

Draft minutes GRADE Working Group meeting, Washington DC, 10 - 11 January 2008

Participants: Alessandro Liberati, Andy Oxman, Art Sedrakyan, Benjamin Djulbegovic, Bill Lawrence, David Atkins, David Rind, Eduardo Ortiz, Elie Akl, Elise Berliner, Geert van der Heijden, Gerald Gartlehner, Gordon Guyatt, Gunn Vist, Hans de Beer, Hilda Bastian, Holger Schunemann, Ian Shemilt, Jan Brozek, Jane Sisk, Joanne Lord, John W Williams, Karen Lee, Lawrence H Schott, Luke Vale, Magali Remy-Stockinger, Mark Helfand, Massimo Brunetti, Nicola Magrini, Peter Tugwell, Philip Alderson, Philipp Dahm, Regina Kunz, Roman Jaeschke, Russel Harris, Signe Agnes Flottorp, Silvia Pregno, Stephanie Chang, Steve Pearson, Susan Norris, Suzanne Hill, Yen-Ping Chiang, Yngve Falck-Ytter.

1. Minutes from Sao Paulo

Point 2, Grading the quality of evidence when there are no events

We recapped the discussion in Sao Paulo about the situations when you have studies with many patients and no events. In SP we agreed that it is informative and to use the absolute effect and make footnotes. The chapter in the Cochrane Handbook and GRADEpro need to be updated to include this.

Action: Holger

Point 3, Guidance for reporting SMDs

We agreed that it is almost impossible to understand SMDs, three suggestions for presentation of SMD in SoF tables were attached to the agenda. None of these current suggestions have strong support and all have pitfalls. Andy suggested to keep working to further develop the presentation and to also consult with the Cochrane statisticians group.

Action: Andy

Point 9, Examples of 'good practice' guidelines

The issue is 'motherhood recommendations', some examples were attached to the agenda. Some of these guidelines are nonspecific recommendation to take history or physical examination or ethical such as "treat your patients with respect". Guideline panels include motherhood recommendations because they think it serves some purpose.

The GRADE option is either to recommend against using motherhood recommendations or to define circumstances when it is ok to do so. Where there are no alternatives, there are no choice and no point in making recommendations as they would make no difference to practice. An alternative would be to separate general statements from making recommendations like the WHO who do not apply GRADE to ethical and principal issues, but otherwise use GRADE. We still need examples of useful motherhood statements.

Action: All

Point 10, Practicalities, shortcuts and modifications

The definitions for recommendations in the minutes was confusion and needed clarification. It now reads:

"We agreed that in order to state that GRADE was used, the GRADE definitions of quality of evidence should have been used. And where recommendations are made, the GRADE definitions of strength of recommendations should have been used. Moreover, the quality

criteria that we suggest should have been used without substantive modifications, deleting or adding criteria. “

It was confirmed that the minutes from Bilbao should be altered so that for an organisation to say that they use GRADE they should ideally have produced an Evidence profile.

Point 11, Evidence profiles for observational studies

The last paragraph needs a change, Holger will send the correction

Action: Holger

2. Incorporating economic considerations

There was no comment or objections to the fifth BMJ paper on resource use.

We split into small groups to discuss suggestion for guidance we can give to people who consider using GRADE for resources use issues, including definitions of resource use.

The small group discussion was followed by plenary reporting and discussion:

The examples by Group 1

Group 1 focused on how to proceed with including resource use as outcome, and the importance of this outcome. The definition need to define the perspective and the level of aggregation of resource use. The group also discussed whether to include costs or not, in situations when resource use not reported, or when detailed table provided in addition. Costs can easily be misused or misinterpreted or as aggregate summary of really long resource use lists. When it comes to quality and reporting, costs are surrogate for resource use and therefore indirect. Information from trials where resource use is higher than usual practice is also indirect evidence.

Checklist for identifying resource consequences by Group 2

Group 2 expect a tendency to downgrade more for economic studies. They see the need for more detailed instructions. If the intervention doesn't pass through effectiveness bias, there is no point in evaluating resource use. For RCTs, the allocation issues are the same as for other outcomes. However, it is not possible to blind for resource use, is it less important?

