Interpreting results of Cochrane reviews and Summary of Findings Tables: GRADE and SoF Workshop

Monday 22 September 2014
Cochrane Colloquium Hyderabad, India
0900-1700

GRADE

Introductions

Nancy Santesso
Elie Akl
Alonso Carrasco
Maria Ximena Rojas Reyes
Madison Zhang

on behalf of the
Cochrane Applicability and Recommendations Method Group

GRADE

Agenda

09.00 Introductions
09.10 Overview: GRADE in the systematic review process
09.30 Background to statistics
10.15 Coffee break
10.30 Example: working in pairs
10.45 How to GRADE the evidence: interactive examples
12.00 Lunch break
12.45 Using the GRADE criteria: small groups
13.45 Hands on use of GDT
15.00 Coffee break
15.15 Hands on use of GDT
15.45 GRADE in unique situations
16.45 Summary and Questions
17.00 Close

GRADE

Overview

GRADE

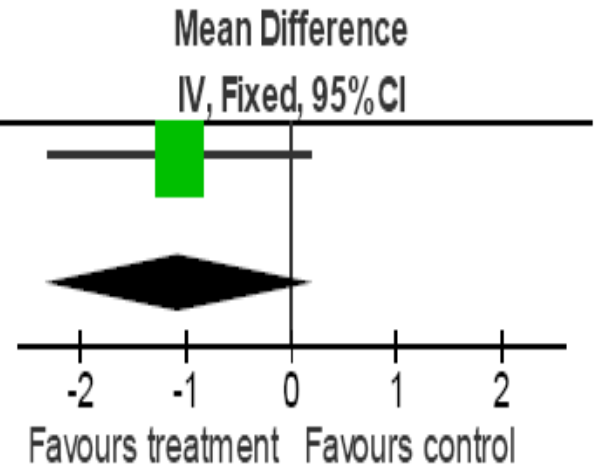
One study: Effect of drug A on pain

1.1 Pain (0-10) 3 weeks follow up

Study or Subgroup	Treatment Mean	SD	Total	Control Mean	SD	Total	Weight	Mean Difference IV, Fixed, 95% CI
Kjeldsen-Kragh 1991	2.64	1.98	17	3.7	1.75	17	100.0%	-1.06 [-2.32, 0.20]
Total (95% CI)			17			17	100.0%	-1.06 [-2.32, 0.20]

Heterogeneity: Not applicable

Test for overall effect: Z = 1.65 (P = 0.10)



GRADE

Narrative synthesis

Acupuncture versus sham in people with chronic back pain

- Two studies measured pain.
- One study (85 people) reported ‘no significant difference’, in number of persons who reported improvement of pain.
- One study (34 people) reported a difference of 4 points on a scale of 24.

GRADE

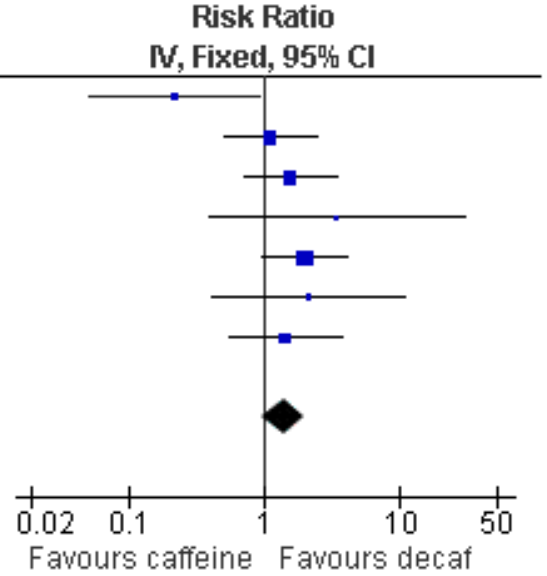
Forest Plot

Effect of caffeine on headache at 24 hours

Study or Subgroup	Caffeinated coffee Events	Total	Decaffeinated coffee Events	Total	Weight	Risk Ratio IV, Fixed, 95% CI
Amore-Coffea 2000	2	31	10	34	6.6%	0.22 [0.05, 0.92]
Deliciozza 2004	10	40	9	40	21.9%	1.11 [0.51, 2.44]
Mama-Kaffa 1999	12	53	9	61	22.2%	1.53 [0.70, 3.35]
Morrocona 1998	3	15	1	17	2.9%	3.40 [0.39, 29.31]
Norscafe 1998	19	68	9	64	26.4%	1.99 [0.97, 4.07]
Oohlahlazza 1998	4	35	2	37	5.1%	2.11 [0.41, 10.83]
Piazza-Allerta 2003	8	35	6	37	14.9%	1.41 [0.54, 3.65]
Total (95% CI)		**277**		**290**	**100.0%**	**1.38 [0.96, 2.00]**
Total events	58		46			

Heterogeneity: $Chi^2 = 8.58$, $df = 6$ ($P = 0.20$); $I^2 = 30\%$

Test for overall effect: $Z = 1.73$ ($P = 0.08$)



GRADE

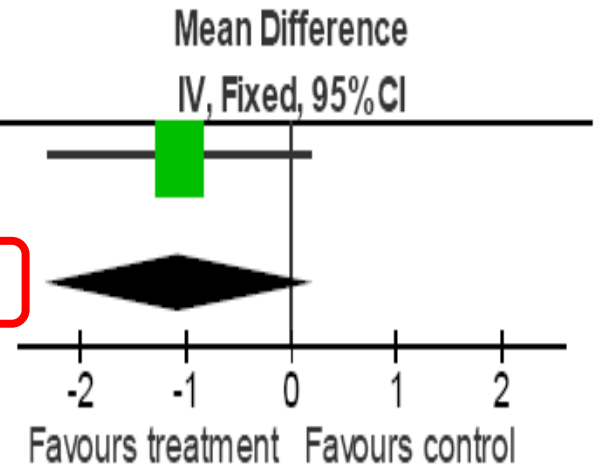
Conclusion?

1.1 Pain (0-10) 3 weeks follow up

Study or Subgroup	Treatment Mean	SD	Total	Control Mean	SD	Total	Weight	Mean Difference IV, Fixed, 95% CI
Kjeldsen-Kragh 1991	2.64	1.98	17	3.7	1.75	17	100.0%	-1.06 [-2.32, 0.20]
Total (95% CI)			17			17	100.0%	-1.06 [-2.32, 0.20]

Heterogeneity: Not applicable

Test for overall effect: Z = 1.65 (P = 0.10)



Drug A reduces pain. Do you believe it?

GRADE

Systematic review process

1. define the question
2. plan eligibility criteria
3. plan methods
4. search for studies
5. apply eligibility criteria
6. collect data
7. assess studies for risk of bias
8. analyze and present results
9. **interpret results and draw conclusions**
10. improve and update review

GRADE

Two main concepts when interpreting results and drawing conclusions

Magnitude of Effect

Confidence in effect

Quality of evidence

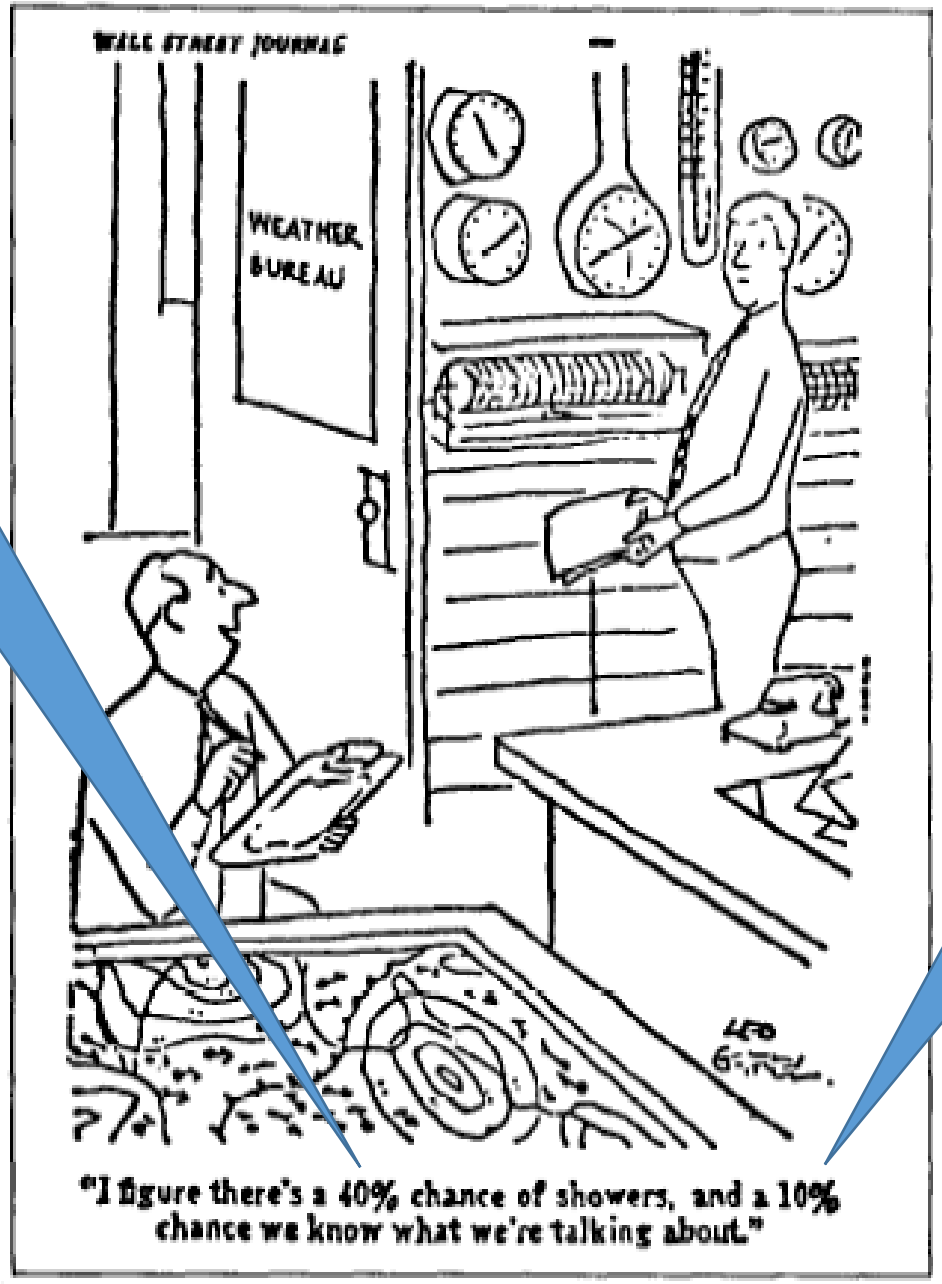


Figure 1. *Belief and confidence: a two-dimensional weather report. (Reprinted by permission from the Wall Street Journal).*

How do we….

…interpret results and draw conclusions?

GRADE criteria *(MECIR standards: mandatory)*

….present results to reader/users?

Summary of Findings Tables *(MECIR standards: highly desirable)*

GRADE

What should I conclude?

Should I believe the effect that I found?

