

## **Minutes GRADE Working Group**

Helsinki, April 28 – 29, 2004

Present: Pablo Alonso, Gerd Antes, David Atkins, Jan Brozek, Francoise Cluzeau, Benjamin Djulbegovic, Yngve Falck-Ytter, Cindy Farquhar, Signe Flottorp, Paul Glasziou (day two), Gordon Guyatt, Mark Helfand, Roman Jaeschke, Eeva Ketola, Ilkka Kunnamo, Regina Kunz, Nicola Magrini, Merce Marzo, Ed Mills, Jacek Mrukowicz, Andy Oxman, Bob Phillips, Vivian Robinson, Holger Schunemann, Anja Tuulonen (day two), Helena Varonen, Jeremy Wyatt, Jelka Zupanj

### **Examples: preparation of evidence profiles and grading**

#### **1. RCTs**

- **Active management of third stage of labour**

We agreed that the evidence profile should be divided into two tables, for women at high and low risk of postpartum haemorrhage (PPH). There was disagreement about whether there was sparse data for the outcome “maternal dissatisfaction”. It was argued that there was because there was only one study, even though the study included 1466 women and the confidence interval excluded no difference (RR = 0.56, 95% CI 0.35 to 0.90). Two previous arguments were repeated: We previously agreed that there should be a lower threshold for downgrading due to sparse data when there is only one study, and it has previously been argued that wide confidence intervals should be taken into account when grading the strength of recommendation rather than when grading the quality of evidence. It was further argued that the quality of evidence is always lower when there is only one study because of concerns about the generalisability of the results. No agreement was reached. We agreed that the quality of evidence for all of the other outcomes was high.

We agreed that relative risks should routinely be used in the summary of findings, when possible, rather than odds ratios or other summary statistics.

It was suggested that better guidance is needed for judging the importance of outcomes and deciding whether an outcome is critical. This judgement depends on the setting. In this example the risk of death due to PPH is larger in low-income countries and a critical outcome there, but this may not be the case in high-income countries.

- **Endoscopic treatment of bleeding peptic ulcers**

We discussed whether the different types of endoscopic methods were similar enough to be lumped together without reaching agreement. It was agreed that an explanation was needed in the quality assessment to explain “important inconsistency” and reconfirmed that any deviation from the default in the quality assessment needs an explanation.

It was suggested that all the columns in the quality assessment table should be reported in the same way; e.g. use strong data instead of sparse data, so that a yes answer would always mean raising the quality of evidence. (See item 5 below.)

Mark pointed out problems with hospital mortality and 30-day mortality as outcome measures because of different ways in which it is reported and bias. Possible ways of addressing this would be not to include this outcome and only include total mortality (survival) as an outcome or to lower the quality of evidence for this outcome either based on study quality (biased outcome assessment for hospital mortality or 30-day mortality) or directness (uncertainty about directness for survival).

Judgements about sparse data were discussed again, particularly with respect to adverse events. It was suggested that for rare events we should consider the confidence interval around the risk difference (in this example the RD  $\approx$  0.00, 95% CI  $\approx$  -0.01 to +0.01) rather than the confidence interval around the relative risk (in this example the RR  $\approx$  1.00, 95% CI  $\approx$  0.33 to 3.09). For this and similar examples the confidence interval around the RD suggests that the data are not sparse, whereas the confidence around the RR suggests that the data are sparse.

It was agreed there was not inconsistency or sparse data for the outcomes other than hospital mortality and, thus, high quality evidence.

## 2. Observational studies and harm

- **Benedectin**

This example, like many others, was based on a systematic review from which it was difficult to produce an evidence profile. It was noted that in real life it would have been necessary to go back to the original studies to prepare an evidence profile. The numbers of participants in the studies were not reported in the review. We reconfirmed the earlier decision that it is better to report the absolute effects as the number of events per 100 (or other appropriate denominator) rather than as NNTs and NNHs.

This example was used to follow-up on discussion from the meeting in Birmingham regarding raising the quality of evidence for an outcome when there is consistent evidence from multiple studies of no effect, parallel to raising the quality of evidence when there is an effect and all plausible confounders would have reduced the effect. It was agreed that when there are multiple studies without serious limitations, the upper boundary of the confidence interval for the RR or OR is close to one, and all plausible biases would have increased the RR or OR, that the quality of evidence should be increased. It was suggested that this should be called the “Helsinki rule”.

