# GRADE for overviews:
# In systematic reviews we trust?

Joanne McKenzie[1] and Sue Brennan[2]

[1] School of Public Health and Preventive Medicine

[2] Cochrane Australia, School of Public Health and Preventive Medicine

MONASH University

# Declaration
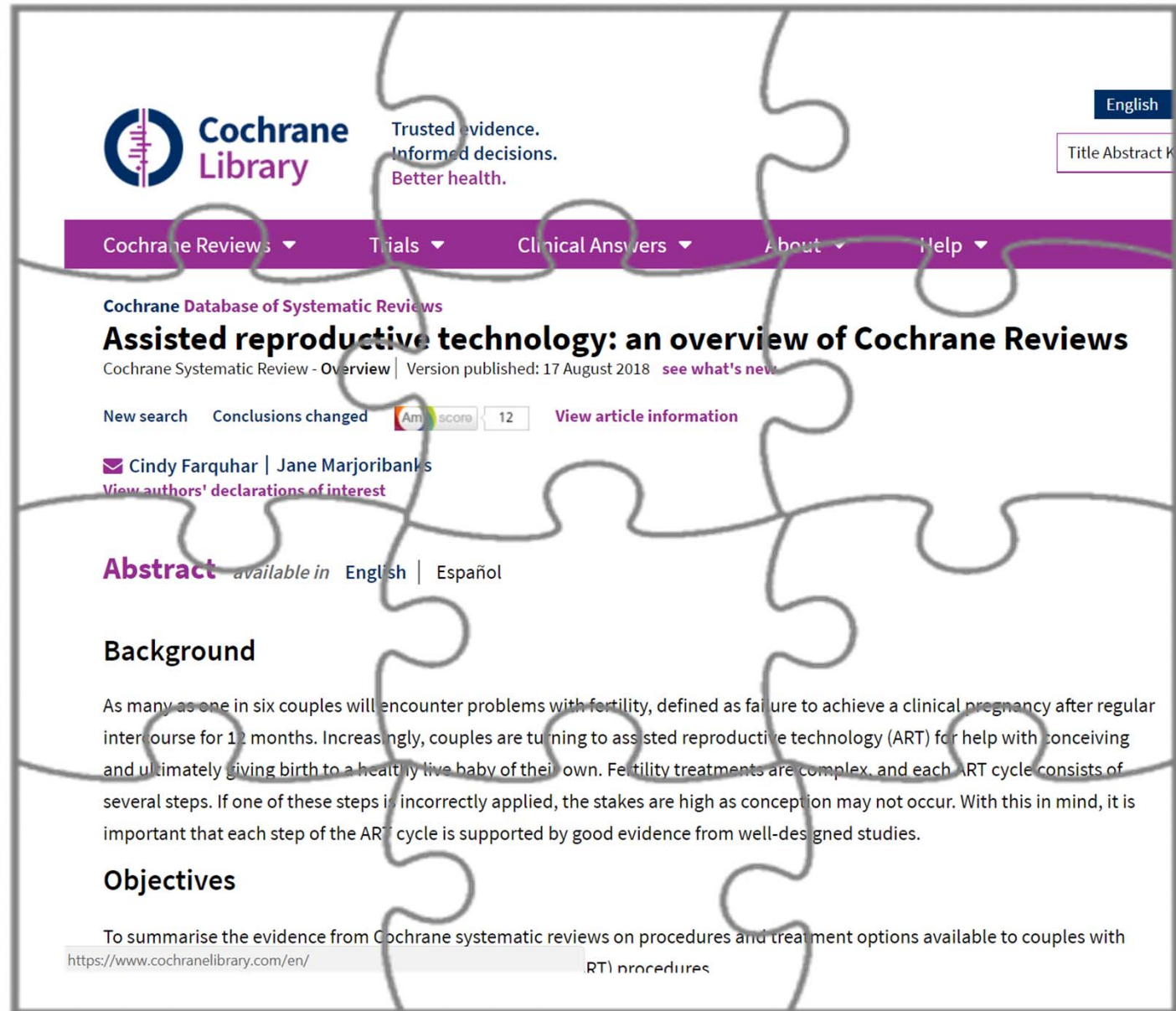
MONASH
University

# Why so much interest in overviews?

**Accessible:** front end for decision makers to access findings of multiple, related reviews

**68 Cochrane reviews on fertility treatments**

# Health advice

We provide Australians with the best available evidence-based advice about improving health and preventing disease.

We develop specific guidelines on various aspects of health, health care, health research and environmental health. We also review the latest research to provide information and advice about specific issues that might concern Australian communities.

Feedback

| Nutrition | Public health | Environmental health | Water quality and health | Alcohol |

## The Review of the Australian Government Rebate on Private Health Insurance for Natural Therapies

To synthesis evidence about the effects of natural therapies, overview methods were used to aggregate findings from

- **348** systematic reviews (**675** component trials, >75,000 participants)
- **68** conditions

Why overviews?

**Efficiency**: expedite the review of large and complex bodies of evidence

**Reduce waste**: capitalise on existing reviews rather than duplicating (41,563 SRs on PROSPERO)

# A compelling case for overviews, but …

**Unique challenges** arise in synthesising review evidence, for which guidance is yet to be agreed

These challenges have ramifications for **assessing certainty of the evidence** in an overview using GRADE or other approaches

For example, authors must decide:

- whether they will consider limitations of the included SRs in their assessment and, if so, how,

- how they will deal with overlapping data (e.g. where the same primary studies contribute to meta-analyses in multiple SRs), and

- how they will deal with missing or discordant data needed to assess certainty (e.g. conflicting risk of bias (RoB) assessments).

