CMIMG Meeting, Stream 2 (Statistical issues in NMA)

16 & 17 July 2013, Bristol, UK

Minutes of the meeting

Attendees:

Julian Higgins, Debbi Caldwell, Christopher Schmid, Georgia Salanti, Tianjing Li, Anna Chaimani, Tony Ades, Nicky Welton, Jeroen Jansen, Ian White, Sarah Donegan, Chris Cates, Tanya Walsh, Cinzia del Giovane, Ludovic Trinquart, Zarko Alfirevic

Session 1: Models for NMA and evaluation of assumptions

- *Jeroen Jansen presented "Overview of models and software"*

Explaining the methodology of NMA to non-statisticians using graphical ways might be useful. For instance, bar plots can be employed to present the absolute effects of treatments in studies, and then combined graphically they can be used to exemplify the notion of the transitivity assumption (Jeroen showed some very educative examples).
Recent updates on the available software for NMA include a) hierarchical models in R programmed by Mark Simmons and b) a front end package for *mvmeta* in STATA programmed by Ian White. Some features of this package in STATA (which is under preparation) is setting up automatically the parameterization of inconsistency models, producing graphs, etc.

- *Georgia Salanti presented "Assumptions of NMA models"*

The consistency assumption can be evaluated statistically by comparing direct and indirect evidence. In case that only indirect evidence is available, the model assumptions can be still evaluated by employing non-statistical techniques (e.g. comparing the distribution of effect modifiers). However, many Cochrane reviewers believe that in the absence of direct evidence the plausibility of consistency cannot be tested. Also, the transitivity assumption is related with the exchangeability of treatment effects.

- *Tony Ades presented "Does the evidence satisfy the assumptions required by the analysis?"*

The procedures by which published and unpublished studies are identified in a Systematic Review are well understood. The question was asked "What aspect of these procedures, if any, are intended to guarantee, or happen to guarantee, that the exchangeability assumptions of either pair-wise or network meta-analysis are satisfied?" . No such aspect of the procedures was identified. It was therefore not surprising that empirical studies of "consistency" in evidence networks, for example Song (2011), showed that Cochrane reviews tended to generate inconsistent evidence. Cochrane reviewers therefore need to *explicitly* consider study inclusion and exclusion criteria with a view the the key assumptions required by quantitative synthesis, otherwise inferences from Cochrane Reviews would remain insecure.

Session 1 General Discussion

Criticism of Cochrane reviews with respect to checking and reporting of assumptions

Cochrane reviews do not sufficiently address the assumptions required for principled analysis of data. For example, Cochrane editors require clear inclusion/exclusion criteria, but it is often unclear how the data have been collected. In diagnostic test accuracy studies, each review requires that primary studies used need to be applicable for each review. Studies that do not address the right question should not be included in a network of interventions. Another issue is that there are old Cochrane reviews that include informal indirect comparisons. In these studies, the authors make assumptions about the comparability of studies, treatments and their effects (sometimes without realizing it), however they do not state it in the publication. It is also important to determine the possible effect modifiers and compare them across all studies to make sure that the data are consistent.

The adaptation of NMA to Cochrane reviews needs careful consideration. Usually, Cochrane reviews using pairwise meta-analysis serve as reviews of the evidence that go beyond a focus on effect sizes. Since NMA focus on effect sizes, then the evidence has to be carefully constructed to make sure that it is appropriately targeted to the right question. Combining previous pairwise meta-analyses into a NMA has the risk of incomparability of studies. In cases that Cochrane authors have presented several pairwise meta-analyses separately which they cannot put together, this should be stated in the text.

Performing NMA requires the inclusion of the appropriate studies. In addition, the plausibility of the assumptions should be discussed transparently. The authors should think carefully the construction of the network with respect to the inclusion of legacy treatments. The idea that if transitivity holds then treatments are jointly randomizable is an important concept equivalent with the idea that missing arms are missing at random. It might be also helpful in order to explain the notion of consistency.