For this example it was agreed that the quality of evidence was moderate, rather than low in light of the plausible biases we could think of for the following outcomes: all malformations, cardiac defects, and oral cleft defects. For the other adverse effects the confidence intervals were sufficiently wide (including an increased risk of 20 to 48%) that we considered the quality of evidence low.

Based on the previous discussion, we would not lower the quality of evidence because of wide confidence intervals around the RR or OR, provided that the confidence intervals were sufficiently narrow around the RD.

In discussing the consistency of the studies in this example, it was pointed out that if 2 or 3 studies out of 100 show an effect, this does not mean that the results are inconsistent.

It was also pointed out that the number of people included in the studies is not a reason for increasing the quality of evidence. If the studies were biased, this would simply increase the precision of a biased estimate.

- **MMR vaccination**

We discussed whether there is an important distinction between evidence of safety (confidence that there is not a purported adverse effect) and evidence of harm (confidence that there is a purported adverse effect), and when a purported adverse effect is critical. It is generally impossible to have high quality evidence of safety (to confidently rule out a purported adverse effect) for rare outcomes. Evidence from observational studies indicating no effect cannot be higher than moderate, using the GRADE criteria. (One possible exception would be consistent evidence from well designed observational studies with an upper 95% confidence interval close to one and studies with an extremely high exposure showing no effect – see the chlorinated drinking water example below.)

For this example we agreed that it would be reasonable to argue that there is moderate evidence of safety (consistent observational studies with an upper limit of the 95% confidence interval close to one) and very low quality evidence of harm.

We debated whether autism is a critical or an important outcome. If it is considered critical and the quality of evidence for autism (safety) is moderate, the quality of evidence for recommending MMR vaccination would be moderate. On the other hand, if autism is considered important, but not critical, the quality of evidence would be high, assuming there is high quality evidence for the benefits of MMR vaccination.

It was agreed that for examples such as this, there needs to be a transparent and consistent way of determining the overall quality of evidence that does not inappropriately lower the overall quality of evidence because of the unlikelihood of having high quality evidence that can rule out purported, but implausible rare adverse effects.

It was agreed that there should only be one row for each outcome and that the second outcome (increased rates of specific subtypes of autism) was not helpful and could be left out of the evidence profile. It was also agreed that only options directly corresponding to the quality assessment criteria should be included in the quality assessment.

- **ASA in the third trimester**

We discussed the consistency of the evidence and reiterated previously agreed guidance, including that when there is a compelling explanation for inconsistency, this should be taken into account when preparing an evidence profile by either leaving out

low quality evidence or by preparing different profiles for different people or interventions.

It was pointed out that there is not a strong association for pyloric stenosis even though the OR was 2.24, because the 95% CI rules out a p value less than 0.01.

We agreed that when there is evidence from both RCTs and observational studies and a compelling reason to include both types of evidence, we should start out by basing the quality of evidence as high, if a majority of the data come from RCTs, and as low, if most of the data come from observational studies. In this case nearly all the data came from observational studies.

There was uncertainty about how much of the data came from case control studies and how much from cohort studies, whether there were important limitations of the studies (this was not adequately reported in the review), and what the baseline risks were. Given these uncertainties, no conclusions were reached about the quality of evidence for this example. It was pointed out that the percents of newborns with heart defects were misleading (3.4% versus 5.4% with an overall OR of 1.01). There was uncertainty about the correctness of the event rates because the included studies were a mix of cohort and case control studies. We discussed alternatives regarding what information to include in the two event rate columns. It was suggested that we need some examples of alternatives. This discussion was followed-up in the small group discussion (see item 6 below).

- **Chlorinated drinking water**

There was insufficient information in the review to adequately prepare the evidence profile, including missing information about the levels of chlorination, how well the exposure be measured, and whether there was a dose-response effect within or across studies.

We agreed that “evidence of a dose response gradient” needs to be based on within study data. As with all deviations from the default, an explanation should be given as a footnote, if there is evidence of a dose response gradient.

