

Minutes GRADE working group meeting

Birmingham, December 1-2, 2002

Present: Nima Asgari, David Atkins, Peter Briss, Martin Eccles, Signe Flottrop, Gordon Guyatt, Robin Harbour, Suzanne Hill (day 1), Roman Jaeschke, Alessandro Liberati, Nicola Magini, James Mason, Andy Oxman, Holger Schunemann, Tessa Tan-Torres, Gunn Vist

Regrets: Dana Best, Dianne O'Connell, David Henry, Robert Phillips, John Williams, Deborah Zarin, Stephanie Zaza

1. Introductions and approval of the agenda

Agenda approved. James reported on reimbursement (300 pounds/per attendant). Reimbursement will be provided upon provision of original receipts.

2. Minutes from Oxford (July 2, 2002) and Stavanger (August 4, 2002)

Alessandro asked to revisit the statement in the minutes from Stavanger related to "For real guidelines ... This is not important for the reliability study." He pointed out that this may indicate that the reliability study has not been done seriously and he asked to change the statement. The minutes were otherwise approved. **Action: Andy**

3. Pilot study – general decisions

3.1. Aims of the pilot study

The aims of the pilot study were clarified as testing whether the approach makes sense relative to the included examples, agreeing on changes to the approach if needed, agreeing on decision rules, agreeing on changes to how the evidence summaries and balance sheets are constructed, and agreeing on specific changes for each of the included examples. It was further clarified that in doing this, the approach should be applied without second guessing the information presented in each example or the approach, but noting problems that are encountered and when following the approach or relying on the information presented in an example results in judgements that do not make sense.

3.2. Specification of an outcome in the question

Outcomes should generally not be specified in the question posed in examples, since all important outcomes should be considered.

3.3. Specification of the setting and baseline risk in examples

We need to clarify specific important characteristics of the setting that should influence judgements in the examples, such as the availability of laboratory facilities for monitoring anticoagulation in the atrial fibrillation example. Baseline risk also needs to be specified. **Action: Gunn, Andy**

3.4. Need for more information in the evidence summaries

Many of us felt that the information presented in the evidence summary was lacking in detail. After some discussion we agreed to reverse a previous decision and to include footnotes explaining the basis for judgements about study quality, consistency and directness. So, for example, when there is a serious flaw we will

say what it is and describe the flaws across studies. Similarly, if there is important inconsistency or uncertainty about the directness of evidence, we will describe the reason for this. **Action: Gunn**

3.5. Inclusion of outcomes in evidence summaries and balance sheets

If disease specific mortality is not important because the intervention has no effect on it then it should not be included in the evidence summary and balance sheet as an outcome. Similarly, if disease specific mortality contributes marginally to all cause mortality, then all cause mortality should not be included. For example, all-cause mortality should not be included in the atrial fibrillation example. In keeping with this guideline all-cause mortality would, for example, also not be included as an outcome for breast cancer screening.

Outcomes for which there are no data should not be included in the evidence summaries and balance sheets (see 4.8 below). If possible, evidence should be found and included for critical outcomes (particularly adverse effects), even when this evidence is not included in the systematic review that is used. However, this may be difficult to do, and we will need to explore the feasibility of doing this.

3.6. Mixed study designs or quality

If the majority of evidence is from high quality studies one can ignore the low quality evidence. If an estimate of effect for an outcome is dependent on lower quality studies, this should be reflected in grading the quality of evidence for the outcome. The overall judgement of flaws and quality should be presented, rather than reporting the quality for each individual study in the evidence summaries.

3.7. Consideration of publication bias in evidence summaries

It was agreed that publication bias (and other reporting biases) should be routinely included as one of the dimensions of quality and that this consideration should be made explicit in the evidence summary by adding a separate column instead of including this under “quality”.

