BMJ

BMJ 2013;346:f2914 doi: 10.1136/bmj.f2914 (Published 14 May 2013)

RESEARCH METHODS & REPORTING

Demystifying trial networks and network meta-analysis

Networks of randomized clinical trials can be evaluated in the context of a network meta-analysis, a procedure that permits inferences into the comparative effectiveness of interventions that may or may not have been evaluated directly against each other. This approach is quickly gaining popularity among clinicians and guideline decision makers. However, certain methodological aspects are poorly understood. Here, we explain the geometry of a network, statistical and conceptual heterogeneity and incoherence, and challenges in the application and interpretation of data synthesis. These concepts are essential to make sense of a network meta-analysis.

Edward J Mills *associate professor*[1 2], Kristian Thorlund *associate professor*[2 3], John P A Ioannidis *professor*[2 4]

[1]Faculty of Health Sciences, University of Ottawa, 35 University Drive, Ottawa, ON, Canada, K1N 6N5; [2]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Humanities and Sciences, Stanford, CA, USA; [3]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada; [4]Department of Health Research and Policy, Stanford University School of Medicine, and Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA

Introduction

When multiple interventions have been used and compared for the same disease and outcomes, network meta-analysis (also commonly referred to as a multiple treatment comparison meta-analysis or mixed treatment meta-analysis) offers a set of methods to visualize and interpret the wider picture of the evidence and to understand the relative merits of these multiple interventions.[1] Network meta-analysis has advantages over conventional pairwise meta-analysis, as the technique borrows strength from indirect evidence to gain certainty about all treatment comparisons and allows for estimation of comparative effects that have not been investigated head to head in randomized clinical trials.[2] For this reason, network meta-analysis is quickly gaining popularity among clinicians, guideline developers, and health technology agencies as new evidence on new interventions continues to surface and needs to be placed in the context of all available evidence for appraisals.[3] For example, over the past two decades more than 20 randomized clinical trials have investigated the long term (>12 months) effects of several variants of warfarin and aspirin as well as other drug treatments for the prevention of stroke in patients with non-rheumatic atrial fibrillation. This accumulation of evidence on multiple treatments has resulted in a network of interventions and comparisons (such as the resulting treatment network, fig 1⇓) that constitutes the randomized evidence between all interventions. In contrast to conventional pairwise meta-analysis, network meta-analysis can provide estimates of relative efficacy between all interventions, even though some have never been compared head to head. For many comparisons, the network meta-analysis may yield more reliable and definitive results than would a pairwise meta-analysis.

In spite of the increasing popularity and widespread use network meta-analysis, certain methodological and interpretational aspects are poorly understood. The strength of evidence and risk of bias for each of the involved comparisons and in the treatment network as a whole[4]; the analytical challenges, tools, and opportunities in detecting and exploring heterogeneity within and between comparisons[5]; and the interpretation of widely used statistical models and effect measures are all matters that deserve further elucidation to ensure high quality synthesis of evidence in the setting of multiple interventions.[6] Here, we aim to demystify these key challenges and opportunities offered by trial networks and network meta-analyses in the context of a working example of interventions for preventing stroke in patients with non-rheumatic atrial fibrillation.[7]

Part 1: network geometry

A key element to understanding a treatment network is the evaluation of its geometry.[8] That is, which of the considered treatments (nodes) have been compared head to head in randomized controlled trials, which of the considered treatments are connected indirectly through one or more "common comparators," and what is the level of evidence informing each comparison. By examining the connections between interventions in a graphical way, as in figure 1⇓, a reader can determine how strong the evidence is for the treatment network as a whole and for the individual comparisons, whether specific comparisons are over-represented or under-represented, and

Correspondence to: E J Mills Edward.mills@uottawa.ca.

For personal use only: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe

BMJ: first published as 10.1136/bmj.f2914 on 14 May 2013. Downloaded from http://www.bmj.com/ on 9 April 2024 by guest. Protected by copyright.