We also agreed that if there is evidence of no effect from studies with an extremely high exposure that this would raise the quality of evidence, if the overall effect (OR) suggests no increased risk (analogously to a “reverse dose-response gradient”). Conversely, this could lower the quality of evidence, if the overall effect suggests an increased risk. We need examples of “reverse dose-response gradients”, high exposures not being associated with a purported adverse effect or lack of a dose response gradient increasing the strength of evidence.

**Action: All**

We concluded that the evidence is very low (observational studies with uncertainty about the directness of the exposure), if there is not a dose response gradient and that would be low if there is evidence of a dose response gradient.

- **Proton pump inhibitors in the first trimester**

The review did not provide information about limitations of the included studies. It was agreed that there was not evidence of a reporting bias.

Sparse data was discussed again in relationship to this example. Jan indicated there was sparse data in the evidence profile based on the wide confidence interval around the RR (0.72 to 1.94) rather than the RD (as discussed above). We debated again whether the quality of evidence should be lowered only when data are sparse (few observations) or when the results are imprecise (wide confidence intervals). It was pointed out that if we exclude consideration of imprecision from judgements about quality, we would need to change our definition of the quality of evidence from “confidence that an estimate is correct” to something like “confidence that an estimate is unbiased”. We were reminded that precision is taken into consideration when judging the balance between benefits and harms and is perhaps better taken account of there than when judging the quality of evidence.

### **3. Diagnostic tests**

- **Ottawa ankle rules**

Paul argued that there are three different decisions about diagnostic tests that might require different considerations when grading the quality of evidence and strength of recommendations: whether a test should be used to triage patients (like this example: determining who needs an x-ray), whether a test should be used as an add-on test, and whether a test should be used as a replacement for another test.

We agreed that the best evidence for all three of these questions would be an RCT comparing one strategy to another and measuring patient important outcomes. When we only have evidence of test accuracy there is some degree of uncertainty about the directness of the test results as an indication of patient important outcomes.

We discussed what outcomes should be included in evidence profiles and agreed that, when possible, there should be six outcomes: true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), positive and negative likelihood ratios. We debated whether the denominators should be the numbers of people with and without the condition (sensitivities and specificities + their inverses), the numbers of people with and without a positive test result (predictive values) or the total number of people tested. It was agreed that the total number of people tested should be used and that the prevalence of the condition needs to be specified for the evidence profile.

David argued that the primary questions for the Ottawa ankle rules were: Do they reduce unnecessary x-rays and are clinically important fractures missed? It was agreed that the evidence was direct for unnecessary x-rays (TNs). They would be treated the same with or without the x-ray. There was some uncertainty about directness for missed fractures (FNs) because the fractures that were missed might not be the same as the ones that were not missed and the ones that were missed might not be clinically important (i.e. it might not make any difference in terms of patient important outcomes). The TPs would be treated the same (have an x-ray and be treated

accordingly) and the FPs would also be treated the same (have an x-ray and be treated accordingly), so the evidence could be considered direct for these outcomes.

- **Diagnosis of acute maxillary sinusitis**  
This example was not discussed.

### **Small group discussions**

Items 4 to 8 were discussed in small groups and reported back.

#### **4. Definitions of sparse data (vs imprecision) and quality**

It was agreed to keep the current definition of “quality” and to change the term from “sparse data” to “Imprecise or sparse data”. A working definition for “imprecise data” is “wide confidence intervals”. The following arguments were considered in favour of this decision in the small group discussion:

- It is likely to be more sensible and understandable to most people considering “quality” to include consideration of imprecision (using the current definition: “confidence that an estimate is correct”), as opposed to excluding consideration of imprecision (using a definition such as “confidence that an estimate is unbiased”).
- It is difficult to change the definition of quality now, given that the BMJ article is in press and we have been using this definition of quality for the past few years.

Counter arguments that were considered were:

- It may be easier to incorporate consideration of imprecision in judgments about trade-offs, and by incorporating imprecision in judgments about quality it needs to be considered both in relationship to quality and trade-offs.
- It may be confusing to label high quality studies with imprecise results as “moderate quality”.