Additionally, trials often distort resource use by paying for it all, which makes the outcome indirect. Study limitations are considered as standard for risk of bias. How to deal with the studies where consumption of drugs are reported only in methods as doses and duration as opposed to how much was actually used? That would then introduce double standard (not ITT, just measure). Is there a global component for resource use? Is there confidence in aggregate resource use? Inconsistency is also here a difficult judgement. 5 perfect trials in different places and with similar health outcomes, but with large variation in resource use, mark down for inconsistency? Subgroup or select the direct results. The group also see a challenge with global focus regarding resource use.

Quality criteria for economic evidence by Group 3

The type of evidence that will be relevant may be different because of study design differences. The group suggested considering using different study designs as high quality. The overall issues are similar to effectiveness, but suggest maybe different criteria. Group 3 also suggests aggregating into a cost number, there are often only a few resource issues that determines or drives the costs. There may be missing data for resource use, might miss gaps

in relevant resource issues and there is a need to evaluate how important they are. Each economic model has parts that go into the model and each of those parts should be evaluated.

Balance sheets (evidence profiles) versus economic models by Group 4

Group 4 thinks that it is often a waste of time to consider resource use. They suggest to rate the importance of outcomes in a two step approach considering the clinical outcomes first. Importance rated as category rather than aggregate. The group thinks the attached was a good list, although there was some disagreement about perspective and non-health care costs.

Other issues related to incorporating economic considerations by Group 5

It is always useful to have resources listed. The group liked the magnesium sulphate example stratified by severity and income level of country. The group is unsure about the usefulness of the risk of bias table and balance sheet as tools for resource use. Group 5 also think the importance should be assessed first.

Outline for JCE article on economic considerations by Group 6

Group 6 agree that same criteria could be used because anything that could bias health outcomes could bias resource use. But there may be differences with internal and external criteria. In addition there are challenges with natural units. Is the inclusion of all the contributors that drives the costs an appropriate measures? Assessment of utilisation and blinding of assessor. The group states that randomisation is good but trials are sometimes not applicable, they suggest that observational data are better than low quality.

The persons reporting from the small group discussions should send notes to Andy

Action: Elie, Mark, Yngve, John, Sue and David

Plenary discussion about Quality criteria

The level of evidence starting point for observational resource use data are (because it is observational data) Low. This relates to the extent to which we believe the trials. A suggestion was made to upgrade for the 'hard data' that we believe in more than downstream things. We already have rules for upgrading of observational data. If we are provided with examples (evidence profiles) that is convincing, we will reopen discussion and consider adding another upgrading rule. Until then, observational data starts as low quality.

Additional information from other sources may help you avoid downgrading for indirectness or sparseness. We discussed the issue of blinding, does it matter enough to downgrade the study? Should the evidence profile for resource use be presented differently?

The pros and cons of one stage versus a two stage process about importance of outcomes

Preference for initial judgement about what issues are likely to be important, the second stage judgement will be influenced by the data that you do have. It needs to be disaggregated to make sense in your own setting. Because the evidence profiles have a limited number of outcomes, and these include resources use outcomes, you might want to aggregate the resources for the EP table (secondary table). Even where there are no differences in effect, it may be useful to look at resource use if one drug is much more expensive or require additional resources.

The worry that we will be constantly downgrading the evidence because we do not really know how to model downstream costs depends on when those outcomes are critical outcomes. The decision about whether the outcome is critical is a subjective judgement.

Jane pointed out that the perspective of health system (excluding sick leave, patient time etc) should be used in BMJ paper which makes it a societal perspective. This has implications for the dimensions included. And the time horizon needs to be specified.

Action: Gordon

Plenary discussion about the outline for the detailed paper 9

Define the pathway of care. The opportunity costs could be very important, this challenge the current traditional considerations of what is important. When groups are making decisions for specific contexts, then costs are important. Time horizon is important because the flow of costs are not consistent (ex surgery) Luke will send a few examples to Andy.

Action: Luke

Framework. Health outcomes can be used as a proxy for resource use. We were warned that the composition of the guideline panel may influence what is considered to be important and that the jurisdiction of the guideline may also be an additional challenge.

Appraising the quality of evidence. There have been suggestions for change, but nothing has been brought forward that convince us to make change. To evaluate if there is an effect of lack of blinding will always be a judgement. Consistency, extrapolation and time horizon may introduce bias both ways. There was some uncertainty regarding at what level to assess consistency, aggregated data? The criteria is largely ok also for economic studies, but Sue suggests to start with directness, go on to design and then inconsistency. Larger variation is expected in resource use data.

