European Heart Journal (2015) **36**, 80–82
doi:10.1093/eurheartj/ehu343

EDITORIAL

Network meta-analysis: introduction and an example that compares devices for PFO closure

Khalid Benkhadra[1]*, Zhen Wang[2], and Mohammad Hassan Murad[1]

[1]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Division of Preventive, Occupational and Aerospace Medicine, Mayo Clinic, Rochester, MN, USA; and [2]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Division of Health Care Policy and Research, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Received 3 July 2014; revised 16 July 2014; accepted 4 August 2014; online publish-ahead-of-print 24 September 2014

This editorial refers to 'Percutaneous closure of patent foramen ovale in patients with cryptogenic embolism: a network meta-analysis'†, by S. Stortecky *et al.* on page 120

Over the last 20 years, the number of published meta-analyses has exploded and, therefore, clinicians have become more familiar with this study design. The results of different trials are statistically combined to produce a more precise estimate [i.e. an estimate with a narrower confidence interval (CI)]. When a meta-analysis is performed after a credible systematic review, the results become more reliable because this precise estimate represents the totality of relevant evidence and can apply to a wider range of patients.

Clinicians have become familiar with these traditional meta-analyses that compare two interventions. However, what if we have three or more potentially effective interventions? A newer type of meta-analysis called network meta-analysis (NMA; also called multiple treatment comparison meta-analysis) addresses such comparisons. We will explain the NMA approach and evaluate a recent example published in this issue of the *European Heart Journal*.

When multiple experimental treatments have not been compared in head-to-head trials or only compared in just a few head-to-head trials—which is a very common scenario—one can infer the effect of intervention A on intervention B from comparing the effects of A and B on a common comparator C (usually placebo). Therefore, we can indirectly compare A and B even if they were not compared in a trial. If direct evidence and indirect evidence exist, they can both be combined to provide a final estimate; this is called NMA.[1] This is demonstrated in *Figure 1*.

The output of NMA includes estimates of effect sizes for all possible pairwise comparisons (e.g. relative risks of A vs. B, A vs. C, C vs. B, etc.) and can include a probability ranking (e.g. probability of intervention A being most effective). The analysis can be done using the most frequent (classic) statistical methods (i.e. the well-known methods of hypothesis testing and CIs) or using Bayesian statistics (i.e. updating a prior probability with newer evidence to produce a posterior probability). Commonly Bayesian meta-analysis uses a vague (i.e. non-informative) prior, which leads to the results of both methods being very similar.

Appraising a network meta-analysis

Readers of an NMA should take a step back and evaluate how the studies included in the NMA were identified. Was the NMA preceded by a credible systematic review? A credible systematic review addresses a sensible clinical question, follows an *a priori* established protocol, searches all relevant databases with all relevant search terms and synonyms, is reproducible, and presents data sufficient for readers to make their own judgement about the risk of bias of the individual studies and the overall confidence in the estimates.[2]

Once a review is deemed credible, readers can shift focus to the analysis part of an NMA. There are three major issues to consider. First, trials have to be sufficiently homogeneous to be combined for each intervention. Broader eligibility criteria may enhance the generalizability of the results. However, it can be misleading if participants are not similar and, therefore, heterogeneity is large. Combining different interventions may also be misleading (e.g. pooling results from different doses or different agents in the same drug family). Secondly, across the trials involved in all interventions, are the studies sufficiently similar, with the exception of the intervention (e.g. similar in populations, design, or outcomes)? Thirdly, when both direct and indirect evidence is available, are their results consistent? If any of these three assumptions are not met (homogeneity for each intervention, similarity across all interventions, and consistency of direct and indirect evidence), the results may not be valid.[3]

Once readers judge the systematic review to be credible and the NMA assumptions are met, the results can be interpreted in a manner similar to a traditional meta-analysis. The effect size produced by an NMA is commonly a relative effect (e.g. relative risk or odds ratio) that is hard to communicate to patients and needs to be transformed to a more intuitive absolute effect to help patients trade benefits and harms. The ranking probabilities are particularly challenging to interpret because they have no clinical meaning and

The opinions expressed in this article are not necessarily those of the Editors of the *European Heart Journal* or of the European Society of Cardiology.

† doi:10.1093/eurheartj/ehu292.