GRADE

Example:

Intravenous magnesium sulfate for treating adults with acute asthma in the emergency department (Review)

GRADE

Analysis

Outcome: 1 Hospital admissions

Study or subgroup	Magnesium n/N	Placebo n/N	Odds Ratio M-H,Fixed,95% CI	Weight	Odds Ratio M-H,Fixed,95% CI
Bilaceroglu 2001	10/40	17/41		6.3 %	0.47 [0.18, 1.21]
Bloch 1995	17/67	24/68		8.9 %	0.62 [0.30, 1.31]
Boonyavorakul 2000	3/17	4/16		1.7 %	0.64 [0.12, 3.46]
Bradshaw 2007	49/62	52/67		5.3 %	1.09 [0.47, 2.52]
Goodacre 2013	279/394	278/358		42.6 %	0.70 [0.50, 0.97]
Green 1992	13/58	11/62		4.1 %	1.34 [0.55, 3.29]
Matusiewicz 1994	45/64	47/67		6.8 %	1.01 [0.48, 2.13]
Porter 2001	5/18	5/24		1.6 %	1.46 [0.35, 6.08]
Silverman 2002	39/122	41/126		13.8 %	0.97 [0.57, 1.66]
Singh 2008	2/30	9/30		4.2 %	0.17 [0.03, 0.85]
Skobeloff 1989	7/19	15/19		4.7 %	0.16 [0.04, 0.66]
Total (95% CI)	**891**	**878**		**100.0 %**	**0.75 [0.60, 0.92]**

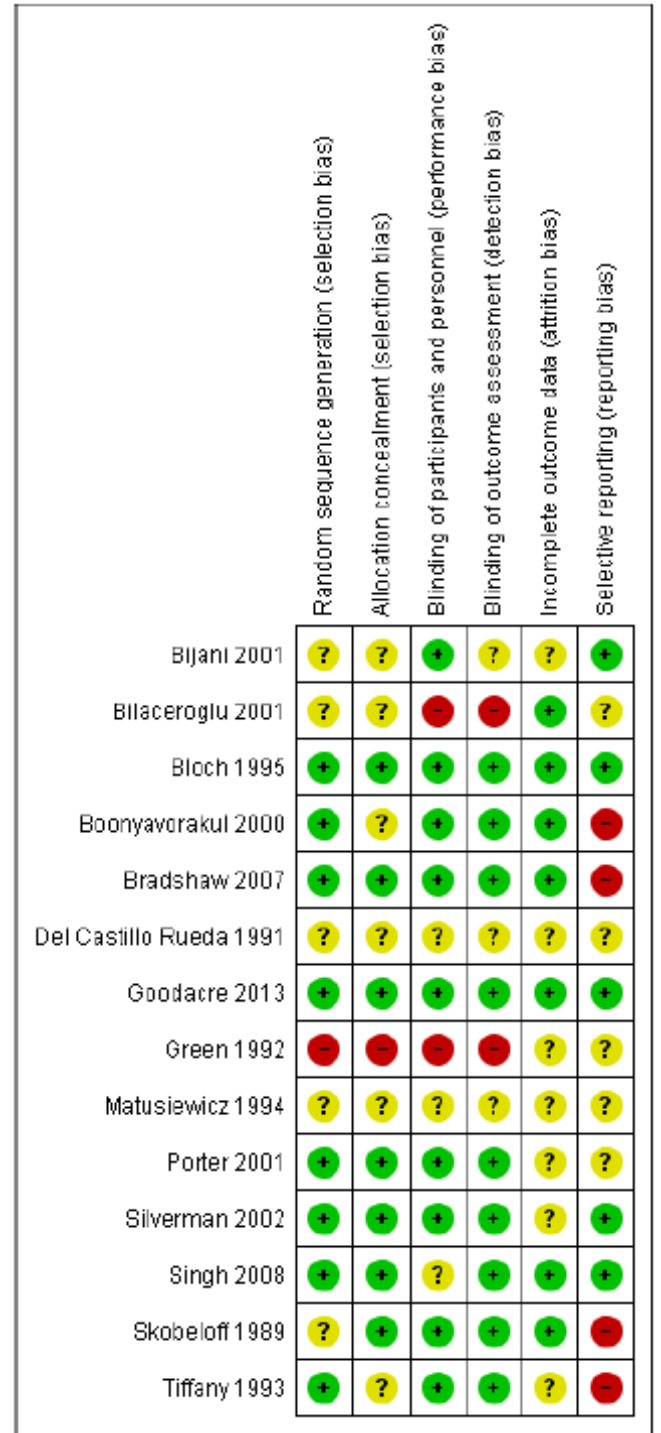
Total events: 469 (Magnesium), 503 (Placebo)

Heterogeneity: $Chi^2 = 13.93$, df = 10 (P = 0.18); $I^2 = 28\%$

Test for overall effect: Z = 2.72 (P = 0.0066)

GRADE

Risk of bias assessment



GRADE

Results section:

Hospital admissions

Combining 11 studies (n = 972) revealed a significant reduction in hospital admissions compared with placebo (OR 0.75, 95% CI 0.60 to 0.92; high-quality evidence; Analysis 1.1). Some heterogeneity that was not statistically significant was observed (I2 = 28%; P value 0.18). In absolute terms, this odds ratio translates to a reduction of seven hospital admissions for every 100 adults (95% CI two to 13 fewer) treated with IVMgSO4 (Figure 3). There was no reason to downgrade for any of the five domains in GRADE (risk of bias, inconsistency, indirectness, imprecision, publication bias). Specifically, risk of bias was generally low or unclear across trials, heterogeneity was not significant, trials matched the research question well, confidence intervals were relatively narrow and almost all studies contributed data to the analysis.

GRADE

Abstract: Results

Intravenous MgSO4 reduced hospital admissions compared with placebo (odds ratio (OR) 0.75, 95% confidence interval (CI) 0.60 to 0.92; I2 = 28%, P value 0.18; n = 972; high-quality evidence). In absolute terms, this odds ratio translates into a reduction of seven hospital admissions for every 100 adults treated with IV MgSO4 (95% CI two to 13 fewer).... Sensitivity analyses in which unpublished data and studies at high risk for blinding were removed from the primary analysis did not change conclusions.

GRADE

Abstract conclusions and Plain language summary

This review showed that IV MgSO4 reduces hospital admissions…

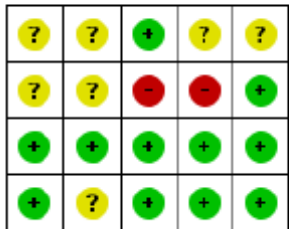
GRADE

Summary of findings table

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	No. of participants (studies)	Quality of the evidence (GRADE)
	Assumed risk	Corresponding risk			
	Control	IV $MgSO_4$			
Hospital admissions	569 per 1000	498 per 1000 (442 to 549)	OR 0.75 (0.60 to 0.92)	1769 (11 studies)	⊕⊕⊕⊕ high[1,2]

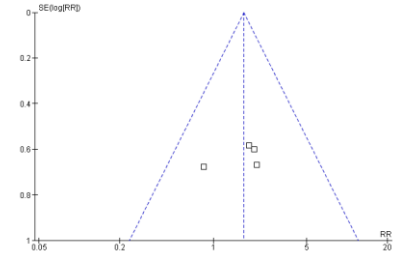
[1] One study introduced risk of bias, but the rest of the studies were generally well conducted.

GRADE



Agnelli 2009

Methods	Randomized clinical trial
Participants	1150 patients with metastatic or locally advanced lung, breast, gastrointestinal (stomach, colon-rectum, pancreas), ovarian or head and neck cancer undergoing chemotherapy.
Interventions	Subcutaneous low molecular weight heparin (nadroparin calcium, one injection/day) vs. placebo for the overall duration of chemotherapy or up to a maximum of 4 months
Outcomes	• Survival, (4 months and 12 months follow-up) • Response to chemotherapy (4 months follow-up) • For patients with central venous catheters (CVC), complications of possible thrombotic origin, such as malfunction or requirement of CVC removal (4 months follow-up) • Superficial thrombophlebitis of lower limbs (4 months follow-up) • Asymptomatic thromboembolic events diagnosed during tests performed for other purposes (4 months follow-up) • Safety (major bleeding, minor bleeding, other adverse events) (4 months follow-up)



Results & Discussion

Abstract & PLS

GRADE: high, moderate, low, very low quality

We found that wearing hip protectors probably leads to 5 fewer hip fractures (from 2 to 7 fewer).

Comparison: 1 Acamprosate versus placebo

Outcome: 3 Return to heavy drinking

Study or subgroup	Acamprosate n/N	Placebo n/N	Risk Ratio M-H,Random,95% CI	Weight	Risk Ratio M-H,Random,95% CI
Anton 2006	211/303	226/309		26.7 %	0.95 [0.86, 1.05]
Chick 2000	246/289	242/292		53.4 %	1.03 [0.96, 1.10]
Kiefer 2003	25/40	30/40		3.0 %	0.83 [0.62, 1.12]
Mason 2006	143/341	119/260		8.1 %	0.92 [0.76, 1.10]
Morley 2006	40/55	43/61		5.1 %	1.03 [0.82, 1.30]
Namkoong 2003	43/72	42/70		3.7 %	1.00 [0.76, 1.30]
Total (95% CI)	**1100**	**1032**		**100.0 %**	**0.99 [0.94, 1.04]**

Total events: 708 (Acamprosate), 702 (Placebo)

Heterogeneity: $Tau^2 = 0.0$; $Chi^2 = 4.15$, df = 5 (P = 0.53); $I^2 = 0.0\%$

Test for overall effect: Z = 0.38 (P = 0.71)

Probiotics compared to no treatment for preventing acute upper respiratory tract infections (URTI) in healthy people

Patient or population: healthy people
Intervention: Probiotics
Comparison: no treatment

Outcomes	Illustrative comparative risks[a] (95% CI) Assumed risk: No treatment	Corresponding risk: Probiotics	Relative effect (95% CI)	No of Participants (studies)	Quality of the evidence (GRADE)	Comments
Number of people who have an URTI (at least 1 URTI) Follow-up: 3-8 months	Low[1] 10 per 100	7 per 100 (6 to 10)	RR 0.73 (0.55 to 0.97)	1836 (6 studies)	⊕⊕⊕⊖ moderate[2,3]	
	High[1] 50 per 100	36 per 100 (28 to 48)				
Duration of URTI (number of days) Follow-up: 3-8 months	The mean duration of URTI ranged across control groups from 6-9 days	The mean duration of URTI in the intervention groups was 0.29 lower (3.71 lower to 3.13 higher)		620 (2 studies)	⊕⊕⊖⊖ low[1,2]	Two studies not in meta-analysis: one reported duration was significantly lower with probiotics, another reported unclear effect.
Days missed from school or work - not measured	See comment	See comment	Not estimable	-	See comment	
People prescribed antibiotics Follow-up: 4-8 months	Moderate 23 per 100	16 per 100 (10 to 23)	RR 0.67 (0.45 to 0.98)	1104 (3 studies)	⊕⊕⊕⊖ moderate[3]	
Complicated episodes of acute lower respiratory infection - not measured	See comment	See comment	Not estimable	-	See comment	
Adverse events (gastrointestinal symptoms) Follow-up: 5 months	Moderate 20 per 1000	18 per 1000 (8 to 45)	RR 0.92 (0.38 to 2.24)	956 (2 studies)	⊕⊕⊕⊖ moderate[3]	Most studies reported minor adverse events with probiotics.

[a] The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

CI: Confidence interval; RR: Risk ratio; OR: Odds ratio;

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] Number of people with URTI in the placebo groups ranged from 5-67% in different populations; therefore control group risk of 10 and 50% shown for illustration.
[2] Few studies reporting this outcome - selective outcome reporting likely. Allocation concealment unclear in most studies.
[3] Moderate heterogeneity that was not explained but not downgraded (considered with risk of bias).
[4] Inconsistent results from studies, and imprecise effect which includes reduction and increase in duration.
[5] Imprecise results due to few events and risk of bias due to unclear allocation concealment in studies.

Determinants to make conclusions

5 factors to consider to evaluate the quality of the evidence

1. Risk of bias
2. Inconsistency *(or heterogeneity)*
3. Indirectness *(PICO and applicability)*
4. Imprecision *(number of events and confidence intervals)*
5. Publication bias

Plus additional factors for observational studies

Dose response, size of effect, confounding

GRADE

Self management for patients with chronic obstructive pulmonary disease

Patient or population: patients with chronic obstructive pulmonary disease
Settings: primary care, community, outpatient
Intervention: self management[1]
Comparison: usual care

Outcomes	Illustrative comparative risks* (95% CI)		Relative effect (95% CI)	No of Participants (studies)	Quality of the evidence (GRADE)	Comments
	Assumed risk **usual care**	Corresponding risk **self management**				
Quality of Life St George's Respiratory Questionnaire. Scale from: 0 to 100. (follow-up: 3 to 12 months)	The mean quality of life ranged across control groups from **38 to 60 points**	The mean quality of Life in the intervention groups was **2.58 lower** (5.14 to 0.02 lower)		698 (7)	⊕⊕⊕O **moderate**[2]	Lower score indicates better quality of life. A change of less than 4 points is not shown to be important to patients.
Dyspnoea Borg Scale. Scale from: 0 to 10. (follow-up: 3 to 6 months)	The mean dyspnoea ranged across control groups from **1.2 to 4.1 points**	The mean dyspnoea in the intervention groups was **0.53 lower** (0.96 to 0.1 lower)		144 (2)	⊕⊕OO **low**[3,4]	Lower score indicates improvement
Number and severity of exacerbations[5]	See comment	See comment	Not estimable[5]	591 (3)	See comment	Effect is uncertain
Respiratory-related hospital admissions (follow-up: 3 to 12 months)	**Low risk population**[6]		OR 0.64 (0.47 to 0.89)	966 (8)	⊕⊕⊕O **moderate**[7]	
	10 per 100	**7 per 100** (5 to 9)				
	High risk population[6]					
	50 per 100	**39 per 100** (32 to 47)				
Emergency department visits for lung diseases (follow-up: 6 to 12 months)	The mean emergency department visits for lung diseases ranged across control groups from **0.2 to 0.7 visits per person per year**	The mean emergency department visits for lung diseases in the intervention groups was **0.1 higher** (0.2 lower to 0.3 higher)		328 (4)	⊕⊕⊕O **moderate**[4]	
Doctor and nurse visits (follow-up: 6 to 12 months)	The mean doctor and nurse visits ranged across control groups from **1 to 5 vists per person per year**	The mean doctor and nurse visits in the intervention groups was **0.02 higher** (1 lower to 1 higher)		629 (8)	⊕⊕⊕O **moderate**[8]	

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).

