

Minutes GRADE Working Group meeting

Lyon, 7-8 December 2005

Present: Abdul-Hameed Hassan (Wednesday), Alessandro Liberati (Wednesday), Anne Bataillard (Wednesday), Andy Oxman, Beatrice Fervers (Thursday), Bernard Burnand (Wednesday), Craig Whittington, David Tovey, Francoise Cluzeau, Gordon Guyatt, Gunn E Vist, Hans de Beer, Holger Schunemann, Illka Kunnamo, Jeff Andrews, Katharine Jones, Margaret Haugh (Thursday), Mario Tristian (Thursday), Mariska Tuut, Melissa Brouwer (Wednesday), Merce Marzo, Michelle Kho, Monika Lelgemann, Nicola Magrini, Pablo Alonso Coello, Patrick Bossuyt (Wednesday), Regina Kunz (Thursday), Sarah Rosen, Vigdis Underland, Yngve Falck-Ytter

1. The minutes from Melbourne 26 October 2005 were approved.

2. Use of GRADE & partnerships with other organisations

We reviewed the use of GRADE by various organisations (Attachment 1). We then discussed what information to put on our web pages about organisations that are using GRADE. For example, the Ontario Ministry of Health recently asked to be included in the list of organisations using GRADE. Adding them was not a difficult decision, since they were clearly using GRADE. However some people expressed concerns about listing organisations that ask to be listed but are not actually using GRADE. We discussed the benefits and risks of listing organisations, including the administrative burden.

We agreed that it is worthwhile to list organisations using GRADE for several reasons and that we need simple criteria to decide whether to include an organisation and a short form to capture this information, including, for example: 1) whether evidence profiles are being produced, 2) whether GRADE is being used to grade the quality of evidence, and 3) whether a modified approach is being used. It should be made clear that we are not endorsing organisations that are listed. There was agreement that the minimum criterion would be use of GRADE to grade the quality of evidence. There was not agreement about the need to produce evidence profiles.

Yngve will draft criteria and guidelines for listing organisations on the GRADE web pages, which he will circulate for comments. The following volunteered to help Yngve with the web page list: Holger, Nicola, Pablo, Gordon, Ilkka, Merze and David T.

Action: Yngve

We revisited the discussion from Melbourne regarding a partnership between GRADE and GIN and considered four options:

1. Become an independent organisation
2. Become part of GIN
3. Become part of Cochrane Applicability and Recommendation Methods Group
4. Remain as we are: independent and informal

We agreed on option 4 and agreed to explore with GIN a possible partnership on that basis, including promoting GRADE on the GIN website, incorporating GRADE in GIN tools, and the possibility of annual GRADE meetings at the GIN annual meetings with support from GIN in return for GRADE contributing to the GIN conference, as was done in Lyon.

3. Unorthodox use of GRADE (Nicola & Alessandro)

Nicola and Alessandro have been using GRADE with oncology panels for developing recommendations for new drugs, often in situations without a systematic review or with just one RCT and no head-to-head comparisons. Nicola presented two examples (herceptin as adjuvant breast cancer treatment and oxaliplatin).

They found consideration of consistency may be problematic when there is only one trial; a need for operational rules for minor and major flaws, and sparse data; and a need for guidance for making judgements about which are important and which are critical outcomes. They argued that there is a need to define the critical outcomes up front, whether or not these are measured and reported, and there was agreement with this suggestion. They also sought guidance regarding how to report disagreement regarding the balance of benefits and downsides, and regarding recommendations when these occur, and how to handle disagreements in groups. This led to a discussion of group processes and the use of formal consensus methods. We agreed that this may be outside the scope of the GRADE Working Group, but at the same time it may be important to address these considerations to some extent. Holger and Alessandro will bring back to next meeting those considerations that they feel need discussion in relationship to GRADE.

Action: Holger & Alessandro

We then discussed changing “do it” and “probably do it” to “strong” and “weak” recommendations. This led to a discussion of how we define strong and weak recommendations. A suggestion was made to tie the recommendation wording to behaviour. This discussion was continued in the small group discussions of guidance for making judgements about trade-offs and recommendations.

4. Judgements about trade-offs and recommendations

Holger summarised the Melbourne discussion and the resulted table included in the minutes from that meeting (a modified version of a table in the paper describing the ACCP adaptation of GRADE). The small groups were given three tasks:

1. Decide on the choice of terms and definitions for different categories of recommendations
2. Develop guidance for making judgements about recommendations, building on the discussion from Melbourne

A suggestion was made in Group 1 not to grade recommendations, but it was agreed that we should continue to do so. Group 1 suggested using the terms “Definitely worth doing”, “Probably worth doing”, “Probably not worth doing”, and “Definitely not worth doing”. It was decided that this wording would not be acceptable because of difficulties with translation and because of economic connotations of the word “worth”. Group 2 suggested using “strong” and “weak” + symbols. Group 3 suggested using “recommend” and “suggest” OR “should” and “might”. It was noted this would not work well in French and German. We concluded that we would suggest using “recommend”, “recommend against”, “suggest” and “suggest not doing” in the text of a recommendation. We also agreed that “Strong” and “Weak” could be used in the margin. This is consistent with what was concluded at the meeting in Melbourne.