# This project

**Objective**: to inform methods guidance by examining **current practice** and **expert perspectives** on assessing the certainty of the evidence in overviews

# Current practice: systematic review of methods used to assess certainty of the evidence

Medline search for overviews
(01 Jan 2015 –  11 Mar 2017)
Validated search filter, Lunny et al JClinEpi 2016;74:107–18

Randomly sorted citations

Applied eligibility criteria to select **up to 50 overviews**
that reported on certainty of the evidence
- GRADE (RoB, indirectness, inconsistency, imprecision, publication bias, +/-)
- Other tool or method (e.g. EPC strength of evidence)
- A mix (as used in the SR)
- Other explicit criteria

```
┌─────────────────────────────────┐
│  1707 citations retrieved and    │
│  screened by 2 authors           │
└─────────────────────────────────┘
               ↓
┌─────────────────────────────────┐
│  381 full text articles screened │
│                                  │
│  Excluded:                       │
│  250 – not an overview           │
│  90 of 131 overviews did not     │
│  assess certainty (formally)     │
└─────────────────────────────────┘
               ↓
┌─────────────────────────────────┐
│  41 overviews met all            │
│  eligibility criteria and were   │
│  included                        │
└─────────────────────────────────┘
```

What tool was used?
- GRADE methods most common (19 overviews)
- 15 reported another tool or method, 1 a mix, 6 used own criteria

Who did the assessment?
- Most overview authors did their own assessment (35 overviews)
- 3 extracted assessments from included SRs
- 3 were unable to complete the assessment due to missing data for GRADE

Was anything done differently from an assessment of primary studies in a SR?
- No - GRADE was applied using the same approach as in a SR of primary studies
- Concerns about SR process were not integrated into the assessment

# Expert perspectives: interviews and working meeting with methodologists (n=23)

Participants (at interview and meeting) were asked to:

- consider how they would GRADE the evidence presented in each of four scenarios commonly encountered by overview authors

- describe the process and reasoning behind their GRADE assessment

Working meeting: asked participants to critique key considerations and approaches identified from interviews

# Scenario 1
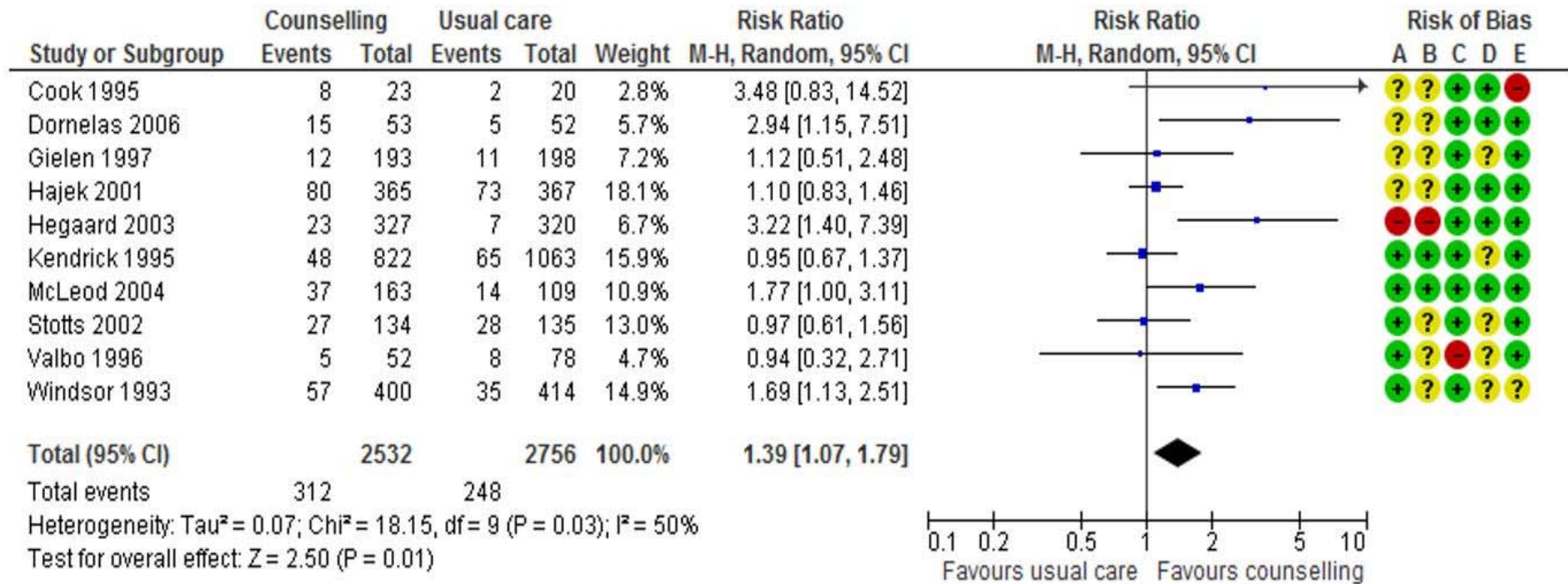
# Scenario 2



SR1
P. Women during pregnancy, any age

ROBIS (high risk 2 domains)

Concerns about methods used to select studies, collect data and appraise studies (one author screened full text, extracted data, assessed RoB)

AMSTAR: 10/11

SR1 2010
4 trials

6 trials

SR2 2015
3 trials

SR2
P. Women during pregnancy, ≥18 yr

ROBIS (high risk 2 domains)

Concerns about specification of eligibility criteria and synthesis (no a priori design; no assessment of likelihood of reporting bias)

