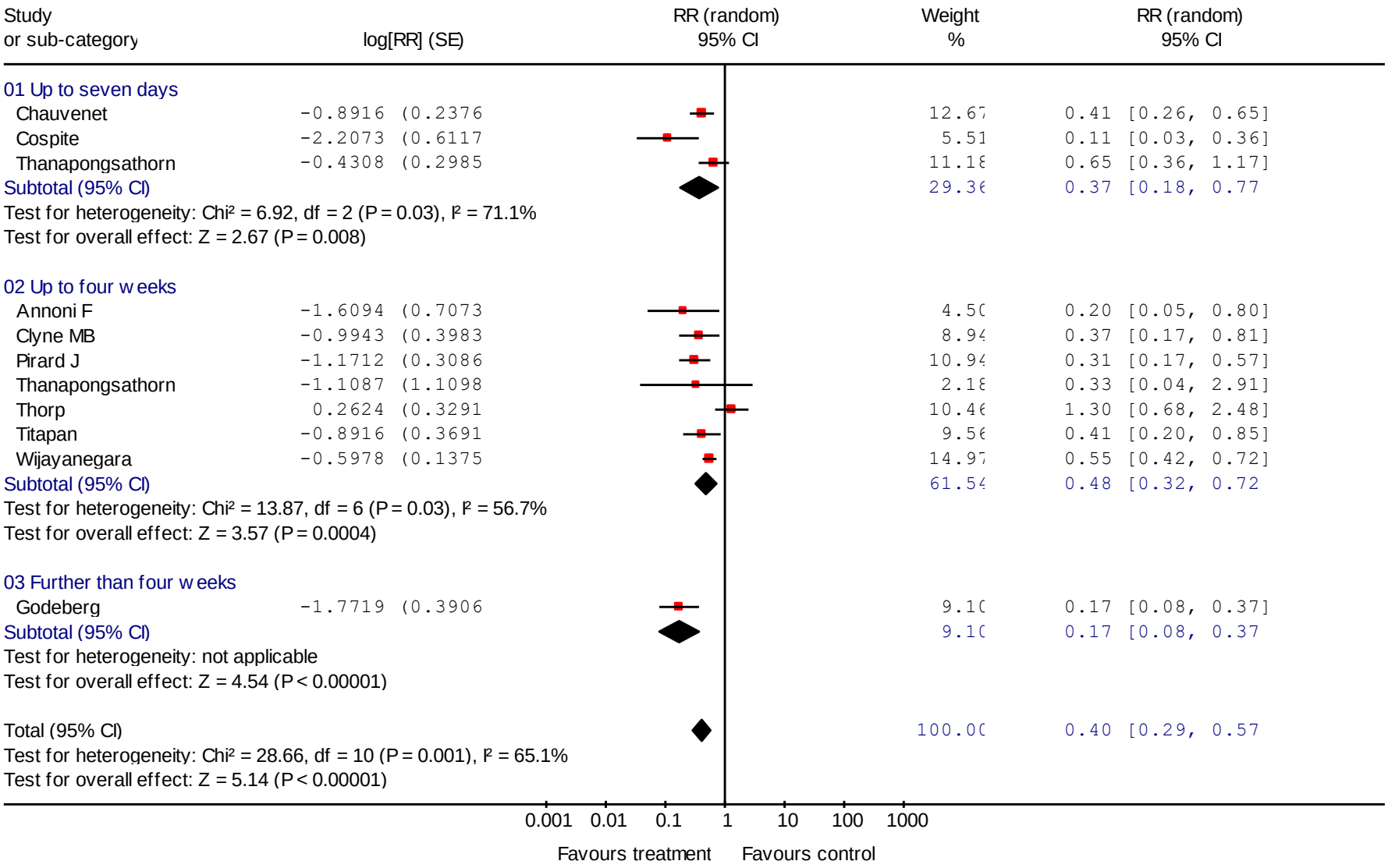
Systematic review

- 14 trials, 1432 patients
- Key outcome
  - Risk not improving/persistent symptoms
  - 11 studies, 1002 patients, 375 events
  - RR 0.4, 95% CI 0.29 to 0.57
- Minimal side effects
- Is France right?
- What is the certainty of evidence?

What can lower confidence?

