

Minutes GRADE Working Group meeting ROME, April 27 & 28, 2005

Attendants: Andy, Oxman, Holger Schunemann, Jan Brozek, Yngve Falck-Ytter, Gunn Elisabeth Vist, Lise Bosquet, Margaret Haugh, Francoise Cluzeau, John Williams, Mariska Tuut, Signe Flottorp, Regina Kunz, Alessandro Liberati, Katharine Jones, David Tovey, Tessa Tan Toores, Jane Thomas, Helena Varonen, Gordon Guyatt, Roman Jaeschke, Jacek Mrukowicz, James Mason, Jeff Andrews, Paul Glasziou, David Atkins.

Apologies: Robin Harbor, Mark Helfand, Pablo Alonso Coello, Bob Phillips, Nicola Magrini.

1. Minutes from Ottawa,

Add Jane to the attendees list. Rephrase the information on NICE in 8. With those changes the minutes from Ottawa was approved.

Action: Gunn

2. GRADEpro

We split into small groups (2-4) and used the GRADEpro to grade the systematic review: Melatonin for the prevention of jet lag by Herxheimer A, Petrie KJ. Issue 3 of the Cochrane Library 2003.

General feedback about the GRADEpro:

Feedback:	Solutions
From when Regina tried to open after installation: "Application has generated an exception that could not be handled"	Report error to technical staff.
GRADEpro cannot be installed on a network	Add this to installation instructions that GRADEpro needs to be installed on your own hard drive.
Conflict with lotus notes	Add to installation instructions

Feedback relating to each of the GRADEpro screens:

New profile screen:

Program an overall rating of the evidence, so that the all in one direction rule can be included
GRADE summary, suggest improved file structure
Suggest that the program open a profile when GRADEpro opens as it is not intuitive how to start the first profile
The question format based on Cochrane titles, Gordon thinks this invites specification of population more than most guideline users would normally do which could influence the outcomes and inclusion criteria. Suggestions for better solutions are welcome.

Outcome screen

Suggestion to have tool tips available instantly	Click boxes on their way
The number of studies was not intuitive	Already improved
How was the outcome assessed, difficult to know what is	Improve help

meant	
The new system is much better, would like to be able to move footnotes manually up and down	Can remove and add, doing so changes #
Suggestion to list the outcomes first before you start on all the details	Add to help
The footnote deletion did not work	Report error to technical staff.
Complicated, to many buttons to click for editing and making footnotes	Explore possibility for improvements
Did not know where to put in economic studies (RCT-Obs-?)	Modifications are on the way for diagnostics and costs
Sometimes outcomes are so different that we may not want to specify all, like the adverse effects: How to merge or skip or add more information	This is likely best dealt with in footnotes. Add instructions to help.
Suggestion to number reports so that it is possible to just add a comment	
Design: A suggestion to starting with a blank, but then you can forget to look at that box, or a default that stops you going on before choosing. A suggestion to keep the default, but would like a prompt. A suggestion for first time with no default with question do you want to have this as default. Efficiency of doing a lot of very similar outcomes is not really an issue for Cochrane SR and guideline panels. A RCT with no issues is very rare. Vote Result: Do not allow people to set options	Avoiding repetitions vs Having to think through

Other consideration screen:

When you have suggestions or suspicions about limitations but no proof, should you downgrade or not? Should people give reason anyway?	Start with the review and assume that all these issues have been considered during the review process. Higgins group are already working with some of these issues: How well do we assess these limitations.
Gordon: this highlights the threshold problem. It is or it is not whereas there is clearly a continuum, rather label where you put the threshold. Need to be explicit.	We should always specify what went into the judgement
More qualitative comments are helpful but for Cochrane we need to avoid more work, there is an extra row entitled: Other comments. We should separate for different users as guideline producers and users are different.	Transparency and explicitness are very important
Would we want a Cochrane systematic review to not have considered all of these issues, a specific detailed GRADEpro would then make it easier for authors to do	

their job. Also easier to identify differences. Some of these issues are already well covered in other mandatory tables.	
A suggestion for different profiles for Cochrane and guidelines with more forced footnotes for guidelines, different version for different audiences. However, the fewer systems the better and there is a need to 'communicate' and to avoid duplication of work. One of the GRADE group goals is to make it easy.	
A post-hoc suggestion by Margaret was to rename footnotes because the term footnote might be understood as something less important. We came up with the word explanations. We should discuss this once more.	
PC, plausible confounders are difficult. Situation with evidence of effect, PC would have reduced the effect No effect and all PC would have increased the chances of biased effect	This was included to help observational studies. Suggest that it is removed when RCT study design.
Several suggestions to include good examples in the GRADEpro.	We will include examples in the GRADEpro