AMSTAR: 9/11

**Overview**: P. Any population
I. Smoking cessation counselling

# Scenario 2

**Forest plot 1**

| Study or Subgroup | Counselling Events | Counselling Total | Usual care Events | Usual care Total | Weight | Risk Ratio M-H, Random, 95% CI | Risk of Bias A B C D E |
|---|---|---|---|---|---|---|---|
| Cook 1995 | 8 | 23 | 2 | 20 | 2.8% | 3.48 [0.83, 14.52] | ? ? + + − |
| Dornelas 2006 | 15 | 53 | 5 | 52 | 5.7% | 2.94 [1.15, 7.51] | ? ? + + + |
| Gielen 1997 | 12 | 193 | 11 | 198 | 7.2% | 1.12 [0.51, 2.48] | ? ? + ? + |
| Hajek 2001 | 80 | 365 | 73 | 367 | 18.1% | 1.10 [0.83, 1.46] | ? ? + + + |
| Hegaard 2003 | 23 | 327 | 7 | 320 | 6.7% | 3.22 [1.40, 7.39] | − − + + + |
| Kendrick 1995 | 48 | 822 | 65 | 1063 | 15.9% | 0.95 [0.67, 1.37] | + + + ? + |
| McLeod 2004 | 37 | 163 | 14 | 109 | 10.9% | 1.77 [1.00, 3.11] | + + + + + |
| Stotts 2002 | 27 | 134 | 28 | 135 | 13.0% | 0.97 [0.61, 1.56] | + ? + ? + |
| Valbo 1996 | 5 | 52 | 8 | 78 | 4.7% | 0.94 [0.32, 2.71] | + ? − ? + |
| Windsor 1993 | 57 | 400 | 35 | 414 | 14.9% | 1.69 [1.13, 2.51] | + ? + ? ? |
| **Total (95% CI)** | | 2532 | | 2756 | 100.0% | 1.39 [1.07, 1.79] | |
| Total events | 312 | | 248 | | | | |

Heterogeneity: Tau² = 0.07; Chi² = 18.15, df = 9 (P = 0.03); I² = 50%
Test for overall effect: Z = 2.50 (P = 0.01)

Favours usual care  Favours counselling

**Forest plot 2**

| Study or Subgroup | Counselling Events | Counselling Total | Usual care Events | Usual care Total | Weight | Risk Ratio M-H, Random, 95% CI | |
|---|---|---|---|---|---|---|---|
| Dornelas 2006 | 15 | 53 | 5 | 52 | 3.7% | 2.94 [1.15, 7.51] | Jadad 3/5 (Dornelas 2006) |
| Gielen 1997 | 12 | 193 | 11 | 198 | 4.9% | 1.12 [0.51, 2.48] | Jadad 2/5 (Gielen 1997) |
| Hajek 2001 | 80 | 365 | 73 | 367 | 20.0% | 1.10 [0.83, 1.46] | Jadad 4/5 (Hajek 2001) |
| Kendrick 1995 | 48 | 822 | 65 | 1063 | 15.6% | 0.95 [0.67, 1.37] | Jadad 4/5 (Kendrick 1995) |
| Lee 2015 | 21 | 140 | 16 | 137 | 7.7% | 1.28 [0.70, 2.35] | Jadad 3/5 (Lee 2015) |
| McLeod 2004 | 37 | 163 | 14 | 109 | 8.6% | 1.77 [1.00, 3.11] | Jadad 5/5 (McLeod 2004) |
| Moore 2002 | 88 | 523 | 108 | 567 | 21.7% | 0.88 [0.68, 1.14] | Jadad 5/5 (Moore 2002) |
| Stotts 2002 | 27 | 134 | 28 | 135 | 11.2% | 0.97 [0.61, 1.56] | Jadad 4/5 (Stotts 2002) |
| Wilkinson 2012 | 9 | 29 | 14 | 31 | 6.6% | 0.69 [0.35, 1.34] | Jadad 4/5 (Wilkinson 2012) |
| **Total (95% CI)** | | 2422 | | 2659 | 100.0% | 1.07 [0.89, 1.30] | |
| Total events | 337 | | 334 | | | | |

Heterogeneity: Tau² = 0.03; Chi² = 12.16, df = 8 (P = 0.14); I² = 34%
Test for overall effect: Z = 0.73 (P = 0.47)

Favours usual care  Favours counselling

# Scenario 3

| Study or Subgroup | Counselling Events | Total | Usual care Events | Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Cook 1995 | 8 | 23 | 2 | 20 | 2.0% | 3.48 [0.83, 14.52] |
| Dornelas 2006 | 15 | 53 | 5 | 52 | 4.1% | 2.94 [1.15, 7.51] |
| Gielen 1997 | 12 | 193 | 11 | 198 | 5.2% | 1.12 [0.51, 2.48] |
| Hajek 2001 | 80 | 365 | 73 | 367 | 13.2% | 1.10 [0.83, 1.46] |
| Hegaard 2003 | 23 | 327 | 7 | 320 | 4.9% | 3.22 [1.40, 7.39] |
| Kendrick 1995 | 48 | 822 | 65 | 1063 | 11.5% | 0.95 [0.67, 1.37] |
| Lee 2015 | 21 | 140 | 16 | 137 | 7.3% | 1.28 [0.70, 2.35] |
| McLeod 2004 | 37 | 163 | 14 | 109 | 7.9% | 1.77 [1.00, 3.11] |
| Moore 2002 | 88 | 523 | 108 | 567 | 13.8% | 0.88 [0.68, 1.14] |
| Stotts 2002 | 27 | 134 | 28 | 135 | 9.4% | 0.97 [0.61, 1.56] |
| Valbo 1996 | 5 | 52 | 8 | 78 | 3.4% | 0.94 [0.32, 2.71] |
| Wilkinson 2012 | 9 | 29 | 14 | 31 | 6.5% | 0.69 [0.35, 1.34] |
| Windsor 1993 | 57 | 400 | 35 | 414 | 10.8% | 1.69 [1.13, 2.51] |
| **Total (95% CI)** | | 3224 | | 3491 | 100.0% | 1.24 [1.00, 1.54] |
| Total events | 430 | | 386 | | | |