3.8. Sparse data

We had the most difficulty in reaching agreement over judgements about sparse data, arising from different thresholds for considering data sparse. The following working definition was proposed for “sparse data”: Data are sparse if they are uninformative. It was also suggested that confidence intervals sufficiently wide that the estimate is consistent with either important harms or important benefits should be considered as sparse data. There was not, however, agreement about either of these suggestions (see 4.9 below).

After repeated discussions in relationship to several of the examples, we agreed on the following:

- We reconfirmed the decision to consider sparse data as a reason for downgrading the grade of evidence, while recognizing that this may be confusing because it is an issue of precision rather than validity and because the precision of estimates of effect is an important consideration when making judgements about trade-offs.
- Sparse data should be routinely considered and reported in evidence summaries by adding another separate column instead of including this under “quality”.

- The threshold for considering data sparse should be lower when there is only one study. A single study with a small sample size (few events) yielding wide confidence intervals with both the potential for harm and benefit should be considered as sparse data.
- Confidence intervals sufficiently wide that, irrespective of other outcomes, the estimate is consistent with conflicting recommendations should be considered as sparse data.
- The importance of sparse data depends on the outcome. For example, sparse data for death as a side effect will receive a much greater weight when making recommendations compared to a less important outcome. As a rule, the less important an outcome is, the less important sparse data are and the less likely it is that sparse data should lower the quality of evidence for the outcome. (See 4.7 below.)
- For now we agreed to acknowledge we have different thresholds for considering data sparse. It may be possible to move towards a consensus and a more precise definition as we work through additional examples and by drawing on other empirical work.

3.9. Little uncertainty about directness

“Little uncertainty” about directness should be replaced with “Direct” to avoid confusion.

3.10. Additional considerations for the quality of evidence for specific outcomes

Clear evidence of a dose response relationship can raise the quality of evidence for an outcome.

When all plausible biases go against the observed results, this could also raise the quality of evidence for an outcome.

3.11. Quality of evidence for harms versus benefits

David expressed concern about grading the quality of evidence for harms the same as evidence for benefits, particularly for questions of safety. He agreed to find specific examples of where this is a problem and come back with specific suggestions of how the quality of evidence for safety should be graded differently from the quality of evidence for benefits or other harmful effects. **Action: David**

3.12. Overall quality of evidence across outcomes

We agreed that if all of the outcomes are going in the same direction (i.e. favouring the same intervention), we should not lower the overall quality of evidence across outcomes if a critical outcome has a lower quality of evidence than other critical outcomes.

We agreed that there may be situations where the pattern of evidence across outcomes may raise or lower the overall quality of evidence across outcomes. Multiple outcomes with consistent results across outcomes can increase the overall quality. Inconsistent results across outcomes can lower the quality of evidence. For example, if the results for different outcomes, including intermediary outcomes, are consistent with what would be expected based on our understanding of biological (or other) relationships, this could raise the overall quality of evidence. On the other hand, if the results for different outcomes are not

consistent with our understanding of biological (or other) relationships, this could lower the overall quality of evidence (see 4.6 below).

3.13. Quality versus validity

After some discussion there was general agreement that we should continue to use the term quality rather than validity, as previously defined and with the same dimensions as previously + dose response relationship and the consistency of results across outcomes.

3.14. Equivalence studies

We discussed whether we should we treat equivalence studies differently. We decided that these studies should not be treated differently, although they may present challenges with regard to judgements about what the smallest important effect is and whether there are sparse data with respect to ruling out an important effect (see 3.8 above).

3.15. NNTs (and NNHs)

In the evidence summary tables, we should always include the time period for NNTs. We should provide a confidence interval around the NNT. We will not provide an NNT when the results are not statistically significant and will indicate this with a question mark. **Action: Gunn**

3.16. Diagnostic tests

The approach can be applied to diagnostic interventions based on studies of their effects. However, we have not yet worked through its application to recommendations about the use of tests based on their diagnostic properties (sensitivity and specificity), such as substitution of a new test for an older test. We decided that we would not address the application of the GRADE guidelines to diagnostic tests in the current paper and that we will address this in a subsequent paper.