Summary of findings:

Suggestion to use words: Fewer or more event because readers are often confused about – or +	For continuous outcomes we already have to click if lower score is better or worse.
Challenge with qualitative summaries	
Clarify Number of patients vs number of patients with outcome	

Report:

Format of the Report can be set to either landscape (one table) or portrait with two tables (one quality assessment part and one summary of evidence part).

It might be a good idea to increase the font size of the footnotes.

Action: Andy, Gunn, Holger, technical staff

3. Cochrane summary of findings of evaluations of GRADE

Cochrane Collaboration Steering Group has agreed that a Summary of Findings table should be included in the next version of RevMan. The current technical solution is to use an altered version of GRADEpro linked to RevMan.

We are planning a pilot test of the use of Summary of Findings in Cochrane systematic reviews and 22 Cochrane Review Groups have agreed to participate in the pilot study that will take place during the summer months of 2005.

We have previously discussed and planned to evaluate GRADE, but do not have funds for this. Françoise knows a guideline group that is using GRADE and will ask them for feedback.

Action: Françoise

Gordon and Holger has new student, Martin O'Donnell, who may be willing to help test GRADE in the next ACCP guideline.

Action: Gordon, Holger

There was a suggestion to have a structured feedback format for testing of the GRADEpro in general as well as for the Summary of Findings table in Cochrane.

We asked for volunteers to help introduce the Cochrane review authors to GRADE during the pilot study. Volunteers: David, Andy, Jacek, Kathrine, Paul, Yngve, Jane, Signe, Francoise, Helena, Jeff, Holger and Gunn

Action: Gunn

4. Equity

We agreed to postpone discussing this until the next meeting

5. Website and discussion list

Yngve agreed to continue to manage the web pages and to improve them. Regarding requests from others to send messages to the discussion list, Yngve will continue to use his good judgement.

John suggest a possibility for people to register to be contacted if there is new information on the web pages

GRADEpro is currently on the password restricted pages. GRADEpro needs more piloting before open release. We should increase the number of people piloting it. Yngve will look into ways of protecting GRADEpro against external non-GRADE-approved changes.

We have not decided about where the copyright will go.

Andy suggests that on the web pages, people who want to register and become a member then they can get the password.

Contact for general questions should be a general email address like: GRADEpro@Cochrane.no

Action: Yngve

6. GRADE workshop/ Teaching package

The group thinks it a good idea to have a standard teaching package.

GRADEpro is very helpful when pc rooms are available. Time is essential for participants to be able to read the review.

Gordon volunteered to take part in the GIN workshop on GRADE.

Jeff Andrews agreed to make a generic presentation of GRADE to audiences. We decided we all would forward presentations about GRADE to Yngve to collate. Jeff would build on these presentations.

Action: Jeff

7. Diagnostic tests

We split into two discussion groups, one diagnostic and one for cost.

Present in the diagnostics group discussion: David A, David T, Francoise, Helena, Jane, Jeff, John (reporter), Margaret, Mariska, Merce, Paul, Regina, Signe and Yngve.

- We want to collaborate with the Cochrane diagnostics group.
- The group started with the example re: shoulder pain by Helena. Suggestion to add: decrease in shoulder pain as an outcome and to include information regarding prevalence.

- Directness is a source of confusion,

The discussion group concluded that some issues had to be worked more on:

- What are the consequences?
- Which are patient important outcomes?
- When is there clear evidence of effect?
- When is it direct or indirect evidence?
- Want a help menu with examples, and separate examples for the diagnostic tests.
- When does evidence become stable?
- What is a strong association; should we use diagnostic odds ratio – LR+ and LR-?
- We need to tackle what is reporting bias in diagnostic tests.
- Dose response – as it gets more positive, this is difficult to label and the group were unsure if you would upgrade if you found that relationship.
- Present vs absent tests can not have DR, although in real life it is never clear cut.
- Some think that PC is irrelevant to diagnostic tests, challenge to find examples, then PC goes back in.