Heterogeneity: Tau² = 0.07; Chi² = 26.17, df = 12 (P = 0.01); I² = 54%
Test for overall effect: Z = 1.92 (P = 0.05)

Risk of Bias: A B C D E

Jadad 3/5 (Lee)
Jadad 5/5 (Moore)
Jadad 4/5 (Wilkinson)

Favours usual care   Favours counselling
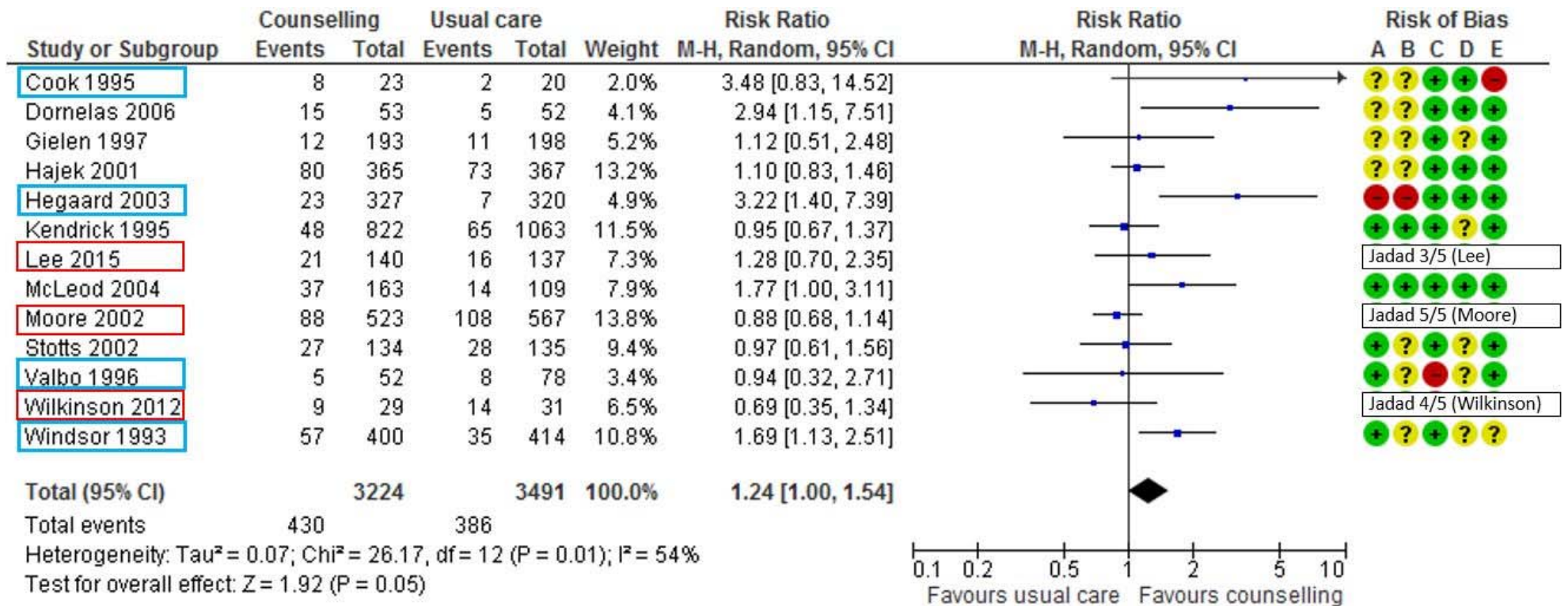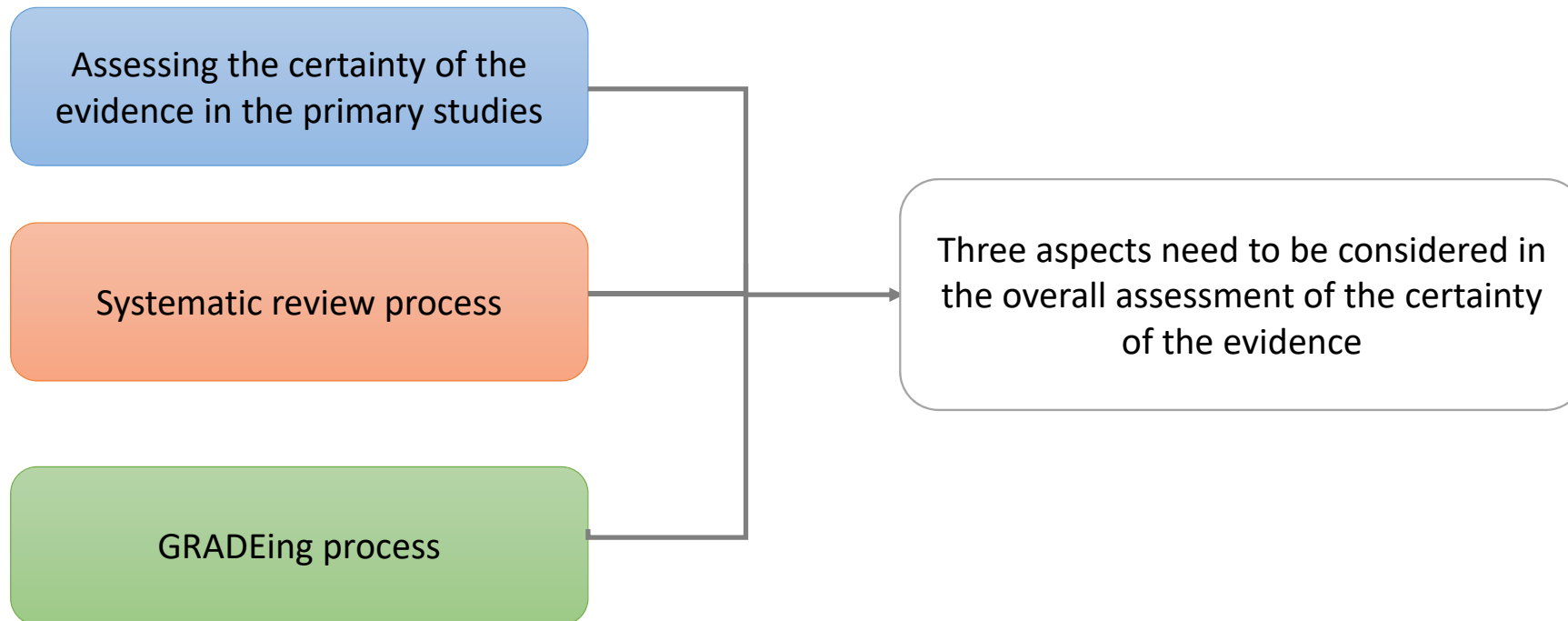
# Results