CI: Confidence interval; OR: Odds ratio;

GRADE

Background: Magnitude of effect

GRADE

What are the effects if older people wear hip protectors to prevent hip fractures?

Outcome: 19 Other fractures

Study or subgroup	Hip protectors n/N	No hip protectors n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
Cristofalo 2013	2/24	0/29		2.0 %	6.00 [0.30, 119.27]
Lucas 1984a	9/76	7/83		29.8 %	1.40 [0.55, 3.59]
Lucas 1984b	15/173	12/170		54.0 %	1.23 [0.59, 2.55]
Schanler 2005	3/88	3/78		14.2 %	0.89 [0.18, 4.26]
Total (95% CI)	**361**	**360**		**100.0 %**	**1.33 [0.79, 2.25]**

Total events: 29 Hip protectors, 22 no hip protectors

Heterogeneity: $Chi^2 = 1.29$, $df = 3$ ($P = 0.73$); $I^2 = 0.0\%$

Test for overall effect: $Z = 1.06$ ($P = 0.29$)

Test for subgroup differences: Not applicable

GRADE

What is the effect?
Do you believe the result?
What other information do you need?
Discuss…

Outcome: 19 Other fractures

Study or subgroup	Hip protectors n/N	No hip protectors n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
Cristofalo 2013	2/24	0/29		2.0 %	6.00 [0.30, 119.27]
Lucas 1984a	9/76	7/83		29.8 %	1.40 [0.55, 3.59]
Lucas 1984b	15/173	12/170		54.0 %	1.23 [0.59, 2.55]
Schanler 2005	3/88	3/78		14.2 %	0.89 [0.18, 4.26]
Total (95% CI)	**361**	**360**		**100.0 %**	**1.33 [0.79, 2.25]**

Total events: 29 Hip protectors, 22 no hip protectors

Heterogeneity: $Chi^2 = 1.29$, $df = 3$ ($P = 0.73$); $I^2 = 0.0\%$

Test for overall effect: $Z = 1.06$ ($P = 0.29$)

Test for subgroup differences: Not applicable

GRADE

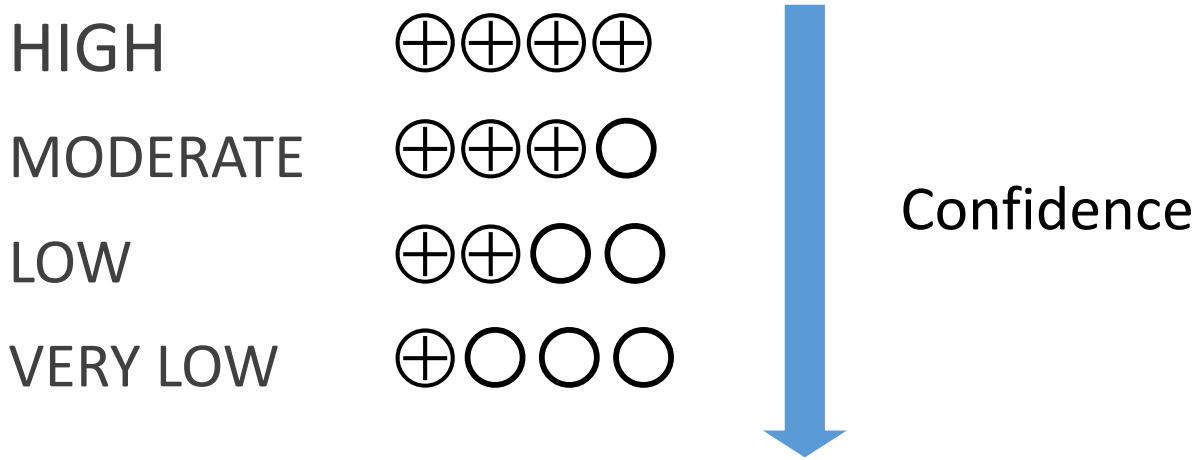
What to consider when making conclusions?

- Do I believe the results from these studies? **Risk of bias**
- Are the results consistent across studies? **Inconsistency**
- How do these results apply to my question? **Indirectness**
- Is this effect size precise? **Imprecision**
- Are these all of the studies? **Publication bias**

GRADE

What about the quality of the evidence?

- quality of evidence or confidence in effect varies from



GRADE

What does this mean?

- **High quality**:
- **Moderate quality**:
- **Low quality**:
- **Very low quality**:

GRADE

What does this mean?

- **High quality**: We are very confident that the true effect lies close to that of the estimate of the effect
- **Moderate quality**:
- **Low quality**:
- **Very low quality**:

GRADE

What does this mean?

- **High quality**: We are very confident that the true effect lies close to that of the estimate of the effect
- **Moderate quality**: 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 quality**:
- **Very low quality**:

GRADE

What does this mean?

- **High quality**: We are very confident that the true effect lies close to that of the estimate of the effect
- **Moderate quality**: 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 quality**: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
- **Very low quality**:

GRADE

What does this mean?

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

GRADE

How to downgrade the quality of the evidence?

HIGH	⨁⨁⨁⨁
MODERATE	⨁⨁⨁◯
LOW	⨁⨁◯◯
VERY LOW	⨁◯◯◯

Lower quality of evidence each time there is a serious concern with

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

GRADE

What to consider when making conclusions?

- Do I believe the results from these studies? **Risk of bias**
- Are the results consistent across studies? **Inconsistency**
- How do these results apply to my question? **Indirectness**
- Is this effect size precise? **Imprecision**
- Are these all of the studies? **Publication bias**

GRADE

Risk of bias criteria: Cochrane tools

- Random sequence generation
- Allocation concealment
- Blinding
- Incomplete outcome data
- Selective outcome reporting
- Other

www.cochrane-handbook.org

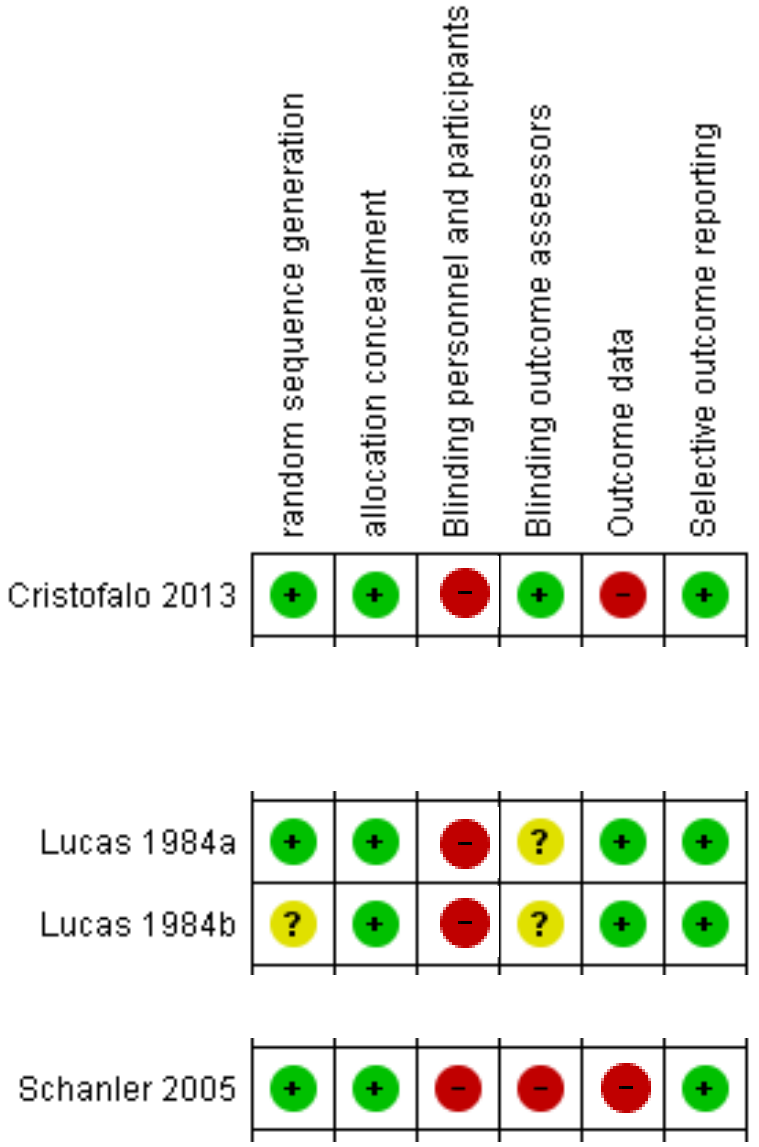
Chapter 8

GRADE

Risk of bias assessment

Is the overall risk of bias…

- ☐ Not serious
- ☐ Serious
- ☐ Very serious



GRADE

What to consider when making conclusions?

- Do I believe the results from these studies? **Risk of bias**
- Are the results consistent across studies? **Inconsistency**
- How do these results apply to my question? **Indirectness**
- Is this effect size precise? **Imprecision**
- Are these all of the studies? **Publication bias**

GRADE

Inconsistency?

Outcome: 19 Other fractures

Study or subgroup	Hip protectors n/N	No hip protectors n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
Cristofalo 2013	2/24	0/29		2.0 %	6.00 [0.30, 119.27]
Lucas 1984a	9/76	7/83		29.8 %	1.40 [0.55, 3.59]
Lucas 1984b	15/173	12/170		54.0 %	1.23 [0.59, 2.55]
Schanler 2005	3/88	3/78		14.2 %	0.89 [0.18, 4.26]
Total (95% CI)	**361**	**360**		**100.0 %**	**1.33 [0.79, 2.25]**

Total events: 29 Hip protectors, 22 no hip protectors

Heterogeneity: $Chi^2 = 1.29$, $df = 3$ ($P = 0.73$); $I^2 = 0.0\%$

Test for overall effect: $Z = 1.06$ ($P = 0.29$)

Test for subgroup differences: Not applicable

GRADE

Inconsistency (heterogeneity)

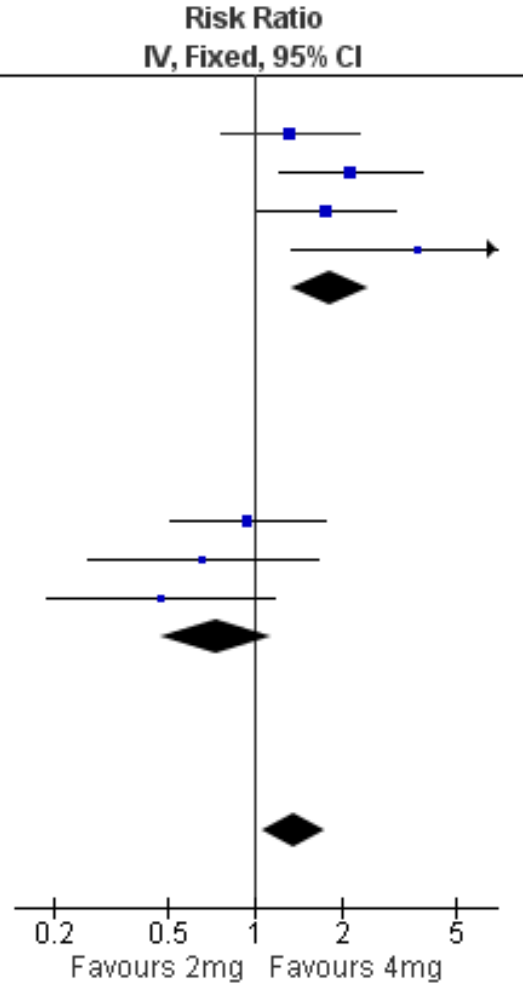
Consider in a meta- analysis

- variation in size of effect
- overlap in confidence intervals
  - If no overlap, then variation between the study results is more than what you would expect by chance
- p value of heterogeneity
- I^2
- Unexplained heterogeneity – did you explore?

GRADE

Can heterogeneity be explained by subgroup analysis?

