Animal Health Research Reviews

cambridge.org/ahr

Systematic Review

Cite this article: Hu D, O'Connor AM, Winder CB, Sargeant JM, Wang C (2019). How to read and interpret the results of a Bayesian network meta-analysis: a short tutorial. *Animal Health Research Reviews* **20**, 106–115. https://doi.org/10.1017/S1466252319000343

Received: 11 September 2019
Revised: 7 December 2019
Accepted: 15 December 2019

Key words:
Network meta-analysis; systematic review; tutorial; veterinary science

Author for correspondence:
A. M. O'Connor, Department of Statistics, Iowa State University, Iowa, United States of America. E-mail: oconnor@iastate.edu

CAMBRIDGE
UNIVERSITY PRESS

How to read and interpret the results of a Bayesian network meta-analysis: a short tutorial

D. Hu[1], A. M. O'Connor[2], C. B. Winder[3], J. M. Sargeant[3] and C. Wang[1,2]

[1]Department of Statistics, Iowa State University, Iowa, United States of America; [2]Department of Veterinary Diagnostic and Production Animal Medicine, Iowa State University, Ames, Iowa, 50010, USA and [3]Department of Population Medicine, University of Guelph, Ontario, N1G 2W1, Canada

Abstract

In this manuscript we use realistic data to conduct a network meta-analysis using a Bayesian approach to analysis. The purpose of this manuscript is to explain, in lay terms, how to interpret the output of such an analysis. Many readers are familiar with the forest plot as an approach to presenting the results of a pairwise meta-analysis. However when presented with the results of network meta-analysis, which often does not include the forest plot, the output and results can be difficult to understand. Further, one of the advantages of Bayesian network meta-analyses is in the novel outputs such as treatment rankings and the probability distributions are more commonly presented for network meta-analysis. Our goal here is to provide a tutorial for how to read the outcome of network meta-analysis rather than how to conduct or assess the risk of bias in a network meta-analysis.

Introduction

Rationale

Network meta-analysis is a common method of analysis in human health and increasingly used in veterinary science (Lu and Ades, 2004; Dias *et al.*, 2014; O'Connor *et al.*, 2014; O'Connor *et al.*, 2016). Network meta-analysis is defined as '*The simultaneous synthesis of evidence of all pairwise comparisons across more than two interventions*' (Coleman, 2010). Although frequently used as a synonym for network meta-analysis, a mixed treatment comparisons meta-analysis is a subset of a network meta-analysis which has '*A statistical approach used to analyze a network of evidence with more than two interventions which are being compared indirectly, and at least one pair of interventions compared both directly and indirectly*' (Coleman, 2010). Direct comparisons of interventions are the observed effect obtained from trials or observational studies that compared the pair of interventions of interest. Whereas indirect comparisons of interventions are calculated based on the results of trials that did not directly compare the pair of interventions of interest. Network meta-analysis offers the advantage of enabling the combined assessment of more than two treatments, and the mixed treatment comparison 'component' of meta-analysis has the additional feature of enabling indirect estimation of treatment comparisons that might not be available in the literature in a formal statistical manner (Lu and Ades, 2004; Dias *et al.*, 2014). Most network meta-analyses are also mixed treatment comparisons meta-analyses. We use the term network meta-analysis throughout this manuscript.

We illustrate the advantage of a network meta-analysis over pairwise comparisons, using two previously conducted meta-analyses. In an previous meta-analysis of treatment of Bovine respiratory disease (BRD) complex two pairwise meta-analyses and corresponding forest plots were reported. One pairwise meta-analysis compared the efficacy of tulathromycin to florfenicol and the other tulathromycin to tilmicosin (Wellman and O'Connor, 2007). Each meta-analysis used only direct comparison of tulathromycin to florfenicol or tulathromycin to tilmicosin from the published literature. This dual approach to pairwise meta-analysis, left readers without an estimate of the comparative efficacy of florfenicol to tilmycosin because no randomized controlled trials were available at the time for that direct effect. The reader was left to try to make a non-statistical naive estimate of the comparison of florfenicol to tilmycosin from the pair of forest plots. The problems with a naive estimate of comparative efficacy are many but include an inability to articulate how differences in the number of study subjects and studies are incorporated into the uncertainty about the naive estimate. Subsequently, a network meta-analysis was conducted and because of the ability to 'borrow' information from the network of evidence, an estimate of the comparative efficacy of florfenicol to tilmycosin was estimated indirectly and the comparative ranking of the three antibiotics was obtained (O'Connor *et al.*, 2014; O'Connor *et al.*, 2016). The network meta-analysis

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press

provided a point estimate of comparative efficacy and a 95% credible interval for the estimate which is an advantage of non-statistical approaches. Given that many producers and veterinarians are interested in comparisons of interventions for which no randomized controlled trials are available, network meta-analysis is very useful and is increasingly being adopted.