3.17. Non-clinical interventions

Peter reiterated his concerns about using the same system for grading clinical and non-clinical interventions. He suggested increasing taxes on tobacco to reduce smoking as an example where our approach would result in counter intuitive judgements about the quality of evidence. **Peter** will identify additional examples and prepare evidence summaries and balance sheets that illustrate his concerns.

4. Pilot study – specific decisions regarding examples

4.1. Depression – SSRIs versus tricyclics

We agreed that because of the short length of the included studies (4 – 12 weeks) depression severity measured within this timeframe was an indirect outcome measure with some uncertainty about its relevance for depression severity of a longer timeframe, given the natural history of depression. The quality of evidence for depression severity was therefore intermediate.

We agreed that there was little uncertainty about the relevance of drop out from treatment as a measure of transient side effects. This needs to be changed to “Direct” in the evidence summary (**Action: Gunn**). This outcome should be changed to “Transient side effects resulting in discontinuation of treatment”. The

quality of evidence for transient side effects (sufficient to stop treatment) was therefore high.

There was discussion about whether the study for poisoning fatalities was seriously flawed. We agreed that ‘population based’ and ‘reporting bias’ should be removed (**Action: Gunn**). Assuming the study did not have a serious flaw, the quality of evidence for poisoning fatalities is intermediate (observational study with a very strong association). James and Martin will provide details regarding whether the study was flawed. If possible, the evidence for this outcome should be updated, since a number of other studies are available. **Action: James, Martin**

We subsequently agreed that the overall quality of evidence was intermediate, that there were net benefits and our recommendation (not considering costs) would be “probably do it” because of the quality of the evidence.

4.2. Atrial fibrillation - aspirin versus oral anticoagulation

There was good agreement on the quality of evidence for stroke and extracranial hemorrhage. We agreed to remove all cause mortality as an outcome. **Action: Gunn**

We agreed to have two separate summary sheets: one for lower risk patients and one for higher risk patients, with NNTs adjusted accordingly (**Action: Gunn**). Following discussion there was general agreement that there are net benefits for high risk patients and the recommendation would be “do it” in a setting in which laboratory facilities for monitoring anticoagulation were readily available.

4.3. Degenerative arthritis - paracetamol versus NSAIDs

Concern was raised that in this example a few relatively small trials receive same grade for quality of evidence as many large trials would. There was complete agreement that the quality of evidence for pain at rest was high. We agreed that pain on motion is an outcome with some uncertainty about its relevance and the quality of evidence for this outcome was therefore intermediate. The quality of evidence for mobility was intermediate because of some uncertainty about directness of the 50 foot walk test as a measure of mobility. The evidence for quality of life was intermediate because of some uncertainty about directness for the Nottingham Health Profile. There was agreement that the quality of evidence for side effects (as measured by treatment drop out due to side effects) was high. We need to update the evidence summary and balance sheet (**Action: Gunn**). More information is needed for serious GI complications before we can determine the quality of evidence and relative importance for this outcome (**Action: Martin**). It was clarified that the study by Silverstein was an RCT, but only data for one arm of the study were used, so that it is correct that it is an observational study in this context. However, it was thought that there is likely more evidence for this outcome than this one study. The numbers in the balance sheet were incomplete and presented incorrectly. Pain at rest, pain at motion and mobility were considered critical outcomes. Quality of life and drop out due to side effects were considered important outcomes. We need more information about the magnitude of the risk for serious GI complications before deciding on the relative importance of this outcome. Before judgements about overall quality, trade-offs and recommendations can be made, we need more information about serious GI side effects.