Study or Subgroup	4 mg dose Events	Total	2 mg dose Events	Total	Weight	Risk Ratio IV, Fixed, 95% CI
5.1.1 High dependency smokers						
Garvey 2000	24	116	18	115	20.9%	1.32 [0.76, 2.30]
Herrera 1995	30	87	13	81	19.3%	2.15 [1.21, 3.82]
Kornitzer 1987	24	73	16	86	21.2%	1.77 [1.02, 3.06]
Tonnesen 1988	12	27	4	33	6.3%	3.67 [1.33, 10.08]
Subtotal (95% CI)		**303**		**315**	**67.8%**	**1.83 [1.34, 2.49]**
Total events	90		51			
Heterogeneity: Chi² = 3.45, df = 3 (P = 0.33); I² = 13%						
Test for overall effect: Z = 3.84 (P = 0.0001)						
5.1.2 Low Dependency Smokers						
Garvey 2000	16	87	17	87	17.0%	0.94 [0.51, 1.74]
Hughes 1990	5	19	8	20	7.5%	0.66 [0.26, 1.66]
Kornitzer 1987	5	17	5	8	7.7%	0.47 [0.19, 1.17]
Subtotal (95% CI)		**123**		**115**	**32.2%**	**0.73 [0.47, 1.15]**
Total events	26		30			
Heterogeneity: Chi² = 1.60, df = 2 (P = 0.45); I² = 0%						
Test for overall effect: Z = 1.36 (P = 0.17)						
Total (95% CI)		**426**		**430**	**100.0%**	**1.36 [1.06, 1.75]**
Total events	116		81			
Heterogeneity: Chi² = 15.96, df = 6 (P = 0.01); I² = 62%						
Test for overall effect: Z = 2.39 (P = 0.02)						
Test for subgroup differences: Chi² = 10.91, df = 1 (P = 0.0010), I² = 90.8%						

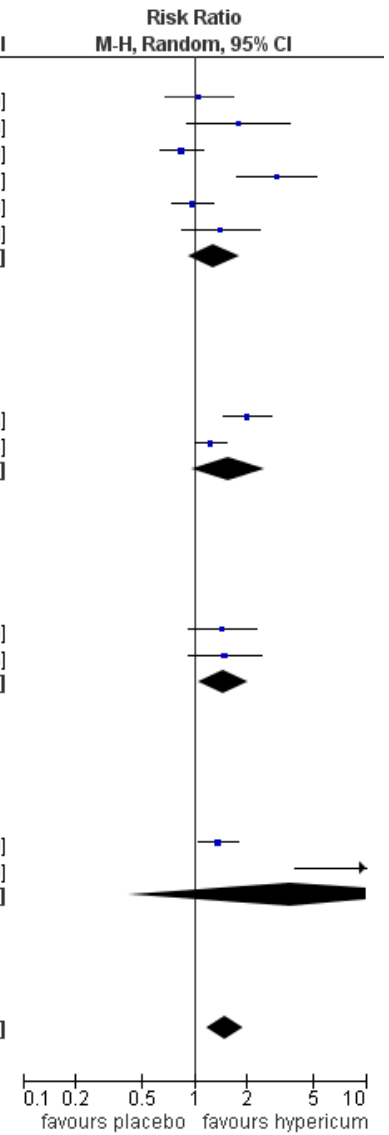


Based on Stead LF, Perera R, Bullen C, Mant D, Lancaster T. Nicotine replacement therapy for smoking cessation. *Cochrane Database of Systematic Reviews* 2008, Issue 1. Art. No.: CD000146. DOI: 10.1002/14651858.CD000146.pub3.

GRADE

Unexplained heterogeneity

Study or Subgroup	hypericum Events	Total	placebo Events	Total	Weight	Risk Ratio M-H, Random, 95% CI
7.4.1 extract LI 160						
Bjerkenstedt 2005	22	54	21	55	8.2%	1.07 [0.67, 1.70]
Fava 2005	17	45	9	43	6.0%	1.80 [0.90, 3.60]
HDTSG 2002	46	113	56	116	10.0%	0.84 [0.63, 1.13]
Hänsgen 1996	35	53	12	55	7.4%	3.03 [1.77, 5.17]
Montgomery 2000	55	123	57	124	10.2%	0.97 [0.74, 1.28]
Shelton 2001	26	98	19	102	7.6%	1.42 [0.84, 2.40]
Subtotal (95% CI)		**486**		**495**	**49.3%**	**1.31 [0.92, 1.86]**
Total events	201		174			
Heterogeneity: Tau² = 0.14; Chi² = 20.94, df = 5 (P = 0.0008); I² = 76%						
Test for overall effect: Z = 1.47 (P = 0.14)						
7.4.2 extract WS 5570						
Kasper 2006	159	243	26	81	9.6%	2.04 [1.47, 2.83]
Lecrubier 2002	98	186	80	189	10.7%	1.24 [1.00, 1.54]
Subtotal (95% CI)		**429**		**270**	**20.3%**	**1.57 [0.96, 2.56]**
Total events	257		106			
Heterogeneity: Tau² = 0.11; Chi² = 6.26, df = 1 (P = 0.01); I² = 84%						
Test for overall effect: Z = 1.79 (P = 0.07)						
7.4.3 extract WS 5572						
Kalb 2001	23	37	15	35	8.2%	1.45 [0.92, 2.29]
Laakmann 1998	24	49	16	49	7.9%	1.50 [0.92, 2.46]
Subtotal (95% CI)		**86**		**84**	**16.1%**	**1.47 [1.05, 2.06]**
Total events	47		31			
Heterogeneity: Tau² = 0.00; Chi² = 0.01, df = 1 (P = 0.92); I² = 0%						
Test for overall effect: Z = 2.26 (P = 0.02)						
7.4.4 extract STW3-VI						
Gastpar 2006	71	131	51	130	10.3%	1.38 [1.06, 1.80]
Uebelhack 2004	41	70	4	70	4.0%	10.25 [3.88, 27.09]
Subtotal (95% CI)		**201**		**200**	**14.3%**	**3.59 [0.41, 31.56]**
Total events	112		55			
Heterogeneity: Tau² = 2.33; Chi² = 18.65, df = 1 (P < 0.0001); I² = 95%						
Test for overall effect: Z = 1.15 (P = 0.25)						
Total (95% CI)		**1202**		**1049**	**100.0%**	**1.51 [1.19, 1.92]**
Total events	617		366			
Heterogeneity: Tau² = 0.13; Chi² = 51.55, df = 11 (P < 0.00001); I² = 79%						
Test for overall effect: Z = 3.34 (P = 0.0008)						
Test for subgroup differences: Chi² = 1.11, df = 3 (P = 0.77), I² = 0%						



Based on Linde K, Berner MM, Kriston L. St John's wort for major depression. *Cochrane Database of Systematic Reviews* 2008, Issue 4. Art. No.: CD000448. DOI: 10.1002/14651858.CD000448.pub3.

GRADE

Inconsistency (heterogeneity)

Consider in a meta- analysis

- variation in size of effect
- overlap in confidence intervals
  - If no overlap, then variation between the study results is more than what you would expect by chance
- p value of heterogeneity
- I^2
- Unexplained heterogeneity – NOT CONFIDENT IN EFFECT

GRADE

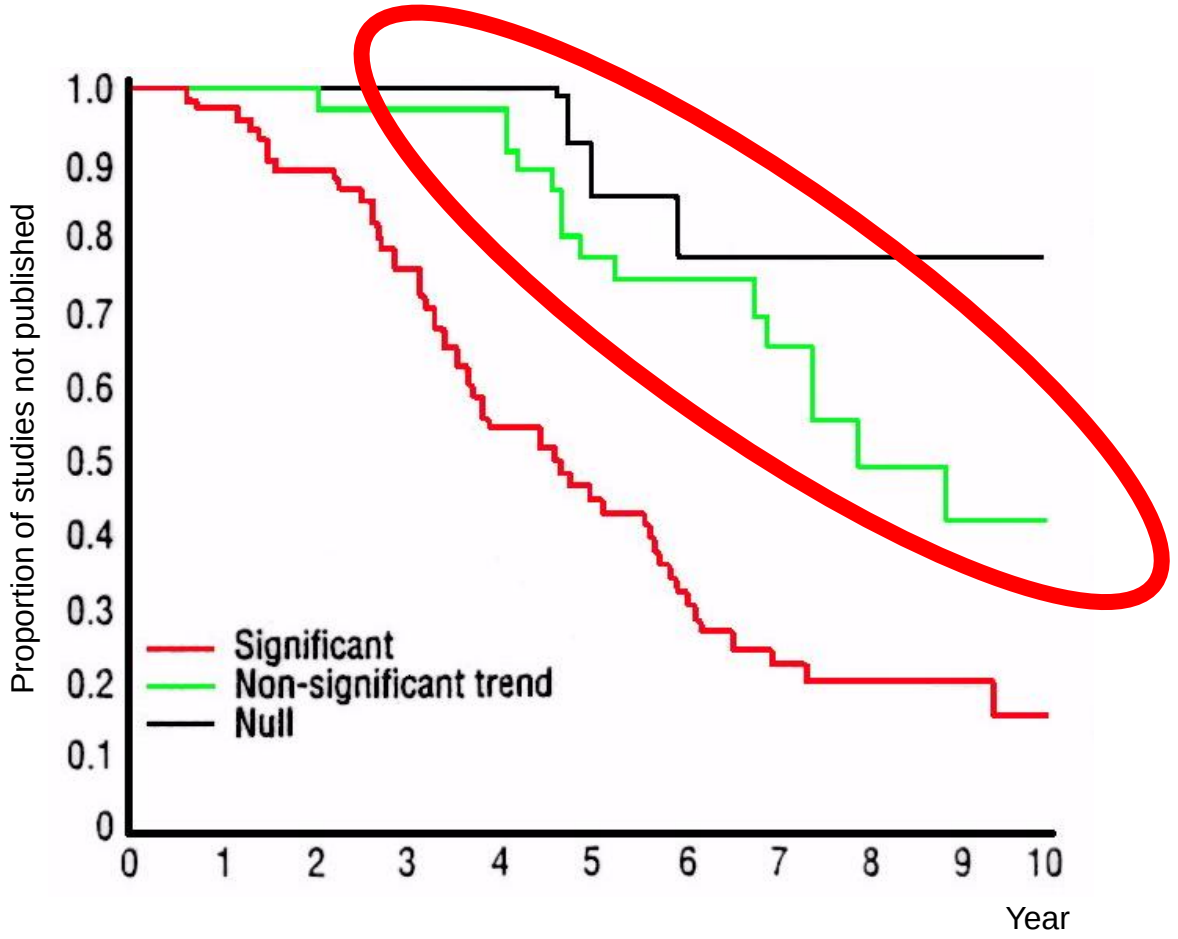
What to consider when making conclusions?

- Do I believe the results from these studies? **Risk of bias**
- Are the results consistent across studies? **Inconsistency**
- How do these results apply to my question? **Indirectness**
- Is this effect size precise? **Imprecision**
- Are these all of the studies? **Publication bias**

GRADE

Reporting/Publication bias: Small studies

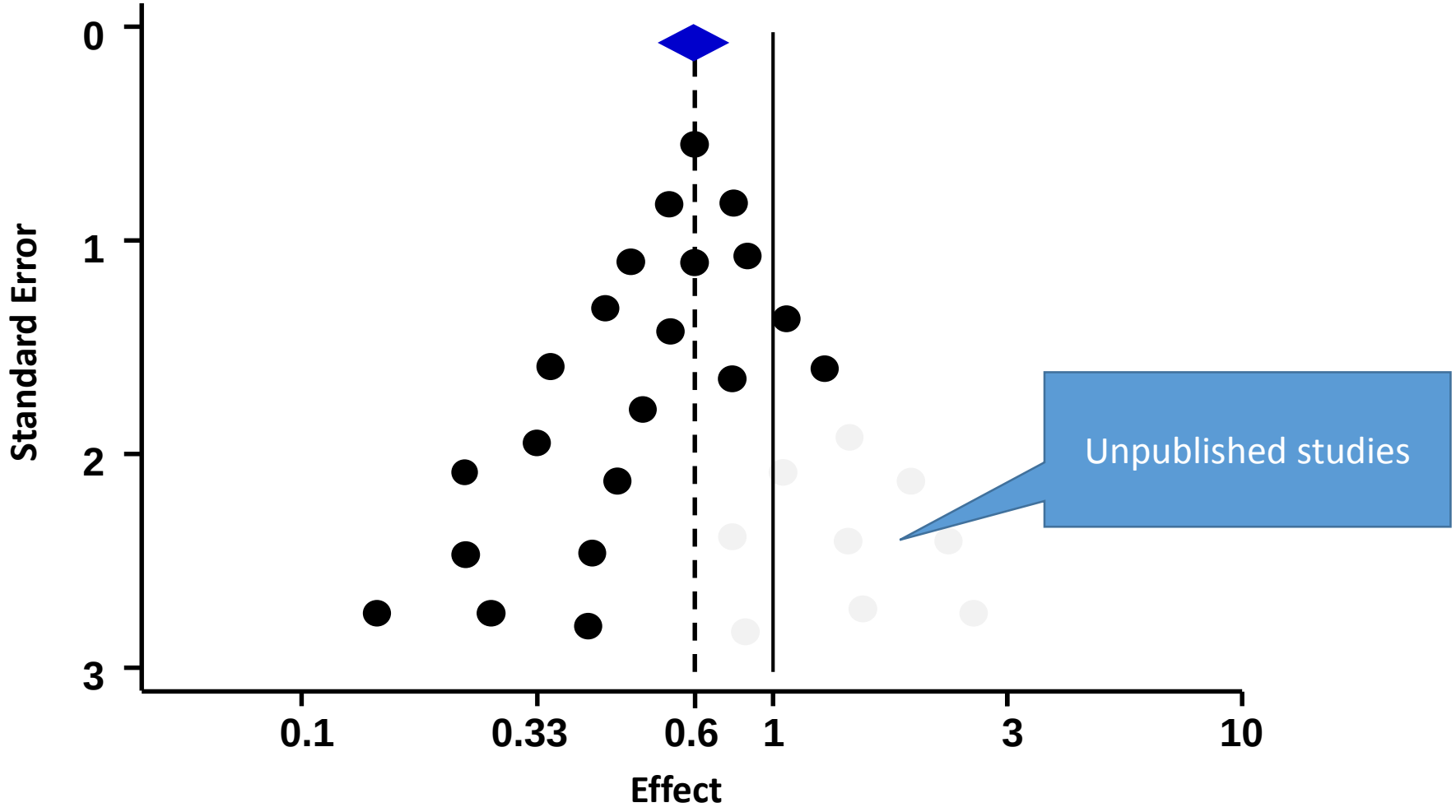
- less precise than large studies
- usually show ‘trends’ or ‘non-significant results’