Group 1 suggested the following definitions:

- Definitely worth doing (“Strong”) = confident that adherence to the recommendation would do more good than harm OR that the net benefits are worth the costs.

- Probably worth doing (“Weak”) = Uncertain that adherence to the recommendation would do more good than harm OR that the net benefits are worth the costs (for specified reasons).

These definitions would have the following implications:

- Definitely worth doing (“Strong”)
 - o Most well informed patients would want the intervention and only a small proportion would not
 - o Most patients should receive the intervention
 - o Decision aids are not likely to be needed
 - o Use of the intervention according to the guideline could be used as a quality criterion
- Probably worth doing (“Weak”)
 - o Most well informed patients would want the intervention, but a large proportion would not
 - o Some patients should receive the intervention
 - o Decision aids are likely to be useful
 - o Offering the intervention and helping patients to make a decision could be used as a quality criterion

We discussed pros and cons of using several definitions to help more people to understand different concepts, implications and to assist with translation (including consideration of different languages, cultural differences, different backgrounds, use in emergency situations versus for chronic conditions), the use of decision aids, and performance indicators.

We did not reach agreement on definitions. A summary of possible definitions or implications associated with strong and weak recommendations include:

Strong recommendations

- Most individuals should receive the intervention
- Just do it
- Most well informed individuals would want the recommended course of action and only a small proportion would not
- Formal decision aids are not likely to be needed to help individuals make decisions consistent with their values and preferences
- Use of the intervention according to the guideline could be used as a quality criterion or performance indicator
- Could unequivocally be used for policy making

Weak recommendations

- Examine the evidence yourself
- The majority of well informed individuals would want the suggested course of action, but a large proportion would not
- Many but not all individuals would follow the suggested course of action
- Decision aids are likely to be useful helping individuals making decisions consistent with their values and preferences
- Offering the intervention and helping individuals to make a decision could be used as a quality criterion or performance indicator
- Policy making will require extensive debates and involvement of many stakeholders

We found that there was some disagreement about definitions due to different thresholds for what people would consider to be a strong recommendation. Some implications would only

apply if there was a relatively high threshold for making a strong recommendation. It was agreed that it is problematic if different groups use different thresholds.

The following volunteered to draft definitions for strong and weak recommendations, including a suggested threshold for making strong recommendations and implications that are consistent with that threshold (across different contexts): Gord, Holger, Nicola, David T.

Action: Gord, Holger, Nicola, David T

There was disagreement about whether we should continue to suggest making explicit judgements about trade-offs before making recommendations. Some people have found the term “trade-offs” confusing and some people found the distinction between making judgements about trade-offs and making the judgement about the strength of recommendation confusing in workshops. Others had not experienced this confusion.

It was suggested that we could suggest starting with a strong recommendation and lowering to weak if there is:

- Low quality of evidence
- Uncertainty about modifying factors
- Uncertainty about the baseline risk
- Important trade offs or uncertain trade offs
- High cost relative to the net benefit
- Inability to reach agreement

It was suggested that a table such as the one below could be used to make these judgements systematically and transparently.

Reasons for going from strong to weak	Low quality of evidence	Uncertainty about modifying factors	Uncertainty about baseline risk	Trade offs	High cost / net benefit	Uncertainty about values
Judgement	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Explanation						

The table would need to be refined, for example for recommendations NOT to do something low quality evidence might not lower the strength of the recommendation and uncertainty about the cost relative to the net benefit would lower the strength rather than high cost relative to the net benefit.

It was suggested that if any of the “Judgement” boxes were “yes”, then the strength of the recommendation would be lowered (i.e. that the threshold for a judgement of ‘yes’ for each criterion is at the point where it would lower the strength of a recommendation). However, it was pointed out that there might be situations in which it is a combination of these factors and not any single factor that would lower the strength of a recommendation.

The table from Melbourne and the one above were seen as complementary, with the one from Melbourne providing explanations and examples and the one above providing a framework for making the necessary judgements systematically and explicitly. We discussed the ordering

of the judgements that go into determining the strength of a recommendation. We agreed that a structured process is desirable to ensure that none of the judgements that should be considered are neglected, but did not reach a conclusion on the order in which the judgements should be considered. Holger together with others agreed to refine the table from Melbourne and the above table to ensure that they are consistent and to bring this back to our next meeting.