whether the network is well connected. The better connected a network is, the more reliable the estimates it provides will be. Figure 1⇓ illustrates an example of the geometry in a treatment network. This figure includes 34 randomized pairwise comparisons, of which warfarin (n=20), aspirin (n=16), and placebo (n=12) have the most links.[7] The most common comparison (n=7 trials) is between warfarin and aspirin (the two most commonly tested treatments), and the network includes four comparisons of each of them against placebo, the third most common comparator. Overall, 45 possible pairwise comparisons can be made between the nine treatments. Of these, 16 comparisons are informed directly by head to head evidence, but six of the direct connections have only one trial. Thus several of the comparisons that have not been directly studied are informed by indirect evidence from only two trials. Nodes in a network that are not well connected, such as indobufen and ximelagatran in this example, should be interpreted with caution. The diversity and strength of a network are determined by the number of different interventions and comparisons of interventions that are available, how represented they are in the network, and how much evidence they carry. Severe imbalance in terms of the amount of evidence for each intervention may affect the power and reliability of the overall analysis,[9 10] as inferences may be driven largely from the evidence on one or a few treatments and comparisons. The treatment network in figure 1⇓ is a fairly diverse treatment network. Some comparisons are informed by several randomized clinical trials (both directly and indirectly), whereas other comparisons are only sparsely informed (either by direct or indirect evidence).

Many pairwise meta-analyses are insufficiently powered,[11] and similar problems may extend to network analyses.[10] Evidence that is procured by small trials tends to be susceptible to greater bias (for example, more prominent publication and selective reporting biases),[12] and small trials may spuriously show larger treatment effects.[13] Combination of such biased results may yield unreliable estimates in a network. Because networks include evidence from both direct and indirect comparisons, power may be better than in simple pairwise meta-analyses that include only direct evidence.[10] However, the uncertainty of the results in networks with limited evidence should not be underestimated and may extend beyond what the results of any traditional data synthesis might show.[10]

Peculiar co-occurrence patterns suggesting comparator preference biases may also exist—for example, most new drugs may be compared against an established inactive comparator (placebo) or a straw man intervention (one that is known to be a poor choice) rather than against the standard of care, or some head to head comparisons may be avoided. For example, in some fields (such as treatment of partial epilepsy with second generation antiepileptic drugs[8] or biologic drugs for rheumatoid arthritis[14]) placebo controlled trials are almost exclusively performed. This may result to some extent from guidance and requirements by regulatory agencies or may represent the choice of industry sponsors. However, sometimes, the lack of specific direct head to head comparisons may simply be due to a lack of attention to important comparisons that need to be looked at in the future. For example, for most neglected tropical diseases, few or no head to head comparisons have been done between the two or three treatments that are recommended by guidelines as main treatments.[15] Trials in this field are rarely sponsored by the industry. This lack of specific informative comparisons cannot be documented robustly unless the whole network of comparisons is visualized. Identification of missing evidence on specific essential comparisons can guide the performance of the most informative trials in the future research agenda.

Part 2: heterogeneity and incoherence

Network meta-analysis offers a unique opportunity to probe whether homogeneity or heterogeneity exists in the results of different trials in each of the pairwise comparisons that it includes and whether coherence or incoherence is present in the results of different trials that inform indirect comparisons versus the respective available evidence from direct comparisons.

Figure 2⇓ summarizes the concepts of statistical and conceptual heterogeneity and incoherence. Statistical and conceptual aspects overlap but are not identical. Statistical heterogeneity is tested by tests such as Cochran's Q and quantified by metrics such as I^2.[16] In the example of interventions for stroke prevention, none of the pairwise comparisons has evidence of statistically significant heterogeneity (P>0.10 for all on the Q test). However, only three of the available pairwise comparisons were informed by more than two trials, so very little power was available to detect any potential heterogeneity statistically. Although not statistically proven, heterogeneity can still be present and can often be checked conceptually. Conceptual heterogeneity refers to differences in methods, study design, study populations, settings, definitions and measurements of outcome, follow-up, co-interventions, or other features that make trials different. In network meta-analysis, such differences are gauged in the same way as they are in conventional pairwise meta-analysis. However, in network meta-analysis one needs to keep in mind that multiple comparisons are involved. For this reason, conceptual heterogeneity should be assessed both within each comparison and between all comparisons. In our example, conceptual heterogeneity between trials is apparent as a varying proportion of the included patient populations had a history of stroke or transient ischemic attack (ranging from 0% to 100%), despite the fact that all trials were designed to evaluate comparative treatment effects in non-rheumatic atrial fibrillation.[7]

Conceptual heterogeneity across comparisons can result in discrepant results from direct evidence and indirect evidence.[17] Such discrepancies are termed incoherence. Incoherence can occur only when both direct and indirect evidence inform the same comparison. For example, for a comparison between treatments A and B, randomized clinical trials must have compared A and B head to head and both interventions with some common comparator, C. This is commonly referred to as a closed loop. Incoherence can exist only in closed loops, and the presence of incoherence can be assessed by comparing the point estimates of the direct and indirect evidence informing the same comparison. This can be done informally by gauging the overlap of the uncertainty intervals accompanying the point estimates, or it can be done formally by statistically testing differences between the direct and indirect point estimate. In the treatment network for stroke prevention interventions, one nominal signal of incoherence was detected for one among the 10 treatment loops in the network—the treatment loop of placebo, aspirin, and adjusted low dose warfarin. Here, the indirect evidence suggested a large, clearly statistically significant reduction in incidence rate ratio (per 1000 person years) of atrial fibrillation events in favor of adjusted low dose warfarin versus aspirin (0.23, 95% confidence interval 0.10 to 0.63), whereas the direct evidence suggested no effect (incidence rate ratio of adjusted low dose warfarin versus aspirin 0.97, 0.47 to 1.94).