Stern 1997

GRADE

Reporting bias: unpublished studies



Source: Matthias Egger & Jonathan Sterne

GRADE

Publication bias

Funnel plot recommended when 10 or more studies

Instead,

- Was the search strategy comprehensive?
- Are foreign language articles missing?
- Mostly small studies?
- …

GRADE

Publication bias

Undetected

Strongly suspected

GRADE

Does the evidence directly answer the question?

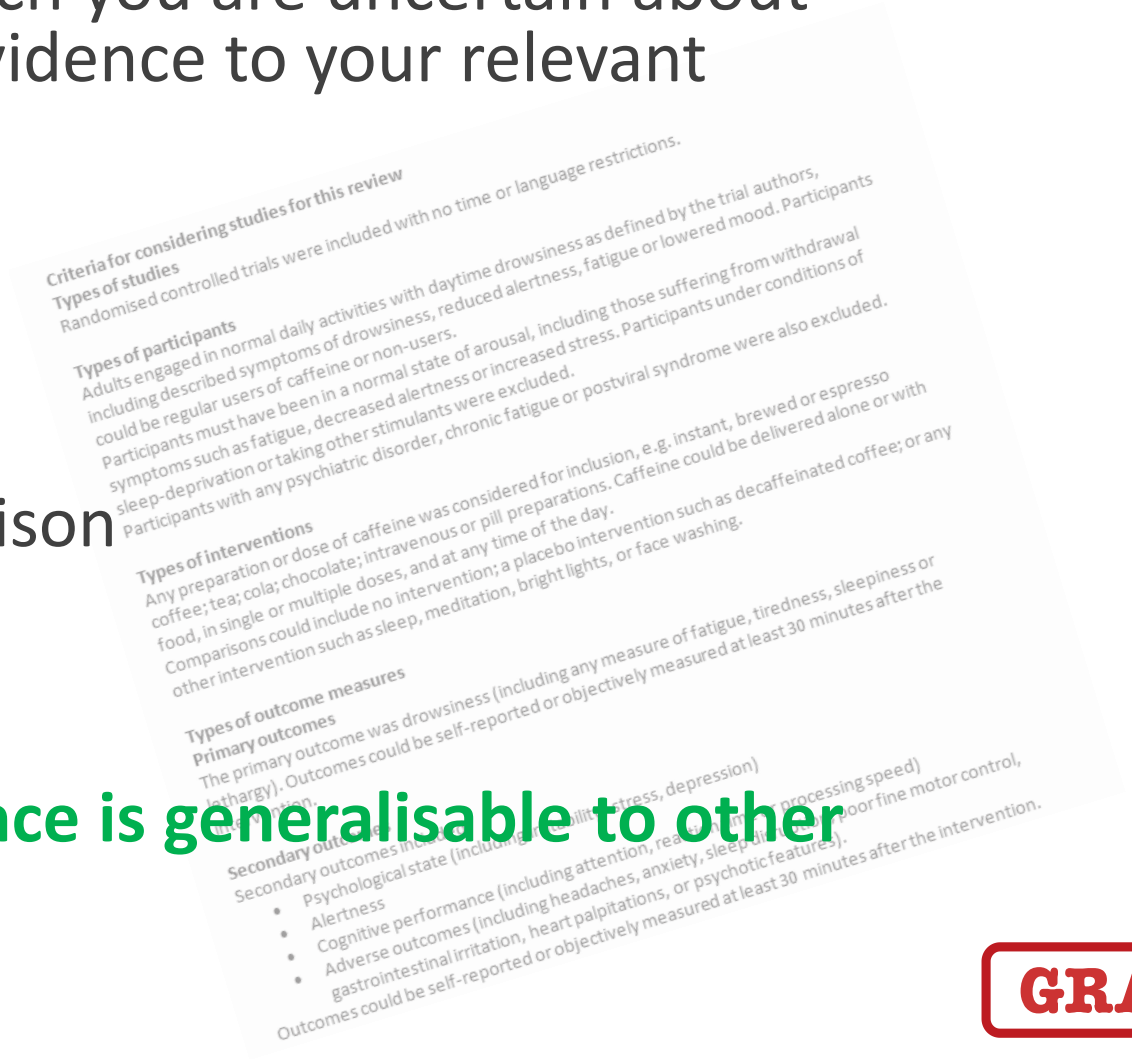
Indirectness

Consider the extent to which you are uncertain about the applicability of the evidence to your relevant question

Consider PICO

- Population
- Intervention and comparison
- Outcome

Not about whether evidence is generalisable to other populations etc.



Criteria for considering studies for this review

Types of studies

Randomised controlled trials were included with no time or language restrictions.

Types of participants

Adults engaged in normal daily activities with daytime drowsiness as defined by the trial authors, including described symptoms of drowsiness, reduced alertness, fatigue or lowered mood. Participants could be regular users of caffeine or non-users.

Participants must have been in a normal state of arousal, including those suffering from withdrawal symptoms such as fatigue, decreased alertness or increased stress. Participants under conditions of sleep-deprivation or taking other stimulants were excluded.

Participants with any psychiatric disorder, chronic fatigue or postviral syndrome were also excluded.

Types of interventions

Any preparation or dose of caffeine was considered for inclusion, e.g. instant, brewed or espresso coffee; tea; cola; chocolate; intravenous or pill preparations. Caffeine could be delivered alone or with food, in single or multiple doses, and at any time of the day.

Comparisons could include no intervention; a placebo intervention such as decaffeinated coffee; or any other intervention such as sleep, meditation, bright lights, or face washing.

Types of outcome measures

Primary outcomes

The primary outcome was drowsiness (including any measure of fatigue, tiredness, sleepiness or lethargy). Outcomes could be self-reported or objectively measured at least 30 minutes after the intervention.

Secondary outcomes

Secondary outcomes included [illegible] (anxiety, depression)

- Psychological state (including [illegible] stress, depression)
- Alertness
- Cognitive performance (including attention, reaction time, memory, processing speed)
- Adverse outcomes (including headaches, anxiety, sleep disturbance, poor fine motor control, gastrointestinal irritation, heart palpitations, or psychotic features).

Outcomes could be self-reported or objectively measured at least 30 minutes after the intervention.

GRADE

Example: Hip protectors for older people

- Older people
  - Most studies included frail elderly
- Hard or soft protectors
  - Most studies used soft
- Hip fracture
  - Some studies radiologically confirmed

GRADE

Characteristics of included studies *[ordered by study ID]*

Birks 2003

Methods	Randomisation of individual participants by a telephone randomisation service
Participants	366 community-dwelling individuals recruited while recovering from a hip fracture in orthopaedic wards of York District Hospital, UK, or volunteers from general population who had sustained a hip fracture in the past Mean age: 80 years Proportion male: 12.6% Inclusion criteria: aged 70 years and over; have sustained one hip fracture; had to have one hip intact; able to give informed consent Exclusion criteria: bed or chair-bound; had bilateral hip replacement; a clothing size of 18 or above
Interventions	Allocation to wear hip protectors. "Intervention group participants were issued with three pairs of hip protectors and general advice (in the form of a leaflet) on how to reduce fracture risk" Controls: "people in the control group received only the leaflet" Hip protectors were Safehip (www.tytex.com our_products/hip_protection/)
Outcomes	Length of follow-up: mean 14 months (range 6 - 41 months) All outcomes were self-reported by post "The main outcome was a second hip fracture." Secondary outcomes were: Number of other fractures

What to consider when making conclusions

- Do I believe the results from these studies? Risk of bias
- Are the results consistent across studies? Heterogeneity
- Are these all of the studies? Reporting bias
- **Is this effect size precise? Imprecision**
- How do these results apply? Applicability, directness

GRADE

Are your results precise?

Outcome: 19 Other fractures

Study or subgroup	Hip protectors n/N	No hip protectors n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
Cristofalo 2013	2/24	0/29		2.0 %	6.00 [0.30, 119.27]
Lucas 1984a	9/76	7/83		29.8 %	1.40 [0.55, 3.59]
Lucas 1984b	15/173	12/170		54.0 %	1.23 [0.59, 2.55]
Schanler 2005	3/88	3/78		14.2 %	0.89 [0.18, 4.26]
Total (95% CI)	**361**	**360**		**100.0 %**	**1.33 [0.79, 2.25]**

Total events: 29 Hip protectors, 22 no hip protectors

Heterogeneity: $Chi^2 = 1.29$, df = 3 (P = 0.73); I^2 =0.0%

Test for overall effect: Z = 1.06 (P = 0.29)

Test for subgroup differences: Not applicable

It is **NOT**
about statistical significance

GRADE

Why we need confidence intervals

Altman DG. World J Surg 29, 554–556 (2005)

The CI obtained provides a **range of uncertainty**

ference between two such estimates. The CI is a range of values either side of the estimate between which we can be 95% sure that the true value lies. A series of identical studies carried out

In a comparative study such as an RCT, a common, serious mistake is to conclude from a nonsignificant result (i.e., with $p > 0.05$) that the groups are "the same." Yet this serious error is

GRADE

Why we need confidence intervals

Altman DG. World J Surg 29, 554–556 (2005)

laparoscopy group [3]. The relative risk is 0.49 with the 95% CI from 0.34 to 0.70. We can interpret this finding as saying that our best estimate is that the risk of recurrence is about halved in the laparoscopy group (relative risk reduction 51%) but that the results are compatible with a reduction in risk of recurrence between 30% and 66%. (The authors cited the odds ratio, which is

GRADE

Imprecision

Consider

- Sample sizes and number of events
  - assess according to effect size, control event rates, Optimal information size (OIS)

- Width of confidence intervals
  - Wide confidence intervals indicate uncertainty about the effect
  - Includes null effect and appreciable benefit or harm (rule of thumb: RR<0.75 or >1.25)

