Experiences presenting GRADE to the Guideline Development Group on the NICE Lower Urinary tract Symptoms (LUTS) Guideline

Clare Jones
Senior Research Fellow/Project Manager

The NCGC is a governance collaboration, hosted by the RCP and funded by NICE

Background


- LUTS in men is one of the first official clinical guidelines developed by NICE to pilot using GRADE.

- NICE previously used SIGN system of assessing the quality of evidence
Aim

- To discuss the experiences and challenges of presenting GRADE quality assessment to the LUTS guideline development group (GDG)

Methods

- Results of systematic reviews were presented to the GDG using GRADE
- This was used to develop recommendations for the guideline
- Trialled combination of different methods of presenting results with GRADE to determine best approach for our GDG to interpret
Challenges

Multidisciplinary audience
- GDG made up of clinicians and patient representatives

Results

- Initial presentation of GRADE with systematic review
  - Present quality assessment rating with evidence statements
  - No further information as how rating decided

- Revised presentations of GRADE
  - Present extensive details for the criteria considered when rating evidence quality of outcomes
Criteria considered for rating outcomes

- **Limitations in design:**
  - Consider randomisation method, allocation concealment, blinding and loss to follow-up

- **Inconsistency:**
  - Consider magnitude of statistical and clinical heterogeneity between studies and explain any subgroup analysis

- **Indirectness:**
  - Does the patient population and intervention fit directly with the guideline?

- **Imprecision:**
  - How wide are the confidence intervals around the estimate of effect relative to the clinically important threshold? How many events and patients in total? Is the effect clinically significant?

Limitations in design

- Presented study limitations using a traffic light system to highlight the risks of bias
Inconsistency

Presented and discussed forest plots and heterogeneity

<table>
<thead>
<tr>
<th>Study or Subgroup</th>
<th>Mean</th>
<th>SD</th>
<th>Total</th>
<th>Mean</th>
<th>SD</th>
<th>Total</th>
<th>Mean Difference</th>
<th>IV Random, 95% CI</th>
<th>Mean Difference</th>
<th>IV Random, 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HIV</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>April 2017</td>
<td>2.2</td>
<td>1.6</td>
<td>94</td>
<td>2.2</td>
<td>1.6</td>
<td>94</td>
<td>-1.20 (2.10, -0.30)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gupta 2006</td>
<td>2.2</td>
<td>1.6</td>
<td>59</td>
<td>2.2</td>
<td>1.6</td>
<td>59</td>
<td>-0.90 (1.60, -0.20)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mantini 2014</td>
<td>3.9</td>
<td>3.9</td>
<td>75</td>
<td>3.9</td>
<td>3.9</td>
<td>75</td>
<td>1.40 (2.00, 0.80)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vestergen 2014</td>
<td>3.8</td>
<td>3.8</td>
<td>61</td>
<td>3.8</td>
<td>3.8</td>
<td>61</td>
<td>1.20 (2.00, 0.40)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall (95% CI)</td>
<td>6.5</td>
<td>5.1</td>
<td>283</td>
<td>6.5</td>
<td>5.1</td>
<td>283</td>
<td>1.20 (1.90, 0.50)</td>
<td>0.52 (1.35, -0.33)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Indirectness</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does the patient population and intervention fit directly with the guideline?</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Indirectness

Does the patient population and intervention fit directly with the guideline?
Imprecision

- Confidence intervals, minimally important differences and optimal information size values were presented to determine whether outcomes were imprecise.

Example of GRADE imprecision summary

**IPSS – optimal sample size calculation**

- Minimally important difference patient can detect is 3 points on IPSS score. Paper by Barry et al., (1995) shows for n=347 patients with AUA = slight improvement the absolute difference in IPSS score is 3 ± 5.03.

- Optimal sample size calculation using these parameters gives 88 patients.

**Follow up** | **OIS (both arms)** | **Total from meta-analysis** | **Sample size > OIS** | **Confidence interval cross MID** | **GRADE Imprecise ?**
--- | --- | --- | --- | --- | ---
3 mths | 88 | 213 | Yes | Yes | Yes
6 mths | 88 | 165 | Yes | Yes | Yes
1 yr | 88 | 283 | Yes | Yes | Yes
2 yrs | 88 | 139 | Yes | Yes | Yes
3 yrs | 88 | 165 | Yes | Yes | Yes
+5 yrs | 88 | 89 | No | Yes | Yes

Power 80% at p=0.05 significance

Use for power calculation [http://newton.stat.ubc.ca/~rollin/stats/ssize/n2.html](http://newton.stat.ubc.ca/~rollin/stats/ssize/n2.html)
### Example of presented GRADE tables

**HoLEP vs. TURP – clinical summary of findings**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Number of Studies</th>
<th>Design</th>
<th>Limitations</th>
<th>Inconsistency</th>
<th>Directness</th>
<th>Other considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptom score at 6 months</td>
<td>5</td>
<td>RCT</td>
<td>Serious limitations (a)</td>
<td>Serious heterogeneity (c)</td>
<td>No serious indirectness</td>
<td>No serious imprecision</td>
</tr>
<tr>
<td>Symptom score at 12 months</td>
<td>5</td>
<td>RCT</td>
<td>Serious limitations (a)</td>
<td>Serious heterogeneity (c)</td>
<td>No serious indirectness</td>
<td>No serious imprecision</td>
</tr>
<tr>
<td>Symptom score at 24 months</td>
<td>3</td>
<td>RCT</td>
<td>Serious limitations (a)</td>
<td>Serious heterogeneity (c)</td>
<td>No serious indirectness</td>
<td>No serious imprecision</td>
</tr>
</tbody>
</table>

- **a)** 4 studies did not report allocation concealment or masked outcome assessment. One study did not report randomisation method used.
- **b)** Statistically significant heterogeneity is present.

---

### Example of presented GRADE tables

**Clinical Evidence Statements - Symptom scores**

There is no statistically significant difference between HoLEP and TURP in reducing symptom scores at 6, 12 and 24 months postoperatively. (LOW)
Quality ratings

- Knowledge of literature on topic
- Concern regarding evidence with low quality or very low quality

Quality ratings

- GRADE definition of quality provided to group to illustrate the difference between the GRADE and NICE previous system
Overall quality for outcome evidence

- **High:**
  - further research is very unlikely to change our confidence in the estimate of effect

- **Moderate:**
  - further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate

- **Low:**
  - 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:**
  - any estimate of effect is very uncertain

Key Points

- Critical to include the right level of detail
- Transparency of process important
- Ensure the GDG understand the process and support the quality rating which underpins the decisions and recommendations made
LUTS guideline

- Published May 2010
- http://guidance.nice.org.uk/CG97

THANK YOU!