What does GRADE bring to the table – clarity or confusion?

Emma Newbatt and Zosia Beckles

National Collaborating Centre for Women's and Children's Health (UK)

G-I-N Conference Berlin 2012

Disclosure of Interests (last 3 years)

Emma Newbatt and Zosia Beckles

We certify that, to the best of our knowledge, no aspect of our current personal or professional situation might reasonably be expected to affect significantly our views on the subject on which we are presenting, other than the following:

- Employment at the National Collaborating Centre for Women's and Children's Health, funded by NICE

Background: Grading of Recommendations Assessment, Development and Evaluation

Developed by the GRADE Working Group:

- A system for grading the quality of evidence and the strength of the recommendations that arise from it
- Aims to combine what is best in existing grading systems (e.g. AHA, ACCP, SIGN)
- Rates quality of evidence by outcome rather than by study
- Presents both absolute and relative effects
- Promotes transparent and systematic quality appraisal

(http://www.gradeworkinggroup.org/)

Background: Guideline development in England and Wales

- National Institute for Health and Clinical Excellence (NICE) commissions four National Collaborating Centres (NCCs) to produce clinical guidelines in the UK
- NCCs conduct systematic reviews and present evidence to multidisciplinary guideline development groups (GDGs)
- GDGs draft evidence-based or consensus recommendations for best practice, based on clinical and cost-effectiveness data

Background: Rationale for project

- GRADE has changed how we present evidence to guideline development groups:
  - Less use of forest plots
  - Move almost entirely to the use of GRADE tables as basis for discussion
- Does this mask trends in the data and place too much emphasis on statistical significance?
- Two stages of project:
  - Part I: current methods of presenting data
  - Part II: evaluating the effect of the transition to GRADE

Background: Literature review

- Previous studies have looked at how information framing methods including numerical data presentation to ranking in systems such as GRADE affect clinicians' decisions [1-4]

- Large body of work on how to apply GRADE; previous presentations at G-I-N have discussed GDG views on GRADE and various modifications to GRADE tables [5-8]

- Unable to find studies comparing the effects of different methods of information framing on guideline development and recommendations

Part I: Current methods of presenting data

Variety of methods used across four collaborating centres:

- GRADE tables
- Forest plots
- Summary of findings tables (adapted from GRADE)

Part II: Survey of GDG members

- Designed three surveys of hypothetical case studies – each survey presented the same data but in one of three different formats:
  - GRADE profile
  - Forest plots
  - Summary tables
- GDG members randomised to receive data on primary outcomes in one of the three formats
- Respondents asked to choose the recommendation that they would make, if all other outcomes were non-harmful:
  - Strongly recommend this intervention
  - Weakly recommend this intervention
  - Not recommend this intervention (make a “do not offer intervention” recommendation)
  - Make a recommendation for further research about the intervention
  - Don’t know

Survey A: GRADE profiles

Quality assessment							No of patients		Effect		Quality
No of studies	Design	Risk of bias	Inconsistency	Indirectness	Imprecision	Other considerations	Chocolate	No intervention	Relative (95% CI)	Absolute (95% CI)	
No further hair loss after 2 months											
1 meta-analysis of 8 studies	randomised trials	no serious risk of bias	no serious inconsistency	no serious indirectness	serious[1]	none	238/466 (51.1%)	212/470 (45.1%)	RR 1.13 (1 to 1.28)	59 more per 1000 (from 0 more to 126 more)	Moderate