Paul volunteered to make initial contact with the Cochrane diagnostic group.

Action: Paul

Holger, John, Gordon, James, Margaret, Roman, Paul, Yngve, Jane and Regina want to keep working on the diagnostic test issues and start by writing a journal article. Holger volunteered to lead this effort. We will request plenty of examples.

Action: Holger

- Diagnostic examples for GRADEpro are needed. A suggestion was made to express as per 1000 and use the same profiler for diagnostic tests rather than making a different version. However, it was pointed out that the studies of diagnostic tests do not fit into study design for test accuracy.
- There should be separate rows for the different outcomes and consequences of the test: TP, FP, FN and TN outcomes.

8. Cost

Present in the cost discussion group: Andy (reporter), Tessa, Katharine, Gordon, Jeff, Holger, Gunn, Jan, Jacek, James and Roman.

The group started with the bed nets profile by Katharine. The study did not provide enough information to be useful. We then had a general discussion about resource utilisation which should include important and critical issues of cost utilisation and should report on natural use. James volunteered to produce examples.

Action: James

The group also discussed the issues of international recommendations, when and how it makes sense.

ACCP had suggested to have two recommendations one with and one without costs, the discussion, which raised concerns about abuse by the pharmaceutical industry of the recommendations without cost, may have discouraged it. (post meeting addition: Gordon and Holger are pushing to abolish this recently suggested policy at the ACCP level)

Further discussion in the large group started with the example of Activated Protein C for severe sepsis by James, Roman and Holger. Footnote 8 is missing in the table. James and Roman got more worried the more they evaluated the individual quality criteria. The primary analysis paper does not report the original data or analysis. And the additional FDA results are a subgroup analysis. The resource utilisation data have no added limitations to that of the effectiveness data and it is directly applicable to the patients.

There is a need to identify the main resources. There are many assumptions and uncertainty about the reliability of these, including simplifications of complex patterns extrapolation to the future. Real life decisions often include multiple comparisons, rather than simple comparisons between two alternatives.

GRADE profiles should include a row for each important resource utilization item.

Evidence base for effectiveness and harm are different, maybe also so for costs.

The SSRI example for resource utilisation that James prepared was difficult. Disaggregated costs/resources were not included. James will revise this profile and we will discuss it at the next meeting.

Action: James

9. Progress reports on organizations using or considering using GRADE

ACCP has moved further in the direction of GRADE and has produced a draft paper for publication. ACCP thought the BMJ paper difficult and suggested that a simpler explanation of GRADE is needed. Their draft paper is intended to fill this need. Gordon asked the GRADE Working Group to endorse the ACCP paper.

It was felt that the ACCP paper reads well and is easy to understand with the use of plenty of examples, however, there are still several important differences between the updated ACCP approach and GRADE. For example, the updated ACCP approach has collapsed the low and very low grade of evidence, then collapse high and moderate in the last row.

Andy argued that the implication issues make it more complex and confusing by making two sets of recommendations. Andy disagreed with implications column, but did not get support for this. We decided that we should mention – in the ACCP paper – that the differences between the ACCP

simplified approach are: number of levels of quality of evidence, the use of numbers and letters and the column of implications.

Action: Gord

Francoise warned that we should not underestimate how many people are going to use GRADE, and that it then is very important to distance GRADE from everything else.

It is common for other systems also to include GRADE aspects as they learn about it, and we agreed this is for the better.

Gordon suggested using a visual analogue scale to improve flexibility. There was support for the GRADEpro tool that GRADEpro forces you to make decisions, either limitations or not, reduces all the compromising. We agreed that it is important to be explicit and that GRADE is a lot easier to use with GRADEpro. We also agreed that the use of examples is very helpful, and it has been the plan all along to include lots of examples in GRADEpro.

Use of the GRADE system/approach

ADO: Do we want to endorse modified GRADE approaches?

Other users of GRADE may not ask for endorsement. What the GRADE group should do is push the GRADE approach. A question came up what to do when other groups starts to 'upgrade' and 'improve' it and start using the name GRADE? We decided that we would try to be conscious about what others do with GRADE and that Yngve when he receives requests through the website will bring it back to the group.