Objectives

As reports of network meta-analyses become more common, it is essential to introduce the approach to readers and to provide guidance as to how to interpret the results. Therefore, with this tutorial we provide a simple example of a network meta-analysis and explain how to interpret some of the more common outputs provided by authors of network meta-analyses. We describe a Bayesian approach to network meta-analysis, as reviews using this approach often provide more outputs that require interpretation compared to a frequentist network meta-analysis. The tutorial is not aimed at helping readers critically appraise a network meta-analysis or to conduct a network meta-analysis. Critical appraisal of a network meta-analysis requires an assessment of the data informing the meta-analysis and an understanding of what assumptions are relevant to the model and analysis approach (Lu and Ades, 2009; Reken *et al*., 2016). These topics are beyond the scope of this manuscript and the reader is directed to other sources (Hoaglin *et al*., 2011; Jansen *et al*., 2011). Here the reader should infer that the data used and the assumptions made for the model are valid i.e. the study populations are exchangeable and that the transitivity assumption is valid. The transitivity assumption implies that each enrolled subject in a given study would be eligible for enrollment in the other studies. Our focus is on understanding the output of the network meta-analysis. We make the assumption that the reader is familiar with frequentist approaches to the conduct of pairwise meta-analyses and is familiar with interpreting pairwise meta-analysis and so we build upon that knowledge. Sometimes we compare the output to a pairwise meta-analysis to assist in this interpretation.

Organization

The organization of the tutorial is as follows. We first provide a very basic introduction to the concept of network meta-analysis, followed by a description of the data-set that we have used for illustrative purposes in this tutorial. We then briefly present the Bayesian model used, mainly for completeness. The final section presents the common outputs reported from a Bayesian network meta-analysis using our tutorial dataset and a non-statistical interpretation of those outputs.

A lay explanation of network meta-analysis

Network meta-analysis uses information from a web of studies. The overall concept has been succinctly described by others (Dias and Caldwell, 2019). *The underlying idea is very simple: consider three friends, Anne, Ben, and Charles. If we know that Ben is 7 cm taller than Anne, and that Charles is 10 cm taller than Anne, then we know that Charles is 3 cm taller than Ben, and is therefore the tallest. We can also rank the friends in terms of who is tallest as 1 = Charles, 2 = Ben, and 3 = Anne. So, by taking Anne's height as reference and measuring the heights of the others compared with hers, we know how everyone's height compares to each other and how to order the friends by height. The only assumption being made is that the heights we measured are an accurate reflection of the true heights of the three friends (in other words, we used a sufficiently accurate measuring tool). It is easy to see that the same relative heights and ranks would be obtained if one of the male friends had been the reference, and how the height relationships would extend if more than three friends had been measured. This is exactly how NMA works, although we also take the uncertainty (i.e. the sampling error) in the relative effect estimates into account, as is standard in meta-analysis* (Dias and Caldwell, 2019). If we translate that concept to the reviews of clinical trials for bovine respiratory disease discussed above, instead of friends we are interested in antibiotic treatments and instead of height we are interested in the effect of the antibiotic treatment on the outcome i.e. the risk of being retreated for BRD for each treatment. This means that Anne could be substituted for florfenicol, Ben for tulathromycin, and Charles for tilmicosin. Height is replaced, in this example, as the risk of being diagnosed with BRD after being treated i.e. a treatment failure. In the example above, we had information about how florfenicol compared to tulathromycin and for tilmicosin compared to tulathromycin, and we will use that information to make inference about florfenicol and tilmicosin.