[1] Wide confidence interval

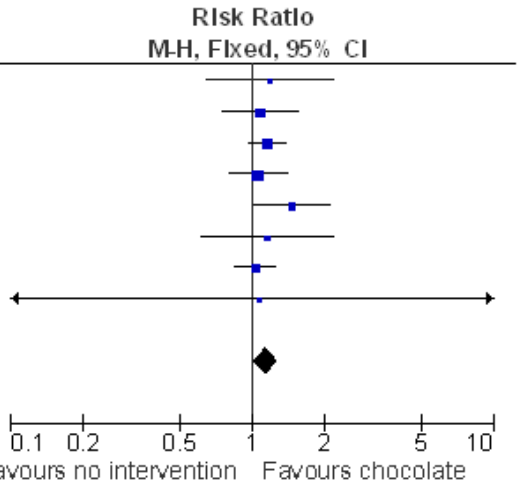
Survey B: Forest plots

Outcome: No further hair loss after 2 months

Study or Subgroup	Chocolate Events	Chocolate Total	No Intervention Events	No Intervention Total	Weight	Risk Ratio M-H, Fixed, 95% CI
Study A	16	47	13	45	6.3%	1.18 [0.64, 2.16]
Study B	42	120	39	120	18.5%	1.08 [0.76, 1.54]
Study E	45	50	39	50	18.5%	1.15 [0.97, 1.37]
Study F	54	113	52	115	24.4%	1.06 [0.80, 1.40]
Study H	34	56	25	60	11.4%	1.46 [1.01, 2.10]
Study I	12	25	10	24	4.8%	1.15 [0.62, 2.15]
Study J	34	40	33	40	15.6%	1.03 [0.85, 1.25]
Study K	1	15	1	16	0.5%	1.07 [0.07, 15.57]
Total (95% CI)		**466**		**470**	**100.0%**	**1.13 [1.00, 1.28]**
Total events	238		212			

Heterogeneity: Chi² = 3.12, df = 7 (P = 0.87); I² = 0%

Test for overall effect: Z = 1.98 (P = 0.05)



Survey C: Summary tables

Outcomes	Number of RCTs (total number of participants)	Imprecision	GRADE quality rating	Absolute difference (per 1000 patients) Chocolate	Comparison event rate (per 1000 patients) No intervention	• Clinical benefit • Not of clinical benefit • Unsure of clinical benefit • Clinical harm • Not of clinical harm • Unsure of clinical harm
No further hair loss after 2 months	8 (n = 936)	serious imprecision[1]	Moderate	59 more per 1000 (from 0 more to 126 more)	451 per 1000 patients	Unsure of clinical benefit

[1] Wide confidence interval

Sample

- 1-3 patient/lay/carer representatives per guideline development group; therefore stratified to ensure spread across survey allocation
- 283 GDG members were contacted including 59 lay members
- Chi-squared test used to analyse data

Results

- 33.9% response rate (range 28.4% - 41.1%)

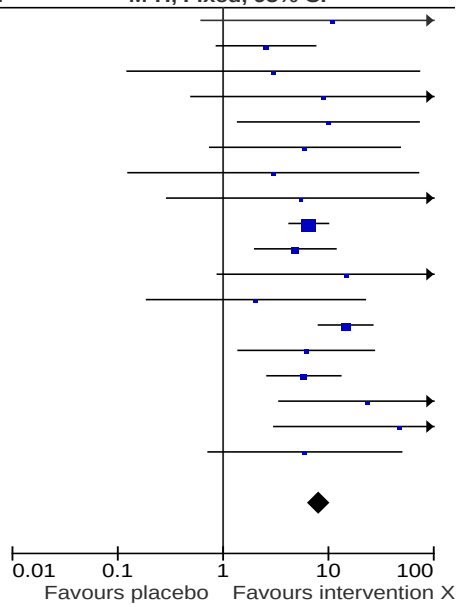
Survey	Responses					
	Clinicians	%	Lay members	%	Total	%
GRADE	21/75	28.0%	6/20	30.0%	27/95	28.4%
Forest plots	32/75	42.7%	7/20	35.0%	39/95	41.1%
Summary tables	21/74	28.4%	9/19	47.4%	30/93	32.3%
Total	74/224	33.0%	22/59	37.3%	96/283	33.9%

Question 1

- Treatment success: RR 8.13 (95% CI 6.26 to 10.55)
- Strong, consistent benefit shown in trials