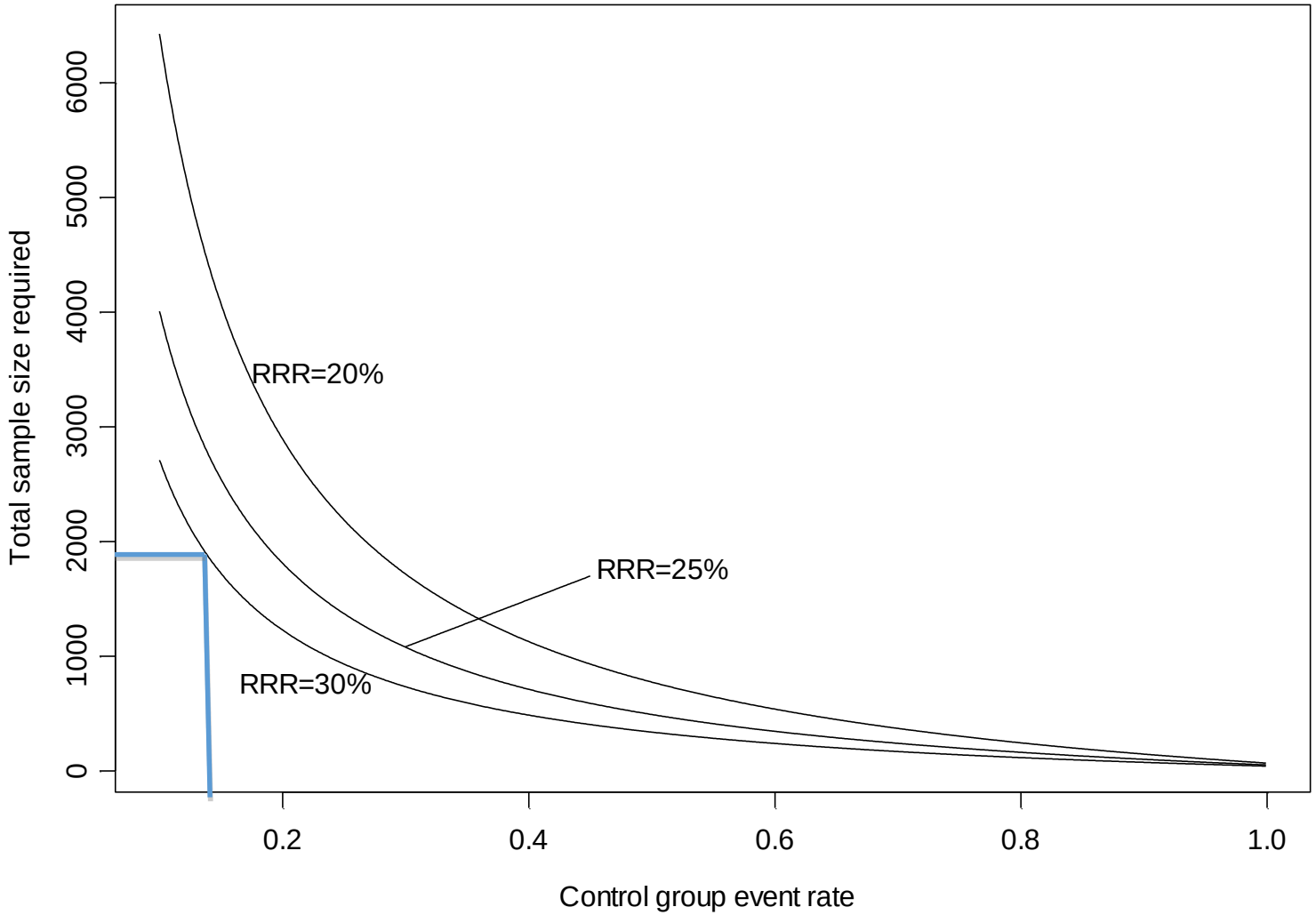
GRADE

Optimal information size (OIS)

- 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

- http://www.stat.ubc.ca/~rollin/stats/ssize/

GRADE


Total sample size required
0
1000
2000
3000
4000
5000
6000
RRR=20%
RRR=25%
RRR=30%
0.2
0.4
0.6
0.8
1.0
Control group event rate

Optimal information size implications: Consider the total number of events

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

Reasonable threshold for rating down for imprecision = 300 events

GRADE

Rules of thumb

Dichotomous outcomes

- 300 events

Continuous outcomes

- 400 people providing outcome measures

GRADE

Imprecision

Consider

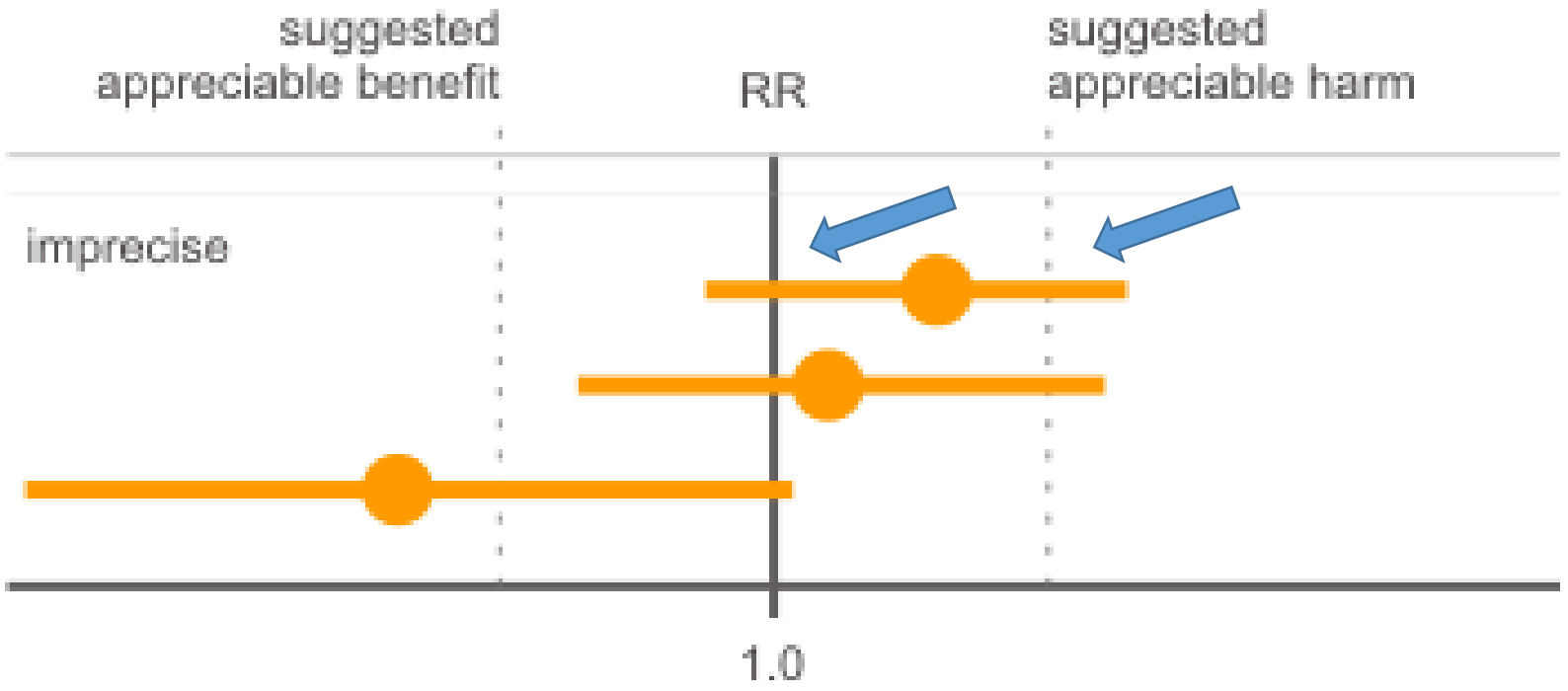
- Sample sizes and number of events
  - assess according to effect size, control event rates, Optimal information size (OIS)

- Width of confidence intervals
  - Wide confidence intervals indicate uncertainty about the effect
  - Includes null effect and appreciable benefit or harm (rule of thumb: RR<0.75 or >1.25)

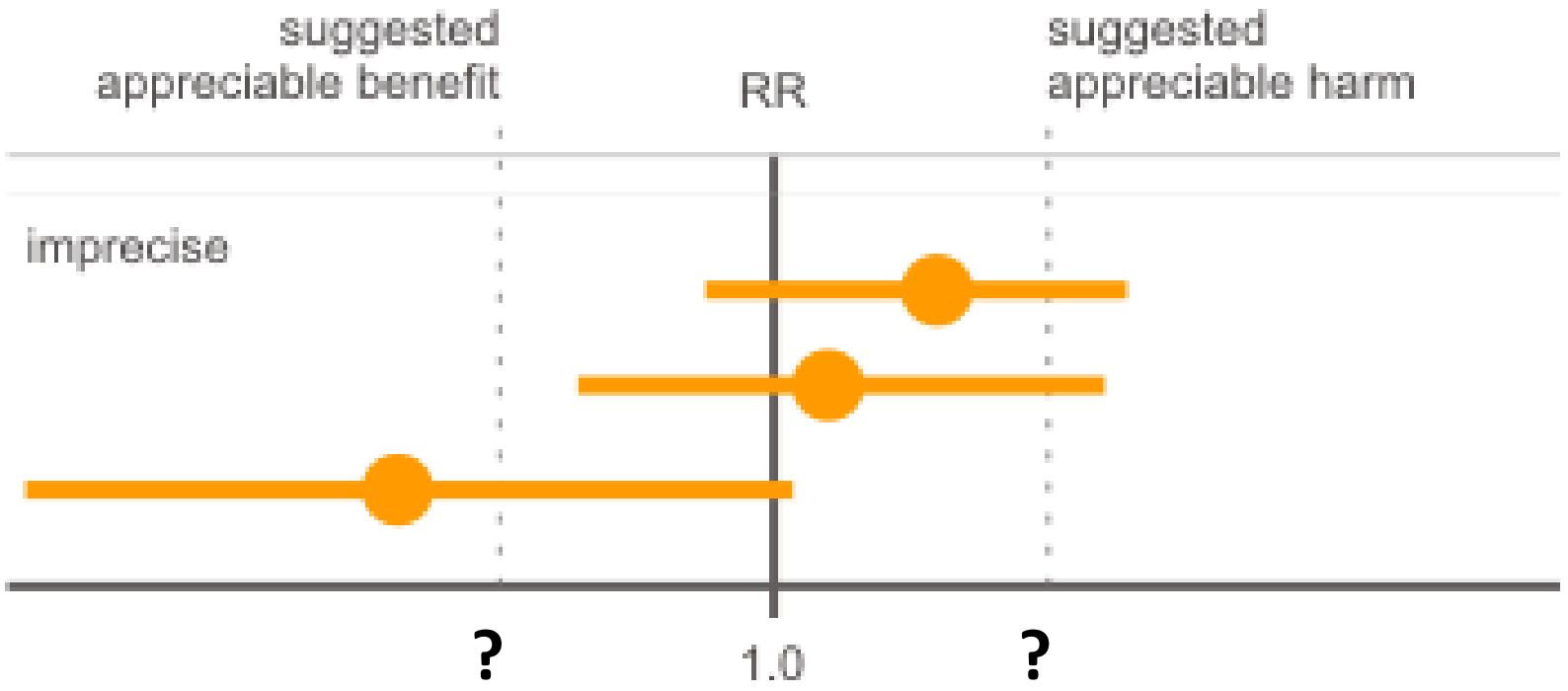
NOT
just null effect

GRADE

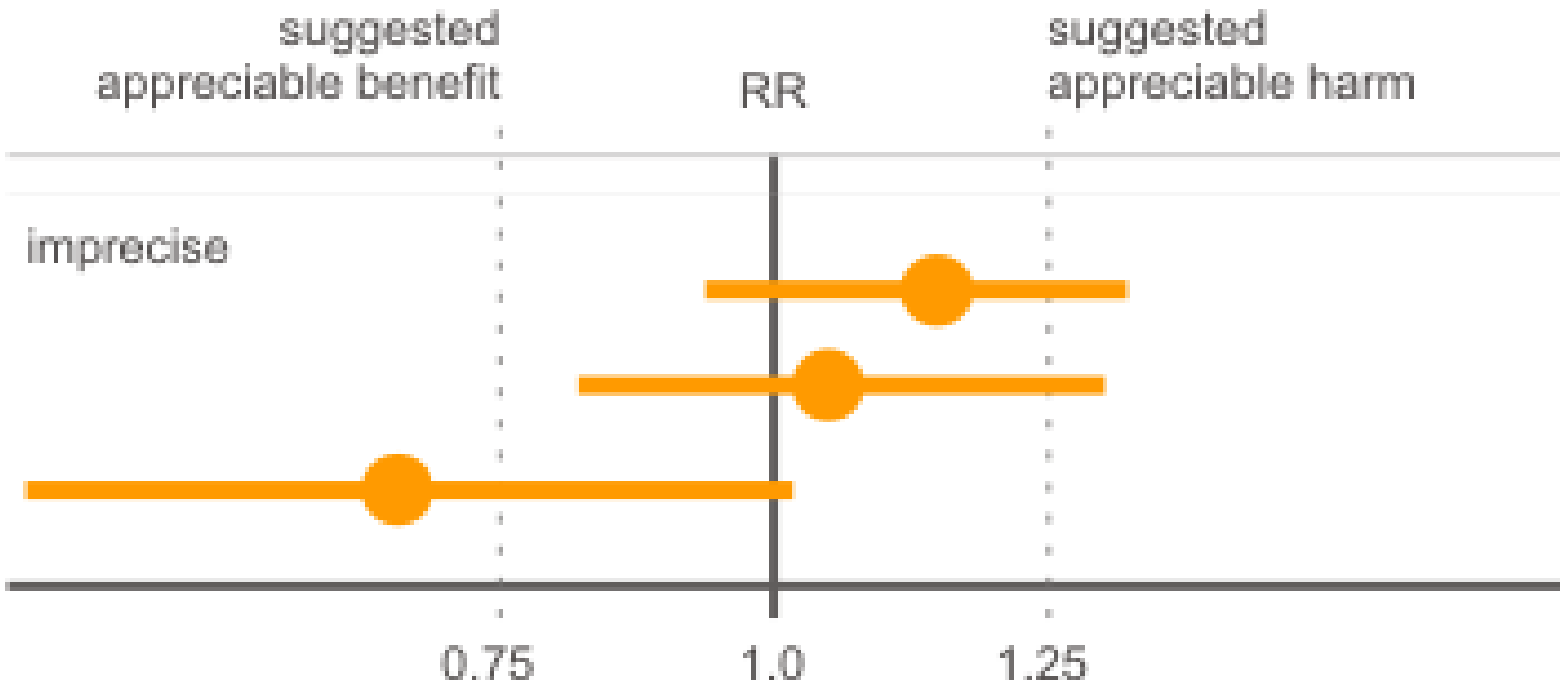
Imprecision:
CI includes null AND appreciable benefit or harm