Considering the apparent conceptual heterogeneity in connection with the limited power to detect statistical heterogeneity and incoherence, results of the network meta-analysis should therefore be interpreted with caution. At the same time, we

BMJ: first published as 10.1136/bmj.f2914 on 14 May 2013. Downloaded from http://www.bmj.com/ on 9 April 2024 by guest. Protected by copyright.

For personal use only: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe

should caution that statistical testing for both heterogeneity and incoherence is subject not only to type II error (lack of power to detect heterogeneity/incoherence, when evidence is sparse) but also to type I error (false positive detection of heterogeneity/incoherence, especially when many tests are performed, as in a very complex network). This means that in most network meta-analyses, finding no nominally significant signals for incoherence does not fully exclude its presence, and finding an occasional nominally significant signal of incoherence may sometimes be a false positive. Neither statistical diagnostics nor conceptual reasoning alone is perfect, but their careful combined consideration may be optimal.[18]

When clear conceptual heterogeneity and incoherence are seen, one has to consider whether synthesizing the results across trials in a network meta-analysis is justifiable. When statistical heterogeneity or incoherence is detected, one needs to think carefully about whether clear conceptual explanations for it exist or whether the signal is a chance finding. Furthermore, if one cannot conceptually explain the detected statistical heterogeneity, incoherence, or both, one has to decide whether combining the data in the same network makes sense and whether the results should be interpreted with extra caution. Random effects meta-analysis models can accommodate unexplained heterogeneity for the available pairwise comparisons and often also make the incoherence signals less prominent.

We should also acknowledge that we lack solid evidence on whether the results of network meta-analyses with evidence of heterogeneity and incoherence have less reliability, and thus have poorer ability to predict the results of a future trial on a comparison of interest. In the largest evaluation to examine the coherence between direct and indirect evidence, Song et al evaluated 112 independent trial networks (including 1552 trials with 478 775 patients) that allowed a test for difference between direct and indirect evidence.[19] Incoherence was statistically significant in 16 cases (14% of tests), yet the direction of treatment effects only differed in two cases.

Part 3: data synthesis

Different models exist for synthesizing data in network meta-analyses.[20] The choice of model may affect the amount of confidence one can statistically put in the point estimates produced. The two most widely used models in network meta-analysis (and conventional pairwise meta-analysis) are the fixed effect model and the random effects model. The fixed effect model assumes that no (or a negligible amount of) heterogeneity exists. This assumption is recognized to be typically unrealistic. When heterogeneity exists and the fixed effect model is applied, uncertainty intervals (for example, 95% credible intervals) become artificially narrow. For this reason, the random effects model, which does assume and account for unexplained heterogeneity, is typically preferred. Returning to our stroke prevention example, some evidence of both statistical and conceptual heterogeneity was identified. For this reason, the random effects model seems the appropriate choice.

One of the most appealing but misunderstood elements of network meta-analysis is the reporting of probabilities of which treatment is the best, followed by next best, and so on. Various methods of displaying probabilities are used.[21] A risk exists that one may incorrectly emphasize the probabilities as being clinically useful when the treatment effects are, in fact, not different from the null beyond chance.[4] Probabilities can be fragile when the network is sparse. The ranking of treatments may change drastically when a new trial is introduced into a network. For that reason, authors should place less emphasis on the probabilities of a network meta-analysis output and greater emphasis on the treatment effects and their uncertainty.

Returning to our example from figure 1⇓, table 1⇓ shows the rate ratio with credible intervals (that is, Bayesian confidence intervals) of each treatment compared with placebo and the associated probability that each treatment is best. Only for four of the eight active treatments do we have sufficient confidence that they are better than placebo. Nevertheless, when we calculated the probabilities of being best, alternate day aspirin was associated with the largest probability (66%) of being the best treatment, even though it is one of the four treatments for which we have no confidence that its effect is any better than placebo. This discrepancy occurs because alternate day aspirin yields the largest point estimate for treatment effect (compared with placebo), and most of the probability mass for its treatment effect is centered around small rate ratios.