Monetary value Discounting is for costs, not for resource use because it is relevant for things that go over time, discounting resource use can give unreliable results. If the evidence profile is to be transportable then you can not discount because different countries have different discounting rates, also need to discount then for preferences. It is not usual to discount but then if results are presented together with economic model that have been then it becomes confusing. However, outcomes long time from now should be discounting to avoid misleading results.

Balance between benefits and costs The attachment was meant as starting point for discussion. QALYs may not be useful in an evidence profile. Economic models as decision aid will always have to be explicit. The more complex the more difficult to interpret, but easy is not always easier. Luke has some papers that he will send to Andy.

Action: Luke

Affordability and equity Affordability and cost effectiveness is often the same thing, but may be different (mammography versus treating rare disease).

JCE series, it is not clear if there will be one or two per issue or a special monograph.

3. Presentation of economic information in NICE clinical guidelines

Joanne presented the internal document for NICE discussion about presentation of economic information. The guideline panels in NICE have enough economists for their key priorities. NICE want to keep the cost and cost effectiveness in the reports, and they want to have them close to the clinical information. They are proposing to make their own version of an economic evidence profile.

There are three parts, health outcomes, resource use (presented same as health outcome), and a third outside of the evidence profile with information from models. We have had agreement not to include models in the GRADE tables.

How confident are you in the estimate is not captured in the limitation section of the proposed model evaluation. It was said that it does not make sense to pool models, only good models will be presented to their panel.

Sometimes economic models are based on observational data that were not considered good enough for the health outcomes. Sometimes, Luke thinks 10-15 % of trials have economic evaluations, these are often the bigger and better health outcomes trials.

In comparison with GRADE – the definitions in the paper is not consistent with GRADE.

4. BMJ GRADE series – update

Gordon informs that all five papers are accepted for publication in the BMJ and should be coming out very soon (within 8 weeks). But he will ask for a few changes to paper 5 based on yesterday's discussions.

5. JCE GRADE series

Recap from previous discussion in Sao Paulo, the first article will have the GRADE Working Group in the by-line, the following articles with the contributing authors in the by-line and with a low threshold for being included as an author.

The list of papers and outline will be circulated again. Phil thinks this series will be a key resource for implementing GRADE, and do not want to wait. It was decided early that a limited amount of details of quality assessments should be presented in this series because we can refer to others' empirical data and Cochrane handbook.

- **Paper 4. Risk of bias**

Publication bias has been split into outcome reporting and lack of publication of study. Selective withholding of information about negative outcomes is the challenge. Some people play with methods of reporting outcomes to present favourably. Difference in reporting between trials may help identify this.

- **Paper 6. Consistency and directness – revised**

The paragraph about Direction of effect on page 4 is confusion and it is suggested to change it. Change figure legend for figure 3.

- **Paper 7. Upgrading and summarising quality and SoF tables**

This paper will be split into two papers.

Group 1. Large effects are not always real examples benefit as the relief of pain from placebo is an example of. There is confusion about the rule of thumb about large and very large effect and the paper refers to another limit – we should be clear that this is only suggestions. It may be a problem for rare events downstream but may be an idea to consider how many events would be enough to switch the decision. We want to warn people to be careful about upgrading. We still need more examples for upgrading.

There is also confusion regarding strong association relates to correlation rather than size of effect. Make it more clear the issues about critical outcomes, critical versus important and not critical outcomes, rephrasing to use important. For the chapter about how to choose the outcomes: For new drugs and devices, think routinely about adverse events even though you probably do have any events. Mark and Andy has several editing suggestions that they will send to Gordon. And Andy will make new figure. Suggest title change for “Exception to the rule” to “Does it still make sense?”

Action: Andy and Mark

Presentation issues:

Group 2, 3, 4, 5 and 6. The content of the Summary of Findings (SoF) table and how it differs from the evidence profiles (EP) should be made explicit, and they should co-exist. We may wish to keep flexibility in EP and SoF tables for presentation, like including publication bias. Then we need to have several versions available. An absolute measure of the difference would be helpful. We need example with observational data, only including the number of patients. We talked about possibilities for user testing in NICE.

