Stream 3: summary of findings tables and GRADE assessments for network meta-analysis

This document provides a summary of discussion topics for the Stream 3 meeting which was held on 18-19 July 2013. The aim of Stream 3 is to produce recommendations and practical guidance for Cochrane review authors on producing Summary of Findings tables (including GRADE assessments) as applied to reviews of multiple interventions. We addressed issues in applying GRADE to estimates of treatment effect from a network meta-analysis (Sessions 2 to 5), followed by issues in applying GRADE ideas to the ranking of treatments from a network (Session 6), and finally ideas for structuring Summary of Findings tables (Session 7).

Indirectness

We started with a star-shaped network with six treatments, in which each treatment has been compared with a common reference treatment (A) but not with each other. In this scenario, the participants of the meeting were in broad agreement that there should be more caution relative to direct evidence (for example, because studies of one comparison may older). However, it was agreed that we should not recommend downgrading for indirectness *by default*. Meeting participants thought it essential to consider carefully the assumption transitivity (also known as consistency) to inform assessment of indirectness. The consensus view was to consider the characteristics of studies rather than statistical output. The relationship between indirectness and transitivity pertains to clinical heterogeneity/inconsistency, but this cannot be judged in a star network.

Inconsistency (versus heterogeneity)

We then discussed a network in which there are multiple sources of evidence for the BC treatment comparison so that network inconsistency (also known as incoherence) is possible. In standard GRADE, the component "INCONSISTENCY refers to variability in the magnitude of effects across studies for a specific comparison that remains unexplained after exploration of such variability (i.e. with subgroup or sensitivity analyses). This variability is more commonly known as heterogeneity. However, meeting participants thought it important to keep terminology as similar as possible to the current GRADE/ Summary of Findings approach. This should avoid confusion and therefore have benefits both for training and in adoption of ideas by those already experienced in applying GRADE within Cochrane reviews. Meeting participants agreed that heterogeneity and inconsistency should be considered together to form a judgment to downgrade but that it is important to distinguish the two concepts and the way they are statistically examined. Meeting participants did not provide solutions to exactly how we should consider heterogeneity and inconsistency together in an assessment of GRADE. However, two presentations (Salanti and Dumville) did provide some thoughts. Participants considered it inappropriate to use statistical thresholds (like the I^2) for

looking at statistical inconsistency. We also discussed the implications of assuming a common between trial variance parameter (tau-squared) in a NMA for the assessment of the inconsistency domain and the implications for GRADE of using an inconsistency model for the NMA. Full solutions to these questions were not reached, and these might be considered as areas for further research.

The meeting participants discussed the need to consider the clinical properties of trials and networks as separate from measures of statistical inconsistency. Meeting participants cautioned that whilst one may wish to downgrade for indirectness due to clinical heterogeneity or for precision given statistical heterogeneity, it was important to avoid downgrading twice for the same underlying problem.

Bias

In standard GRADE, risk of bias (RoB; also called limitations in design and implementation) and publication bias are two of the five domains. In NMA this is an ongoing area of research; several potential directions are still being developed in terms of how RoB assessments are to be incorporated into assessing results from NMA. An important general point was that we need to acknowledge that while we consider RoB of individual studies, we summarize at the comparison level. The implications of this in NMA need to be considered. It is important to be able to "track down" which individual studies were at higher risk of bias. Some participants believed it is highly important to show the impact/contribution of each study while others thought less so.

Meeting participants agreed that assessment of the RoB in each included study is the same for network meta-analysis as it is for a standard meta-analysis – it is done at study level. There was agreement to evaluate trial-level RoB first. In a NMA there may be an increased opportunity to explore the impact of RoB via meta-regression although this requires more data than is often available. There was also agreement that a quantitative approach to incorporating RoB was preferable to a visual "nearest neighbour" approach. Restricting to the highest quality studies alone was considered a moot point, as one way or the other the findings will be downgraded (either through imprecision or study limitations).

Imprecision

In standard GRADE the imprecision of effect estimates can be addressed by interpreting confidence intervals in relation to acceptable thresholds for treatment effects and using the optimal sample size. In the meeting, participants discussed whether precision and quality should be considered together or separately in GRADE for NMA. Participants were generally uncertain whether sample size should be used as a measure of precision in a network meta-analysis and did not make a recommendation. On the issue of whether precision should be considered separately from 'quality of the body of evidence'? It was noted that judgment of precision cannot be considered completely independently from other domains. Since GRADE is grading the confidence in the strength of evidence, it is almost impossible to separate the statistical precision from the quality of evidence. Once labelled as high quality, people may take the estimate at face value. But we should caution the users about the level of confidence that can be placed in the evidence, for example, with the presence of inconsistency and publication bias.

