GRADing quality of evidence and strength of recommendations

Gordon Guyatt
McMaster University
**Why this talk?**

- You will be seeing a lot of GRADE
- Exemplifies three key principles of EBHC
  - Need for systematic reviews of best evidence
  - Hierarchy of evidence
  - Need for values and preferences
- If you understand GRADE, you understand how to use evidence to inform practice
Outline

- GRADE background
- Two steps
  - Quality of evidence (certainty, confidence)
  - Strength of recommendation
- Evidence profiles
- An exercise in applying GRADE
GRADE (Grades of recommendation, assessment, development and evaluation)

International group

- Australian NMRC, SIGN, USPSTF, WHO, NICE, Oxford CEBM, CDC, CC

- ~ 40 meetings over last 18 years

- Started with management
  - Now diagnosis and prognosis
  - This presentation focuses on management
>110 organizations have adopted GRADE
What are we grading?

Two components

NOT GRADING individual studies, bodies of evidence

Strength of recommendation: Strong and weak (conditional)
Determinants of confidence

- RCTs start high
- Observational studies start low
- What can lower confidence?
GRADE assessment of quality/certainty/confidence in evidence

Risk of Bias

Inconsistency of results

Indirectness of evidence

Imprecision of results

Publication bias
GRADE assessment of quality/certainty/confidence in evidence

Risk of Bias
- Failure of blinding
- Incomplete reporting
- Allocation concealment
- Loss to follow-up

Inconsistency of results
- Indirectness of evidence
- Imprecision of results
- Publication bias
GRADE assessment of quality/certainty/confidence in evidence

- Small number industry-funded studies
- Publication bias
- Risk of Bias
- Indirectness of evidence
- Imprecision of results
- Inconsistency of results
GRADE assessment of quality/certainty/confidence in evidence

- Indirectness of evidence
  - Patients, interventions, Comparators, outcomes
- Applicability/generalizability
- Risk of Bias
- Inconsistency of results
- Imprecision of results
- Publication bias
GRADE assessment of quality/certainty/confidence in evidence

- Imprecision of results
- Wide confidence intervals
- Indirectness of evidence
- Inconsistency of results
- Publication bias
- Risk of Bias
GRADE assessment of quality/certainty/confidence in evidence

- Inconsistency of results
- Risk of Bias
  - Indirectness of evidence
- Imprecision of results
- Publication bias
Inconsistency – happy with these results?

Relative Risk (95% CI)

0.73 (0.49, 1.07)

0.74 (0.59, 0.94)

0.76 (0.51, 1.12)

0.71 (0.56, 0.90)

0.73 (0.61, 0.88)
What about these?

Relative Risk (95% CI)

- 0.44 (0.30, 0.65)
- 0.45 (0.36, 0.60)
- 1.25 (0.84, 1.84)
- 1.17 (0.92, 1.49)
- 0.73 (0.61, 0.88)
GRADE assessment of quality/certainty/confidence in evidence

- Inconsistency of results
  - Dissimilar point estimates
  - High $I^2$
  - Non-overlapping confidence intervals
  - Low $p$-value
  - Heterogeneity test

- Indirectness of evidence

- Imprecision of results

- Publication bias

- Risk of Bias
Relative Risk (95% CI)

- 0.44 (0.30, 0.65)
- 0.45 (0.36, 0.60)
- 1.25 (0.84, 1.84)
- 1.17 (0.92, 1.49)
- 0.73 (0.61, 0.88)

If these results what next?
What can raise confidence?

- Large magnitude can rate up one level
  - Very large two levels

- Common criteria
  - Everyone used to do badly
  - Almost everyone does well
  - Quick action

- Dialysis in renal failure
- Insulin in diabetic ketoacidosis
## Certainty assessment criteria

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Confidence in estimates</th>
<th>Lower if</th>
<th>Higher if</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized trials</td>
<td>High</td>
<td>Risk of bias</td>
<td>Large Effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Large</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td>+1 Very large</td>
</tr>
<tr>
<td></td>
<td>Moderate</td>
<td>Inconsistency</td>
<td>Dose response</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 Evidence of a gradient</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td>All plausible confounding</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>Indirectness</td>
<td>+1 Would reduce a demonstrated effect or</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td>+1 would suggest a spurious effect when results show no effect</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td></td>
</tr>
<tr>
<td>Observational studies</td>
<td>Low</td>
<td>Imprecision</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Serious</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very serious</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Very Low</td>
<td>Publication bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-1 Likely</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-2 Very likely</td>
<td></td>
</tr>
</tbody>
</table>
# Beta blockers in non-cardiac surgery

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Number of participants (studies)</th>
<th>Risk of Bias</th>
<th>Consistency</th>
<th>Directness</th>
<th>Precision</th>
<th>Publication Bias</th>
<th>Quality</th>
<th>Relative Effect (95% CI)</th>
<th>Absolute risk difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myocardial infarction</td>
<td>10,125 (9)</td>
<td>No serious limitations</td>
<td>No serious limitations</td>
<td>No serious limitations</td>
<td>No serious limitations</td>
<td>Not detected</td>
<td>High</td>
<td>0.71 (0.57 to 0.86)</td>
<td>1.5% fewer (0.7% fewer to 2.1% fewer)</td>
</tr>
<tr>
<td>Mortality</td>
<td>10,205 (7)</td>
<td>No serious limitations</td>
<td>inconsistent</td>
<td>No serious limitations</td>
<td>Serious Imprecision</td>
<td>Not detected</td>
<td>Low</td>
<td>1.23 (0.98 – 1.55)</td>
<td>0.5% more (0.1% fewer to 1.3% more)</td>
</tr>
<tr>
<td>Stroke</td>
<td>10,889 (5)</td>
<td>No serious limitations</td>
<td>No serious limitations</td>
<td>No serious limitations</td>
<td>No serious limitations</td>
<td>Not detected</td>
<td>High</td>
<td>2.21 (1.37 – 3.55)</td>
<td>0.5% more (0.2% more to 1.3% more)</td>
</tr>
</tbody>
</table>
**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 RR of death, at least 1%
  - 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 1.5% per year
Strength of recommendations