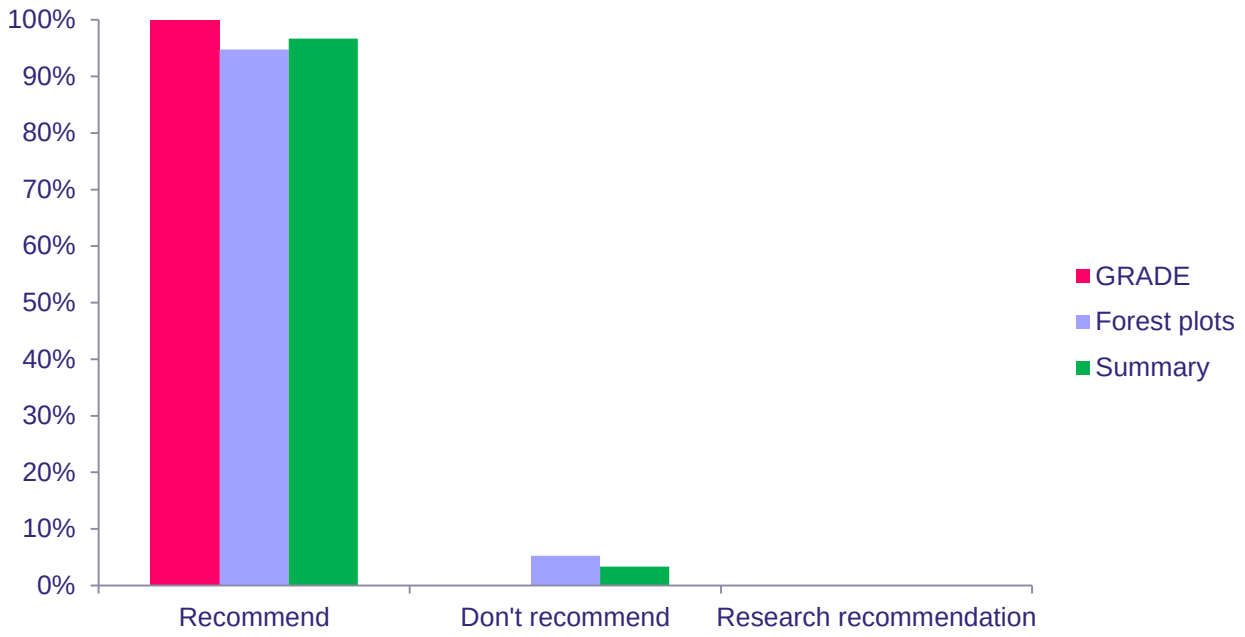
	Intervention X		Placebo			Risk Ratio	Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Fixed, 95% CI	M-H, Fixed, 95% CI
Study A	6	116	0	97	0.9%	10.89 [0.62, 190.88]	
Study B	10	43	4	44	6.5%	2.56 [0.87, 7.54]	
Study C	1	150	0	150	0.8%	3.00 [0.12, 73.06]	
Study D	4	47	0	47	0.8%	9.00 [0.50, 162.62]	
Study E	10	25	1	25	1.6%	10.00 [1.38, 72.39]	
Study F	6	49	1	49	1.6%	6.00 [0.75, 48.01]	
Study G	1	40	0	40	0.8%	3.00 [0.13, 71.51]	
Study H	2	10	0	11	0.8%	5.45 [0.29, 101.55]	
Study J	142	398	22	402	35.7%	6.52 [4.25, 9.99]	
Study K	80	488	5	148	12.5%	4.85 [2.00, 11.75]	
Study L	7	40	0	40	0.8%	15.00 [0.89, 254.13]	
Study M	2	113	1	116	1.6%	2.05 [0.19, 22.33]	
Study N	164	321	11	314	18.1%	14.58 [8.08, 26.32]	
Study O	12	218	2	224	3.2%	6.17 [1.40, 27.23]	
Study P	37	79	6	75	10.0%	5.85 [2.62, 13.06]	
Study Q	20	23	1	27	1.5%	23.48 [3.41, 161.73]	
Study R	35	45	0	30	1.0%	47.85 [3.05, 751.35]	
Study S	6	348	1	347	1.6%	5.98 [0.72, 49.43]	
Total (95% CI)		**2553**		**2186**	**100.0%**	**8.13 [6.26, 10.55]**	
Total events	545		55				