Assessing the quality of evidence for a treatment ranking

A key benefit of network meta-analysis is the ability to rank the treatments according to the probability it is the best, or worst, on a given outcome. Issues discussed in the meeting included the stability of ranking probabilities and how to evaluate the five GRADE items for treatment ranking. For the latter, the participants concluded that the uncertainty in treatment ranking is related to the imprecision of relative treatment effects. It was considered highly important to display the uncertainty of the ranks, for example looking at the distributions of ranking probabilities to summarize the overlap of two distributions showing the confidence intervals of the SUCRAS can also illustrate this uncertainty.

Participants felt it might be helpful for decision making to evaluate the confidence in ranking only for the 'top' treatments. However, the evidence base and the analysis should include all the available treatments. Participants concluded that GRADE for treatment ranking should be included in SoF tables as it might be the most interesting for the reader. For the four GRADE outcomes other than imprecision, the judgment can be derived similarly as for the relative treatment effects but considering the entire network (suggestion in the paper by Salanti et al.). Reporting of treatment ranking might not be mandatory in Cochrane reviews but, if treatment ranking is reported, GRADE assessment for treatment ranking should be reported as well.

Summary of Findings tables

In a Cochrane intervention review, Summary of Findings (SoF) tables present the main findings of a review in a transparent and simple form. They provide key information concerning the quality of the evidence (GRADE), the magnitude of treatment effects and the sum of the available data on the main outcome.

With specific regard to NMA, participants discussed how findings from a network meta-analysis should be presented together, implying a separation of results by outcome with all comparisons being presented together. If this approach were to be taken for SoF tables, then this is contrary to the current guidance for SoF tables in intervention reviews. Of course, SoF tables need not match exactly the analyses performed during the review. It may sometimes be possible to identify the key comparison for the review's user on the basis of the results of the network meta-analysis (e.g. the most effective treatment compared with the most commonly used treatment). This might not always be possible or desirable, however, and new guidance is required on how SoF tables might be constructed after a network meta-analysis. Two possible structures put forward by the GRADE Working Group are included in Tables 1 and 2. These tables preserve familiarity with existing formats, which would speed adoption in Cochrane reviews. Meeting participants discussed these options and agreed on a need for further user testing.

Table 1: Possible Summary of Findings table structure for network meta-analysis – format A (from GRADE Working Group)

Which oral anticoagulant should be used in patients with atrial fibrillation

Patient or population: patients with atrial fibrillation
Settings: primary care, community, outpatient
Interventions: Dabigatran, Rivaroxaban, and Apixaban
Comparison: Warfarin

Outcomes	Effects and confidence in the effects. Main comparator is warfarin unless specifically mentioned							Comments
	Dabigatran (150mg)		Rivaroxaban		Apixaban		Warfarin	
Death	**OR 0.89** (0.78 to 1.01)	**11 fewer per 1000 (22 fewer to 1 more)**	**OR 0.93** (0.83 to 1.04)	**7 fewer per 1000 (17 fewer to 4 more)**	**OR 0.90** (0.80 to 1.00)	**11 fewer per 1000 (22 fewer to 1 more)**	**100 per 1000 (10%)**	**None/xy of the differences between the new anticoagulants were of important magnitude. Therefore, we did [not] include these comparisons in the table.**
					OR 1.04 (0.89 to 1.23) compared to Rivaroxaban	**1 more per 1000 (4 fewer to 6 more)** compared with Rivaroxaban with moderate confidence in estimate)		
	⊕⊕⊕◯ **moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 14143 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 18201 patients (1 study)			
Stroke or systemic embolism	**OR 0.65** (0.52 to 0.81)	**11 fewer per 1000 (6 fewer to 14 fewer)**	**OR 0.88** (0.74 to 1.04)	**3 fewer per 1000 (1 more to 7 fewer)**	**OR 0.80** (0.66 to 0.95)	**6 fewer per 1000 (1 fewer to 16 fewer)**	**30 per 1000 (3%)**	
	OR 1.35 (1.03,1.79) compared to Rivaroxaban	**13 more per 1000 (1 more to 30 more)** compared with Rivaroxaban with moderate confidence in estimate			**OR 1.11** (0.87 to 1.42) compared to Rivaroxaban	**3 more per 1000 (3 fewer to 9 more)** compared with Rivaroxaban with moderate confidence in estimate		
	⊕⊕⊕◯ **moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕⊕ **high** confidence in estimate based on 14143 patients (1 study)			
Major Bleeding	**OR 0.94** (0.82 to 1.08)	**1 fewer per 1000 (3 fewer to 1 more)**	**OR 1.03** (0.89 to 1.190	**0 more per 1000 (2 fewer 3 more)**	**OR 0.70** (0.61 to 0.81)	**5 fewer per 1000 (3 fewer to 6 fewer)**	**16 per 1000 (1.6 %)**	
	OR 1.10 (0.9 to 1.35)	**6 more per 1000 (2 more to 44 more)**			**OR 1.48** (1.21 to 1.82)	**8 more per 1000 (3 more to 13 more)**		

