

## Minutes GRADE working group meeting Birmingham

Birmingham, November 30 – December 1, 2003

Present: Jan Brozek, Francoise Cluzeau (day two), Yngve Falck-Ytter, Signe Flottorp, Gordon Guyatt, Robin Harbour, Margaret Haugh, Roman Jaeschke, Merce Marzo, James Mason, Andy Oxman, Bob Phillips, Holger Schunemann, Helena Varonen, Gunn Vist, John Williams

### 1. The minutes from Barcelona were approved without further comment.

### 2. Evidence summaries, balance sheets and judgments

#### • Breast implants – harm

This example was used as a possible example of there being moderate evidence of no effect (safety) based on consistent results from observational studies that found no evidence of effect for a postulated adverse effect. It was agreed that there is too much missing information for this example to make judgments about the balance between benefits (e.g. quality of life, self-esteem) and harms (e.g. breast cancer, leakage) and the reason for the implants would need to be specified. We will check with Peter Tugwell to determine if it is worth completing this example.

**Action: Andy O**

#### • NHCT vs IVP for suspected urolithiasis – diagnosis

Before a decision can be made based on this example, we need more information about the setting; particularly the volume of people seen and the availability of equipment for NHCT. We also need more information about the consequences of false negative tests. The additional outcomes in the balance sheet should also be presented in the evidence summary. There is no information about costs so this outcome should be removed from the tables. Assumptions about costs that are made should then be noted down when making a recommendation.

Consensus:

Consequences of positive tests:	Low quality
Consequences of negative tests:	Low quality
Severe allergic reactions & anaphylaxis from contrast (IVP):	?
Diagnosis of extra-urethral pathology:	?
Overall quality:	Low quality
Balance of benefits and harms:	Net benefits/ Uncertain net benefits
Recommendation:	Do it/ Probably do it depending on costs

#### • CDT vs GGT for detecting problem drinkers – diagnosis

Consensus:

Consequences:	Very low quality
Costs:	?
Overall quality:	Very low

Balance of benefits and harms:      Uncertain net benefits  
Recommendation:                      Don't do it

• **Counseling in primary care – costs**

This example requires additional information about the setting and the seriousness of the depression or emotional problems being treated. We had difficulty interpreting the score and were unsure about the importance of an average change of 1 on the Beck Depression Inventory. In this example it was agreed that the cost per unit on the Beck Depression Inventory was not necessary and could be misleading.

Consensus:

Improved mental health 6 months:      High quality  
Improved mental health 12 months:      High quality  
Direct costs, 6 months:                      High quality  
Direct costs, 12 months:                      High quality  
Overall quality:                                  High  
Benefits and harms:                              Net benefit

(John W noted there were, no data on adverse effects)

Recommendation:                                  ?

• **Dietary fat – equity**

The question in this example needs to be more specific, including whether it is targeted at high risk patients (secondary prevention) or low risk (primary prevention). For the purpose of working through this example, we focused on secondary prevention, since the baseline risk in the control group corresponded to high-risk patients. Since the typical diet in the Philippines is likely different from that of participants in the study, the quality of the evidence would be lowered because of indirectness. However, it was pointed out that disadvantaged people in the US would likely have high fat diets, which might increase the benefit of reducing dietary fats. NNTs and NNHs should only be presented for statistically significant results and should always include time frame.

Consensus (for secondary prevention):

General (developed country)

Total mortality:                                  High quality  
Combined cardiovascular events:      High quality  
Overall:    High

Philippines (disadvantaged)

Total mortality:                                  Low quality  
Combined cardiovascular events:      Low quality  
Overall:    Low

Balance: Uncertain net benefits  
Recommendation: No consensus

### **3. Observational studies with finding of no apparent association to suggest harm – summary of conclusions from examples + further discussion**

We agreed that consistent evidence alone was not sufficient to raise the quality of evidence, but that if all plausible residual confounding, if present, would increase the postulated effect, this could raise the evidence to moderate quality. This would be a logical extension of how plausible confounding is taken into consideration when there is an observed effect. We noted that since bias tends to create spurious results rather than obscure them, the threshold for declaring that plausible confounders are all in the direction of increasing the effect may be lower than the threshold for an analogous conclusion in the presence of an effect. We noted that even taking into consideration these factors the highest evidence we would get of no harmful effect from observational studies would be moderate quality evidence.

We need to change the wording regarding this (which currently only applies when there is an observed effect) and find some examples that properly illustrate this. It also needs to be made clear that the quality of evidence should only be raised on this basis if there are sufficient data to support it.