Absolute effect was taken out of the SoF tables because of the need to minimise, and user testing showed that people misinterpreted the absolute risks. However, some feels that presenting it stops people from miscalculating. Andy still thinks it is important to minimize the number of columns. The expectation peoples have will influence their interpretation.

Confidence Interval (CI) of the frequency around the estimate, and CI around the NNTs as well (NNT to NNH or x fewer to y more when necessary). It was considered not optimal for guideline panels but useful for clinicians when there are few outcomes. A suggestion to not allow NNT when results are not significant was put forward, but that would focus the presentation on significance. Andy suggest small working group to make examples and test it.

Action: ?

Criteria for assessing quality, the group wished for them to be standard with easy flexibility to make modifications to include publication bias, strong associations, dose response and plausible confounders where it is used. Holger assured us this should not be a problem to share and change while exchanging xml files of GRADEpro.

We discussed about options of reordering the columns in the SoF table. We were reminded that user testing and journalist logic has determined the order of presentation.

It was suggested to get rid of the words “illustrative and corresponding” in the headings in the SoF table. But others thought that illustrative is the most correct word for the use. We could have an option to leave out or change. Cochrane statisticians suggests median rather than mean as is explained in the Cochrane Handbook. And it was suggested from the group that even with 0 event it would be beneficial to see estimate of baseline risk and confidence interval.

We agreed to continue this debate at a later meeting. Please send good ideas

Action: All

Modifications of GRADE by AHRQ

Overall, based on Mark's presentation, we felt that the proposed EPC approach to grading the quality of evidence was quite similar to the GRADE approach. This is great, but it is confusing and unfortunate for end users that they will have yet another grading system to cope with.

There may be good reasons for making minor modifications and using different terminology, and we would value the opportunity to discuss these. However, there is a risk that the use of different terminology and minor modifications by different groups will confuse people. The GRADE Working Group provides a forum where organisations like the EPCs can discuss and try to harmonise the terminology that they use and reduce the risk of minor differences adding to the confusion and dismay that there is over the growing number of grading systems being used. We would welcome discussion with the EPCs about these differences and encourage anyone who is interested to join the working group.

In addition to potentially causing confusion, some minor modifications could turn out to be substantive differences (creating even further confusion). Examples of what appear to be minor modifications that we discussed include:

- It was not clear exactly how quality of evidence or different levels of quality are defined by the EPCs, but it appeared that they were essentially the same as what GRADE uses. What GRADE calls 'very low' quality is called 'insufficient' by the EPCs.
- Applicability is split off from what GRADE calls "directness" and is addressed in a separate chapter. This could be a substantive difference and a problem if there is not a transparent way of integrating this into judgements about the quality of evidence when it is relevant; e.g. if studies use a comparator that is not applicable for the question being addressed.
- Making some criterion "optional" is likely to generate some confusion, although it appears that this is not a substantive difference, given how optional is defined by the EPCs.
- A global assessment of the overall quality of evidence across criteria is made rather than using categories for each criterion (e.g. downgrading by one category). We acknowledge the problems with using categories and address this explicitly in the detailed guidance that we are developing. However, using categories forces groups to be explicit about their judgements. It is not clear how this will work in the EPC approach.

In addition to reducing unnecessary confusion due to differences in terminology and minor modifications, there is a huge advantage to collaborating and having some degree of consistency across grading systems for those responsible for applying them. We can reduce unnecessary redundancy and improve our work by jointly developing or sharing tools, training materials and training. It also makes it easier to use evidence tables prepared by others, and we can all improve our work by sharing experience and collaborating on methodological research.

We would welcome the opportunity to discuss more substantive differences between the proposed EPC and GRADE approaches. This would help GRADE and all the organisations involved in the working group and using the GRADE approach to make modifications, if there are compelling reasons for them. The EPCs might also benefit from discussion of these differences and the collective experience of others in the working group.

Based on the discussion we had we identified two substantive differences.

1) In the EPC system observational studies can start at moderate.