The data-set

For the rest of the tutorial, we use a data-set that is not specific to any particular treatment of species or disease. The data-set used for the analysis consists of five treatments labeled A, B, C, D, and E. The outcome in all the trials is the same binary variable i.e. diseased or not diseased. The event of interest is disease, and as the data set relates to treatments, the goal of treatment is to prevent disease. In this example, the most effective treatment has a lowest risk of the disease. The data-set has 25 two arm trials and 1 three arm trial and therefore 53 study arms. The data used in this example are reported in Table 1. This data-set and analysis uses arm level data, in the form of frequency counts, because these data were presented by the author of the original studies. Some network meta-analyses use a relative (or comparative) measure such as the odds ratio (OR) because the authors of the original study did not report the arm-level frequency count data or the arm level frequency count data are not valid because of adjustment for covariates.

The network of studies

The network of studies created by this data-set is provided in Fig. 1. This plot is often referred to as the network plot. This network plot is a common approach to displaying the evidence included in the network meta-analysis. This plot documents the number of treatment and the number of treatment arms that are included in the data-set. Each treatment is a node. In our plot, the size of the node is a relative descriptor of the number of arms available for the meta-analysis. In our plot the number of arms is also included in parentheses. In Fig. 1 we can see that the largest node is for A. Treatment A has 19 treatment arms compared to 13, 11, 1, and 9 for B, C, D, and E respectively. The network plot also attempts to illustrate to the reader how the network is connected. When there are lines between the nodes, this indicates that a direct comparison is available in the network of studies. If the width of line differs between nodes, this is usually indicative of the relative size of the study populations for each comparison. In this data-set, we have arm-level frequency count

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press

Table 1. The arm level data for the 26 studies included in the network meta-analysis of five treatments

Study	Number of event in arm 1	Number of event in arm 2	Number of event in arm 3	Total. number in arm 1	Total. number in arm 2	Total. number in arm 3	Total	Arm 1	Arm 2	Arm 3	Number of arms	Arm 1	Arm 2	Arm 3
1	25	17	20	41	84	100	225	A	B	C	3	1	2	3
2	36	32		41	84		125	A	B		2	1	2	
3	19	7		25	25		50	A	B		2	1	2	
4	20	5		25	50		75	A	B		2	1	2	
5	41	47		50	100		150	A	B		2	1	2	
6	122	69		160	314		474	A	E		2	1	5	
7	236	53		402	399		801	A	E		2	1	5	
8	23	15		27	52		79	A	E		2	1	5	
9	175	166		281	274		555	B	E		2	2	5	
10	57	20		119	118		237	B	E		2	2	5	
11	19	12		100	100		200	B	E		2	2	5	
12	19	7		100	100		200	B	E		2	2	5	
13	16	21		258	254		512	B	E		2	2	5	
14	42	15		50	100		150	A	B		2	1	2	
15	64	34		154	154		308	A	C		2	1	3	
16	34	15		53	106		159	A	C		2	1	3	
17	70	42		130	129		259	A	C		2	1	3	
18	92	31		121	121		242	A	C		2	1	3	
19	35	20		45	90		135	A	C		2	1	3	
20	41	62		59	117		176	A	C		2	1	3	
21	37	15		43	85		128	A	C		2	1	3	
22	16	21		18	35		53	A	C		2	1	3	
23	70	35		122	123		245	A	B		2	1	2	
24	204	71		300	300		600	A	D		2	1	4	
25	111	66		523	526		1049	C	E		2	3	5	
26	60	50		305	297		602	B	C		2	2	3	

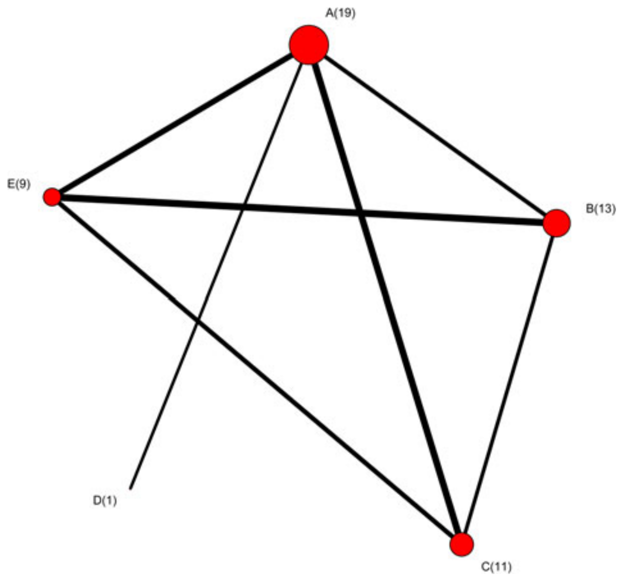

Fig. 1. The network of treatment arms used in network meta-analysis. The size of the dot is a relative indicator of the number of arms and the width of the lines is a relative indicator of the number of direct comparisons (number of arms).