Aspirin after MI – do it

Anticoagulants vs ASA in low risk Afib
-- probably do it
-- probably don’t do it
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
Venotonic agents
- increase venous return

Popularity
- 90 venotonics commercialized in France
- none in Sweden and Norway
- France 70% of world market

Possibilities
- French misguided
- rest of world missing out
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 quality 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 placebo
Outcome: 08 Overall improvement: no improvement/some improvement

<table>
<thead>
<tr>
<th>Study or sub-category</th>
<th>log[RR] (SE)</th>
<th>RR (random)</th>
<th>Weight %</th>
<th>RR (random)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 Up to seven days</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chauvenet</td>
<td>-0.8916 (0.2376)</td>
<td></td>
<td>12.67</td>
<td>0.41 [0.26, 0.65]</td>
</tr>
<tr>
<td>Cospite</td>
<td>-2.2073 (0.6117)</td>
<td></td>
<td>5.51</td>
<td>0.11 [0.03, 0.36]</td>
</tr>
<tr>
<td>Thanapongsathorn</td>
<td>-0.4308 (0.2985)</td>
<td></td>
<td>11.16</td>
<td>0.65 [0.36, 1.17]</td>
</tr>
<tr>
<td><strong>Subtotal (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>29.36</strong></td>
<td>0.37</td>
<td><strong>0.18, 0.77</strong></td>
</tr>
<tr>
<td>Test for heterogeneity: Chi² = 6.92, df = 2 (P = 0.03), I² = 71.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 2.67 (P = 0.008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02 Up to four weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annoni F</td>
<td>-1.6094 (0.7073)</td>
<td></td>
<td>4.50</td>
<td>0.20 [0.05, 0.80]</td>
</tr>
<tr>
<td>Clyne MB</td>
<td>-0.9943 (0.3983)</td>
<td></td>
<td>8.94</td>
<td>0.37 [0.17, 0.81]</td>
</tr>
<tr>
<td>Pirard J</td>
<td>-1.1712 (0.3086)</td>
<td></td>
<td>10.94</td>
<td>0.31 [0.17, 0.57]</td>
</tr>
<tr>
<td>Thanapongsathorn</td>
<td>-1.1087 (1.1098)</td>
<td></td>
<td>2.18</td>
<td>0.33 [0.04, 2.91]</td>
</tr>
<tr>
<td>Thorp</td>
<td>0.2624 (0.3291)</td>
<td></td>
<td>10.46</td>
<td>1.30 [0.68, 2.48]</td>
</tr>
<tr>
<td>Titapan</td>
<td>-0.8916 (0.3691)</td>
<td></td>
<td>9.56</td>
<td>0.41 [0.20, 0.85]</td>
</tr>
<tr>
<td>Wijayanegara</td>
<td>-0.5978 (0.1375)</td>
<td></td>
<td>14.97</td>
<td>0.55 [0.42, 0.72]</td>
</tr>
<tr>
<td><strong>Subtotal (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>61.54</strong></td>
<td>0.48</td>
<td><strong>0.32, 0.72</strong></td>
</tr>
<tr>
<td>Test for heterogeneity: Chi² = 13.87, df = 6 (P = 0.03), I² = 56.7%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 3.57 (P = 0.0004)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>03 Further than four weeks</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Godeberg</td>
<td>-1.7719 (0.3906)</td>
<td></td>
<td>9.10</td>
<td>0.17 [0.08, 0.37]</td>
</tr>
<tr>
<td><strong>Subtotal (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>9.10</strong></td>
<td>0.17</td>
<td><strong>0.08, 0.37</strong></td>
</tr>
<tr>
<td>Test for heterogeneity: not applicable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 4.54 (P &lt; 0.00001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>100.00</strong></td>
<td>0.40</td>
<td><strong>0.29, 0.57</strong></td>
</tr>
<tr>
<td>Test for heterogeneity: Chi² = 28.66, df = 10 (P = 0.001), I² = 65.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test for overall effect: Z = 5.14 (P &lt; 0.00001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Publication bias?

- Size of studies
  - 40 to 234 patients, most around 100
- All industry sponsored
What can lower confidence?

- Risk of bias
  - Lack of detail re concealment
  - Questionnaires not validated

- Inconsistency
  - Almost all show positive effect, trend
  - Heterogeneity p < 0.001; I² 65.1%

- Indirectness

- Imprecision
  - RR 0.4, 95% CI 0.29 to 0.57

- Publication bias
  - 40 to 234 patients, all industry sponsored
Is France right?

- **Recommendation**
  - yes
  - no against use

- **Strength**
  - strong
  - weak
Clinicians, policy makers need summaries
- Confidence in evidence
- Strength of recommendations

Explicit rules
- Transparent, informative

GRADE
- Complex assessment
  - simple as possible, transparent, systematic
- Increasing wide adoption
- Captures all key elements of EBM approach
@EBCPMcMaster
Follow us on Twitter