Summary

Treatment networks and network meta-analysis of randomized trials offer an exceptional opportunity to understand how much evidence is available for each treatment and treatment comparison, where and why more evidence is needed, where and why heterogeneity and incoherence exist, and what the best available treatments are, as well as the uncertainty surrounding such assessments. As network meta-analyses become more popular and influential, familiarity with these opportunities and challenges will be necessary for providing transparent and reliable evidence synthesis.

We are grateful to Steve Kanters for statistical assistance and Georgia Salanti for providing available trial level data.

Contributors: All authors conceived, drafted, and revised the paper for important critiques and approved the final submission.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organization for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med* 2004;23:3105-24.
2 Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* 2005;331:897-900.
3 Sutton A, Ades AE, Cooper N, Abrams K. Use of indirect and mixed treatment comparisons for technology assessment. *Pharmacoeconomics* 2008;26:753-67.
4 Mills EJ, Ioannidis JP, Thorlund K, Schunemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA* 2012;308:1246-53.
5 Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. *Value Health* 2011;14:417-28.
6 Coleman CI, Phung OJ, Cappelleri JC, Baker WL, Kluger J, White CM, et al. Use of mixed treatment comparisons in systematic reviews. [Publisher?], 2012.
7 Cooper NJ, Sutton AJ, Lu G, Khunti K. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. *Arch Intern Med* 2006;166:1269-75.
8 Salanti G, Kavvoura FK, Ioannidis JP. Exploring the geometry of treatment networks. *Ann Intern Med* 2008;148:544-53.
9 Mills EJ, Ghement I, O'Regan C, Thorlund K. Estimating the power of indirect comparisons: a simulation study. *PLoS One* 2011;6:e16237.
10 Thorlund K, Mills EJ. Sample size and power considerations in network meta-analysis. *Syst Rev* 2012;1:41.
11 Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6: rating the quality of evidence—imprecision. *J Clin Epidemiol* 2011;64:1283-93.
12 Ioannidis JP, Cappelleri JC, Lau J. Issues in comparisons between meta-analyses and large trials. *JAMA* 1998;279:1089-93.
13 Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 2012;308:1676-84.

For personal use only: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe

BMJ: first published as 10.1136/bmj.f2914 on 14 May 2013. Downloaded from http://www.bmj.com/ on 9 April 2024 by guest. Protected by copyright.

Summary points

- Networks of randomized trials and network meta-analysis allow readers to visualize and interpret a wide picture of the evidence for specific conditions and to understand the relative merits of multiple interventions
- The geometry of the network allows one to understand how much evidence exists for each treatment, whether some types of comparisons have been avoided, and whether particular patterns exist in the choices of comparators
- Evaluating heterogeneity in the results of different trials in each of the pairwise comparisons and incoherence in comparisons of direct versus indirect evidence is important
- Both conceptual and statistical heterogeneity and incoherence should be assessed
- Estimates of treatment effects from network meta-analyses should be interpreted with due attention to their uncertainty; although appealing, plain treatment rankings or probabilities can be misleading

14 Thorlund K, Druyts E, Avina-Zubieta JA, Wu P, Mills EJ. Why the findings of published multiple treatment comparison meta-analyses of biologic treatments for rheumatoid arthritis are different: an overview of recurrent methodological shortcomings. *Ann Rheum Dis* 2012; published online 20 Oct.
15 Kappagoda S, Ioannidis JP. Neglected tropical diseases: survey and geometry of randomised evidence. *BMJ* 2012;345:e6512.
16 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
17 Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *CMAJ* 2009;181:488-93.
18 Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP. Critical interpretation of Cochran's Q test depends on power and prior assumptions about heterogeneity. *Research Synthesis Methods* 2010;1:12.
19 Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ* 2011;343:d4909.
20 Dias S, Welton N, Sutton A, Ades AE. NICE DSU technical support document 1: introduction to evidence synthesis for decision-making. 2012. www.nicedsu.org.uk/TSD1Introduction.final.08.05.12.pdf.
21 Salanti G, Ades AE, Ioannidis JP. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. *J Clin Epidemiol* 2011;64:163-71.

Accepted: 13 March 2013

Cite this as: *BMJ* 2013;346:f2914

© BMJ Publishing Group Ltd 2013

BMJ: first published as 10.1136/bmj.f2914 on 14 May 2013. Downloaded from http://www.bmj.com/ on 9 April 2024 by guest. Protected by copyright.

