

Draft minutes GRADE Working Group meeting 6 & 7 May, Rome 2008

Participants: Elie Akl, Jeff Andrews, Pablo Alonso Coello, Hilda Bastian, Hans de Beer, Sergio Bonini (day 1), Patrick Bossuyt, Helena Brändström, Jan Brozek, Massimo Brunetti, Stephanie Chang, Francoise Cluzeau, Siobhan Crowley, Jon Deeks, Ben Djulbegovic, Kristina Eklund, Yngve Falck-Ytter, Bo Freyschuss, Signe Flottorp, Paul Glasziou, Gordon Guyatt, Robin Habour, Margaret Haugh, Mark Helfand, Suzanne Hill, Roman Jaeschke, Katharine Jones, Brice Kitio, Regina Kunz (day 2), Gero Langer, Alessandro Liberati (TBC), Joanne Lord, Nicola Magrini, Alison Martin, Merce Marzo, Paola Muti, Andy Oxman, Silvia Pregno, Nancy Santesso, Rob Scholten, Holger Schünemann, Vijay Shukla, Jane Sisk, Gunn Vist, Craig Wittington, John Williams

1. **The minutes from Washington DC** were approved.

2. **The BMJ GRADE series**

The series is now coming out. Thank you to Gordon.

The open access was not open to all, Andy will follow up.

The short and long versions are not consistently linked, Gordon will follow up.

Action: Andy & Gordon

3. **The JCE GRADE series**

- (4) Rating the quality of evidence – risk of bias – revised (Attachment 3)
- (6) Rating the quality of evidence - inconsistency and indirectness – revised (Attachment 4)
- (7) Upgrading and summarizing the quality of evidence - revised (Attachment 5)
- (8) Preparing summary of findings (SoF) tables (Attachment 6)
 - How to present continuous outcomes?

We agreed that it is ok to use these drafts for guidance and to share them. This series is intended to be a complete series containing all the necessary information and examples for applying GRADE. The main target group is systematic review authors and guideline developers. The BMJ series targeted end-users.

There was a discussion about the need for a separate article in the series focusing on observational studies. Observational studies are already addressed in other articles; including, for example, the current article seven on upgrading the quality of evidence. It was argued that there is still a need for a separate paper focusing on special challenges of applying GRADE to observational studies with relevant examples. We also still need examples of SoF tables that include both RCT and observational data. We agreed to add another article to the JCE series focusing on observational studies. Holger and Sue agreed to draft the outline for this article.

Action: Holger & Sue

A suggestion to add another article on applying GRADE to screening was discussed. For the time being it was decided not to add another article on special challenges of applying GRADE to screening.

Article 4, Rating the quality of evidence – risk of bias

We discussed the usefulness of additional guidance in this article on the assessment of the risk of bias in relationship to losses to follow up. When are losses serious enough to lower the quality of evidence? Mark had two additional examples regarding prognostic adjustments that he will send to Gordon. Other issues of wording and spelling have been or should be forwarded to Gordon.

Action: Mark & Gordon

Article 6, Rating the quality of evidence - inconsistency and indirectness

We discussed variation in the cut-offs/ranges that have been suggested for deciding when there is serious heterogeneity. Julian Higgins in the Cochrane Handbook suggests overlapping ranges: 30 – 60 = maybe moderate, = 50 to 90 maybe substantial, 75 to 100 = considerable heterogeneity. These seem to capture the arbitrariness of using precise cut-offs for which there is little basis.

Article 7, Upgrading and summarizing the quality of evidence

It was suggested that we should delete the rule of basing the overall quality of evidence on the lowest quality for any critical outcome because of there being so many exceptions. We decided to keep the rule and to specify and clarify the exceptions.

When there are observational studies, and dose-response data for different doses than the intervention, this is indirect evidence. In general, this does not provide an adequate basis for upgrading the quality of evidence, although there may be exceptions. We still need examples of dose-response relationships that justify upgrading the quality of evidence. Pablo had an example with RCTs where the quality of evidence was downgraded because of imprecision where it might also be justified to upgrade because of a dose response relationship, he will send the example to Gordon.