We concluded that neither of these were big problems in the small group discussion.

#### **5. Column headings in quality assessments**

It was agreed to change the heading from “Quality” to “Limitations” because the entire table is about quality.

It was also agreed to collapse the columns after directness to “Other considerations”.

We did not reach a conclusion about how to label the considerations and present assessments in the quality assessments. One suggestion was to use zeros and one or two plus or minus signs (or up and down arrows), where zeros represent the default. The range differs for each criterion, so this needs to be taken account of in the presentation, for example by presenting the range for each criterion beneath the column label. Paul agreed to prepare a suggestion for an improved presentation of the quality assessments, which is attached.

#### **6. What should be included in summaries of findings?**

The discussion focused on dichotomous outcomes. Reporting of continuous outcomes was not discussed.

**Intervention and comparison columns:** For RCTs or cohort studies the pooled numbers of events and participants should be reported with the proportion in brackets underneath. The denominator should be clearly identified (e.g. whether it is the total number of participants or person years). When there is evidence from case control and cohort studies, the control event rate should be derived from the cohort studies and the intervention column should be left empty.

**Relative effect:** When possible, the relative risk (RR) should be reported. For case-control studies the odds ratio (OR) should be reported.

**Absolute effect:** When possible, this should be reported as frequencies with positive and negative sign; e.g # per 100. When there are both cohort and case control studies, the pooled estimate of the relative effect and the pooled baseline risk from cohort studies (or other data) should be used to estimate the absolute effect. When this is done, it should be flagged and explained in a footnote.

If the pooled RR or absolute effect differs from the experimental event rate (EER) and control event rate (CER) data in the intervention and comparison columns, a footnote should explain the reason for the difference.

## 7. Observational studies

It was suggested that cohort studies should start out as low quality and case control studies should start out as very low quality, but this suggestion was rejected. Problems with the use of the term “quasi-randomised trials” were discussed. It was agreed not to use this term and to take account of limitations due to quasi-randomisation (risk of bias due to inadequate concealment of allocation) when considering study limitations.

## 8. Diagnostic tests

The criteria for study design and limitations for studies of diagnostic accuracy were discussed in a workshop on Tuesday 27 April, before the meeting, and further developed in the small group discussion. The main concerns for study design are the choice of an appropriate population and an appropriate criterion standard, and these should be captured under study design. Other concerns are captured under study limitations, as summarised in the following table.

## Quality assessment criteria for studies of diagnostic test accuracy

Quality of evidence	Study design	Lower if *
High	Cross-sectional (or cohort) studies of patients with diagnostic uncertainty with direct comparison	<b>Study limitations</b> (including representativeness of population, choice of gold standard, incomplete performance of tests, independence of test interpretation) -1 Serious limitations -2 Very serious limitations  -1 <b>Important inconsistency</b>  <b>Directness</b> -1-Some uncertainty -2-Major uncertainty  -1 <b>Sparse or imprecise data</b>  -1 <b>High probability of reporting bias</b>
Moderate		
Low	Anything else	
Very low		

It was reiterated, that recommendations should be made based on a test's usefulness (based on patient important outcomes) rather than its accuracy, so that the other considerations for the quality of evidence are generally the same as they are for other types of interventions and that the directness of the evidence is particularly importance when translating test results into patient important outcomes.

A modified template is needed for evidence profiles for diagnostic interventions based on evidence of test accuracy.

### 9. Summaries of findings in Cochrane reviews

The Cochrane Collaboration Steering Group has agreed that the next generation of the Information Management System should support the inclusion of summaries of findings. Specifications for these tables need to be developed this year.

Challenges that were discussed in adapting GRADE summaries of findings include:

- In Cochrane reviews there may be problems with judgements about directness (if quality assessments are included in the summaries of findings) because the reviews are written for an international audience without specifying a setting.
- Summaries of findings require specification of the baseline risk if they include an absolute effect.

Some Cochrane reviews are likely to require multiple (or interactive) tables to accommodate different settings and baseline risks.

There is a need to balance simplicity with precision in deciding what to present in summaries of findings in Cochrane reviews: simpler tables are likely to be more understandable (and desirable) for users, but with simplicity there is a loss of precision. Summaries of findings in Both clinicians and consumers should be able to understandable summaries of findings in Cochrane reviews. However, it is possible that different presentations would be optimal for clinicians and patients.