* Corresponding author. Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Division of Preventive, Occupational and Aerospace Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA. Tel: +1 507 266 4800, Fax: +1 507 284 1731, Email: Benkhadra.khalid@mayo.edu

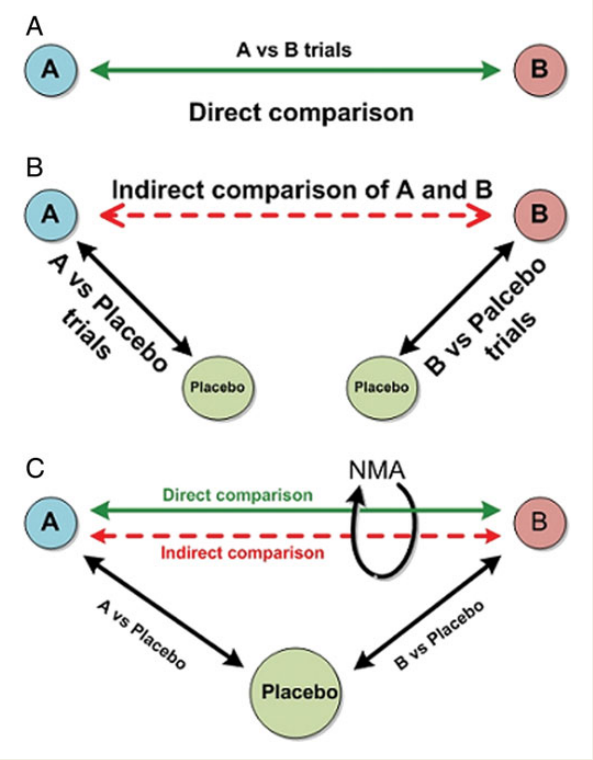

Figure 1 (A) Direct comparison. (B) Indirect comparison. (C) Combining both direct and indirect comparisons in network meta-analysis. Here we are interested in A vs. B.

do not convey the magnitude of difference. Probabilities can also be quite unstable when the number of studies in the NMA is small or a new intervention is added in the NMA. Although some statistical approaches have been proposed to make the ranking more robust, such as presenting a rankogram (a histogram of probabilities of all possible ranks, as opposed to showing only the probability of being ranked best) or estimating the surface under the cumulative ranking curve,[4] ranking remains less helpful in decision-making. A more intuitive approach is to determine if the differences between the treatments are clinically meaningful.

The confidence in estimates (quality of evidence) from network meta-analysis

Decision-making requires that before acting on evidence, a determination (rating) of confidence in this evidence is to be made. This confidence depends on five domains: (i) the risk of bias; (ii) indirectness (are the population, intervention, and outcomes similar between the available studies and the clinical question at hand); (iii) imprecision (are the CIs sufficiently narrow that decision-making would not change if either boundary represented the truth); (iv) consistency of the results across studies; and (v) the possibility of publication bias.[5] A simplified approach for making this judgement about an NMA is first to evaluate the direct evidence (from head-to-head trials). If the direct evidence is of high quality (e.g. derived from large, consistent, low risk of bias trials that clearly address the question of interest), then maybe we do not need to consider indirect evidence. If not, then we should carefully apply all of these criteria to rate our confidence in the indirect evidence and the combination of direct plus indirect evidence.

The case of devices for closure of a percutaneous patent foramen ovale in patients with cryptogenic embolism

An association is found between cryptogenic stroke and the prevalence of patent foramen ovale (PFO); however, randomized trials failed to demonstrate superiority of percutaneous PFO closure over medical therapy. An NMA was conducted by Stortecky *et al.*[6] to combine these trials and determine the effectiveness and safety of PFO closure with different devices compared with medical therapy. Four trials compared three devices: Amplatzer (AMP), STARFLEX (STF), and HELEX (HLX). The results for the main outcome, stroke, showed that AMP was significantly better than medical therapy [rate ratio (RR) 0.39; 95% CI 0.17–0.84]. Conversely, the other two devices were not significantly better than medical therapy (STF, RR 1.01; 95% CI 0.44–2.41; and HLX, RR 0.71; 95% CI 0.17–2.78). The probability of AMP being best in preventing strokes was highest (77.1%) and the authors concluded that the effectiveness of PFO closure depends on the device used and that AMP is superior to medical therapy. We will evaluate this conclusion following the aforementioned steps.

