

DRAFT minutes GRADE working group meeting

Krakow, April 1-2, 2003

Present: Maigorzata Bala, Jan Brozek, Yngve Falck-Ytter, Signe Flottorp, Piotr Gajewski, Gordon Guyatt, Robin Harbor, Margaret Haugh, Roman Jaeschke, Wiestaw Latuszek-Lukasiewicz, Wiktoria Lesniak, Jacek Mrukowicz, Andy Oxman, Holger Schunemann, Helena Varonen, Gunn Vist.

1. Piotr Gajewski welcomed us on behalf of the Polish Institute for EBM. Jacek Mrukowicz informed us that syphilis did not originate in ‘the new world’ but had been present in Poland prior to Christopher Columbus’ return to the old world. Typical lesions identified by Polish dermatologists on wooden statues in Krakow’s main church provide high quality evidence of this.
2. The **minutes from Birmingham** were approved without further comments.
3. **Guidelines for grading paper:** We went through the manuscript page by page. A number of changes were discussed and agreed. These will be incorporated in the next version of the manuscript. Especially the newcomers felt that they wanted more explanation and examples. Jacek pointed out that the system first comes alive when examples are discussed. He suggested including a summary of judgements like the ones Gunn has prepared for the past two meetings.

It was agreed that there should be two articles. The current manuscript, which we previously agreed should be submitted to BMJ, and a more detailed technical document that addresses all of the important nuances we have discussed, with clear instructions, and many examples, targeted at people interested in applying our system. It was agreed that the technical document should also include more detailed explanations of concepts and a glossary. Andy suggested that this should perhaps be electronic (including printable documents). The current manuscript should not attempt to address all of the nuances that we have discussed. This should be made clear in the introduction.

Andy will circulate a revised manuscript.

Action: Andy

4. **Observational studies:** It was felt that the revised manuscript adequately reflected the comments by Peter Briss. It was agreed to add two of the examples of evidence of harm from observational studies provided by David Atkins. We discussed Peter’s suggestion to split observational studies into subcategories and decided not to do this.

Action: Andy

5. **Costs** (resource utilisation/ allocation): Robin and Gord volunteered to prepare a draft of guidelines for taking account of costs in grading recommendations, together with James Mason and Tessa Tan-Torres, who had expressed an interest in this previously. They will attempt to circulate a draft prior to our next meeting.

Action: Robin, James, Gord, Tessa

Andy suggested doing an exercise at the next meeting focusing on taking into account costs when making a recommendation. We would consider one or more examples with high and low costs paid for either out of pocket or by a third party, specifying the income (for out of pocket costs) or country for third party payments.

Action: Andy, Gunn

6. **Equity:** Tony Dans, Peter Tugwell and others have expressed an interest in incorporating guidance for taking into account equity when grading. Gord pointed out that we already specify populations in our approach and that this could easily accommodate considerations of equity. Andy noted that issues may need to be addressed elsewhere in the guidelines development process and not, or not only, in grading evidence and recommendations. Andy, Signe and Gunn volunteered to follow-up on this with Tony Dans, Peter Tugwell et al. They will attempt to prepare draft guidelines to discuss at our next meeting.

Action: Tony, Peter, Signe, Andy, Gunn

7. **Diagnostic tests:** It was agreed that additional guidance is needed to take account of evidence of diagnostic test accuracy, when this is relevant to a recommendation. This was illustrated in two of the examples that we considered where it was agreed that the quality of studies should not be scored down for being observational (cross-sectional studies are appropriate), but that in some cases it may be appropriate to score down the directness of the evidence due to the fact that test accuracy is a surrogate outcome (e.g. for the consequences of false positive tests) with some or major uncertainty. Gord volunteered to prepare a draft and circulate this prior to the next meeting.

Action: Gord

8. **Evidence summaries, balance sheets and judgements**

The evidence summaries and balance sheets need to be updated so that they contain all of the information agreed upon at the past two meetings. Authors of the reviews upon which the summaries are based will be contacted to obtain the additional information needed to complete the tables. This includes information about any scales that were used for continuous outcome measures and the direction of results. Important inconsistencies in results will be examined and explained in terms of statistical heterogeneity and clinical importance.

Action: Gunn

Helena pointed out that “intermediate” quality is confusing, since it is not clear what it is between and suggested “moderate” as an alternative term. We agreed to make this change.

Action: Gunn, Andy

Several who had not participated in earlier discussions questioned including “maybe do it” as a recommendation. After further discussion we agreed not to include “maybe do it / toss up” as a recommendation and to replace this with the option of no recommendation.

Action: Gunn, Andy

We agreed to replace question #6 in the judgement form, regarding the quality of the tables, with an open question about suggested changes.