Heterogeneity: Chi² = 16.55, df = 17 (P = 0.49); I² = 0%
Test for overall effect: Z = 15.76 (P < 0.00001)



Question 1: results

- As expected, almost all respondents recommend intervention
- No significant difference between proportions recommending intervention with any of the methods



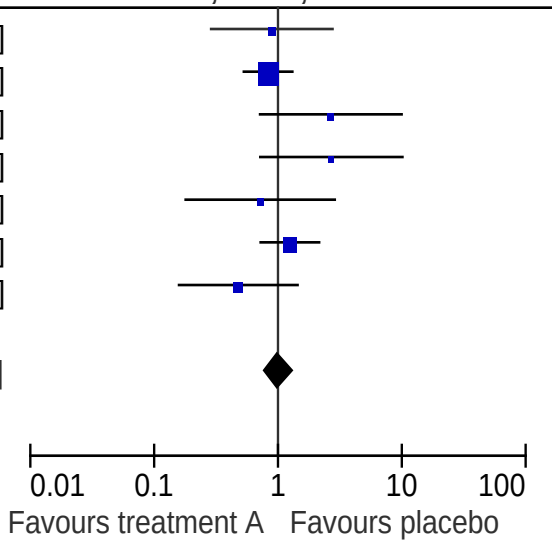
Question 3

- Continuing symptoms at 48 hours: RR 1.02 (95% CI 0.76 to 1.36)
- Relative risk close to no effect – consistent lack of effect in trials

	Treatment A		Placebo			Risk Ratio	Risk Ratio
Study or Subgroup	**Events**	**Total**	**Events**	**Total**	**Weight**	**M-H, Fixed, 95% CI**	**M-H, Fixed, 95% CI**
Study 1	5	56	6	60	8.2%	0.89 [0.29, 2.76]	
Study 2	25	100	30	100	42.5%	0.83 [0.53, 1.31]	
Study 3	8	297	3	298	4.2%	2.68 [0.72, 9.99]	
Study 4	6	10	2	9	3.0%	2.70 [0.72, 10.14]	
Study 5	3	25	4	24	5.8%	0.72 [0.18, 2.89]	
Study 6	19	60	18	71	23.4%	1.25 [0.72, 2.16]	
Study 7	4	56	10	67	12.9%	0.48 [0.16, 1.44]	
Total (95% CI)		**604**		**629**	**100.0%**	**1.02 [0.76, 1.36]**	
Total events	70		73				

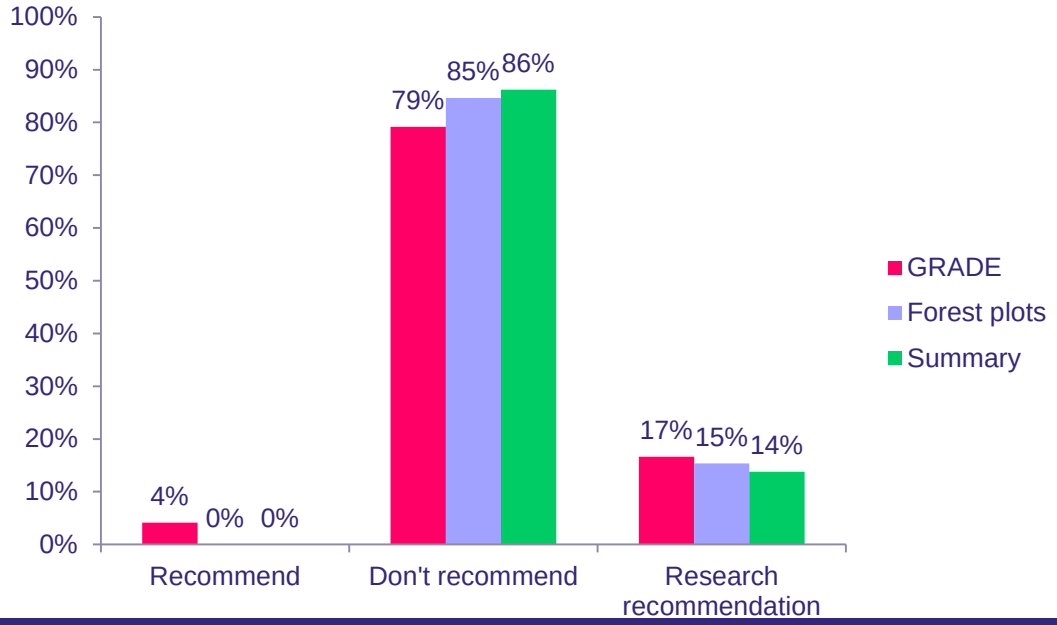
Heterogeneity: Chi² = 7.53, df = 6 (P = 0.27); I² = 20%