Discussion on the guidance needed for Cochrane reviews

The NMA chapter in the Handbook should be short and should not provide statistical details, since reviewers will ask statisticians to do the statistical part of the review. Details can be provided in the appendix. Otherwise, the readers of the Handbook will not be able to follow the chapter considering also that the methodology is still being updated. It should include only simple approaches and not complicated models. One question that needs discussion is whether there is enough consensus in the statistical community regarding the different approaches. The CMIMG received money in order to train reviewers in NMA. It is necessary to produce a document as a basis for such training.

The Handbook needs to emphasize the importance of starting with a clear scientific question and the need to describe the validity of methods for combining direct and indirect evidence. Although there is an already available checklist, developed by ISPOR, for decision making in NMA, it evaluates only the actions after performing a NMA.

It is necessary to develop a protocol defining the analysis to be carried out. The purpose of the protocol is to ensure the conduct of the analysis prospectively in an appropriate manner and is not related to the practicality of doing the analysis with the available data. A fully worked example would be useful to explain the steps and the methodology of NMA. The analysis should be dependent on the research question. For example, is the focus of the review to get a treatment hierarchy or to compare A vs. B incorporating indirect evidence to gain precision? Possibly this idea should infiltrate the whole Handbook (i.e. even pairwise comparisons can be improved by using indirect comparisons); hence, writing something self-contained is an option.

All suggested methods (e.g. subgroups, meta-regression) should be implemented in software, otherwise the reviewers will not adopt them. The level of support that the CMIMG will provide to statisticians needs consideration. Understanding the data is time-consuming and can be done by non-statisticians. Then it can be fed to statisticians. This is at interface of statistics and non-statistics. There is a steep learning curve (12 months already) for clinicians who are fully committed.

Undertaking a NMA is worthwhile if the clinical question suggest it. However, Cochrane reviews usually focus on pairwise comparisons. Pairwise analyses can be useful for indicating when NMA would be informative, but this implies re-doing everything from scratch. Cochrane protocols and reviews should include a section where authors will need to defend the transitivity assumption. A list of questions as a guide might be helpful for authors (like asking about possible effect modifiers). This checklist can be based on the ISPOR checklist and the NICE ground document. The Oral health RG has a checklist as well.

<u>Discussion on technical issues of NMA and recommendations</u>

Data with adjusted results are becoming more frequent and as a result binomial models may not be possible. It is important for statisticians to carry out the analysis without requiring specific software. WinBUGS and *mvmeta* can incorporate most models (e.g. shared parameter models that can use all the availbale data). Informative priors on variances are also necessary. Probably the best practice is to outline all strengths and weaknesses of the different methods without recommending anyone.

Session 2: Presenting data from the network and results from NMA

- *Anna Chaimani presented "Presenting data and results"*
- *Nicky Welton presented "Presenting data and results"*

Various numerical and graphical options have been presented (see Background document). Then participants formed three discussion groups and provided detailed feedback which is summarized below:

Session 2 General Discussion

Feedback on network graphs

1. They should have an option to put one node in the middle as alternative of a polygon with all nodes outside.
2. They should allow adding nodes without changing the shape of the graph and changing colors or width of the lines.
3. They should include numbers of studies and patients or as different sizes of nodes and edges
4. In case that the network is not too large, they should be able to indicate the multi-arm trials.
5. They should be analysis/outcome-specific.
6. They should address missing outcomes.
7. Textual matrix of data indicating which studies reported which treatments is a good supplement to the network graphs. The empty cells in this matrix (missing treatment arms) could be replaced with something informative like number of patients.

Feedback for relative treatment effects and ranks (Results)

1. The primary goal is to demonstrate how much difference there is between the treatments.
2. Mean rank plots should including shading to indicate the credible interval.
3. Boxes reporting the basic parameters contain all the necessary information. Boxes with functional parameters could be presented as supplementary material. [where is this leading ????]
4. Maybe ranking and estimates can be provided in separate graphs
5. Providing predictive intervals is an option.
6. Pie charts should be avoided as a bad graphical practice.
7. The problem is that the ranks may vary by endpoint, so it is challenging to show more than one outcome at once. Each treatment median rank could be in stacked bars or bars next to each other.
8. Median ranks or SUCRAs can be reported as secondary information. [not sure if sucras have CIs]
9. The probability of a treatment being better than another could also be provided along with the respective relative treatment effect.
10. Forest plots showing the relative treatment effects are nice. They can include all the pairwise relative effects and the respective summary effects from pairwise meta-analysis or the relative effects of all treatments vs. a reference. Treatments can be sorted according to the magnitude of the effect.
11. Rankograms are easier to interpret compared to cumulative rankograms

12. SUCRA plots could have confidence bars around them. Mean rank is a linear combination of SUCRA and so credibility intervals for mean ranks can be translated to credibility intervals for SUCRA.
13. The summary can include the forest plots.
14. Predictive intervals need to account for correlations.
15. Most of the presented plots do not show the individual study data or heterogeneity.