Ilkka offered to check what information could be transferred automatically from sources such as Cochrane reviews into the evidence profile using the XML structure and to coordinate this with related work by GIN. Cindy and Regina volunteered to offer suggestions about how to simplify summaries of findings.

**Action: Andy, Cindy, Regina, Ilkka**

#### **10. Challenges faced by organizations considering using GRADE**

- UK NICE: Francoise and Jeremy reported that there is an ongoing process to improve NICE grading of evidence and recommendations. The GRADE approach is being considered, but there are concerns about it being too complicated.
- Poland: Jan and Jacek explained that in Poland it is not possible to discuss recommendations without taking costs into account. It was noted that the extent to which costs are considered (or excluded from consideration) varies from country to country. The GRADE paper has been translated into Polish. Jacek, Jan and Roman have proposed using the GRADE approach as part of drug approval (for reimbursement) in Poland.
- US FDA: The FDA has its own system for evidence grading. The GRADE approach was considered in revising this system, but will not be used.
- Scotland SIGN: SIGN is reconsidering its approach to grading. The GRADE approach is being considered. There were also concerns about it being too complicated for sign, as well as concerns about how to make a transition from its current system, which is well established, to a new approach.
- WHO: Jelka said it is difficult to speak from the point of view of the whole organization. Generalisability is a major challenge for WHO guidelines given that they are intended for a wide range of settings. The GRADE approach is included in the WHO guidelines for guidelines, but so far those guidelines are not being used. Jelka is attempting to use the GRADE approach in considering a proposed change in guidelines for birth spacing.
- US AHRQ Evidence-based Practice Centers (EPCs): AHRQ does not dictate what approach the EPCs should use. They are using similar approaches, but the GRADE approach is unlikely to be formally adopted at present.
- Finland Current Care Guidelines: Anja reported that they may use the GRADE approach in revising their glaucoma screening guidelines.
- Finland EBM and Current Care Guidelines: Ilkka reported they use a system for study quality that is similar to the GRADE approach. They are considering adopting the GRADE definitions of quality and rules for upgrading and downgrading the quality of evidence. The GRADE approach could help make the Finish approach more explicit, but there are concerns about the workload. There are currently more than 4000 evidence summaries. Also, there is a desire to keep the current evidence statements as an essential part of the Finnish evidence summaries.
- Clinical Evidence has started to consider diagnostic tests in their topics. As a member of editorial board Ilkka offered to keep Clinical Evidence informed about the GRADE approach.
- US ACCP: Gordon reported that the American College for Chest Physician (ACCP) will consider adopting the GRADE approach, though it is uncertain whether it will be adopted.
- Pablo explained that in Spain there is a national system guidelines. To adopt GRADE the first step is to translate it the judgements into Spanish and disseminate it.

- Germany: Gerd will try to move guidelines groups to use GRADE, but complexity of GRADE is a challenge.
- Up-to-Date is gradually moving towards including grades. They may decide to grade recommendations and not the quality of evidence. They seem unlikely to adopt GRADE at this point, although it might be possible given that most of the work for Up-to-Date is done by a small number of paid staff.
- New Zealand: An 18-month process developing guidelines methodology has just been completed. After that Cindy doubted that guideline producers would be likely to change to the GRADE approach. AGREE has been adopted in New Zealand and the AGREE checklist is used for guidelines. Cindy also raised concern about the need to validate the GRADE approach.

Andy summarised by concluding that the experience to date in different countries suggests that the GRADE approach is too complex. A manual or support tool is clearly needed before groups are likely to begin to use the GRADE approach. Gordon suggested that most of the complexity is in relationship to grading the quality of evidence. We could have a simple and a more complex version of guidelines for grading the quality of evidence. Andy offered to work on simplifying the process and presentation.

**Action: Andy**

### **11. The GRADE Working Group**

Most queries to the GRADE Working Group are currently coming directly to Andy. This has been manageable and no one has complained so far. However, the BMJ paper may change that, as there may be an increase in the number of queries following publication of that paper.