Test for overall effect: Z = 0.11 (P = 0.91)



Question 3: results

- As expected, almost all respondents do not recommend intervention
- No significant difference between proportions recommending intervention with any of the methods



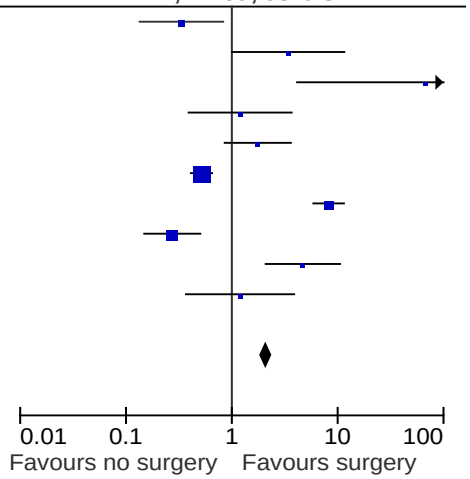
Question 4

- No need for hearing aid: RR 2.13 (95% CI 1.84 to 2.46)
- Consideration of point estimate alone suggests surgery should be recommended
- Forest plot illustrates variability in demonstrated trial effects

	Surgery		No surgery			Risk Ratio
Study or Subgroup	Events	Total	Events	Total	Weight	M-H, Fixed, 95% CI
Study A	6	116	15	97	7.9%	0.33 [0.13, 0.83]
Study B	10	43	3	44	1.4%	3.41 [1.01, 11.55]
Study C	33	150	0	150	0.2%	67.00 [4.14, 1083.53]
Study F	6	49	5	49	2.4%	1.20 [0.39, 3.67]
Study H	8	10	5	11	2.3%	1.76 [0.86, 3.61]
Study J	45	100	87	100	42.3%	0.52 [0.41, 0.65]
Study K	300	488	34	455	17.1%	8.23 [5.91, 11.45]
Study O	12	218	45	224	21.6%	0.27 [0.15, 0.50]
Study Q	20	23	5	27	2.2%	4.70 [2.10, 10.52]
Study S	6	348	5	347	2.4%	1.20 [0.37, 3.88]
Total (95% CI)		**1545**		**1504**	**100.0%**	**2.13 [1.84, 2.46]**
Total events	446		204			