4.4. Myocardial infarction - antiplatelet therapy

The title should be changed to long-term antiplatelet treatment and the dose of the intervention should be specified (**Action: Gunn**). This example is a good illustration of the dose having an impact on directness (of the intervention). Evidence about harm (bleeding risk) is missing from this example (**Action: Martin**). There was agreement that the evidence for all cause mortality was intermediate because of important inconsistency. The inconsistency needs to be explained in a footnote (**Action: Gunn**). There was complete agreement that the quality of evidence for non-fatal stroke and non-fatal MI was high. All of the outcomes were considered critical. We agreed that the overall quality was high, even though the quality of evidence for all cause mortality was intermediate, because of the new rule about overall quality (see 3.12 above). There was complete agreement that there were net benefits and that the recommendation should be “do it”.

4.5. Myocardial infarction – rehabilitation

There was good agreement on this example. We agreed that the quality of evidence for both outcomes was high and both were critical. We need the baseline rates for mortality in this example (**Action: Martin**), but in the absence of this information agreed that there are net benefits and (without considering costs) the recommendation would be “do it”. No harms were identified.

4.6. Deep venous thrombosis - low molecular weight heparin versus IV heparin

The outcome should not be specified in the question. The overall judgement of flaws and quality should be presented, rather than reporting the quality for each individual study (**Action: Gunn**). We agreed that the evidence for all of the outcomes was high, all were critical and the overall quality was high. The majority concluded that there are net benefits and the recommendation, without considering costs, was “do it”, but there was uncertainty about this. There was discussion about whether the overall quality of evidence should be lower because the magnitude of the effect for overall mortality was greater than for the two outcomes which could be expected to have an impact on overall mortality and the confidence intervals for these two outcomes were wide, including a potential harmful effect as well as a potential benefit. However, the confidence intervals were overlapping. It was also suggested that because of the wide confidence intervals for major bleeding and recurrent thromboembolism it might be reasonable to conclude that there are uncertain net benefits and, consequently, a recommendation of “probably do it” might be reasonable.

4.7. Maxillary sinusitis - antibiotics

After some discussion there was general agreement that there was not sparse data for clinical cure (looking across the evidence for amoxicillin and penicillin, which suggests an overall “statistically significant” effect) or dropout due to adverse effects (there were only a total of four dropouts among 460 patients across studies), even though the confidence intervals were wide. Some of us combined the evidence and results for amoxicillin and penicillin, whereas others considered the evidence for these separately, as presented in the evidence summary and balance sheet. We need to clarify whether the evidence for amoxicillin and penicillin should be lumped or not and whether there is important inconsistency or not (**Action: Gunn, John**). Pending these clarifications, we agreed that the quality of evidence for clinical cure was intermediate (because of “important

inconsistency” and “sparse data should be removed from the evidence summary), the quality of evidence for dropout due to side effects was high (and “sparse data” and “important inconsistency” should be deleted from the evidence summary), and the quality of evidence for relapse was intermediate because of sparse data (**Action: Gunn**). There was some uncertainty about relapse, but the majority considered this sparse data because there was only one study with wide confidence intervals including both an important benefit and an important harm. Clinical cure and dropout due to side effects were considered critical outcomes. Relapse was considered important. The overall quality of evidence was considered intermediate. There was general agreement that the balance between benefits and harms was uncertain and that the recommendation was “probably do it”, because of uncertainty about the downsides of treatment and pending clarification whether the evidence for amoxicillin and penicillin should be pooled.

4.8. Tuberculosis – BCG vaccine

There was poor agreement about the quality of evidence for TB and TB death. After discussion, we agreed that the quality of evidence for TB was intermediate because of important inconsistencies. The inconsistency should be explained in a footnote, and the reason for including observational studies (which was thought to be because of longer follow-up and different locations of studies) needs to be clarified (**Action: Gunn**). The quality of evidence for TB death was considered high, although there was some concern about the directness of the evidence since the studies were reported between 1948 and 1974). The quality of evidence for TB meningitis was considered intermediate (observational studies with a strong association (OR 0.36), (presumed) consistent results and (presumed) no plausible confounders. These assumptions need to be confirmed and made explicit in the evidence summary (**Action: Gunn**). Data for adverse effects are missing and are needed before judgements about trade-offs and recommendations can be made. For very serious adverse events, even low quality data might tip these judgments. We need to either find and include data for adverse effects or delete these outcomes from the evidence summary and balance sheet (**Action: Gunn**). Ignoring adverse effects, the overall quality of evidence was considered high (since all of the outcomes go in the same direction), there are net benefits and the recommendation would be “do it”.