Action: Pablo

We also discussed the use of logic, biologic reasoning or plausibility as a basis for upgrading – or downgrading for the lack thereof. We agreed this rarely if ever provides an adequate basis for upgrading the quality of evidence. Gordon will attempt to add something about this to the article based on the discussion and input from others.

Action: Gordon, Paul, Mark, Holger, Andy, Regina, David Atkins

We split into groups for discussions of the remaining articles

Article 8, preparing summary of findings (SoF) tables

We reconfirmed the decision from the meeting in Washington DC that there should be the option of using different presentations for different audiences.

There was agreement that the presentation of continuous outcomes remains difficult. Most of the available alternatives are difficult to understand, based on uncertain assumptions, or open to misinterpretation.

It was also suggested that NNTs should be discussed in the paper, including their limitations when presented for several outcomes or for outcomes with confidence intervals that cross no difference.

4. JCE Article 5, rating the quality of evidence – random error

There are uncertainties about when to downgrade for imprecision. Additionally, some statisticians are worried about the use of the term narrow confidence interval. It was suggested that a step-by-step description in the article of what to look for would be useful.

5. JCE Article 10, special challenges - resource use

The two issues raised by the group that discussed this article were the need for more practical tips related to the implementation of the approach outlined in the paper and further consideration of the practicality of including economic evidence in a summary of findings table versus using an economic model.

6. JCE Article 9, the GRADE approach for diagnostic tests and strategies

Jon Deeks gave a presentation about the Cochrane diagnostic test accuracy group including the methods of meta-analysis. He also informed us that the first diagnostic test accuracy systematic reviews are expected in the October Issue of the Cochrane Library this year.

Nancy, Jan and Holger prepared a series of examples of SoF Tables for diagnostic tests to work through the way that SoF Tables and evidence profiles should be presented. These examples were based on different reviews including one draft Cochrane Review.

We split into small groups to discuss GRADE for diagnostic questions and test accuracy, focusing on presentation issues around summary of findings.

Example 1

A two stage process was suggested. The group suggested that the presentation of stage 1 could be clearer about which test is better in terms of accuracy. They also suggested presenting the five quality criteria. Indirectness needs to be considered at two levels. The group also suggested reducing the number of questions in QUADAS to a shorter list of criteria for evaluating study limitations.

Directness is assessed in stage 2 in relationship to the extent that test accuracy is associated with patient important outcomes. Questions were raised how prevalence and a cut-off value for the test should be chosen, which depends on the balance of benefits and harms and can therefore not be guessed in advance.

Example 2

The group considered the evidence profile in relationship to its use by guideline panels and the SoF table in relationship to its potential use by guideline panels in deciding on recommendations and for communicating with end users. The group also discussed the importance of being clear about the patient important outcomes that are considered and judgements about the association between test results and those outcomes. They suggested that the directness column should perhaps be expanded to include separate columns for the different types of directness.

The group suggested that it would be better to sort the table by prevalence (pre-test probability) rather than by outcomes with multiple prevalences for each outcome (TP, TN, etc.), perhaps with separate tables for different prevalence's. The group also suggested eliminating redundant information, such as repeating the study design, participants and publication bias in each row. This needs to be balanced against the desirability of having consistent presentations for intervention profiles and diagnostic test profiles.

The group considered cost, complications and accuracy in phase 1. They thought that maybe we should rename CI because the test accuracy CI does not reflect confidence. A suggestion was made to add a relative effect column, and to stratify by test threshold (rather than selecting one).

Example 3

The group suggested listing the true positive, false positive, true negative and false negative values in columns instead of rows. They also suggested that the importance column should be changed to consequences.

There is a need for a clearer distinction between estimates derived directly from test accuracy studies and those based on assumptions (about the relationship between test results and patient important outcomes).