Action: Holger

Two of the small groups also discussed the use of letters and numbers versus symbols for grades of evidence and recommendations. We agreed that GRADE would endorse words and the symbols that we previously agreed on, but that the use of symbols is optional (as is anything we suggest). We also agreed that we would discourage the use of letters and numbers, but if others are going to use letters and numbers, we suggest to use numbers for the strength of the recommendation and letters for the quality of evidence.

5. Cochrane Summary of Findings and GRADE Profiler

The programming of GRADEpro will now be done in Rome instead of Oslo. Following consideration of the results of the pilot test of Cochrane Summary of Findings using GRADEpro and discussion of modifications based on this and other experience with GRADEpro, it was decided to completely reprogram GRADEpro. The next version will have a different interface and will function with RevMan (extracting relevant data from reviews and incorporating the Summary of Findings tables in the review in RevMan). The proposed new interface was shown and there was unanimous support for this. We agreed on three primary options when GRADEpro is first installed and set up:

- Guideline option (GRADE evidence profile)
- Cochrane option (Cochrane Summary of Findings with link to RevMan)
- Other options (modifications, such as the one being used by ACCP, if these are requested, funded and incorporated in GRADEpro)

We agreed to require users to select an option from the pick list for each quality criterion (rather than having the default options shown initially), and to continue to require people to include a footnote whenever anything other than the “default” is selected.

Action: Holger

The results of the Cochrane pilot test were presented at the Melbourne meeting and were briefly summarised again. The majority of negative feedback was related to the software and specially the footnotes. The review authors who participated in the pilot successfully prepared Summary of Findings using GRADEpro and were generally positive. Many suggestions for improvements were made and these will be used in developing the next version of GRADEpro and revising the detailed guidance which is incorporated in the GRADEpro help file.

6. Quality of evidence for single RCTs

We have had several discussions regarding possible problems with GRADE for single RCTs since our meeting in Helsinki. Based on these discussions, considerations of several possible methodological studies that might inform these discussions, and the discussion we had in Lyon, we concluded that those who believe that there is a problem should identify examples that demonstrate this, rather than continuing to pursue the various methodological studies that have been proposed, and not to make any changes until there is evidence of a problem

(consistent with our general approach to making changes). Nonetheless, the methodological studies that have been suggested were considered interesting in their own right and Gunn will pursue a simplified approach to this that she has proposed: using a sample of Cochrane reviews to find out if there are times when the first RCT in a review would result in a misleading judgement of high quality evidence.

Holger presented the results of the survey he conducted. The aim of the survey was to determine how frequently GRADE members would be uncomfortable with accepting a strong recommendation for or against an action based on high or moderate quality evidence. Only eight people responded. There were large differences in what the respondents indicated would be acceptable. Several people responded that they found it difficult to answer the questions.

7. Future meetings

We reconfirmed the decision taken in Melbourne that we should not have GRADE meetings at the same time as other meetings, such as Cochrane or GIN, but only as satellite meetings before or after these. The following meetings were suggested:

- May 2006, Bologna (Nicola and Alessandro)
- October 2006, Dublin (before or after the Cochrane Colloquium)
- February or March 2007, Spain (Pablo and Merce)
- August 2007, Toronto (before or after the GIN Conference)
- October 2007, Sao Paulo (before or after the Cochrane Colloquium)

We check regarding preferred dates for meetings by email.

Action: Gunn, Holger

8. Quality of evidence for interrupted time series analyses

Currently all observational studies are lumped in the low quality group except for case series and case reports that start at very low. This approach does not seem sensible to some. Andy suggested that sometimes evidence from time series (well done) provide more than low confidence in the results. The low quality stamp is then non-intuitive. It may be important to consider ITS analyses differently from other observational studies. Andy suggested moving ITS analyses up to moderate quality. There was opposition from the group, in part reflecting different understandings of what ITS analyses are. It was agreed that we needed good examples where use of the current approach is problematic and a clear definition and common understanding of what ITS analyses are in order to take this discussion further. Andy will circulate a definition and examples, including the full reviews, with the next meeting agenda. If we pursue this, we may need to re-consider all the types of observational studies and not just ITS analyses. There was general agreement that we need to be cautious about making changes such as this, and that we should only make changes to our approach if there is good evidence of a problem (and a solution that does not create new problems).

Action: Andy

9. GRADE PowerPoint presentations

Jeff showed the PowerPoint presentation he had produced drawing on other presentations that are on the website. Several positive comments were made and Jeff was thanked for the work he had put into this. Some suggestions for improvement were given and Jeff invited people to send him more detailed feedback by editing the presentation (which will be put on the website) or sending him suggestions by email. It was suggested that we need both a short and

a longer presentation. It was agreed to keep the slide presentations in the members only area of the website.