<u>Feedback on specific plots presented by Nicky Welton and outlined in a document distributed to the participants</u>

1. Plotting in the same graph more than one outcome may provide too much information for clinicians (i.e. triangle plot in 1st page). Different outcomes could be plotted separately.
2. For very large number of treatments a summary forest plot could be used such as those in pages 10 and 12 or 2 and 4.
3. For smaller number of treatments matrix format is preferred. Ordering the treatments according to the median ranks would be confusing for more than one outcome. Therefore, the order of the treatments should be the same across all the outcomes. Putting usual care at the top makes the graph easy to use. If the best treatment is placed at top left, the relative effects need to be reported as row minus the column.
4. Prediction intervals could be added but this is at the expense of causing confusion by giving too much information.
5. One option is to add the pairwise probability of being better for each contrast in the boxes in the matrices.
6. Double-clicking a diamond could take you to a picture of the underlying forest plot.
7. There is a great deal of information in this matrix plot.
8. Larger letter font is needed.
9. The matrix on page 2 or 4 is nice. The diagonals could be presented in different colors.
10. The heterogeneity estimate could be in bigger font.
11. The summary forest plot on Page 22 is nice, but may be too busy if there are a lot of individual studies.
12. There should be some user testing of the graphical displays.
13. Different views were expressed on cumulative or simple rankograms. We can take advantage of our predominantly electronic format.
14. The summary plot with clustered trials by benefit and adverse events was nice.

Session 3: Selecting the appropriate effect size for NMA

- *Anna Chaimani presented "Overview of empirical evidence"*
- *Tony Ades presented "Choice of outcome in meta-analysis"*

Choice of effect measure to report should be distinguished from choice of statistical model. The appropriate model (linearity in log odds, log risk, log hazard, or risk), should be determined by goodness of fit, minimum heterogeneity, or maximum consistency. Where the dataset is too small to distinguish the different possibilities, other literature in the clinical area can be consulted. It is likely that the appropriate model varies across clinical areas and types of outcome measure.

Session 3 General Discussion

The discussion was widened to include consideration of continuous outcomes, additive vs proportional effects, binomial outcomes observed at several time points, and ordinal outcomes. Each Cochrane review group needs to take statistical advice on the appropriate measures and models for the trials in their area, and and advise the reviewers accordingly.

Session 4:Evaluation of inconsistency

- *Sarah Donegan presented "Overview of methods to estimate inconsistency"*
- *Anna Chaimani presented "Graphs for inconsistency"*

Session 4 General Discussion

Usually, statistical tests have low power to detect inconsistency. Many analysts use inappropriate methods to evaluate the presence of inconsistency. For example, they compare the network estimates with the direct estimates rather than direct and indirect estimates. The global tests for evaluating inconsistency are useful but they have low power. They can suggest no important inconsistency although it might be present. Local test are can help to understand the differences in the estimates.

The node-splitting approach is a local test and it is a form of fixed effects with one additional heterogeneity pattern. Often, it occurs that a single trial appears to be in disagreement with the rest of the network, particularly if it is the only study on an edge of a loop. Predictive intervals could be built to test whether a study that is far away from the center of the data lies within the predictive interval. The aassumption of a single heterogeneity variance in the entire network (or within each closed loop) may have a large effect on consistency results. Inconsistency can be also a result of poor data extraction.

It would be useful to construct a table that classifies tests as local/global and specify which of those tests deal successfully with multi-arm studies. Gert van Valkenhoef's idea of splitting multi-arm trials (presented at Providence SRSM meeting) should be considered.