Other organizations using GRADE

- The American Endocrine Society is posting that they use GRADE on their website.
- An executive committee of the major Urological Societies asked Andy and Holger to present GRADE. Holger presented to the 10 presidents/executive officers of these organisations. They liked the idea of an international standard approach. They plan to establish a task force of urologists that will get together with 5 GRADE representatives to help develop their guidelines. And they wish help to develop two papers. The Urologists want one paper with 10 urology examples and one with methodological details.
- Oncology guidelines in Italy. Holger is in the group trying to develop national guidelines, this was previously a mix of grading approaches but the next version may be based on GRADE.
- The Norwegian Health Services Research Centre has decided to adopt the use of GRADE in addition to their modified version of SIGN. There is an increasing interest in GRADE, and a Norwegian translation of the GRADE method will be included in the Centres manual.
- Regina reported some interest from her organisation. They may move towards using GRADE for their guidelines.
- There will be a GRADE workshop at the iHTA annual conference this year.
- BMJ Clinical Evidence has tried using GRADE against quite a bit of resistance, but this changed when introduced to GRADEpro.
- Mariska thinks that the Dutch guideline committees may be willing to modernise soon.
- Yngve is working on a translation to publish a German version of GRADE.
- There are 2 guideline producers in Finland and large interest in GRADE and GRADEpro.
- Some in NICE already use GRADE, informally they like GRADE.

- The French oncology societies are starting to review the way they grade their guidelines. There are plans of a French translation of the BMJ paper. French organisations are being reorganised, but they are positive to GRADE. They have an evidence table group.
- The Cochrane Collaboration is pilot testing Summary of Findings tables that are based on GRADE.
- ACP Journal Club/EBM produces summary tables and Paul will think about GRADE in that context.
- Polish guidelines will use the GRADE approach with GRADEpro. They are pilot testing GRADEpro with the guideline producers.
- In WHO “the experts rule and they are not overly concerned by evidence”. Tessa would be happy to introduce a GRADE member who is willing to join the drug guideline group.
- Katharine has tried to use GRADE in the WHO malaria guidelines group through the Cochrane Infectious Diseases group. There was not a very positive response to GRADE, but this was in a group that was negative to EBM.
- Merce is translating the BMJ paper. Different guideline groups in Spain are organised differently but there is guidelines cooperation group. Many groups use the SIGN approach. There has been a GRADE workshop. There is a need for a manual for understanding the process of grading. There is a need for more training. Translation is a challenge.
- James has used both GRADE and another approach simultaneously and plan to write up that experience.
- Andy has had some contact because he is the contact person on the BMJ paper. There has been some interest from Australia but the interest has seemed to dissipate. Some interest from HTA in Denmark, where he gave a workshop. A US oncology group led by Ben Djulbegovic is using GRADE. The WHO Advisory Committee on Health Research (ACHR) has established a subcommittee chaired by Andy that will likely advocate using GRADE as part of its recommendations to WHO on how to improve the use of research evidence in its recommendations. A challenge for WHO is how to make guidelines that are international or can easily be adapted to a specific setting. Many WHO recommendations focus on public health and health systems, which raise additional challenges regarding what evidence to include and how to grade this evidence and these recommendations.
- EB Guidelines does not understand GRADE, Jeff has written a 3 part paper series.
- The Vanderbilt Center has adapted GRADE.
- David A report that the PS and USPSTF and 13 EBM practice centers have asked to use GRADE domains, there is spontaneous interest in GRADE.

GRADE presentations

All of us should send our GRADE presentations to Yngve who will put them on the web. Jeff will draft a standard presentation template based on these. This will be presented at the next meeting where we will discuss this further.

Action: Jeff, ALL

The GRADEpro help file (built into GRADEpro) is a start for a manual. All should give feedback to Holger who will keep working on the manual.

Action: Holger, ALL

10. Publications, applications, funding and future meetings

Margaret has secured some funding and will host a GRADE meeting in Lyon in conjunction with the next GIN Conference. The GRADE meeting is set for the 8th & 9th of December. (8 December is the light festival in Lyon). There is currently 3 slots of 90 minutes where GRADE workshops could be offered at the GIN conference.

Action: Margaret

We agreed that it is a good idea to have a GRADE meeting at The Cochrane Colloquium in Melbourne. Holger volunteered to organise that GRADE meeting.