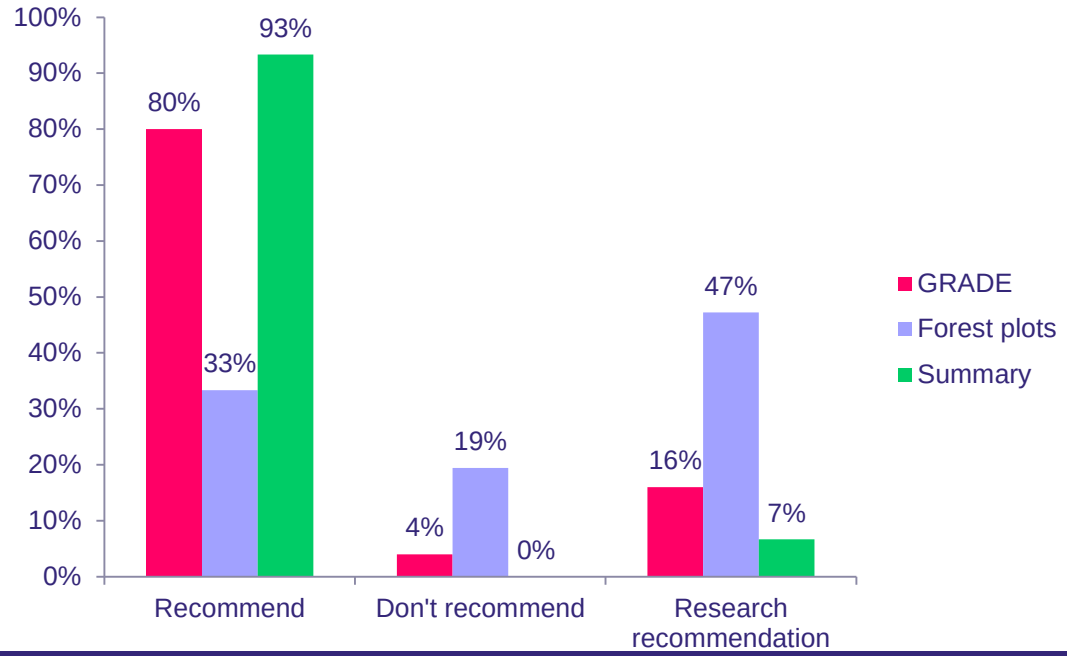
Heterogeneity: Chi² = 281.84, df = 9 (P < 0.00001); I² = 97%

Test for overall effect: Z = 10.10 (P < 0.00001)



Question 4: results

- GRADE vs. forest plots: p < 0.0005
- Summary table vs. forest plots: p < 0.0005
- GRADE vs. summary table: p = 0.139

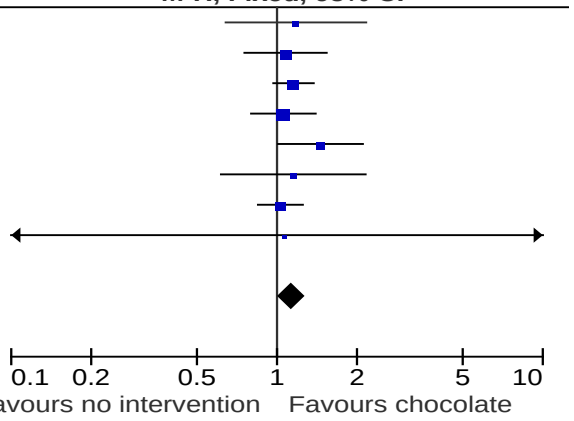


Question 6

- No further hair loss after 2 months: RR 1.13 (1.00 to 1.28)
- Relative risk touching unity; consistent (but often non-significant) benefit shown on forest plot
- Summary table states “unsure of clinical benefit”