The potential of downgrading for the same thing more than once needs to be addressed and clarified. The assumptions that are made about prevalence (pre-test probability) should be explicit and ideally should refer to a relevant research.

The reporting back from the small groups was followed by discussion of the issues that were raised.

There was broad agreement that a two stage approach should be clarified, with the first stage focusing directly on test accuracy results and the second stage focusing (transparently) on the evidence that is used and the judgements that are made about the relationship between test results and patient important outcomes. Further discussion and clarification of different types of directness is needed for each stage.

It was suggested that the QUADAS criteria should be shortened. Three of the QUADAS criteria are dropped from the Cochrane diagnostic test accuracy group criteria because they address reporting rather than bias.

It was suggested that SoF tables should have columns with sensitivity and specificity.

The SoF tables are the bases for decision making – the rows should indicate patient important outcomes. Questions were raised about the usefulness of stage 2 if stage 1 shows that a test is useless. We need more examples of evidence profiles and SoF tables for diagnostic tests.

Patrick suggested that we should be cautious about promoting GRADE for diagnostic tests before we have more experience. Some of those present also thought that for interventions we started in 2000 and met many times over 4 years before we were ready to publish the first GRADE article in the BMJ. However, it was clarified that we have been working on diagnostic examples for more than four years now. Furthermore, there is a greater potential for harm if we do not provide any advice for how to apply GRADE for diagnostic tests. The BMJ paper is out and users are asking for advice. Another suggestion was to wait until Cochrane diagnostic test reviews become available, to see what the authors of those reviews do and seek more input from them. We all agreed that we need to include more examples in the paper. Examples and comments on the paper should be sent to Holger.

Action: All

7. Alternative expressions for “strong” and “weak”

Elie and others have evaluated the use of different presentations in randomized studies. A comparison of numbers and symbols found that symbols were better than numbers for conveying the strength of recommendations and there was little difference between letters and symbols. Some concerns were raised about strong recommendations to do something, whereas the biggest problem has been with the term “weak” recommendations. There were different perceptions of the extent of the problem. Gordon suggested that there has not been a problem with strong and weak in North America. Sue reported that WHO guideline panels have

revolted against the terms “strong” and “weak”. Suggested alternatives for “weak” include “conditional” and “qualified”. Francoise reported that reflecting the strength of recommendations is new for NICE. They are using a modified version of GRADE and as GRADE they incorporate cost-effectiveness and patient considerations. It was noted that NICE makes recommendations for funders within a legally binding framework, whereas many guideline developers make recommendations for clinicians without the same authority or mandate that NICE has.

It was agreed that we need an alternative term for “weak”. Three alternatives were proposed and voted on:

Conditional, 11

Qualified, 5

Discretionary, few

Gordon suggested for weak to remain as the default, 11 voted for and 2 against.

8. Good practice guidelines

This issue was revisited from Washington DC, we still need a way forward. After a short discussion it was agreed that a draft proposal would be prepared that could be discussed at our next meeting.

Action: Elie and John W and Gordon

9. What is necessary to claim to be using GRADE?

We do not have a formal process for deciding what we would accept as a “modified GRADE” approach versus what we would consider to be fundamentally different from GRADE and prefer should not be called a “modified GRADE” approach. If, however, we are asked this by a group that is using or proposes to use a modified GRADE approach, we should be consistent. Our main concern is to avoid confusion that can arise from having multiple versions of “GRADE”, particularly ones that are fundamentally different from what we suggest.

AHRQ has proposed an approach to grading that is largely based on GRADE, but with some differences. These were discussed at our meeting in Washington D.C. and in a subsequent teleconference with members of the AHRQ working group and GRADE. Some of the differences were in the wording that is used. Overall the proposed AHRQ approach is quite similar to the GRADE approach and AHRQ wants to continue working with GRADE. They have submitted a paper for publication that describes their approach. Their approach does not use “points” for downgrading and upgrading. They had initially dropped publication bias as a criterion for downgrading, but have since included that. They call criteria that often are not relevant “optional”, but may change that term. An important difference is “applicability”, which they consider separately from the quality of evidence and not as a reason for downgrading (as a type of indirectness). For a second important difference is that observational studies can start as either moderate or low. They have not specified when observational studies would start out as moderate. They are still considering this and whether to change this.