- Findings of key considerations, methods / approaches are presented

# Aspects to assessing the certainty of the evidence: interview participants

Assessing the certainty of the evidence in the primary studies

Systematic review process

GRADEing process

Three aspects need to be considered in the overall assessment of the certainty of the evidence

# Options proposed for integrating concerns about the SR process

Option 1: separate GRADE assessment of the primary studies and of the SR process

# Options proposed for integrating concerns about the SR process

Option 2: add new domain(s) to GRADE to assess the SR process
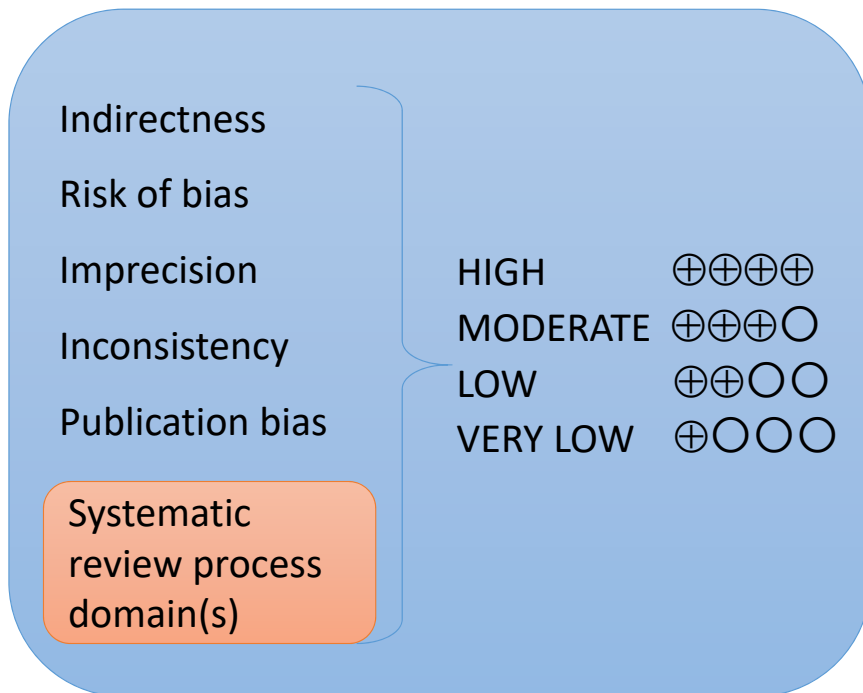
Indirectness

Risk of bias

Imprecision

Inconsistency

Publication bias

Systematic review process domain(s)

HIGH ⊕⊕⊕⊕
MODERATE ⊕⊕⊕◯
LOW ⊕⊕◯◯
VERY LOW ⊕◯◯◯

# Options proposed for integrating concerns about the SR process

Option 3: map concerns in the SR process to current GRADE domains



"In terms of **review indirectness at SR level it is ok**, but may be a **problem at the overview level**."
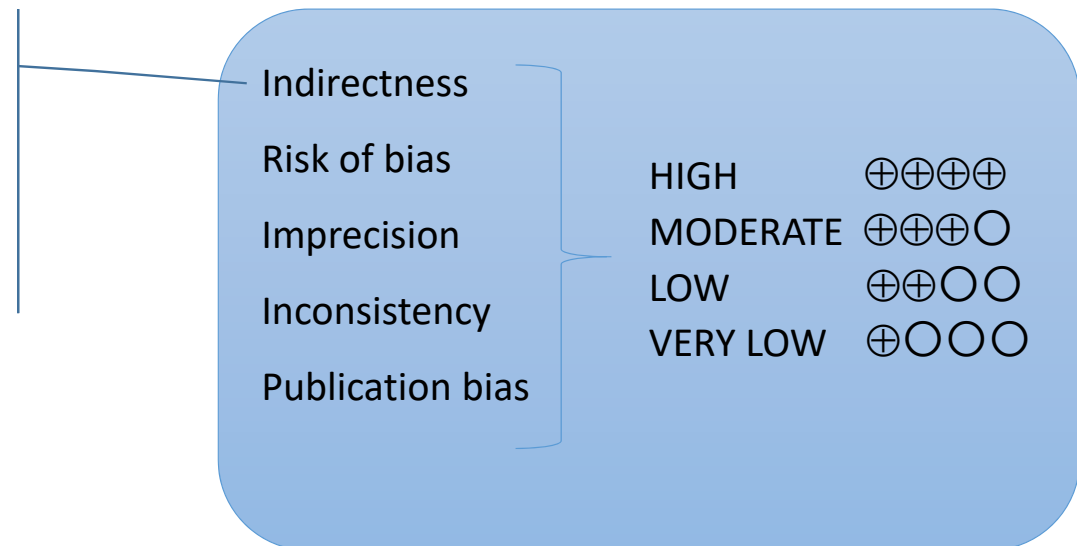
Indirectness