4.9. Sciatica – surgical disectomy

Following discussion the majority considered that there are sparse data for “condition unchanged over one year” and “no success at two years” and this should be added to the evidence summary (**Action: Gunn**), but there was irresolvable disagreement about this. This is an example where we clearly had different thresholds for what we considered to be “sparse data”. Those arguing that there are sparse data based this on there being few, small studies with wide confidence intervals including both important harms and benefits. Those arguing that there are not sparse data based this on the judgement that the data are informative and, therefore, are not sparse. Accordingly, the quality of evidence for these two outcomes was considered either high or intermediate, because of sparse data. The quality of evidence for “poor outcome at one year” (surgeon rated) was considered intermediate, because of some concern about the directness of the outcome. It was agreed that more information was needed to decide if these studies had a serious flaw for this outcome (**Action: Gunn**). If the surgeons were

not blinded, this would be a serious flaw and the quality of evidence would then be low, or this might even be considered a fatal flaw and the outcome would not be included. The quality of evidence for “a second procedure within one year” was considered intermediate because of uncertainty about the relevance of the outcome (due to the decision likely being made by surgeons not blinded to the initial treatment and with a bias in favor of surgery). It was suggested that this outcome might better be expressed as the proportion of patient having surgery within one year (which would than be 100% for surgical disectomy).

This is an example of indirect evidence between studies (surgical disectomy versus chymopapain and chemopapain versus placebo to draw conclusions about disectomy versus no surgery). However, because there was considered to be little uncertainty about the relevance (based on the assumption that chemopapain was found to be better than placebo), the decision (pending further information) was not to replace “Indirect with little uncertainty about relevance” with “Direct” in the evidence summary and not to lower the quality of evidence because of this.

Action: Gunn

We decided that condition unchanged at one year (patient rated) and no success at two years (independent observer rated) were critical outcomes. Surgeon-rated poor outcome at one year was considered important, if there was not a serious flaw, and unimportant, if there was a serious flaw (lack of blinding). The relative importance of a second procedure needed within one year was considered either important or unimportant, depending on more information from the original studies about how decisions about a second procedure were made and who was making these decisions. Risks and side effects of surgery should be dropped, if data cannot be found (**Action: Gunn**). The overall quality of evidence would be either high or intermediate, depending on the judgement about sparse data (and ignoring the lack of data for risks and side effects of surgery). We agreed that there are either uncertain net benefits or trade-offs, depending on whether the risks and side effects of surgery are considered. The recommendation would be probably don't do it or “toss up”, if adverse effects are taken into consideration.

4.10. Dental caries – water flouridation

Peter explained that all of the studies included in the table were considered to be well designed observational studies with concurrent comparisons and it was reasonable to include mixed study designs because the better study designs were only available for earlier time periods with important differences in baseline risk and conditions. There was not agreement about the quality of evidence, but the majority considered the quality of evidence for all of the outcomes to be very low (observational studies with important inconsistency) and, accordingly, the overall quality to be very low. This is different from the conclusion of the US Task Force on Community Preventive Services, which considered the evidence to be strong. More information is needed in the evidence summary to explain the inconsistencies and better data are needed in the balance sheet. **Action: Peter**

Changes in the rates of dental caries following starting or stopping water flouridation were considered critical outcomes. There was some uncertainty about the importance of dental fluorosis, but Peter explained why the panel did not consider this an important outcome (it is not noticable without close examination). Bone fractures were considered either important or critical, depending on whether indirect evidence about the relationship between high doses of fluoride and bone

fractures were taken into account (suggesting that the relationship is plausible). We were not sure about the plausibility of the relationship between fluoride and cancer, but generally considered cancer mortality to be a critical outcome because of concerns about this outcome and the severity of the outcome.