KDIGO has decided to use something different from GRADE, although they have been calling it GRADE. Gordon, Holger, Regina and Andy drafted a letter expressing concerns about KDIGO’s approach to recommendations, which includes three rather than two grades of recommendations and are not consistent with GRADE’s two grades and definitions. It was agreed that this was a fundamental difference.

We had a discussion about which organisations should be listed on our web pages. A key reason for listing users of GRADE on our web pages is to give people a place to find examples of how GRADE has been implemented. Including organisations that are using approaches with fundamental differences is likely to be confusing. We agreed that we need some text on our web pages to explain the criteria we use for deciding which organisations to include.

Action: Yngve

10. Clinical Evidence

BMJ Clinical Evidence has decided to use GRADE. In order to apply GRADE to all of the existing chapters they found it necessary to take some shortcuts. This has been done for over half of Clinical Evidence’s content (250 systematic reviews). Clinical Evidence has found this difficult and have identified a number of issues that required pragmatic decisions, including how to address harms. They have seven editors, a limited budget, no in-house statisticians, and they search for systematic reviews and RCTs only (not observational studies), while still trying to ensure that harms are assessed and included.

Several organizations are now beginning to use GRADE and aim to keep reviews and recommendations up-to-date. It would be possible to reduce unnecessary duplication of efforts if there was a mechanism for sharing evidence profiles. UptoDate has elected to introduce GRADE slowly. The Cochrane Collaboration has also

started to include Summary of Findings tables. This entails a lot of work because of the need to go back to all the included studies. Silvia reported a pilot study including 10 systematic reviews from different Cochrane groups. Challenges included the amount of time required to go back to the original studies, particularly for reviews with large numbers of studies or without electronic versions of study reports, and reviews with lots of outcomes.

11. Should long-term effects (e.g. harms) be discounted relative to effects that occur much earlier (e.g. benefits)? and other considerations about values and preferences

For example, people screened for colorectal cancer may decrease their risk of dying from colorectal cancer (a long-term benefit) while increasing the risk of dying from a complication of screening (a short-term harm). We agreed that discounting needs to be addressed in the JCE series. It is currently addressed in the paper on resource use. It was suggested that it should also be addressed in the paper on strength of recommendation in relationship to health outcomes. The emphasis should be on transparency in terms of whether and if so how discounting was taken into account in judgements about the balance between desirable and undesirable effects.

Action: Gordon

12. GRADE like criteria for prevalence and risk factor studies

John reported that one of the US Evidence-based Practice Centres is working on methods for summarising and grading evidence on prevalence and risk factors. Andy suggest to encourage them to come to one of our meetings, and also to suggest that they might want to collaborate with the Cochrane Prognosis Methods Group.

Action: John

13. Should randomised trials not be considered high quality evidence for quality improvement?

We agreed that there was nothing new or substantive in this article (a commentary by Donald Berwick published in JAMA in March) or a strong argument for not considering randomised trials as high quality evidence for quality improvement.

14. Cochrane Summary of Findings tables

Andy reported that the Cochrane Collaboration is facing substantial challenges in incorporating SoF tables in Cochrane reviews, particularly in reviews that are already published in the Cochrane Library. Julian Higgins and others are organising a workshop that will address training needs.

15. GRADE Profiler

GRADEpro is available on the Cochrane Website for download – officially since March 2008. Holger urged everyone to go to <http://www.cc-ims.net/gradepr> and try GRADEpro. There is a feedback form on the website which everyone should complete.

Action: All

Holger, Jan and the rest of the team were thanked for all of the hard work that they have put into developing GRADEpro.