Action: Holger

Merce and Tessa said that it may be possible for them to organise meetings next year

Action: Merce, Tessa

Holger suggested submitting the protocol on the reliability of the 1st RCT for 80000 euro funding.

Action: Holger

11. Quality of evidence for single RCT's

David A summarised previous discussions on this topic. We agree that a single, large, well-done trial is high quality evidence, but there is concern about smaller and less replicable studies. There is uncertainty about whether the current criteria (with additional guidance) adequately address concerns about single small trials or if an additional criterion is needed.

The Ioannidis group has published several studies showing variation in effect size that is reduced with a larger number of patients, but they did not look at the quality of these studies, so we do not know which factors causes the variation. Susan has written a draft protocol to look at the frequency of false results from the 1st high quality RCT.

Holger, Pablo and Gordon have made two different protocols. One is similar to Susan and David's protocol. The other one looks in more depth, also looking for the potential explanatory reasons: what makes a first trial different? They tried to develop a sampling frame to answer these questions.

Issues raised: Random sample (for example Cochrane systematic reviews or guidelines, or comparison of those two)? Should we also consider the effect sizes for the 2nd and 3rd or each RCT compared to the overall estimate from a meta-analysis? How should prematurely stopped trials be handled? What quality criteria should be used? Multi-centre and single centre trials should be considered separately. It makes more sense to focus on the number of events than the number of patients. What difference is important? Is it only important when a difference in effect size would lead to a different recommendation? Low quality information about harms might reduce the overall quality and then a recommendation might not change, despite an important difference in effectiveness. What is the risk of a wrong or misleading answer. To what extent are the findings of the 1st RCT replicated by following trials. What are the determinants of misleading results? What is the maximum incidence of misleading 1st RCTs (i.e. misclassified as high quality evidence, when subsequent trials led to different conclusions) with which we would not worry about this being a problem? The papers by the Ioannidis can help inform the discussions regarding what is "sparse data". We could also use a simulator to help address this question.

Jane described another relevant study that compared effect sizes with quality, when the trial was done, etc. She will find the reference and pass this on to Holger.

Action: Jane

Andy emphasised the importance of keeping this study feasible and suggested starting simply by asking: How often is the 1st RCT graded as high quality high quality evidence using the current criteria?

Holger, David A, Gordon, Pablo, Roman and Gunn want to be involved in this work. Regina would like to be kept informed. Holger and David will take the lead.

Action: Holger, David

12. Guidance for balancing benefits and harms, and making recommendations.

What should be considered, and how? We need to come up with domains that should be considered for strength of recommendation. We split to 2:

1. How to begin the development of detailed guidance
2. How to formulate the recommendations, actual wording on the formulation

When formulating recommendations it is important to use standard language:

- Do we need different recommendations for different audiences?
- The population should be specified in recommendations.
- The action or intervention should be specified in recommendations.
- We should collect or develop examples that illustrate different formats for recommendations.

Current practises include:

- Standards and options, if suitable for all = standard, if not = options (used by oncology groups in France).
- Polish recommendations do not translate well according to Roman, according to Jacek they are: We suggest and we recommend.
- Should be offered, consideration should be given to, = choice rather than force (Jane), but (should be given in emergency)
- Should be or should probably be used in reference to (Katherine)
- Strongly recommend or recommend or insufficient evidence (David A)
- Have not used phrasing of recommendations in Finland
- German Cochrane extracted all the formulations and pilot tested these different phrases = surprised how different interpretations were obtained.
- Not formulated in a standardised way in Holland
- BMJ Clinical Evidence does not make recommendations

We agree that to not make a recommendation can be a problem, and that strong and weak recommendation are not good for phrasing. We are left with two suggestions:

1. Should or might
2. Suggest and recommend

How should recommendations be formulated when there are two equally effective treatments?

- ACCP recommends either A or B over nothing, then A over B or opposite

- The GRADE BMJ paper does not address this. We previously had a “Toss up” option but this was voted out. It was argued that this is not helpful to clinicians and that it was extremely rare that two treatments or interventions are exactly equal, range of outcomes, harms, effect sizes, etc.

14. Any other business

Alessandro and others are revising the QUORUM statement and want to circulate a questionnaire to the GRADE Working Group. We agreed this would be ok.

Action: Alessandro