There was not agreement, but the general conclusion was that there were uncertain net benefits, based on the information in the evidence summary and balance sheet with the additional information noted above. We did not reach agreement about a recommendation (in contrast to the Task Force, which made a strong recommendation to “do it”).

4.11. Use of child safety seats and hormone replacement therapy

The last two examples were not discussed due to a lack of time.

5. Reliability and sensibility protocol

The following changes have been made in the protocol since the last meeting: The protocol has been shortened (with most of the information from the appendices now included in the three draft articles) and updated, reflecting decisions made at the meetings in Oxford and Stavanger. There are still three categories of judges, but the main analysis will be between groups of judges, who will work together in groups of three to reach a consensus on all of the judgements for six examples. The sample size calculations done by Richard Cook support the rough estimate of sixty examples from earlier drafts of the protocol. The main challenge will be identifying sixty examples and preparing evidence summaries and balance sheets for these. Gunn and Andy estimate that it will not be possible to prepare all sixty examples before the end of 2003. Everyone should read the protocol and send feedback to Gunn and everyone should send suggestions for examples (and offers to help prepare evidence summaries and balance sheets!) to Gunn. Gunn will follow-up on this with specific requests. **Action: Gunn + All**

6. Guidelines for grading article

Page 5. Peter suggested changing the wording to emphasize caution about the types of interventions to which the guidelines apply. He will provide a suggestion.
Action: Peter

Page 6 + 14. Gord suggested we should omit the judgement about the balance between benefits and harms, etc and go straight to the recommendation. The pros and cons of doing this were discussed, but no clear decision was reached. We agreed to reconsider this after the paper is revised.

Page 7. Peter suggested “beefing up” the discussion about the validity of RCTs versus observational studies. He will send suggestions to Andy. **Action: Peter**

Andy suggested we may want to clarify the distinction between “observational studies” (which might include cohort studies, case control studies, interrupted time series analyses, and controlled before-after studies) and “other” (for example, unrolled before-after studies).

Alessandro raised a concern about the term “reasonable” threshold in the last sentence of the second paragraph.

Page 10. The paragraph on additional considerations needs to be edited in light of the decisions summarised above, and something needs to be added about “sparse data”.

Page 13. James suggested adding a sentence clarifying that further work needs to be done regarding the consideration of costs and that this is not addressed in this paper.

Page 15. We voted about whether to delete the rules of thumb for making judgements about recommendations based on the proportion of people likely to do something. Seven voted against keeping them, five voted to keep them, and three were undecided. Those who voted against keeping them were asked to provide alternative suggestions for a guideline for making this judgement. **Action: All**

Page 17. We decided not to seek endorsements and to drop this from the paper.

Page 24 + 25 (Box 1). It was agreed not to pick on surgeons and public health advocates in the first bullet point. Peter wanted to change the wording of the last bullet point. He will send suggestions to Andy. **Action: Peter**

Page 26 (Table 1). We agreed to use an alternative presentation for table 1 (quality of evidence for each main outcome) based on Roman's suggestion:

Table 1. Quality of evidence across studies for each main outcome		
RCT	Quality of the evidence	Observational studies
No serious flaws in study quality	High	Extremely strong association and no threats to validity
Serious flaws in design or execution or quasi-experimental design	Intermediate	Strong, consistent association and no plausible confounders
Very serious flaws in design or execution	Low	No serious flaws in study quality
Very serious flaws and at least one other serious threat to validity	Very low	Serious flaws in design and execution
Additional factors that lower study quality are: important inconsistency of results; some uncertainty about directness; high probability of reporting bias; and sparseness of data. Major uncertainty about directness can lower the quality by two levels.		
Additional factors that may increase quality of observational studies are: all plausible residual confounding, if present, would reduce the observed effect; and evidence of a dose-response gradient		

(Following the meeting, The McMaster group suggested that footnotes should be added to this table and that we should label or provide examples of the strength of association that would move the quality of evidence up. This should also be spelled out in the manuscript. Possible examples might be insulin for diabetic ketoacidosis and the association of smoking with lung cancer (Odds ratio = 10) as examples for an extremely strong association, and coumadin for atrial fibrillation in patients with mitral valve disease (RRR of ~ 68% in patients without atrial fibrillation in RCTs) as an example of a strong association.)