16. Database of evidence profiles

Yngve demonstrated how to export evidence profiles as xml files and upload them onto www.gradeworkinggroup.org. Please also send Yngve the GRADEpro file for double checking that files are uploaded correctly. Yngve also demonstrated the search options that are currently available. Please send feedback and suggestions for improvements to Yngve. We decided that the listed profiles should be available for reimport to GRADEpro and that we need to ensure that modified GRADE profiles and summary of findings tables should be labelled as modified and include the source when they are uploaded again.

Action: Yngve

Thank you to Yngve.

A question was raised regarding whether it would be possible to include evidence profiles from Cochrane reviews since the SoF tables are published (and not open access). It was pointed out that ACP Journal Club uses information from articles to produce new tables and references the original article, without any problem, and that might be an option, if it turns out not to be possible to include evidence profiles from Cochrane reviews in our database of evidence profiles.

17. GRADE website

We need to change the tense of the wording in our aim on the web pages.

Action: Yngve

18. Publications, workshops, applications

- Holger has a series of articles for Allergy. He noted that it is helpful to translate articles to German because of the language barrier. Gero will work with Holger on this. Peter Tugwell has said that it will be ok to translate the JCE series. A GRADE workshop will take place on the day following this meeting in Rome.
- Mark reported that AHRQ is planning its second annual meeting in Washington D.C. in September and there will be an all day session on GRADE. A workshop will take place in May in Washington at AHRQ - a series of papers on diagnostic tests, going from accuracy to outcomes will be distributed at that meeting. GRADE will be represented at that meeting.
- Regina is writing a paper for an internal medicine journal.
- Elie is planning to submit the presentation trial paper for publication.
- Brice reported that there is interest in France in using GRADE and there are plans to translate the BMJ series.
- Benjamin reported that guideline panels with which he has worked love seeing the evidence profiles.
- Kristina reported that GRADE is now being used by the National Board of Health and Welfare for guidelines.
- John reported that the Veterans Administration guideline producers are considering changing and will consider using GRADE.
- Signe reported plans for a GRADE workshop at the GIN Conference and that they are using GRADE at the Norwegian Knowledge Centre for the Health Services.
- Vijay reported that CADTH is using GRADE a lot and presenting results in publications and at conferences.
- Rob reported on a workshop for authors of systematic reviews of diagnostic test accuracy.
- Pablo reported on a GRADE workshop in Spain next week for the national guideline producer. The Spanish Cochrane Centre is providing workshops on GRADE.
- Jan reported that the Allergy papers are soon finished and he has been asked to produce another series for another journal.
- Gero reported using GRADE for guidelines for nurses .

GRADE workshops for industry representatives

A few members have been approached by industry about GRADE workshops. We discussed how this may affect us and how it might have implications for publications. Suggestions were made that workshops are generally open to the public, including people from industry and that if they are funded by industry they should be open to the public. It was noted however, that if you take money from industry this may be perceived as a conflict of interest. We agreed that we should add to our web pages acknowledgement of organisations that have sponsored GRADE Working Group meetings, but not necessarily sponsors of workshops.

Action: Yngve

Andy summarized the discussion by saying that decisions about funding for workshops are personal decisions since GRADE is not a formal organisation and it is not possible to give money to GRADE. We should, however, remain cautious about accepting funds from industry for GRADE Working Group meetings.

19. Future meetings

- 8 October 2008, Freiburg. Holger has organized this. People should register to attend this meeting on the web pages for the Cochrane Colloquium.
- It was suggested that we should apply to the Rockefeller Foundation for a meeting at Bellagio in May 2009. Gordon is eager to do this and offered to write an application together with Sue, Holger and Regina.

Action: Gordon, Sue, Holger, Regina

- SIGN may be willing to host a meeting next year.
- NICE may also be willing to host a meeting next year.

Andy will follow-up on these possibilities.

Action: Andy

- Holger suggested that if there is not funding for a meeting, we might still want to have a meeting if someone were able to provide a meeting room and we paid our own travel expenses.
- Vijay will check into whether CADTH might be able to fund a meeting.

Action: Vijay