Action: All

We agreed that examples of evidence profiles should be collected and made available on the web as well.

Action: Andy, Vigdis, Gunn

It was suggested and agreed that the logo that Yngve made for the web pages should become our logo, and Yngve was thanked for this.

10. GRADE website

The web pages should have links to information about GRADE and the GRADE profile instructions in languages other than English. Francoise will send a set of guidelines they have compiled for translating the AGREE Instrument to Yngve.

Action: Francoise

Pablo and Merce will send a Spanish version to Yngve who will put it on the web. Others interested in translating material to other languages should contact Yngve for the translation guidelines and information about which pages to translate.

Action: All

11. Guidance for judgements about limitations and other quality criteria

We will work through developing detailed guidance with examples for making judgements about this and each of the other quality criteria at the next and subsequent meetings.

12. Consideration of equity

Andy briefly reported on some work that has been done on this. Some examples from guidelines developed in the Philippines could be discussed at the next meeting.

Action: Andy, Tony Dans

13. Diagnostic tests

No one from the Cochrane Screening and Diagnostic Tests Methods Group was able to make it to the meeting to discuss collaboration between GRADE and that group. We will invite someone to come to a future meeting and will continue to work on writing up a description of our approach.

Action: Holger

14. Costs

We will work through some examples at next and subsequent meetings. Joanne Lord will bring examples from NICE to the next meeting.

Action: Francoise, Andy

15. Publications, applications

- Editorial in ACP Journal Club – In press
- ACCP publication including GRADE in Chest – In press
- Surviving sepsis - There may be a publication from this experience.
- Nicola and Holger are working on an editorial for BMJ.
- Holger is working on the paper on diagnosis.
- Gunn is working on a paper from the Cochrane Summary of findings pilot study.

- Holger is working on an urology paper.
- UpToDate grading paper is with UpToDate – likely to go on their website
- ATS paper in final draft format

We need to update the manual. Please use track changes and send suggested improvements to Holger.

Action: All

Funding would be helpful for the further development of GRADEpro and other projects, including an agreement study, comparisons of different presentations, and evaluating alternative methods for converting continuous outcomes. Possible funding sources that were suggested include the NHS R&D HTA program, the EC and AHRQ. We all need to be on the look out for good opportunities for funding.

Action: All

Thank you to Margaret and Sarah and SOR guideline program of FNCLCC (French Federation of Comprehensive Cancer Centres) for hosting the meeting!

Attachment 1

Use of GRADE by various organisations:

- EB Urology Task Force has had two meetings already and they realise that they should collaborate better and want advice on using GRADE.
- American Thoracic Society is interested in adopting GRADE, will decide at a meeting in the spring. An editorial suggesting GRADE was published in November.
- The Ontario Ministry of Health is using GRADE.
- Some NICE guidelines groups have used and are using GRADE (e.g. within mental health for depression in children and dementia). A group in NICE is working on assessing all the approaches used in NICE, including GRADE and is considering how to address economic considerations and diagnostic tests.
- Cancer Ontario is exploring the use of GRADE.
- Swiss federal office of medical care is considering using GRADE for submissions for new procedures by specialty societies.
- Nicola and Alessandro are using GRADE with oncology guidelines panels in Italy for new oncology drugs.
- Am J Obs & Gyn approached Jeff A.
- The Spanish Primary Care J is publishing something on GRADE.
- ACCP adopted a modified GRADE approach.
- UpToDate is adopting the modified GRADE approach used by the ACCP.
- Roman is helping the Surviving Sepsis Campaign to use GRADE in revising their guidelines.
- All the Finnish guideline users are grading the quality of evidence now and summary of findings tables (based on GRADE approach) are being produced.
- Spain Primary Health Group Clinical Practice guidelines are trying to introduce GRADE as well as SIGN and are starting some seminars with only GRADE. A publication is planned based on this experience.
- Kunnskapssenteret in Norway is using GRADE for the quality of evidence and the first report using GRADE in Norwegian has been published.
- Bone & Joint Surgery J published an article about GRADE the day before the meeting.
- GRADE will be used in some Italian regions and in national guidelines starting next year.
- Guidelines for WHO Guidelines include using GRADE, but few groups have done so so far.
- A kidney association is using a modified GRADE (advert in GIN p-16)
- A review of different approaches by the Canadian Coordinating Office for HTA selected GRADE & SIGN after assessing different approaches (https://www.ccohta.ca/compus/compus_pdfs/COMPUS_Evaluation_Methodology_final_e.pdf)
- ?Asco? were interested but at one of the official meeting someone said that they should not be making recommendations. They rarely do SRs themselves
- World Gastroenterology Society is interested in copying the EB Urology approach, using GRADE.