Unanticipated serious adverse effects are an important outcome and this should generally be considered as a possible outcome in evidence profiles, but would usually not be a critical outcome if there were no evidence. Gord G suggested that the following reference supports routinely considering the risk of unanticipated serious adverse effects for new drugs: Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. Timing of new black box warnings and withdrawals for prescription medications. JAMA. 2002; 287:2215-20.

We discussed the Bross article suggested by Paul Glasziou and found it helpful. There was a consensus that an  $RR > 2$  is a strong association and an  $RR > 5$  is a very strong association; the Bross table provides some support for this, as did consideration of examples. For a strong association there must be consistent results from at least two studies with no plausible confounders. For both the association has to be unlikely to be due to chance (e.g.  $p < 0.01$ ) Andy will check with the Cochrane Statistical Methods Group regarding this.

**Action: Andy O**

Possible examples for observational studies, adverse effects and safety:

MMR vaccine and autism and back to sleep (SIDS)

**Action: Bob P to look for a systematic review**

SSRI side effects

**Action: John W to check his systematic review**

An analgesic that was taken off the market in several countries

**Action: Helena V to look for a systematic review**

Oral contraceptives

**Action: Signe F to look for a systematic review**

Bendectin for nausea in pregnancy

**Action: Roman J to look for a systematic review**

Bicycle helmets, seat belts, and tobacco taxation

Psychotropic drugs and falls, drug interactions

**Action: John W to look for a systematic reviews**

Troponin?

**Action: Holger S to look for a systematic reviews**

Other examples, including ones that can be used to discuss consideration of the quality of evidence for estimating baseline risk??

**Action: All**

#### **4. Diagnostic tests – summary of conclusions from examples + further discussion**

It was agreed that recommendations regarding use of the test should be based on the consequences of the test results in preference to data on test accuracy. Test results are generally surrogate outcomes for patient important outcomes and the directness of diagnosis on clinical consequences needs to be taken into consideration in grading the quality of the evidence. This may lower the quality of data on test accuracy (i.e. high quality data on accuracy into lower quality data on consequences).

It was agreed that when evidence of test accuracy is used the results/ consequences should be reported in separate rows for true positives, false positives, false negatives and true negatives per 100 (or 1000) consecutive patients, with the consequences specified.

It was agreed on the following classification for study designs for test accuracy:

- High: i) Both tests in the same patients at (close to) the same time with ii) an independent criterion standard, and iii) an appropriate spectrum of patients (similar to those in whom the test would be used).
- Low: Any of the above not met

Other quality criteria (e.g. completeness of the follow-up, handling of indeterminate tests, quality of the criterion standard, independent assessment of the tests) should be considered when assessing the quality of the studies, as with assessments of effect studies.

If two tests (a and b) were compared to a criterion test in different studies, this would lower the quality of evidence due to it being an indirect comparison.

Possible examples for recommendations based on test accuracy:

HIV retroviral test

**Action: Helena V to look for a systematic review**

Ultrasound in pregnancy?

Huntington's disease?

JAMA physical examination series?

John W to check with Bob Sentor (?) who has developed software for ROC curve analyses, diagnosis of sinusitis (systematic review by J Lau), screening for causes of hypertension

**Action: John W**

Whisper test for hearing

**Action: Bob P to look for a systematic review**

H Pylori diagnostic tests

**Action: Helena V to look for a systematic review**

Screening for Down's syndrome

**Action: Margaret H to look for a systematic review**

Helio CT for PE

**Action: John W to look for a systematic review**

Diagnosis of pulmonary lesions

**Action: Gordon G to look for a systematic review**

Diabetic retinopathy (example of indirect comparisons)

All to look for other good examples.

**Action: All**

## **5. Costs – summary of conclusions from examples + further discussion**

We agreed that the disaggregated costs should always be presented. The extent to which utilization of different types of resources should be aggregated and valued (in monetary terms) depends on the guideline panel. However, the disaggregated resource utilization information that was used should always be made available either in the GRADE tables or in additional tables – to ensure transparency, facilitate judgments about transferability, and to facilitate use of the information in different jurisdictions. In general it is preferable to report resource use in natural units, but it may sometimes be easier for guidelines panels to be provided with cost information in the local currency.

It should be noted clearly when important cost (resource utilization) information is missing.

Published cost-effectiveness analyses are rarely likely to be helpful because of the assumptions that are made – unless detailed information is available and it is possible to check the complete model and, if necessary, modify it. Guidelines panels may want to conduct their own analyses, using appropriate assumptions and testing the sensitivity of the results to these assumptions.

We agreed that the categories of studies for resource utilization are the same as those for other effects (randomized trials = high and observational data = low).