- Risk of bias
  - Lack of detail re concealment
  - Questionnaires not validated
- Indirectness – no problem
- Inconsistency, need to look at the results

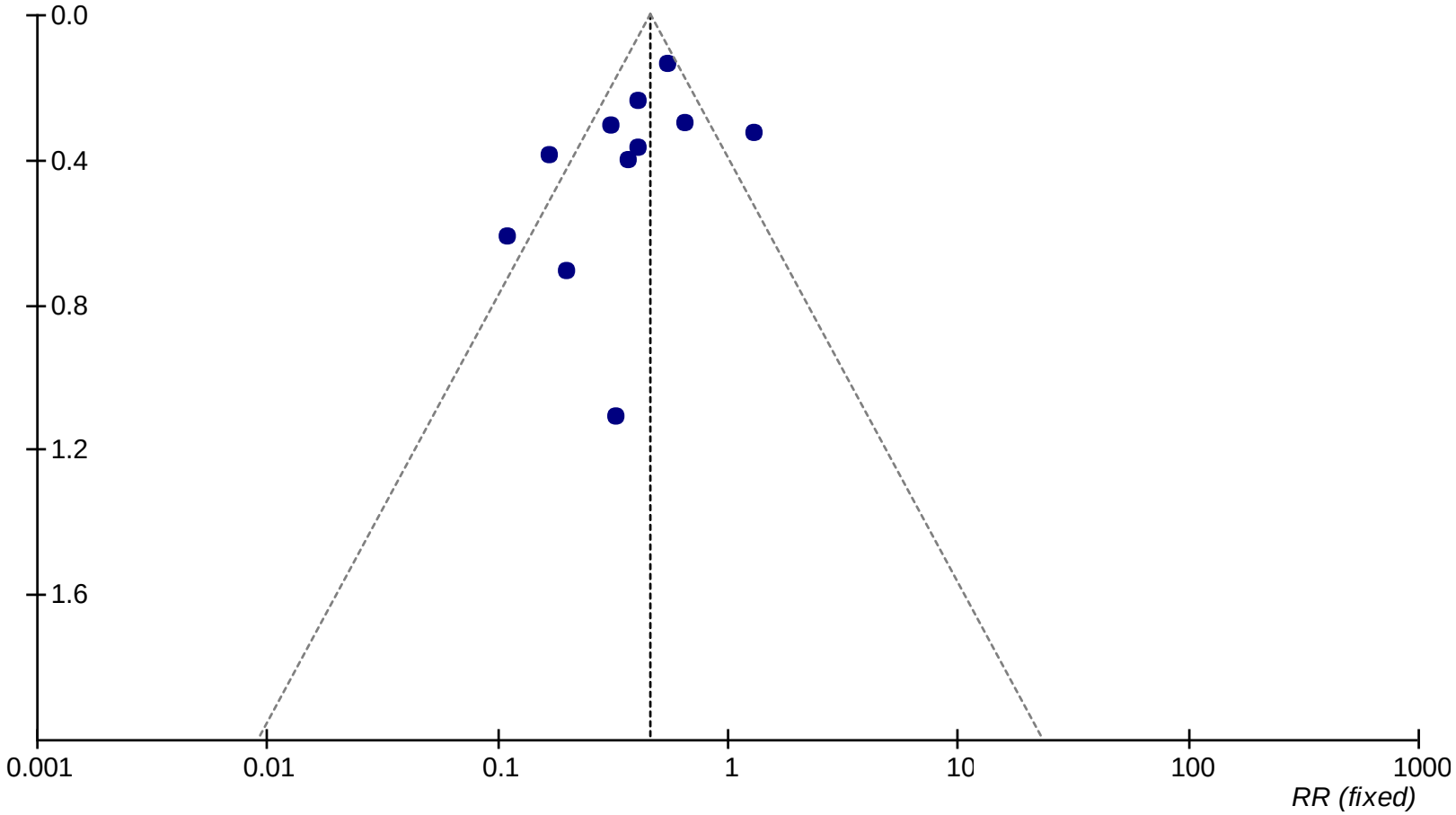
Review: Phlebotonics for hemorrhoids
Comparison: 01 Venotonics vs placebp
Outcome: 08 Overall improvement: no improvement/some improvement



Publication Bias

- Size of studies
  - 40 to 234 patients, most around 100
- All industry sponsored

Review: Phlebotonics for hemorrhoids
Comparison: 01 Venotonics vs placebp
Outcome: 08 Overall improvement: no improvement/some improvement



What can lower certainty?

- Risk of bias
  - Lack of detail re concealment
  - Questionnaires not validated
- Inconsistency
  - Almost all show positive effect, trend
  - Heterogeneity $p < 0.001$; I^2 65.1%
- Indirectness
- Imprecision
  - RR 0.4, 95% CI 0.29 to 0.57
- Publication bias
  - 40 to 234 patients, most around 100

Is France right?

- Recommendation
  - Yes
  - No against use
- Strength
  - Strong
  - Weak

Conclusion

- Systematic review, HTA need quality evidence
- Guideline need recommendation strength
- GRADE very widely increasingly used
- Transparent, explicit to quality, strength

- Do you tweet?

@EBCPMcMaster

Follow us on Twitter