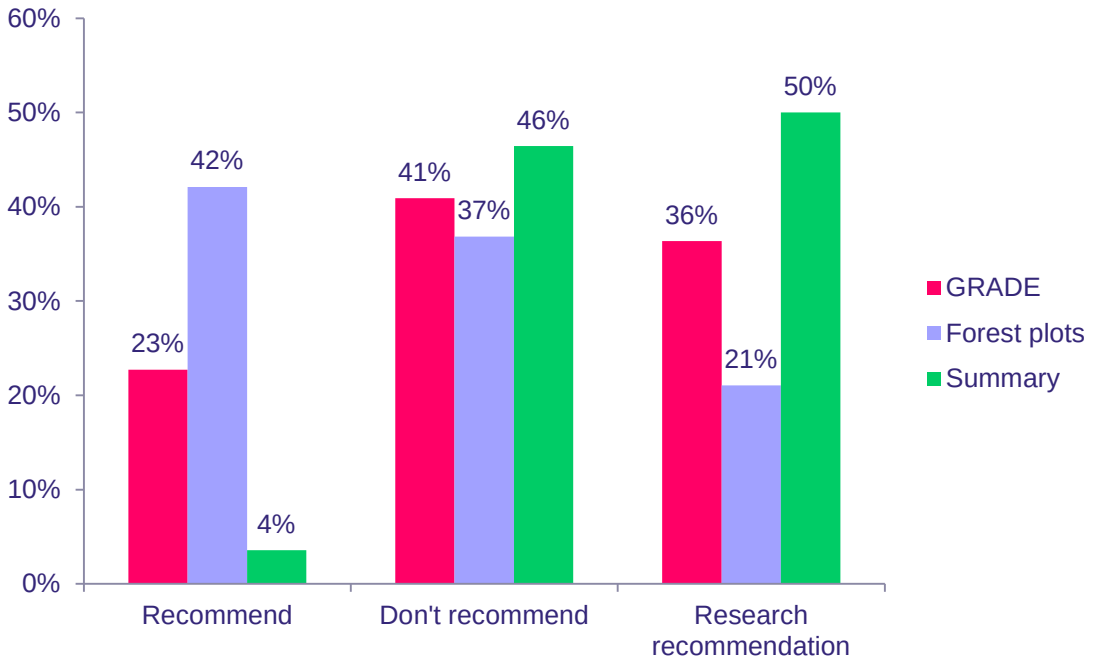
Study or Subgroup	Chocolate Events	Chocolate Total	No intervention Events	No intervention Total	Weight	Risk Ratio M-H, Fixed, 95% CI
Study A	16	47	13	45	6.3%	1.18 [0.64, 2.16]
Study B	42	120	39	120	18.5%	1.08 [0.76, 1.54]
Study E	45	50	39	50	18.5%	1.15 [0.97, 1.37]
Study F	54	113	52	115	24.4%	1.06 [0.80, 1.40]
Study H	34	56	25	60	11.4%	1.46 [1.01, 2.10]
Study I	12	25	10	24	4.8%	1.15 [0.62, 2.15]
Study J	34	40	33	40	15.6%	1.03 [0.85, 1.25]
Study K	1	15	1	16	0.5%	1.07 [0.07, 15.57]
Total (95% CI)		**466**		**470**	**100.0%**	**1.13 [1.00, 1.28]**
Total events	238		212			

Heterogeneity: Chi² = 3.12, df = 7 (P = 0.87); I² = 0%
Test for overall effect: Z = 1.98 (P = 0.05)



Question 6: results

- Chi-squared overall: p = 0.00728



Summary

Significant difference in proportion recommending intervention?						
	GRADE vs. forest plot		GRADE vs. summary table		Summary table vs. forest plot	
Q1	**No**	(p = 0.3)	**No**	(p = 0.4)	**No**	(p = 0.7)
Q2	**Yes**	(p = 0.04)	**Yes**	(p = 0.008)	**No**	(p = 0.4)
Q3	**No**	(p = 0.2)	**No**	(p = 0.3)	**No**	(p > 0.5)
Q4	**Yes**	(p < 0.0005)	**No**	(p = 0.1)	**Yes**	(p < 0.0005)
Q5	**No**	(p = 0.3)	**Yes**	(p = 0.002)	**Yes**	(p = 0.02)
Q6	**No**	(p = 0.1)	**Yes**	(p = 0.04)	**Yes**	(p < 0.0005)

- Significant difference in proportion choosing strong or weak recommendation in 8/18 (44%) cases

Unexpected answers

- "Don't know"
  - GRADE: 11% of responses
  - Forest plots: 2% of responses
  - Summary tables: 2% of responses
- Recommended opposite to expected:
  - GRADE: 14% of responses
  - Forest plots: 18% of responses
  - Summary tables: 12% of responses
- 0 - 33% of "don't know" or incorrect answers were from people unfamiliar with the mode of presentation (mean 18%)