Session 5:What to do with inconsistent networks

- *Jeroen Jansen presented "Network meta-regression models"*

- *Ian White presented "Random and fixed inconsistency models"*

Session 5 General Discussion

Although studies may initially have allowed for joint randomization, after they are completed their populations may differ on some effect modifier; hence although in principal the network should be consistent, the data does not fit together. Imbalance in effect modifiers can cause inconsistency. However, sometimes we can account for this by using estimates adjusted within study.

Using inconsistency models to account for the possible inconsistency can result in increased standard errors. Heterogeneity can be evaluated by looking at clinical and design differences with respect to effect modifiers. In case that there are implications for important heterogeneity in the network this need to be considered in the analysis. For example, it might be necessary to perform meta-regression or subgroup analysis before using a random effects model.

In fixed effects models node-splitting should preferred compared to individual parameters approaches. The magnitude of the treatment effect in relation with the heterogeneity variance is an important statistic. Also, node-splitting can be considered as a sensitivity analysis to find out how much the estimates change when excluding a comparison.

There are cases that direct evidence might be preferred to indirect evidence. In the presence of sponsorhip bias, publication or selective reporting bias direct evidence may be vulnerable. However, differences in effect modifiers may have a large effect on indirect evidence but not on direct. Additionally, direct evidence may be the most applicable for the population in one trial, but not the populations in other trials in the network; hence indirect evidence is useful as well.

It would be very helpful to provide guidance about how to deal with inconsistency in a similar way to that for pairwise meta-analysis in the presence of heterogeneity. Reviewers should be cautious when they state what causes inconsistency in the network. In sparse networks, there is a limited ability to identify effect modifiers and the power to detect inconsistency might be low. It might be useful to fit different models for different levels of effect modifiers or to stratify networks. Enough trials are required within comparisons to provide enough information.

Established covariate-by-treatment interactions are difficult to interpret and can lead to different recommendations for the treatments according to the values of the covariate. Whether it is worthwile or not to perform a meta-regression depends on the sparseness of the network. In the produced guidance, random inconsistency models could be mentioned, but without making recommendation.

It is important to provide a strong warning message for people not to perform network MA unless they have carefully considered all the assumptions. Each Cochrane review group should specifically guide the authors to the possible effect modifiers in each particular field.

General discussion of the meeting and summary: What do you expect to see reported in the methods section of a Cochrane protocol and review.

- *Debbi Caldwell presented "PRISMA extension for network meta-analysis"*
- *Jeroen Jansen presented "ISPOR checklist"*

Two discussion points are whether Cochrane is likely to adopt the PRISMA extension for NMA, given that several Cochrane members (including CMIMG) are involved in this initiative and whether the bar should be higher for NMA with respect to quality of studies. A 'good practice' and 'good reporting' checklist is necessary to guide people while they perform NMA. This checklist should be fairly similar to the list used to evaluate NMA once it is finished (e.g. ISPOR checklist). One possibility could be to extend MECIR in order to account also for reviews that compare multiple interventions. If NMA reviews become popular this might be a need to address. Checklists are useful but do not solve everything. For example, current checklists assume that the proper question has already been asked. A problematic issue is that the assumptions of NMA can usually not be assessed from a published abstract. Cochrane Reviews have often a rather broad scope in terms of PICO. However, in decision making context, researchers need to start with a very specific question first. For example, if the dose has no effect, then they can proceed as usual, but only if they have no concerns about this. In many cases, reviews are combining studies with different characteristics. Therefore, justification and transparency in reporting are important.

<u>The following considerations were made for the sections of a Cochrane Review and protocol</u>

1. Objectives' section should include the decision to rank and compare treatments irrespectively of their number.
2. The section for the inclusion criteria of the studies does not need any changes.
3. Searching for trials should not start from systematic reviews. Rather, reviewers should do a new search based on the research question. They may need to re-define the eligibility criteria to account for legacy treatments or other treatments that provide indirect evidence.
4. Caution is needed when including connector treatments – the larger the network, the more difficult to defend transitivity.
5. The data extraction should include effect modifiers that might affect the network.
6. The methods section of reviews needs an additional heading related to the assessment of clinical similarity across comparisons.

7. A new heading for inconsistency should be added.
8. It should include local, global, graphical approaches to explore heterogeneity/consistency is important.