As we understand it, there are no clear criteria, examples or guidance for when observational studies would be considered moderate. While there may be compelling reasons for this difference, introducing it in this way seems to conflict with what we understand to be key goals for the EPC system - predictability and reliability (as do some of the other differences). The working group has discussed the possibility of some observational studies starting at moderate numerous times in the past and we continue to have discussions about this. Up to now we have not identified compelling arguments or examples for suggesting that some specific types of observational studies provide moderate quality evidence. On the other hand, no approach is likely to address every possible reason for upgrading or downgrading. There may be compelling reasons in specific circumstances to upgrade (or downgrade) the quality of evidence, which are not covered by the GRADE approach. We suggest that this should be done explicitly with a clear explanation of the rationale.

2) Coherence is an additional criterion in the EPC system.

There was some confusion about how coherence is defined by the EPCs, but as it was explained to us, it is largely a judgement about how compelling the evidence is for a subgroup analysis. This would be a helpful addition to the detailed guidance that we are developing for applying GRADE, but it would not be a reason for upgrading the quality of evidence once the decision to base a conclusion or recommendation on a subgroup analysis was taken.

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

We agreed to delay this discussion until the meeting in Rome in May when Elie & Holger will have more results to present.

7. GRADE Profiler

Everyone please send feedback.

8. Cochrane Summary of Findings tables

There was nothing new to report since the meeting in Sao Paulo.

9. Database of evidence profiles

This will be an open, free, easy access, trusted (then we need quality checks), secure, flexible database that will accommodate collaboration. This project is in planning and early pilot phase.

10. GRADE website

There have been 11361 unique visitors to the GRADE Working Group web pages, from 123 different countries. The web pages are moving from static to data driven web pages which will result in better membership management, database project management, etc-

11. Publications, workshops, applications

The demand for Workshops is increasing, We plan to collect more teaching material on the web pages. Suggestion was made to focus on training the trainers.

- Holger reported on an asthma guideline together with Jan. He has written a GRADE methods paper. The production of 45-50 evidence profiles for a guideline. Holger has also written Thoracic paper on diagnostics. Holger and Jan have written a series of papers about allergic disease. They have a workshop approved by the Canadian Cochrane network where Nancy has an abstract. They will hold a workshop in Vienna,

and one in Rome following the GRADE Working Group meeting in May focusing on GRADEpro.

- Joanne reports that NICE will need training, would like train the trainers help.
- Yngve and Regina held a workshop in Freiburg. Yngve and Regina has written a paper for the Journal of Gastroenterology.
- Geert is trying to organise a Dutch workshop in September
- Signe reports that GIN would like GRADE people at their meetings and workshops, and she reports on interest in Norway for guideline advice
- Gerald reports that in Austria the Ministry of Health is planning to evaluate new devices using GRADE.
- David R reports that all editors in Uptodate have to learn the ACCP modification of GRADE
- Eduardo reported that the National Institute of Health USA are interested in more workshops
- Regina has written a paper for the Haematology Journal in Germany.
- Nicola reported on a workshop with Sue for the Italian national guideline who has accepted to use GRADE. He keeps using GRADE for recommendations.
- Sue reported that they have workshop waiting lists for GRADE and that they want more. Sue is working on the WHO handbook, she finds a challenge for policy interventions and observational data that is not synthesized well. WHO are building network of Centres who can produce GRADE evidence profiles at short notice.
- Hilda is producing patient information evidence profiles and developing data visualisation tools, they are hiring someone to produce the profiles and setting out a tender for people to produce GRADE evidence profiles.
- Roman reports that the Canadian Critical Care groups 12 point system has accepted to use GRADE for the future. Roman finds it challenging that GRADEpro stops with evidence profile/SoF table, he would have liked another 3 buttons to move to recommendations
- David A will collect examples and will change methods chapter at AHRQ. David will also look into funding to train people to produce evidence profiles. Training grants in comparative effectiveness reviews.
- Elie reported about a trial with Gordon and Holger on alternative ways of phrasing recommendations.
- Andy has given workshops to the UK Cochrane network and to SBU (Swedish HTA)

Questions for further consideration:

- Should we seek funds to develop a training module on GRADE?
- Should we do a study to compare the guideline results for GRADE vs a less vigorous approach? Should we compare costs of different approaches?

12. Future meetings

- May 6 & 7 2008 in Rome.

Focus for this meeting will be diagnostics. It is planned to circulate diagnostic examples beforehand. Diagnostic accuracy studies will be discussed in small groups. Survey of different versions probably after Rome. Wording of recommendation.

Holger will decide if the meeting will be 1,5 or 2 days.

- October 2008, Freiburg (8 October)