Imprecision:
CI includes null AND appreciable benefit or harm

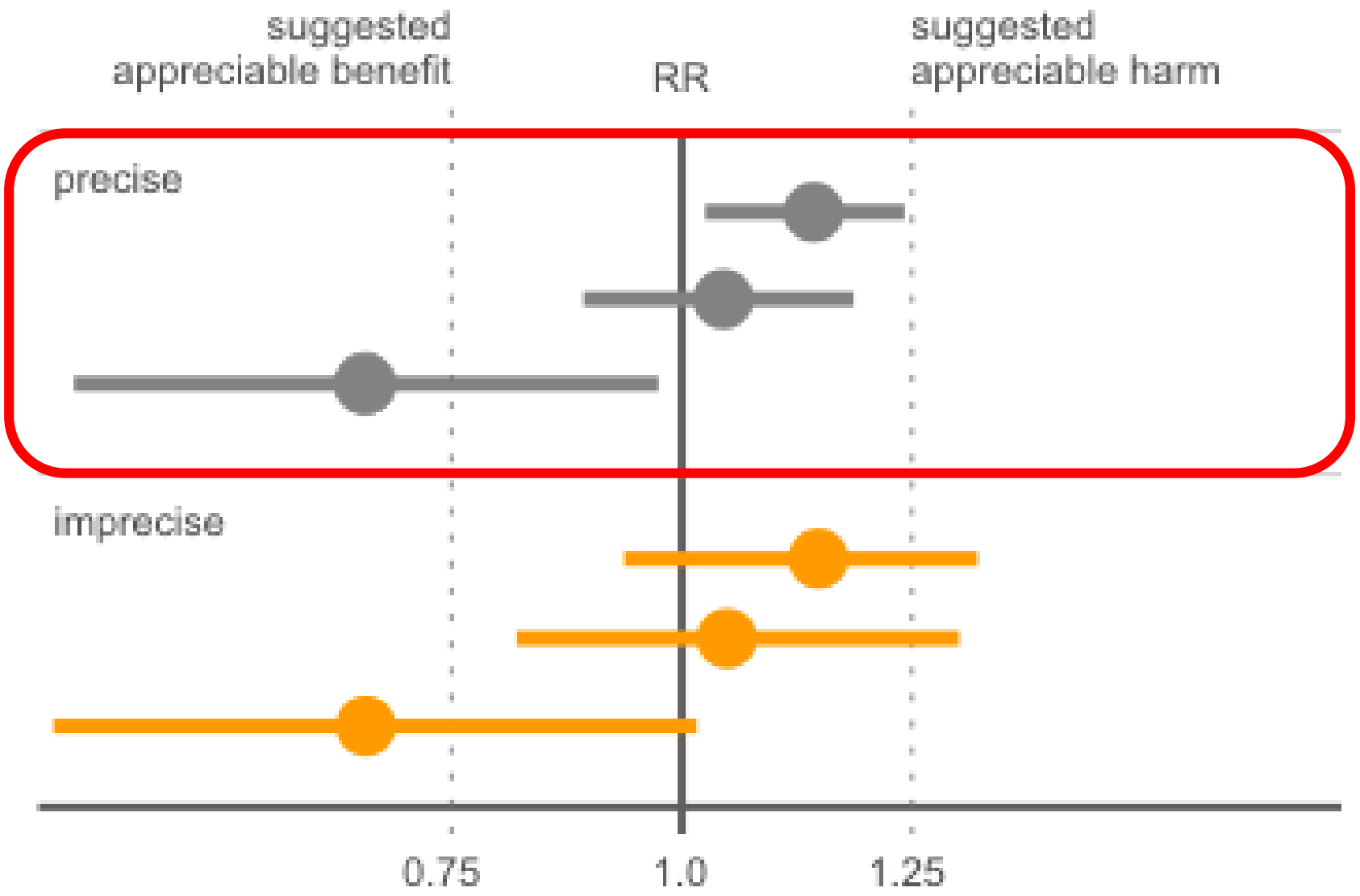


Imprecision:
CI includes null AND appreciable benefit or harm



rule of thumb: RR <0.75 or >1.25

Precise:


suggested
appreciable benefit
RR
suggested
appreciable harm
precise
imprecise
0.75
1.0
1.25

Exception to the rule

Small absolute effects

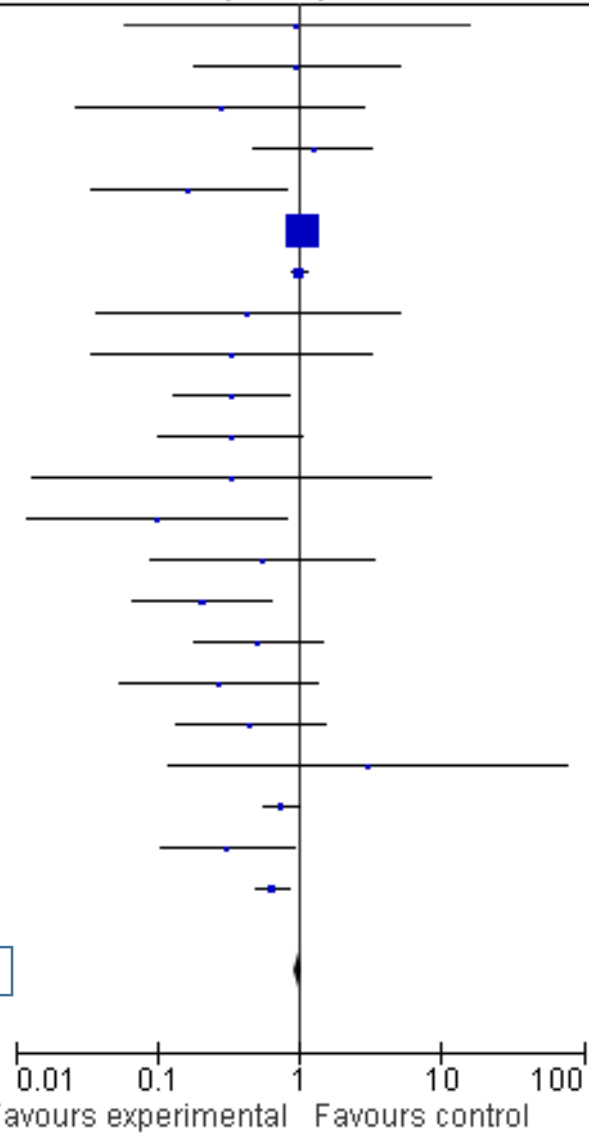
When event rates are very low, 95% confidence intervals around relative effects can be very wide, but 95% confidence intervals around absolute effects may be narrow. Under such circumstances one may decide not to downgrade the quality of evidence for imprecision.

GRADE

Study or Subgroup	Magnesium Events	Total	Placebo Events	Total	Weight	Odds Ratio M-H, Fixed, 95% CI
Abraham 1987	1	48	1	46	0.0%	0.96 [0.06, 15.77]
Bhargava 1995	3	40	3	38	0.1%	0.95 [0.18, 5.00]
Ceremuzynski 1989	1	25	3	23	0.1%	0.28 [0.03, 2.88]
Feldstedt 1991	10	150	8	148	0.3%	1.25 [0.48, 3.26]
Gyamlani 2000	2	50	10	50	0.4%	0.17 [0.03, 0.81]
ISIS-4 1995	2216	29008	2103	29038	71.6%	1.06 [1.00, 1.13]
MAGIC 2000	475	3113	472	3100	14.8%	1.00 [0.87, 1.15]
Morton 1984	1	40	2	36	0.1%	0.44 [0.04, 5.02]
Nakashima 2004	1	89	3	91	0.1%	0.33 [0.03, 3.27]
Raghu 1999	6	169	18	181	0.6%	0.33 [0.13, 0.86]
Rasmussen 1986	4	56	14	74	0.4%	0.33 [0.10, 1.06]
Santoro 2000	0	75	1	75	0.1%	0.33 [0.01, 8.20]
Shechter 1990	1	50	9	53	0.3%	0.10 [0.01, 0.82]
Shechter 1991	2	21	4	25	0.1%	0.55 [0.09, 3.37]
Shechter 1995	4	96	17	98	0.6%	0.21 [0.07, 0.64]
Singh 1990	6	81	11	81	0.4%	0.51 [0.18, 1.45]
Smith 1986	2	92	7	93	0.3%	0.27 [0.06, 1.35]
Thogersen 1995	4	130	8	122	0.3%	0.45 [0.13, 1.54]
Urek 1996	1	31	0	30	0.0%	3.00 [0.12, 76.58]
Woods 1992	90	1150	118	1150	4.0%	0.74 [0.56, 0.99]
Wu 1992	5	125	12	102	0.5%	0.31 [0.11, 0.92]
Zhu 2002	101	1691	134	1488	4.9%	0.64 [0.49, 0.84]
Total (95% CI)		**36330**		**36142**	**100.0%**	**0.99 [0.94, 1.04]**
Total events	2936		2958			

Heterogeneity: Chi² = 57.78, df = 21 (P < 0.0001); I² = 64%

Test for overall effect: Z = 0.48 (P = 0.63)



Imprecision?

☐ Not serious ☐ serious ☐ very serious

Outcome: 19 Other fractures

Study or subgroup	Hip protectors n/N	No hip protectors n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
Cristofalo 2013	2/24	0/29		2.0 %	6.00 [0.30, 119.27]
Lucas 1984a	9/76	7/83		29.8 %	1.40 [0.55, 3.59]
Lucas 1984b	15/173	12/170		54.0 %	1.23 [0.59, 2.55]
Schanler 2005	3/88	3/78		14.2 %	0.89 [0.18, 4.26]
Total (95% CI)	**361**	**360**		**100.0 %**	**1.33 [0.79, 2.25]**

Total events: 29 Hip protectors, 22 no hip protectors

Heterogeneity: $Chi^2 = 1.29$, df = 3 (P = 0.73); I^2 =0.0%

Test for overall effect: Z = 1.06 (P = 0.29)

Test for subgroup differences: Not applicable

GRADE

What is the conclusion?

Outcome: 19 Other fractures

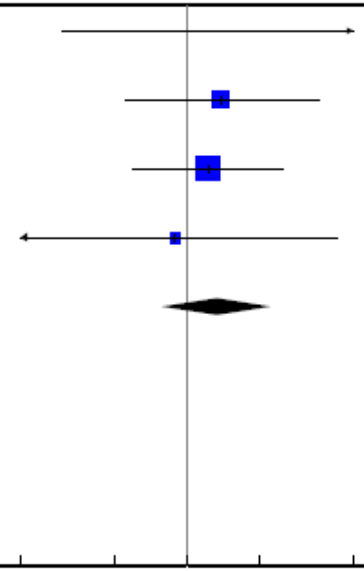
Study or subgroup	Hip protectors n/N	No hip protectors n/N	Risk Ratio M-H,Fixed,95% CI	Weight	Risk Ratio M-H,Fixed,95% CI
Cristofalo 2013	2/24	0/29		2.0 %	6.00 [0.30, 119.27]
Lucas 1984a	9/76	7/83		29.8 %	1.40 [0.55, 3.59]
Lucas 1984b	15/173	12/170		54.0 %	1.23 [0.59, 2.55]
Schanler 2005	3/88	3/78		14.2 %	0.89 [0.18, 4.26]
Total (95% CI)	**361**	**360**		**100.0 %**	**1.33 [0.79, 2.25]**

Total events: 29 Hip protectors, 22 no hip protectors

Heterogeneity: $Chi^2 = 1.29$, $df = 3$ ($P = 0.73$); $I^2 = 0.0\%$

Test for overall effect: $Z = 1.06$ ($P = 0.29$)

Test for subgroup differences: Not applicable



GRADE

Overall quality of evidence?