This NMA appears to be based on a credible systematic review. The authors pursued a sensible and clinically important question, searched multiple relevant databases with the relevant terms, and performed study selection, data extraction, and quality assessment in duplicate; and presented data needed for readers to judge the confidence in estimates. The assumptions of an NMA seem to be met. There is homogeneity in the treatment effect on stroke because of the low τ^2, which is a measure of heterogeneity (a low value indicates low heterogeneity and that differences in results across studies are probably due to chance). There was also similarity in the trials included in various comparisons (although one can argue that the distribution of antithrombotic treatments varied). Direct and indirect estimates were consistent. NMA is justified since the quality of evidence from direct comparisons is quite low due to severe imprecision (a single trial with a low number of events and wide CIs).

The first step in rating the quality of evidence is to evaluate the risk of bias. Although the trials concealed allocations and three of the four blinded outcome assessors, there was large loss to follow-up that is not remedied by using the intention to treat analysis. The loss to follow-up ranged from 0% to 13.3%, and averaged 6.7%, compared with a low event rate (2.3% for the outcome of most importance, stroke). A rationale for rating confidence due to the risk of bias exists. A second important domain to evaluate is precision. With only 68 strokes (across all four trials), the results are imprecise.[7] Indeed, when visually inspecting a traditional forest plot of the three devices compared with medical treatment (*Figure* 2), it is clear that the credible intervals reported by the authors greatly overlap and it is unlikely that one treatment is better than the other. An interaction test with *P*-value of 0.27 verifies that, and a

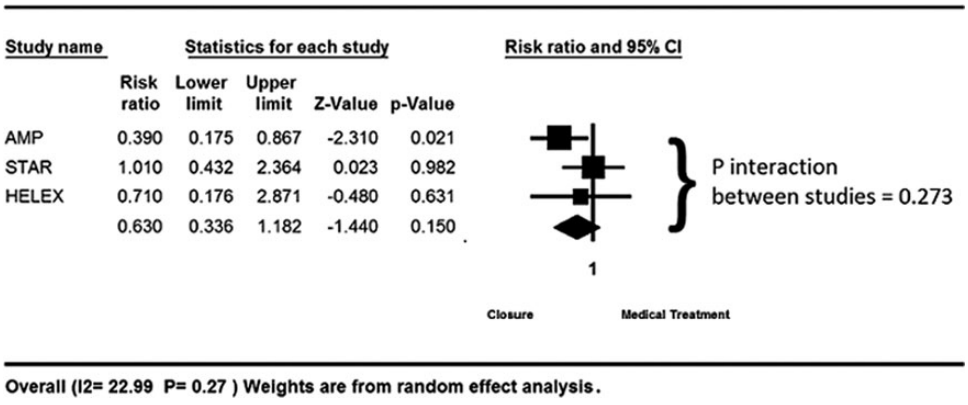

Figure 2 Risk ratios for experiencing a stroke (among all four trials) after undergoing the following interventions: (i) AMP, (ii) STAR, (iii) HELEX (data derived from the abstract) compared with medical treatment.

random-effects meta-analysis of the three devices combined shows that they are not better than medical therapy.

In summary, the confidence in the available estimates comparing the various medical devices for PFO closure with each other and with medical therapy is low, due to imprecision and possible bias. A high probability of 77.1% for one device to be superior hides this confidence rating and is misleading. Although future trials may show otherwise, which is what we expect in the setting of low quality evidence, we cannot at the present time conclude that these devices are better than medical therapy or conclude that one device is better than the rest.

Conflict of interest: none declared.

References

1. Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. *BMJ* 2013;**346**:f2914.
2. Murad MH MV, Ioannidis J, Jaeschke R, Devereaux PJ, Prasad K, Neumann I, Carrasco-Labra A, Agoritsas T, Hatala R, Meade M, Wyre P, Cook D, Guyatt G. How to read a systematic review and meta-analysis and apply the results to patient care. Users' guide to the medical literature. *JAMA* 2014:**312**:171–179.
3. Mills EJ, Ioannidis JPA, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA* 2012; **308**:1246–1253.
4. Jansen JP, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R, Salanti G. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. *Value Health* 2014;**17**:157–173.
5. Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S, Guyatt GH. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;**64**:401–406.
6. Stortecky S, da Costa BR, Mattle HP, Carroll J, Hornung M, Sievert H, Trelle S, Windecker S, Meier B, Juni P. Percutaneous closure of patent foramen ovale in patients with cryptogenic embolism: a network meta-analysis. *Eur Heart J* 2015; **36**:120–128.
7. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams JW Jr, Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J, Schünemann HJ. GRADE guidelines 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol* 2011;**64**:1283–1293.