Risk of bias

Imprecision

Inconsistency

Publication bias

HIGH        ⊕⊕⊕⊕
MODERATE  ⊕⊕⊕○
LOW         ⊕⊕○○
VERY LOW   ⊕○○○

# Options proposed for integrating concerns about the SR process

Option 3: map concerns in the SR process to current GRADE domains

"To me it seems kind of intuitively sensible to try to come up with some kind of combined assessment based on RoB across the two if possible. If **high RoB in two ROBIS domains, plus potential moderate RoB within the included studies, might say overall concerned about high RoB**, in terms of the overall assesment of RoB."

Indirectness

Risk of bias

Imprecision

Inconsistency

Publication bias

HIGH          ⊕⊕⊕⊕
MODERATE   ⊕⊕⊕○
LOW           ⊕⊕○○
VERY LOW   ⊕○○○

# Options proposed for integrating concerns about the SR process

**Option 3:** map concerns in the SR process to current GRADE domains

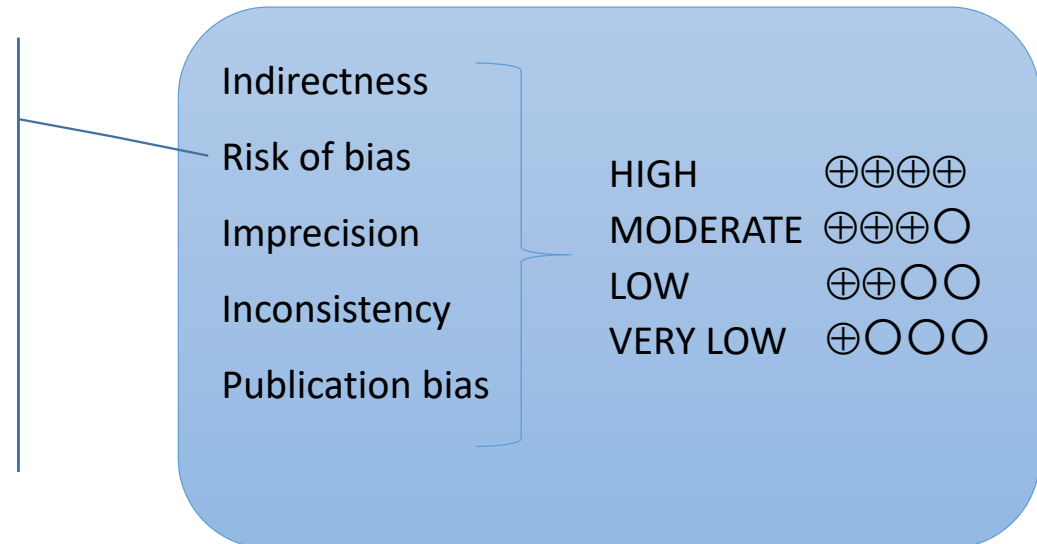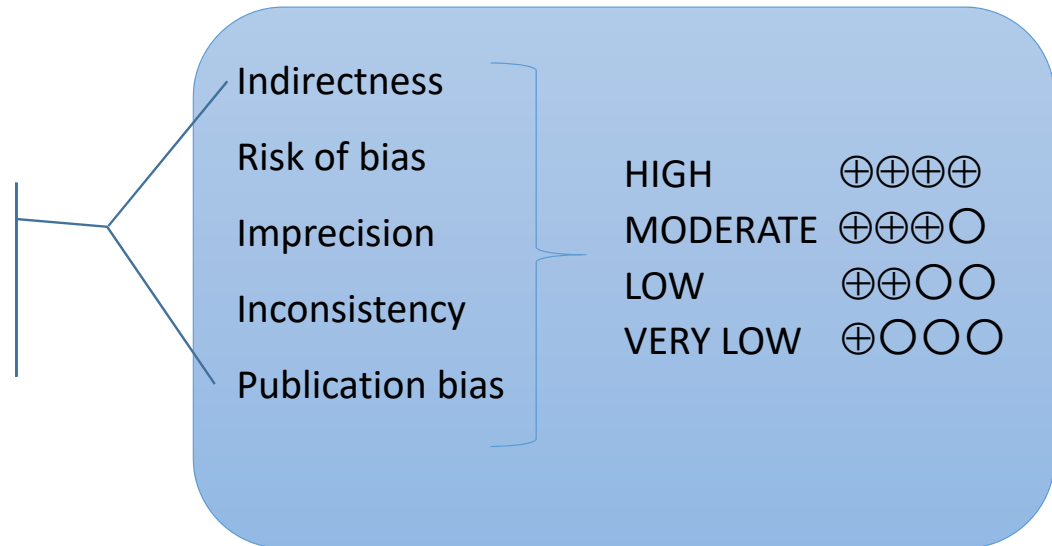"**If search is not comprehensive** maybe fits onto indirectness. Or **publication bias, which we always say look at [the] search strategy**."