Action: Gunn

We discussed sparse data several times, again, when working through the examples. We were not able to reach agreement about a definition of what constitutes sparse data. Several people routinely judged one study with relatively few patients as sparse data, even when the results were statistically significant. Others argued that if the results are statistically significant, this cannot be considered sparse data. We reconfirmed that it may not be possible to reach agreement on a definition of sparse data and that the key message is that reviewers and guideline developers should make these judgements transparent. We agreed that if the concern is about publication bias, this should be indicated, rather than arguing that there is sparse data. Similarly, if the concern is about the quality of the study or the generalizability (directness) of the results, this should be indicated, rather than arguing that there is sparse data.

A lack of information about harms is an important source of uncertainty and disagreement when judging the balance of benefits and harms and making recommendations. This led to a discussion about risk aversion and we agreed to include this in a new box in the article. It was also agreed to include plausible harms in the tables when no information is available (as was done for adverse effects of Chlamydia screening) and to grade the quality of evidence in such cases as “no evidence”.

Action: Gunn, Andy

Many of the examples have reported results as weighted mean differences or standardised mean differences, which most of us find uninformative. Gord has a way of converting these into NNT's and he offered to help with this.

Action: Gunn, Gord

We discussed the decision not to grade down the overall quality of evidence when the main outcomes go in the same direction and one outcome is of lower quality, but in the same direction. It was agreed that this should only apply when where there is significant benefit in one or more of the other (higher quality) outcomes and it is very unlikely that the lower quality outcome goes in the opposite direction. This should be clarified or dropped in the article, and clarified in the technical document.

Action: Andy

We agreed that the accuracy of screening as an outcome for a screening test should not be downgraded because of study design. The accuracy of screening can provide information about the risk of false positives and false negatives, as an indirect outcome measure for the consequences of this. It will sometimes, therefore, be appropriate to include test accuracy in the evidence summaries and balance sheets for screening tests.

Examples:

8.1 Should patients with acute low back pain be treated with NSAIDs?

There was good agreement for this example, although the information provided generated some uncertainty about the results and their direction. Gunn will contact authors of the review to get more information about the scales that were used and confirm the direction of results. It needs to be made clear that for pain intensity there was statistically significant heterogeneity that could not be explained. Gord suggested that in this and other examples the alternative should be more clearly specified. Bob in his comments raised concern about rare serious adverse effects from NSAIDs and wondered to what extent these were adequately addressed. Others raised uncertainty about the severity of the side effects. This was not discussed, but should be considered in revising this example.

Action: Gunn

Consensus:

Pain intensity	Moderate quality	Critical
Global improvement	High quality	Critical
Additional analgesic use	High quality	Important/Critical (no consensus)
Side effects	High quality	Critical
Overall quality	Moderate	
Trade-offs	Net benefits	
Recommendation	Do it	

8.2 Should patients with non-specific low back pain be treated with acupuncture?

Because there was only one study with a small number of patients, many considered this sparse data and because of this there was some degree of disagreement about the quality of evidence for all four outcomes. Gord argued that there were statistically significant results that are informative because they at least rule out a harmful effect. Margaret raised concerns about the quality of global improvement and functional status as outcome measures. It was agreed that the note about sham studies should be removed because those studies address a different question. There was no information about harms. Because of this it was agreed that there were uncertain net benefits. There was disagreement about the recommendation, although there might have been consensus to recommend don't do it after considering costs.

Consensus:

Pain intensity	Moderate quality*	Critical
Global improvement	Moderate quality*	Critical
Functional status	Moderate/High quality* (no consensus)	Critical
Return to work	Moderate quality*	Critical
Overall quality	Moderate/Low* (no consensus)	
Trade-offs	Uncertain net benefits	
Recommendation	Probably do it (ignoring plausible harm)/ Probably don't do it (considering uncertainty + harms) (no consensus)	

8.3 Should women be screened for chlamydia infection?

The outcome chlamydial infection is a marker for risk of sterility and hence indirect and it was agreed that there was not a strong association. The importance of asymptomatic chlamydial infection depends on why it is assessed, however, it is a public health interest to stop the spread and after discussion we agreed that

this is a critical outcome. The adverse effects are not known, but critical because the intervention is being offered to otherwise healthy women.

Consensus:

Pelvic inflammatory disease	High quality	Critical
Chlamydia infection	Very low quality	Critical
Adverse effects	No evidence	Critical
Overall quality	Very low	
Trade-offs	Uncertain net benefits	
Recommendation	“Probably do it” (i.e. should be considered/offered)	

8.4 Should pregnant women at high risk of preterm delivery be screened for bacterial vaginosis?

There was uncertainty regarding the extent to which the question and the evidence was for asymptomatic or symptomatic women?