Alessandro suggested adding an example to which readers could apply the guidelines.

We tentatively decided that BMJ would be a good venue for publishing the paper and that the paper should be written with this venue and audience in mind. There were mixed views about the need for a writing committee. It was agreed that Andy will chair the writing committee and arbitrate final decisions. Peter, Roman, Gord, Holger, and David volunteered to be on the writing committee. **Action: Andy**

7. Critical appraisal and pilot study + six systems papers

We did not have time to discuss these draft papers. Everyone should send corrections and suggestions regarding these papers to Gunn. **Action: All**

8. Presentation protocol

Holger asked everyone to send him comments on the draft protocol. **Action: All**

We discussed the pros and cons of using the web to do this study, and problems with obtaining representative samples over the web. Peter suggested that there would be an advantage of obtaining different small samples from specific sampling frames for this study. Nima suggested using a British commercial website that offers access to all British doctors as a sampling frame. A response rate of 70% was achieved in a previous study using this website. Gord suggested offering CME credit as an incentive to get physicians to complete the exercise. Andy and Alessandro suggested using meetings, such as rounds and conferences, to recruit participants. Having an entertaining presentation might increase response rates at meetings. This may require randomisation by groups, to facilitate discussion at meetings. Tessa suggested that an INCLIN meeting to be held in China in February could provide an opportunity to get responses from clinicians from a number of different countries and diverse cultures.

Action: Holger

Martin offered to look into funding opportunities in the UK if a proposal that includes a budget is sent to him. **Action: Holger, Martin.**

(Holger, Andy, Gord and Roman discussed the protocol further following the meeting. It was suggested that we need to be more inventive and include more symbols. Andy gave several suggestions to Holger. It was also suggested that we should drop several of the questions, including ease of comprehension, succinctness, and literacy. For the final project, but not the preliminary evaluation, we will also drop questions about clear number of levels, clear direction and clear limits. The question about interpretation should be one about understanding for one's home language and culture.)

9. Endorsements, involvement of others, funding

We decided not to seek endorsements. There was agreement that future meetings and activities should be open to others. Although we will not seek endorsements, we will seek input from other organizations. After there is agreement about next revision of the draft guidelines these will be circulated to others who have expressed an interest in what we are doing and to individuals in selected organisations with which each of us have contacts. **Action: All.**

The involvement of others might pose problems with funding for future meetings and, potentially the size of meetings. Martin will look into the possibility of funding for another meeting in connection with the next NICE conference and David will look into the possibility of funding from AHRQ, such as a large conference grant. We further discussed how a large conference might also be used as an opportunity to collect data for the reliability and sensibility study. **Action: Martin, David**

We agreed that while it is likely not possible to completely avoid coming back to issues that we have previously discussed as more people become involved, the system as described in the draft guidelines for grading article will be considered a starting point for involvement.

10. Future meetings

Possibilities for future meetings include Poland (**Roman**), the US (funded by AHRQ) (**David**), in connection with the next NICE conference (**Martin**), in connection with the next Scientific Basis of Health Services conference hosted by AHRQ (**David**), in connection with the next Cochrane Colloquium (**Andy**).

11. Any other business

We thanked James and Martin for hosting the meeting and asked them to convey our thanks to Janice and NICE (**Action: James and Martin**). Gunn was thanked for all her hard work on the examples and Andy for chairing the meeting.