Qualitative responses: overall

- Combination of data display methods preferred

  "The combination of figures, narratives and grades help understanding of the overall results"

- Mixed feelings about GRADE

  "I really struggle with following the GRADE tables"

- Narrative summaries are useful, especially for non-RCTs

  "...narrative summaries help me to understand better"

- Forest plots generally easy to understand and many liked the visual representation of the direction of evidence

  "Forest plots provide strong pictorial sense of what is effective or not"

Qualitative responses: lay members

- More training/explanation of data display methods required

  "I find GRADE tables confusing and not particularly helpful without extra explanation"

- Combination of data display methods preferred

  "GRADE tables without narrative summaries are very difficult to interpret"

- Need for data on patient experience – best presented through narrative summaries

  "Need to have qualitative as well as quantitative data to fully understand how patients have responded"

Limitations

- Artificial situation:
  - Lack of cost-effectiveness data
  - Normally multiple outcomes would be considered
- Familiarity with method:
  - Each centre does something different
  - Some people randomised to GRADE may never have seen it before
  - Attempt to control/evaluate this
- Data collection:
  - Low response rate

Conclusions

- Different methods of data display may lead to different recommendations where evidence is variable
- GRADE is a useful tool for transparently appraising evidence; however, compared to other tools many people do not find it intuitive
- GDG members generally prefer a combination of data display methods
- Narrative summaries are an important adjunct to visual data representations

Implications for guideline developers

- Consider presenting data using a combination of methods to ensure decision-makers have the best chance of understanding research findings
- Allow sufficient time and resources for training guideline developers in different data display methods
- Further consideration and more formal research in this area would be valuable to inform methodology and help developers avoid inadvertently introducing bias through the way that evidence is presented

References

1. Cuello Garcia CA, Pacheco Alvarado KP, and Perez GG. Grading recommendations in clinical practice guidelines: randomised experimental evaluation of four different systems. *Archives of Disease in Childhood* 2011; 96:(8)723-8.
2. Covey J. A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making* 2007; 27:(5)638-54.
3. Rosenbaum SE, Glenton C, Nylund HK *et al.* User testing and stakeholder feedback contributed to the development of understandable and useful Summary of Findings tables for Cochrane reviews. *Journal of Clinical Epidemiology* 2010; 63:(6)607-19.
4. Rosenbaum SE, Glenton C, and Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. *Journal of Clinical Epidemiology* 2010; 63:(6)620-6.
5. Jones CN, Chong LY, Fenu E, Hill J, and Homer K. [S36] - Experiences presenting GRADE to the Guideline Development Group on the NICE Lower Urinary Tract Symptoms (LUTS) Guideline. [Abstract] Otolaryngology: Head and Neck Surgery 2010; 143(1S1):29-30.
6. Tan TP, Siddiqui F, and Thornton J. [S10] - GRADE: Presentation of evidence profiles in clinical guidelines. [Abstract] Otolaryngology: Head and Neck Surgery 2010; 143(1S1):14-15.
7. Thornton J, Sharma T, Kelly V, Tan T, Nyong J, Siddiqui F, and Ayiku L. [S29] - Attitudes of guideline development groups to use of GRADE in evidence evaluation and development of recommendations. [Abstract] Otolaryngology: Head and Neck Surgery 2010; 143(1S1):25.
8. Vandvik PO, Santesso N, Akl EA, You J, Mulla S, Spencer FA, Johnston BC, Brozek J, Kreis J, Brandt L, Zhou Q, Schunemann HJ and Guyatt G. Formatting modifications in GRADE evidence profgiles improved guideline panelists comprehension and accessibility to information. A randomized trial. *Journal of Clinical Epidemiology*. 2012; 65:(7)748-755

Acknowledgements

NCGC

NCC Cancer

NCC Mental Health

Colleagues at NCC-WCH

Participating GDG members

Any questions?