Possible examples for recommendations based on test accuracy:

SSRIs versus tricyclics

**Action: James M to send document to Gunn**

Nice appraisals

**Action: Francoise C to send suggestions to Gunn**

James promised to send examples.

**Action: James**

Collaborative care versus usual care for depression

**Action: John W**

All to look for other good examples

**Action: All**

## **6. Equity – summary of conclusions from examples + further discussion**

Tony Dans and the INCLEN group need to revise the dietary fat example and develop other examples.

**Action: Tony D et al.**

Francoise C circulated an article on socioeconomic evidence in clinical practice guidelines from the November 29 issue of BMJ.

## **7. Continuous outcomes – summary of conclusions from examples + further discussion**

Weighted mean differences are difficult, at best, to interpret and standardized mean differences are not informative. Gordon believes it is possible to reliably convert these to risk differences or NNTs. Gunn will send Gordon GRADE examples with continuous outcomes and Gordon will convert these to risk differences or NNTs and we will discuss this at our next meeting.

**Action: Gunn V, Gordon G**

## **8. Revision of summary instructions for grading evidence**

Based on the Barcelona workshop experience and similar experience it was agreed the original table of “Quality assessment criteria’ is easier to understand and should be used for training and explaining the criteria. The current table may still be helpful for presentations to some audiences.

## **9. Guidance for preparing evidence summaries and balance sheets**

We agreed to use “limitations” instead of “flaws” under quality.

Holger, Gunn & Gordon volunteered to work on instructions for preparing the GRADE tables.

**Action: Holger S, Gunn V, Gord G**

It was agreed that it is generally better to report the absolute effects as the # per 100 or per 1000 patients or patient years rather than as NNTs and NNHs. The time frame should always be included.

Francoise suggested we should copyright GRADE based on the AGREE experience. Following some discussion there was a consensus that this was a good idea. Robin H suggested we might want to do what SIGN has done.

**Action: Gunn V**

It was agreed that it would be helpful to produce an electronic support tool to assist groups with preparing evidence profiles and grading evidence and recommendations, similar to the Protocol

Support Tool that has been developed by Practihc (an EC INCO project). Andy will prepare an outline of the contents and a mock up for the next meeting.

**Action: Andy O**

#### **10. Term for ‘balance sheet’**

We discussed the suggestions that were raised in Barcelona and some more suggestions made in Birmingham. There was a clear consensus for calling the two tables together an **evidence profile**, the first table (evidence summary) a **quality assessment** and the second table (balance sheet) a **summary of findings**.

#### **11. BMJ article**

We discussed the BMJ referees comments and agreed that Andy O and David Henry should proceed with amending the paper for resubmission, if possible before the end of the year.

**Action: Andy O, David Henry**

#### **12. Other articles**

We agreed to go ahead as planned and send both papers to BioMed Central. Both papers will be submitted before the end of the year.

**Action: Gunn V**

#### **13. Presentation study**

Holger reported that a draft protocol for the presentation study is ready. This will be a randomized trial of different presentations. There is still a need for more ideas on how to recruit participants. Holger will send the draft protocol to the group.

**Action: Holger S**

#### **14. Including balance sheets in Cochrane reviews**

Before the next Cochrane Colloquium we will have a set of examples of evidence profiles suitable for inclusion in Cochrane reviews. Andy will check about the timing for incorporating evidence profiles in the next version of the Review Manager software.

**Action: Andy O**

#### **15. Website**

We all agreed that it is a good idea for GRADE to have a website. Holger, Yngve, Andy and Jan all said that they could host the GRADE web pages. Following discussion it was decided that the best way forward would be to put the web pages on the German Cochrane Center’s server.

We discussed possible domains. “Grade” is already taken. Yngve and Holger will decide on a suitable name and acquire it on behalf of the group.

**Action: Yngve F, Holger S**

#### **16. Applications, funding and future meetings**

It was agreed that applying to AHRQ for a large conference grant should be a high priority.

**Action: Holger S, Andy O**

Margaret H suggested the European Science Foundation as a possible source of comparable funding.

Future meetings:

Helena will look into possibilities of a GRADE meeting in Helsinki in the spring.

**Action: Helena V**

Roman will look into Krakow possibilities

**Action: Roman J**

James will look into possibilities for further support from NICE

**Action: James M**

We will arranged a meeting at the Cochrane Colloquium in Ottawa next October.

**Action: Gunn V, Andy O**

James was asked to forward our thanks to NICE for sponsoring this meeting and Andy O offered to write a letter on behalf of the group.

**Action: James M, Andy O**