Ilkka suggested that all the information should be placed on the website to make it widely available.

The Working Group has been functioning as an informal group up to now. Andy asked whether the group wished to establish a more formal structure. Nobody wanted this.

We discussed who could represent the GRADE Working Group. It was agreed that any member of the group can speak on behalf of the group (assuming that what is said is consistent with what we have agreed) and that everyone is strongly encouraged to help disseminate the GRADE approach.

### **12. Costs**

We need to flag in evidence profiles and grades whether costs have been considered or not. We need more examples of evidence profiles that include costs. James Mason was supposed to prepare one for SSRIs vs tricyclics. Roman volunteered to prepare one for activated protein-C.

**Action: James, Roman**

### **13. Single RCTs**

David suggested that the grading of RCTs should be modified so that the evidence from a single RCT is considered moderate. A single multi-centre trial or multiple RCTs would then be upgraded to high. Examples of single RCTs that subsequently were found to be misleading include zinc for the common cold and PPI for stroke. There may other concerns with a single study including chance (imprecision), reporting bias (positive studies are more likely to be published sooner than negative studies), and generalisability (directness). Paul supported this suggestion. It also is more consistent with how observational studies are graded. Andy suggested that before we change the approach, we should have evidence profiles that document that the current approach does not work and that a change is warranted. When there is only one single-centre RCT the evidence would almost always be downgraded using the current approach because of the above concerns. Another argument put forward by Holger was that if we were to start at moderate with a single RCT, a single case-control study with a strong effect having considered all critical outcomes would be of the same quality. We agreed to postpone a decision and discuss this at our next meeting when we have some examples of evidence profiles that illustrate the problem with the current approach.

**Action: David, Paul**

### **14. Future meetings**

Cindy suggested that there should be a GRADE workshop at the GIN meeting in New Zealand. She will attempt to secure funds for a couple of people to offer the workshop and send an email to the discussion list inviting people to offer the workshop. Holger volunteered.

**Action: Cindy**

We will have a meeting at the Cochrane Colloquium in October in Ottawa, although the meetings at the Colloquia tend not to be productive because of all of the conflicting commitments that everyone has and the large size of the meetings. Andy will send out a proposal for this meeting.

**Action: Andy**

Two possibilities were discussed for a meeting in early 2005. David said that AHRQ would be happy to physically host a meeting in Washington, but could not fund a meeting. Tessa has been looking into getting WHO to fund a meeting, and this might be another possibility.

**Action: Tessa, Andy**

**Attachment  
Paul's suggestions for presenting quality assessment criteria and quality assessments**

**Quality assessment criteria**

	↑	○	↓
		(default)	
Quality	-	(acceptable)	Serious Limitations
Precision	-	Good Precision	Imprecise or Sparse Data
Directness	-	Direct	Some uncertainty
Full Reporting	-	Good reporting	High probability of reporting bias
Consistency	-	<i>No clear inconsistency</i>	Important Inconsistency
Strong Association	↑ Strong (>2) ↑↑ Very strong (>10)	-	-
Dose Response	Evidence of dose response gradient	No information	-
Plausible Confounders	No plausible confounders (or all would have increased effect)	-	-

**Example of a quality assessment table**

**SSRIs versus tricyclics**

Outcome: <b>Depression severity</b> (measured with Hamilton Depression Rating Scale after 4 to 12 weeks)									
Studies	Design	Quality	Consistency	Directness	SD	SA	RB	DR	PC
8 trials Citalopram 38 trials Fluoxetine 25 trials Fluvoxamine 2 trials Nefazodone 18 trials Paroxetine 4 trials Sertaline 4 trials Velafaxine	RCTs	○	○	↓	○	○	○	○	○
Outcome: <b>Transient side effects resulting in discontinuation of treatment</b>									
8 trials Citalopram 50 trials Fluoxetine 27 trials Fluvoxamine 4 trials Nefazodone 23 trials Paroxetine 6 trials Sertaline 5 trials Velafaxine	RCTs	○	○	○	○	○	○	○	○
Outcome: <b>Poisoning fatalities</b>									
Office for National Statistics (British)	Observational data	↓	○	○	○	↑↑	○	○	○