For personal use only: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe

Table

Table 1| Treatment effect estimates from example network

Network comparator treatments	Rate ratio (95% credible interval)	Treatment	Probability of being best treatment (%)
—	—	Placebo	0
Adjusted standard dose warfarin *v* placebo	0.37 (0.26 to 0.53)	Adjusted standard dose warfarin	3
Adjusted low dose warfarin *v* placebo	0.32 (0.18 to 0.56)	Adjusted low dose warfarin	16
Fixed low dose warfarin *v* placebo	0.76 (0.30 to 1.76)	Fixed low dose warfarin	1
Aspirin *v* placebo	0.62 (0.43 to 0.86)	Aspirin	0
Fixed low dose warfarin and aspirin *v* placebo	0.98 (0.60 to 1.67)	Fixed low dose warfarin and aspirin	0
Ximelagatran *v* placebo	0.35 (0.19 to 0.65)	Ximelagatran	11
Alternate day aspirin *v* placebo	0.17 (0.01 to 1.15)	Alternate day aspirin	66
Indobufen *v* placebo	0.46 (0.19 to 1.14)	Indobufen	5

BMJ: first published as 10.1136/bmj.f2914 on 14 May 2013. Downloaded from http://www.bmj.com/ on 9 April 2024 by guest. Protected by copyright.

For personal use only: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe

Figures

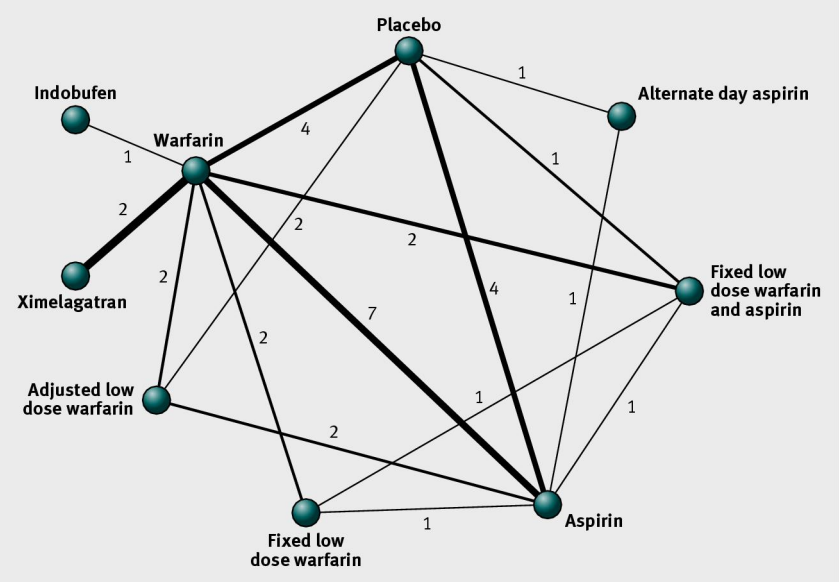

Fig 1 Network geometry of well connected network of randomized controlled trials (RCTs) evaluating stroke prevention among populations with atrial fibrillation. Circles represent the drug as a node in the network; lines represent direct comparisons using RCTs; thickness of lines represents the number of RCTs included in each comparison, also represented by the numbers

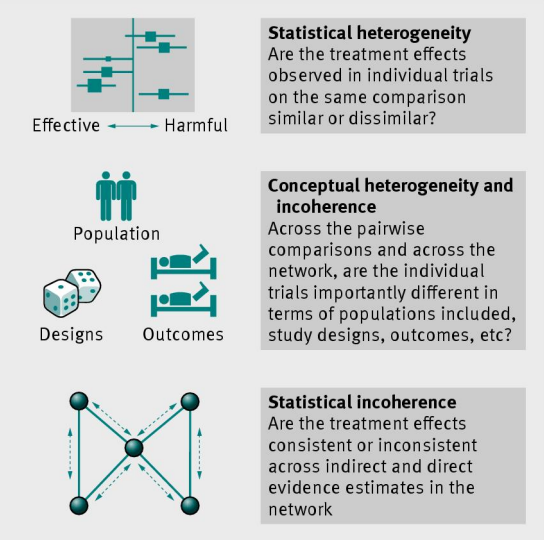

Fig 2 Common considerations of heterogeneity and inconsistency in a network

BMJ: first published as 10.1136/bmj.f2914 on 14 May 2013. Downloaded from http://www.bmj.com/ on 9 April 2024 by guest. Protected by copyright.

For personal use only: See rights and reprints http://www.bmj.com/permissions Subscribe: http://www.bmj.com/subscribe