It took some time to clarify that the trials were of treatment and not screening?

Robin and Gord questioned whether test accuracy should be in the table at all. We agreed that test accuracy could give useful information about the probability of false negative and false positive results.

It was agreed that this example needs to be redone before we can use it.

Action: Gunn

8.5 Should the general population be screened for melanoma?

We agreed there was major (not some) uncertainty about the directness of the evidence for test accuracy (related to the intervention, the patients and the outcome measure). The majority agreed on “don’t do it” as a recommendation because of uncertain benefits and potential (unknown) harms being offered to healthy people.

Consensus:

Accuracy of screening	Very low quality	Critical
Lethal melanoma	Very low quality	Critical
Adverse effects	No evidence	
Overall quality	Very low	
Trade-offs	Uncertain net benefits	
Recommendation	Don’t do it/Probably do it (Holger and Roman)	

8.6 Should patients with low back pain be treated with massage?

There was unresolved disagreement about whether there was sparse data and whether there were net benefits or uncertain net benefits. Even so, there was agreement about a recommendation to probably do it. Some of the disagreement about the trade-offs related to how much we thought a massage is enjoyable, as well as to how concerned we were about there only being one small study.

Consensus:

Pain intensity short term	High/Moderate quality	Critical
Pain intensity long term	High/Moderate quality	Critical
Pain quality short term	High/Moderate quality	?
Pain quality long term	High/Moderate quality	?

	(uncertainty about what pain quality means)	
Function short term	High/Moderate quality	Critical
Function long term	High/Moderate quality	Critical
Overall quality	High/Moderate	
Trade-offs	Net benefits/Uncertain net benefits	
Recommendation	Probably do it	

8.7 Should patients with chronic low back pain be treated with transcutaneous electrical nerve stimulation (TENS)?

We disagreed about whether there was sparse data for patient satisfaction. Gord argued that patients are always satisfied and that when there are no other significant outcomes patient satisfaction is not critical. We discussed whether the outcomes go in the same direction and, therefore, the lowest quality would not apply. Gord suggested to only use the “same direction” rule for overall quality when there are statistically significant results, as was not the case for this example. There was disagreement about the recommendation, in part because some perceived TENS to be inconvenient and others perceived it to be pleasant. It was argued that it may help and there are no plausible harms. We agreed to recommend to probably do it, if you consider it a pleasant experience, otherwise not.

Consensus:

Quality of pain	High quality	Critical
Function	Moderate quality	Critical
Patient satisfaction	High/Moderate quality	Important/Critical
Overall quality	Moderate	
Trade-offs	Uncertain net benefits	
Recommendation	Probably do it	

8.8 Should patients with low back pain be treated with injection therapy?

We briefly discussed this example. It was agreed that we need to distinguish between subacute and chronic low back pain, and to clarify the number of adverse events, the outcomes and direction of results. Andy noted that for short term pain relief there was statistically significant heterogeneity with results going in opposite directions. Gunn will redo this example.

Action: Gunn

8.9 Should patients with acute low back pain be advised to stay active?

We briefly discussed this example and agreed to try to merge similar outcomes. There are few and small studies. There was again disagreement about whether this was sparse data? Helena argued that when the outcomes are similar and in same direction it is not necessary to downgrade for sparse data.

Action: Gunn

8.10 Should patients with non-specific low back pain be treated with exercise therapy?

We briefly discussed this example. It included several studies measuring the same outcome, but with one study for each specific outcome measure. There was again disagreement about sparse data. Gunn will try to merge studies/outcomes, if this is possible. There was a lot of missing information in the balance sheet. More

information is needed. Andy argued that there is a possibility of harm and thereby convinced Gord and Roman that the recommendation should be probably don't do it or not to make a recommendation.

Action: Gunn

8.11 Should distribution of child safety seats and education programs be recommended?

It was quickly agreed that the question had to be changed. We previously agreed not to include an outcome in the question. We worked through this example with the assumption that child safety seats are effective in preventing death and injuries. Given this assumption, the outcomes were considered direct evidence rather than indirect. We agreed that if randomised trials without serious flaws were available, the observational studies should not be included, unless there was a clear reason for doing so. We agreed that since the outcomes go in the same direction and we assumed there was good evidence for child safety seats effectively reducing injuries, the overall quality of evidence was high. We ignored the downside of fighting with crying and kicking kids to get them into the seats as a possible adverse effect.

Consensus:

All fatal & nonfatal injuries	Very low quality	Critical
Correct use early	High quality	Critical
Correct use at follow-up	High quality	Critical
Possession of a seat	High quality	Critical
Overall quality	High quality	
Trade-offs	Net benefit	
Recommendation	Do it	

8.12 Should Hormone Replacement Therapy (HRT) be given to healthy asymptomatic peri/post menopausal women to prevent chronic diseases?