Indirectness

Risk of bias

Imprecision

Inconsistency

Publication bias

HIGH ⊕⊕⊕⊕
MODERATE ⊕⊕⊕◯
LOW ⊕⊕◯◯
VERY LOW ⊕◯◯◯

# Options proposed for integrating concerns about the SR process: meeting participants' views

**Option 1:** separate GRADE assessment of the primary studies and of the SR process

→? Participants' felt that as a minimum an assessment of the SR process should be reported in the SoFs table

**Option 2:** add new domain(s) to GRADE to assess the SR process

→ ? Some participants' felt that uncertainty arising from the SR process was an artefact and should not be integrated with the assessment of primary studies

**Option 3:** map concerns in the SR process to current GRADE domains

→ Rejected. Participants' felt that many concerns about the SR process do not map well to current GRADE domains

# Concerns in the SR process: what were these?

From the <span style="color:red">interviews participants</span>, variability in opinion about which elements of the SR process may introduce bias

# "One author extracted data"

"… it bothers me but it **doesn't bother me that much**, especially if the single person has **expertise** in an area …"
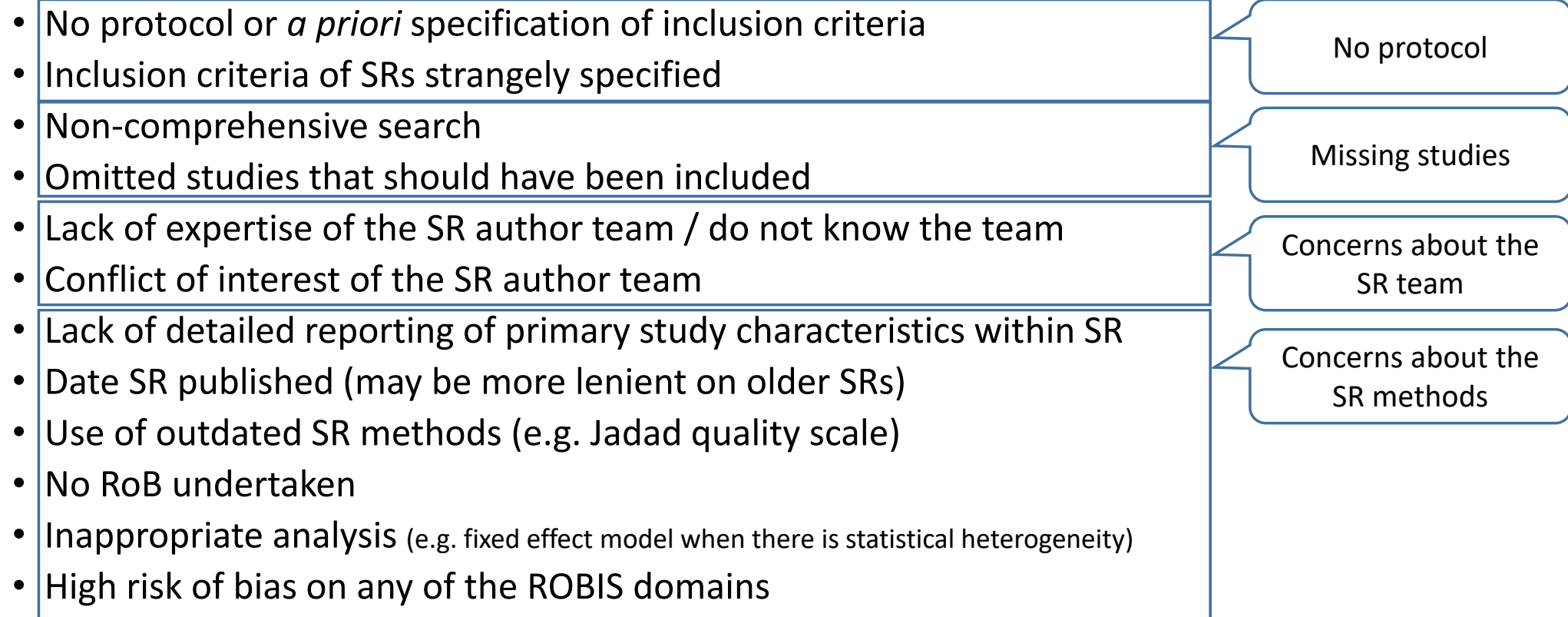
"… **not sure if the best way is for two people [to] do … data extraction** and so on, as **I think it also heavily depends on the experience**, so even though it is a gold standard, then **I'm not sure** …"

"… **Data extraction** and risk of bias assessment are the **most worrying factors described here**…"

Less concerned

More concerned

# SR process: elements of concern

- No protocol or *a priori* specification of inclusion criteria
- Inclusion criteria of SRs strangely specified

> No protocol

- Non-comprehensive search
- Omitted studies that should have been included

> Missing studies

- Lack of expertise of the SR author team / do not know the team
- Conflict of interest of the SR author team

> Concerns about the SR team

- Lack of detailed reporting of primary study characteristics within SR
- Date SR published (may be more lenient on older SRs)
- Use of outdated SR methods (e.g. Jadad quality scale)
- No RoB undertaken
- Inappropriate analysis (e.g. fixed effect model when there is statistical heterogeneity)
- High risk of bias on any of the ROBIS domains