	compared to Rivaroxaban	compared with Rivaroxaban with moderate confidence in estimate			compared to Rivaroxaban	compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		
Intracranial Bleeding	**OR 0.42** (0.28 to 0.60)	**9 fewer per 1000** **(6 fewer to 11 fewer)**	**OR 0.66** (0.47 to 0.92)	**5 fewer per 1000** **(1 fewer to 8 fewer)**	**OR 0.42** (0.30 to 0.58)	**9 fewer per 1000** **(6 fewer to 10 fewer)**	**14.9 per 1000** **(1.49 %)**
	OR 1.58 (0.95 to 2.66) compared to Rivaroxaban	**5 more per 1000** **(0 more to 6 more)** compared with Rivaroxaban with moderate confidence in estimate			**OR 1.56** (0.97 to 2.5) compared to Rivaroxaban	**6 more per 1000** **(0 fewer to 15 more)** compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		
Major GI Bleeding	**OR 1.45** (1.14 to 1.86)	**9 more per 1000** **(3 more to 18 more)**	**OR 1.61** (1.30 to 1.99)	**13 more per 1000** **(6 more to 21 more)**	**OR 0.88** (0.68 to 1.15)	**3 fewer per 1000** **(7 fewer to 3 more)**	**20.9 per 1000** **(2.09%)**
	OR 1.11 (0.8 to 1.53) compared to Rivaroxaban	**4 more per 1000** **(1 more to 25 more)** compared with Rivaroxaban with moderate confidence in estimate			**OR 1.83** (1.30 to 2.57) compared to Rivaroxaban	**28 more per 1000** **(10 more to 53 more)** compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		
Myocardial Infarction	**OR 1.29** (0.96 to 1.75)	**4 more per 1000** **(1 fewer to 9 more)**	**OR 0.80** (0.62 to 1.05)	**2 fewer per 1000** **(5 fewer to 1 more)**	**OR 0.88** (0.66 to 1.17)	**2 fewer per 1000** **(4 fewer to 2 more)**	**12.5 per 1000** **(1.25%)**
	OR 0.63 (0.42 to 0.93) compared to Rivaroxaban	**5 fewer per 1000** **(0 more to 11 more)** compared with Rivaroxaban with moderate confidence in estimate			**OR 0.92** (0.62 to 1.35)	**1 fewer per 1000** **(4 fewer to 4 more)** compared with Rivaroxaban with moderate confidence in estimate	
	Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		Confidence in estimate due to ….. based on XXXX participants (XXXX study)		

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; OR: Odds ratio;

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.

Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] FOOTNOTES

Table 2: Possible Summary of Findings table structure for network meta-analysis – format B (from GRADE Working Group)

Which oral anticoagulant should be used in patients with atrial fibrillation

Patient or population: patients with atrial fibrillation
Settings: primary care, community, outpatient
Interventions: Dabigatran (150mg), Rivaroxaban, and Apixaban
Comparison: Warfarin and/or Rivaroxaban

Outcome		Effects and confidence in the estimate of effects						Comments
		Rivaroxaban		Dabigatran		Apixaban		
Death								
Warfarin Comparator	**100 per 1000 (10%)**	**OR 0.93** (0.83 to 1.04)	**7 fewer per 1000** (17 fewer to 4 more)	**OR 0.89** (0.78 to 1.01)	**11 fewer per 1000** (22 fewer to 1 more)	**OR 0.90** (0.80 to 1.00)	**11 fewer per 1000** (22 fewer to 1 more)	None/xy of the differences between the new anticoagulants were of important magnitude. Therefore, we did [not] include these comparisons in the table. The quality of evidence needs to be evaluated separately for the comparisons if there are differences for it.
Rivaroxaban Comparator	**29 per 1000 (2.9%)**					**OR 1.04** (0.89 to 1.23)	**1 more per 1000** (4 fewer to 6 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on **14143 patients (1 study)**		⊕⊕⊕◯ **Moderate** confidence in estimate **due to risk of bias** based on **12098 patients (1 study)**		⊕⊕⊕⊕ **High** confidence in estimate based on 18201 patients (1 study)		
Stroke or systemic embolism								
Warfarin Comparator	**30 per 1000 (3.0%)**	**OR 0.88** (0.74 to 1.04)	**3 fewer per 1000** (1 more to 7 fewer)	**OR 0.65** (0.52 to 0.81)	**11 fewer per 1000** (6 fewer to 14 fewer)	**OR 0.80** (0.66 to 0.95)	**6 fewer per 1000** (1 fewer to 16 fewer)	
Rivaroxaban Comparator	**38 per 1000 (3.8%)**			**OR 1.35** (1.03,1.79)	**13 more per 1000** (1 more to 30 more)	**OR 1.11** (0.87 to 1.42)	**3 more per 1000** (3 fewer to 9 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯ **Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)		
Major Bleeding								