Gord argued that loss to follow up in the WHI study was a serious flaw, but this was ignored for this example. Everyone agreed that all outcomes were High quality and Critical. The results show benefit for colorectal cancer and hip fractures, and harm for breast cancer, CHD and stroke. No consensus was reached regarding the balance of benefits and harm. However, we agreed that the recommendation should be don't do it. It was agreed that ideally this example should be based on systematic reviews for each outcome (based on randomised trials only). Andy suggested redoing this example based on systematic reviews done prior to the trials, based on the single study (as we have already done) and based on systematic reviews, perhaps for two different questions: should women be started on HRT? and should women already taking HRT be told to stop?

Consensus:

CHD	High quality	Critical
Breast cancer	High quality	Critical
Stroke	High quality	Critical
Colorectal cancer	High quality	Critical
Endometrial cancer	High quality	Critical
Hip fracture	High quality	Critical
Death due to other causes	High quality	Critical

Overall quality	High
Trade-offs	Trade offs/No net benefits (no consensus)
Recommendation	Probably don't do it/Don't do it (no consensus, although the majority agreed with Don't do it)

Gord questioned whether the balance of benefits & harms step is necessary. Roman said that it was helpful as a verbalisation of values. Margaret, who said that it formalises the process and is very useful, supported this. Andy pointed out that the same trade-offs sometimes resulted in different recommendations, even prior to considering costs. He argued that it is also likely to be useful to consider the trade-offs between health benefits and harms (does it do more good than harm) prior to considering costs (are the net benefits worth the costs). We sometimes have disagreed about the trade-offs but ended up agreeing about the recommendations. We ended up agreeing to keep this as an explicit judgment prior to making a recommendation, but not including it in the manuscript.

9. **Comparison of NCI, GRADE and USPST approaches.** There was full support for encouraging David and/or others to go ahead and undertake this comparison. Helena would like to do a similar comparison with the Finish guidelines system and will contact David about doing this collaboratively with the US group.
Action: David, Helena

10. **Reliability and sensibility study.** Andy suggested exploring undertaking this study at conferences, perhaps one in the US and one in Europe. He will explore possibilities of funding for this from AHRQ with David in the US and look into possibilities for funding for a European conference. Margaret suggested the European Science Foundation as a potential funder. Marageret will also consider integrated an evaluation of GRADE into a EU 6th Framework proposal she is preparing for cancer guidelines developers.
Action: Andy, David, Margaret

11. **Presentation of grades.** Holger reported that the manuscript about presenting grades is being revised for publication in CMAJ.

WHO is beginning to use the GRADE grading system and is planning a workshop possibly in September. Other groups are also beginning to, or considering, using the GRADE grading system. When these groups start to make recommendations based on this system, it will be important that we have agreed how to present GRADE grades. We discussed the words that we have been using and agreed to continue to use High, Moderate (instead of Intermediate), Low, and Very low for the quality of evidence. We also agreed to continue to use Do it, Probably do it, Probably don't do it, Don't do it and to replace Maybe do it with no recommendation.

We discussed alternative ways of presenting grades. The majority present (9) preferred using letters and numbers. A minority (3) was strongly opposed to this, given the wide confusion about what letters and numbers mean, because they are used differently by different systems currently in use. We agreed that the presentation should be intuitive and easy to remember, after it is first explained,

and easy to verbalise. It was also important to use something that could easily be typeset and printed without taking a lot of space. We largely agreed that we should avoid undesirable associations, although the majority still preferred using numbers and letters. Holger and Yngve presented a symbol using modified traffic lights, developed over wine and dinner the night before. Andy and they thought this was brilliant, but no one else liked it. In the end we agreed on using two arrows pointing up or down, one or two of which could be filled depending on whether the recommendation was strong or weak, to represent trade-offs; and four circles or other symbols that could be filled or empty to present quality of evidence; for example:

⊕⊕⊕⊕	High quality evidence
⊕⊕⊕○	Moderate quality evidence
⊕⊕○○	Low quality evidence
⊕○○○	Very low quality evidence
↑↑	Do it
↑?	Probably do it
↓?	Probably don't do it
↓↓	Don't do it

However, following the meeting, Gord accused Andy of being undemocratic and called for a new vote, which Andy agreed to.

Action: Gord, Andy

12. **Future meetings:** We agreed to meet at the Cochrane Colloquium in Barcelona on either Thursday, October 30 or Friday October 31, with the preference being for Thursday. Gunn and Andy will make arrangements for this meeting. Financial support is not available for this meeting.

Action: Andy, Gunn

Andy will find out from Martin if NICE will support another meeting at their annual conference, the first week in December.

Action: Andy