Risk of bias?

Inconsistency?

Indirectness?

Imprecision?

Publication bias?

Quality of evidence?

HIGH, MODERATE, LOW, VERY LOW?

GRADE

Other fractures	**Moderate**		**RR 1.33 (0.79 to 2.25)**	**721 (4 RCTs)**	**⨁⨁◯◯ LOW**	
	60 per 1000	80 per 1000 (47 to 135)				

Downgraded for risk of bias, and imprecise results due to few events and participants in trials.

Wearing hip protectors may increase the risk of other fractures (low quality evidence)

GRADE

GDT exercise: Creating an Summary of Findings Tables (SoF)

http://www.gradepro.org

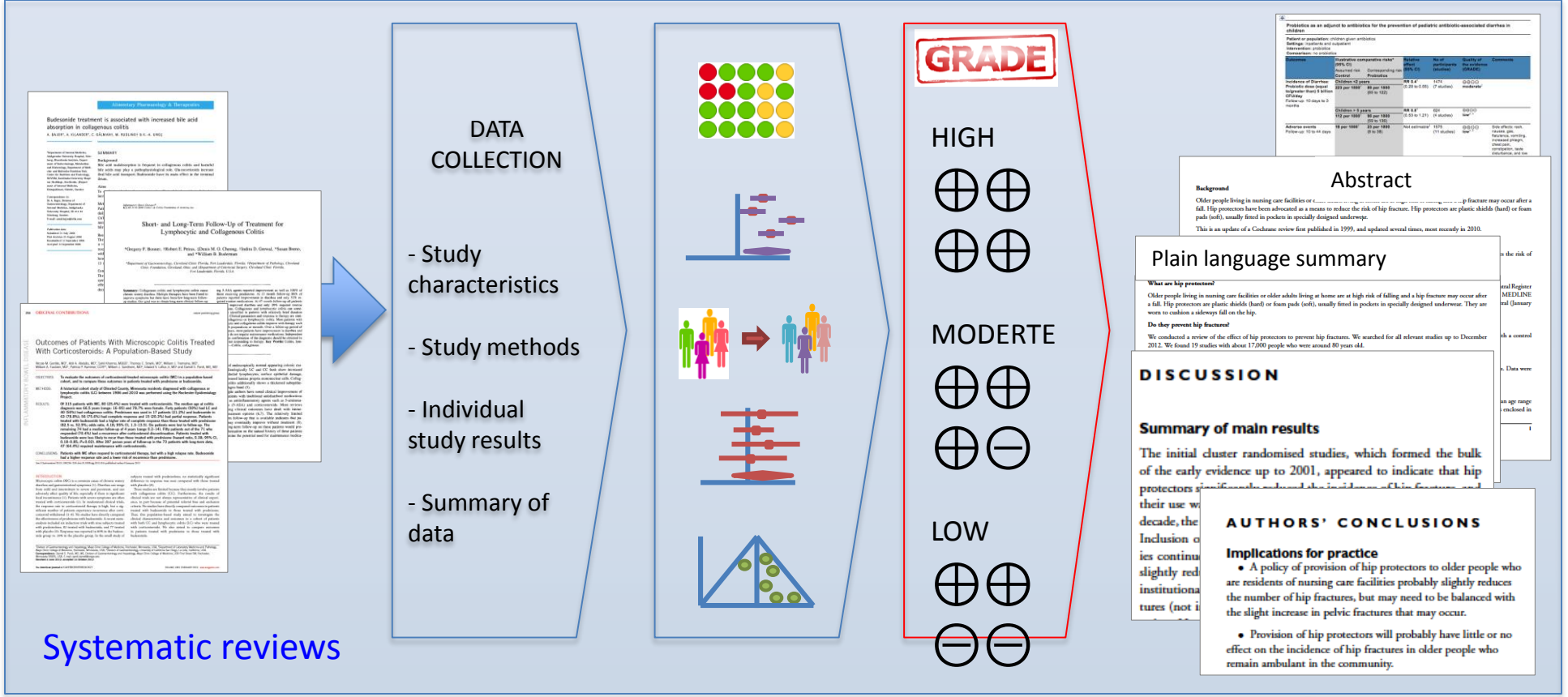
Revman file and example of SoF

http://cebgrade.mcmaster/hyderabad

GRADE

GRADE
for systematic reviews and for clinical practice guidelines

GRADE



Clinical practice guidelines

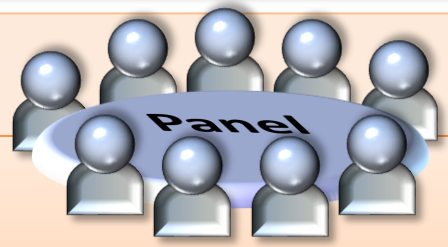
Balance between benefits and harms

Quality of evidence

Patients values and preferences

Resource utilization

Systematic reviews



Formulate recommendations

- For or against (direction) ↓↑
- Strong or weak (strength)

Additional resources

http://www.cochrane-handbook.org: Chapters 11 and 12

http://www.gradeworkinggroup.org

- Publications about GRADE criteria from Journal of Clinical Epidemiology and events

http://cebgrade.mcmaster.ca

- Online training modules for GRADE and Summary of Findings Tables

http://tech.cochrane.org/revman/other-resources/gradepro

- More training materials for GRADEproGDT

Contact us at support@gradepro.org

GRADE

Additional slides

What can raise quality?

Large magnitude

- Very large upgrade two levels (RR > 5 or RR < 0.2)
- Large upgrade 1 level (RR > 2 or RR < 0.5)

- everyone used to do badly but after treatment almost everyone does well

Example: parachutes to prevent death when jumping from airplanes

GRADE

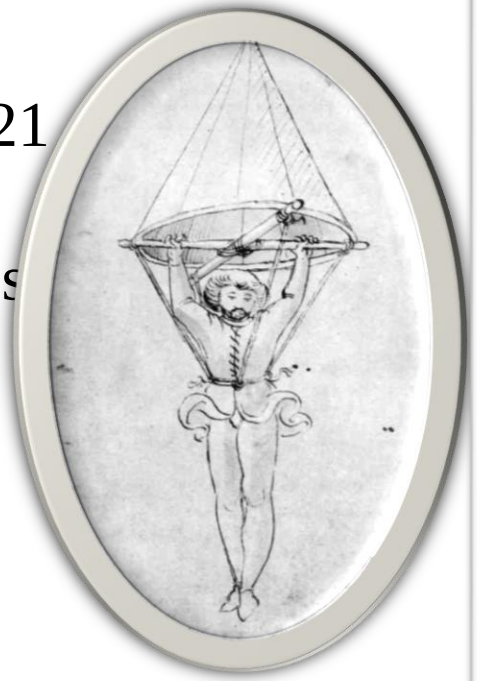
Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials

Gordon C S Smith, Jill P Pell

Relative risk reduction:

….> 99.9 % (1/100,000)

U.S. Parachute Association reported 821 injuries and 18 deaths out of 2.2 million jumps in 2007



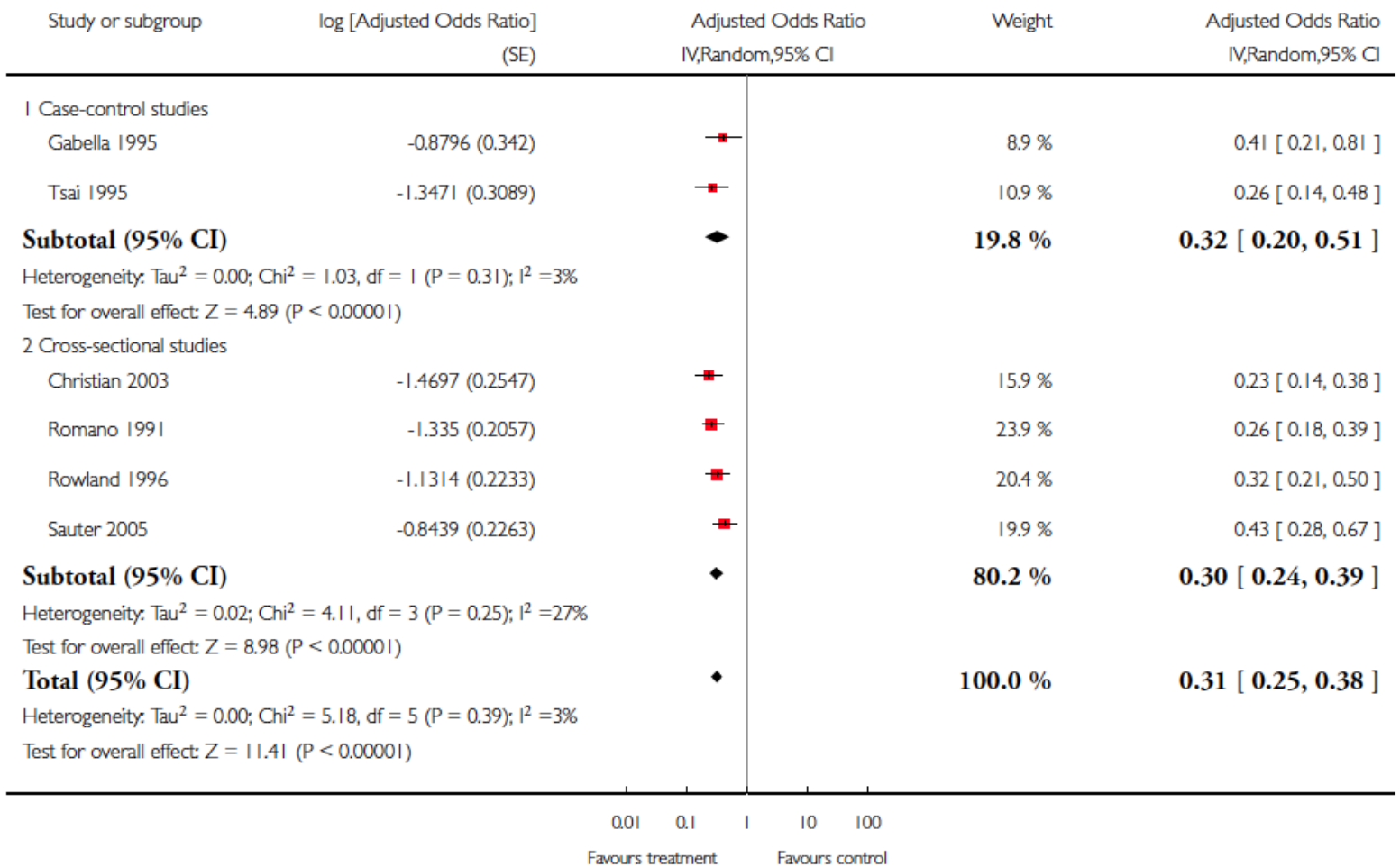
GRADE

Analysis 1.3. Comparison 1 Motorcycle helmet versus no helmet, Outcome 3 Head Injury (adjusted).

Review: Helmets for preventing injury in motorcycle riders

Comparison: 1 Motorcycle helmet versus no helmet

Outcome: 3 Head Injury (adjusted)



What can raise quality?

Dose response relation

Example: 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%)

GRADE

What can raise quality?

Effects of plausible residual confounding

- may be working to reduce the demonstrated effect

or

- increase the effect if no effect was observed

GRADE

All plausible residual confounding would result in an overestimate of effect

Example: Metformin

- Hypoglycaemic drug phenformin causes lactic acidosis
- The related agent metformin is under suspicion for the same toxicity.
- Large observational studies have failed to demonstrate an association even though clinicians would be more alert to lactic acidosis in the presence of the agent

Example: Vaccination and autism

GRADE

Other situations

Can you use GRADE?

GRADE

Narrative synthesis – no meta-analysis

Can still use criteria to evaluate quality of the evidence

Criteria to downgrade the evidence

1. Risk of bias
2. Inconsistency *(or heterogeneity)*
3. Indirectness *(PICO and applicability)*
4. Imprecision *(number of events and confidence intervals)*
5. Publication bias

Plus additional factors to increase the quality of the evidence

1. Response
2. size of effect
3. confounding

GRADE

Observational studies

Can still use criteria to evaluate quality of the evidence

Criteria to downgrade the evidence

1. **Risk of bias – OBSERVATIONAL STUDIES START AS LOW (new NRS tool?)**
2. Inconsistency
3. Indirectness
4. Imprecision
5. Publication bias

Plus additional factors to increase the quality of the evidence

1. Response
2. size of effect
3. confounding

GRADE

Only found 1 study

Can still use criteria to evaluate quality of the evidence

Criteria to downgrade the evidence

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

Plus additional factors to increase the quality of the evidence

1. Response
2. size of effect
3. confounding

GRADE