Warfarin Comparator	**16 per 1000 (1.6%)**	**OR 1.03** (0.89 to 1.190)	**0 more per 1000** (2 fewer 3 more)	**OR 0.94** (0.82 to 1.08)	**1 fewer per 1000** (3 fewer to 1 more)	**OR 0.70** (0.61 to 0.81)	**5 fewer per 1000** (3 fewer to 6 fewer)	
Rivaroxaban Comparator	**56 per 1000 (5.6%)**			**OR 1.10** (0.9 to 1.35)	**6 more per 1000** (2 more to 44 more)	**OR 1.48** (1.21 to 1.82)	**8 more per 1000** (3 more to 13 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯ **Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)		
Intracranial Bleeding								
Warfarin Comparator	**14.9 per 1000 (1.49%)**	**OR 0.66** (0.47 to 0.92)	**5 fewer per 1000** (1 fewer to 8 fewer)	**OR 0.42** (0.28 to 0.60)	**9 fewer per 1000** (6 fewer to 11 fewer)	**OR 0.42** (0.30 to 0.58)	**9 fewer per 1000** (6 fewer to 10 fewer)	
Rivaroxaban Comparator	**8 per 1000 (0.8%)**			**OR 1.58** (0.95 to 2.66)	**5 more per 1000** (0 more to 6 more)	**OR 1.56** (0.97 to 2.5)	**6 more per 1000** (0 fewer to 15 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯ **Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)		
Major GI Bleeding								
Warfarin Comparator	**20.9 per 1000 (2.09%)**	**OR 1.61** (1.30 to 1.99)	**13 more per 1000** (6 more to 21 more)	**OR 1.45** (1.14 to 1.86)	**9 more per 1000** (3 more to 18 more)	**OR 0.88** (0.68 to 1.15)	**3 fewer per 1000** (7 fewer to 3 more)	
Rivaroxaban Comparator	**32 per 1000 (3.2%)**			**OR 1.11** (0.8 to 1.53)	**4 more per 1000** (1 more to 25 more)	**OR 1.83** (1.30 to 2.57)	**28 more per 1000** (10 more to 53 more)	
		⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)		⊕⊕⊕◯ **Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)		⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)		
Myocardial Infarction								
Warfarin Comparator	**12.5 per 1000 (1.25%)**	**OR 0.80** (0.62 to 1.05)	**2 fewer per 1000** (5 fewer to 1 more)	**OR 1.29** (0.96 to 1.75)	**4 more per 1000** (1 fewer to 9 more)	**OR 0.88** (0.66 to 1.17)	**2 fewer per 1000** (4 fewer to 2 more)	
Rivaroxaban Comparator	**14.3 per 1000 (1.43%)**			**OR 0.63** (0.42 to 0.93)	**5 fewer per 1000** (0 more to 11 more)	**OR 0.92** (0.62 to 1.35)	**1 fewer per 1000** (4 fewer to 4 more)	

	⊕⊕⊕⊕ **High** confidence in estimate based on 12098 patients (1 study)	⊕⊕⊕◯**Moderate** confidence in estimate due to risk of bias based on 12098 patients (1 study)	⊕⊕⊕⊕ **High** confidence in estimate based on 14143 patients (1 study)	

*The basis for the **assumed risk** (e.g. the median control group risk across studies) is provided in footnotes. The **corresponding risk** (and its 95% confidence interval) is based on the assumed risk in the comparison group and the **relative effect** of the intervention (and its 95% CI).
CI: Confidence interval; OR: Odds ratio;

GRADE Working Group grades of evidence
High quality: Further research is very unlikely to change our confidence in the estimate of effect.
Moderate quality: Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low quality: Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low quality: We are very uncertain about the estimate.

[1] FOOTNOTES