> Concerns about the SR methods

# Additional challenges with results from multiple SRs



| Study or Subgroup | Counselling Events | Total | Usual care Events | Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Cook 1995 | 8 | 23 | 2 | 20 | 2.0% | 3.48 [0.83, 14.52] |
| Dornelas 2006 | 15 | 53 | 5 | 52 | 4.1% | 2.94 [1.15, 7.51] |
| Gielen 1997 | 12 | 193 | 11 | 198 | 5.2% | 1.12 [0.51, 2.48] |
| Hajek 2001 | 80 | 365 | 73 | 367 | 13.2% | 1.10 [0.83, 1.46] |
| Hegaard 2003 | 23 | 327 | 7 | 320 | 4.9% | 3.22 [1.40, 7.39] |
| Kendrick 1995 | 48 | 822 | 65 | 1063 | 11.5% | 0.95 [0.67, 1.37] |
| Lee 2015 | 21 | 140 | 16 | 137 | 7.3% | 1.28 [0.70, 2.35] |
| McLeod 2004 | 37 | 163 | 14 | 109 | 7.9% | 1.77 [1.00, 3.11] |
| Moore 2002 | 88 | 523 | 108 | 567 | 13.8% | 0.88 [0.68, 1.14] |
| Stotts 2002 | 27 | 134 | 28 | 135 | 9.4% | 0.97 [0.61, 1.56] |
| Valbo 1996 | 5 | 52 | 8 | 78 | 3.4% | 0.94 [0.32, 2.71] |
| Wilkinson 2012 | 9 | 29 | 14 | 31 | 6.5% | 0.69 [0.35, 1.34] |
| Windsor 1993 | 57 | 400 | 35 | 414 | 10.8% | 1.69 [1.13, 2.51] |
| **Total (95% CI)** | | 3224 | | 3491 | 100.0% | 1.24 [1.00, 1.54] |
| Total events | 430 | | 386 | | | |

Heterogeneity: Tau² = 0.07; Chi² = 26.17, df = 12 (P = 0.01); I² = 54%
Test for overall effect: Z = 1.92 (P = 0.05)

Risk Ratio M-H, Random, 95% CI
0.1  0.2    0.5    1    2    5    10
Favours usual care    Favours counselling

Risk of Bias A B C D E

Jadad 3/5 (Lee)
Jadad 5/5 (Moore)
Jadad 4/5 (Wilkinson)

**How to deal with different RoB / quality scales across SRs?** Should we treat trials where quality scales have been reported as 'missing assessments' and downgrade?

Does having knowledge that there were missing studies lead to upgrading (now have some of the missing studies), or downgrading (evidence that authors missed studies, so there may be others)?

Do differences in the data reported and RoB assessments for the same studies across the SRs lead to upgrading (if no differences) or downgrading (if differences)?

# Conclusions: assessing the certainty of evidence

Assessing the certainty of evidence in overviews

- Limitations of the systematic review process need to be considered in assessing the certainty of evidence
- Different options for how this could be done
  - Separate GRADE assessment of the primary studies and of the SR process
  - Add new domain(s) to GRADE to assess the SR process
  - Map concerns in the SR process to current GRADE domains
- Diverse views on the elements of the SR process that are most important to consider when assessing certainty
- Next step is to develop principles for applying GRADE in overviews with project group members

# Conclusions: broader issues

- Meta-epidemiological evidence has informed which elements of trial conduct may lead to biased intervention estimates
    - → Cochrane RoB tool
- More limited empirical evidence to guide which elements of the SR process may lead to biased intervention estimates

# Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records

Phil Edwards[1,*,†], Mike Clarke[2], Carolyn DiGuiseppi[3], Sarah Pratap[4], Ian Roberts[1] and Reinhard Wentz[4]

## Testing the Risk of Bias tool showed low reliability between individual reviewers and across consensus assessments of reviewer pairs

Lisa Hartling[a,*], Michele P. Hamm[a], Andrea Milne[a], Ben Vandermeer[a], P. Lina Santaguida[b], Mohammed Ansari[c], Alexander Tsertsvadze[c], Susanne Hempel[d], Paul Shekelle[d], Donna M. Dryden[a]

## Dual computer monitors to increase efficiency of conducting systematic reviews

Zhen Wang[a,*], Noor Asi[a], Tarig A. Elraiyah[a], Abd Moain Abu Dabrh[a], Chaitanya Undavalli[a], Paul Glasziou[b], Victor Montori[a], Mohammad Hassan Murad[a]

[a]Knowledge & Evaluation Research Unit, Mayo Clinic, 200 First Street SW, Rochester, MN 55904, USA
[b]Centre for Research in Evidence-Based Practice, Faculty of Health Sciences and Medicine, Bond University, Queensland 4229, Australia
Accepted 1 June 2014; Published online 30 July 2014

# Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews

## Disagreements in meta-analyses using outcomes measured on continuous or rating scales: observer agreement study

Britta Tendal, PhD student,[1] Julian P T Higgins, senior statistician,[4] Peter Jüni, head of division,[2,3] Asbjørn Hróbjartsson, senior researcher,[1] Sven Trelle, associate director,[2,3] Eveline Nüesch, PhD student,[2,3] Simon Wandel, PhD student,[2,3] Anders W Jørgensen, PhD student,[1] Katarina Gesser, PhD student,[5] Søren Ilsøe-Kristensen, PhD student,[5] Peter C Gøtzsche, director[1]

Cochrane Database of Systematic Reviews

## Bias due to selective inclusion and reporting of outcomes and analyses in systematic reviews of randomised trials of healthcare interventions

Matthew J Page | Joanne E McKenzie | Jamie Kirkham | Kerry Dwan | Sharon Kramer | Sally Green | Andrew Forbes
View authors' declarations of interest

# Conclusions: broader issues (continued)

- Meta-epidemiological evidence has informed which elements of trial conduct may lead to biased intervention estimates
    - → Cochrane RoB tool
- More limited empirical evidence to guide which elements of the SR process may lead to biased intervention estimates
- Having more empirical evidence of the key elements would provide a better foundation for knowing which SRs we should trust more
- Systematic collation of these empirical evaluations, with continued updating (particularly in the assessment of automation technologies), is required

# Acknowledgements