data so we are aware of the number of animals enrolled in each study (Table 1). Sometimes, a network meta-analysis is conducted based on comparative estimates,(such as the OR or risk ratio) and in that situation the number of participants might be unknown and the line width can be used to represent the number of studies. Sometimes authors might use an equal line width between nodes and do not incorporate additional information visually. In our example, we know that the line from C to E is a single study. As C–E study has more participants (1049 participants, study 25 in Table 1) compared to the number of participants in study A–D (600 participants, study 24 in Table 1). C–E study has a thicker line connecting the two nodes. Once the nodes and lines are organized, we can then compare the information available. For example, from Fig. 1 we can see a line directly between A and D. This direct path is not connected to other nodes and means there is a single study comparing A to D. An indirect path is a line connecting two nodes that goes through another node. Indirect path are used for indirect comparisons. In our example, there are no indirect paths from A to D, but three indirect paths from A to E (E to C to A; E to B to A; E to C to B to A).

Assessing and interpreting the geometry of the network

The network geometry can provide some information and is often assessed visually and with statistical approaches (Salanti *et al.*, 2008). The network geometry does not have an influence on the approach to analysis per say i.e. different network geometries do not mean a different meta-analysis approach. However, it is important to evaluate the network geometry to understand how many comparative estimates from the network will be based on indirect evidence only or a mixture of indirect and direct evidence. For example, from Fig. 1 we can see only one direct path between A and D is available. Therefore, when the meta-analysis is conducted only data from the direct study of A to D will contribute to the estimation of the treatment effect of D, although data from multiple studies will contribute to the estimation of A. However, for the comparison of A to E, there are both direct paths and indirect paths. Direct evidence of the effect of E will come from the three empirical studies that compare A to E. However, there are also three indirect paths from A to E (A–B–E, A–C–E, A–B–C–E) and these will be used to create an indirect estimate. One of those indirect paths is a path from A to C and then from C to E. If the data are consistent, as described above, then we can use these indirect estimates of the treatment effect of E and A to help us estimate the effect of E compared to A. Examination of the network helps clarify which is and is not an indirect estimate.

Based on recommendations for reporting of network meta-analysis, the probability of an inter-species encounter (PIE) index and the C-score test is often reported statistics associated with the network geometry (Salanti *et al.*, 2008; Hutton *et al.*, 2015). The PIE index is a continuous variable that decreases in value as unevenness increases. It has been suggested that the PIE index values of 0.75 or less can be considered to reflect limited diversity (Salanti *et al.*, 2008). Networks with few treatments are not diverse, and when networks have the same number of treatments, a network has less diversity when the treatments are not equally represented. Therefore if we had the same five treatments but the 53 arms were more evenly distributed across all the possible pairwise comparisons, such a network would have a higher PIE score. The C score represents co-occurrence and assesses if particular pairwise comparisons of specific treatments are preferred or avoided. In our example, we see co-occurrence does occur i.e. the comparison of A to B occurs more often than other comparisons. However, if we had the same five treatments, and four studies each of all possible comparisons, which would be 40 arms, there would be no co-occurrence. For our example, the PIE score is 0.75, the maximum PIE score we can obtain given five treatments and 53 treatment arms is 0.81 and the C score was 43.2 with a p value of 0.1. These results can be interpreted as suggesting that our example network is 'reasonably diverse'. These numbers are simply descriptors of the network geometry that formalize what can be evaluated visually from the network plot. They do not have an influence on the decision to conduct a meta-analysis nor do they suggest a particular approach to meta-analysis.

The Bayesian analysis

The Bayesian approach to analysis is described in detail elsewhere (Dias *et al.*, 2010). Here we provide a summary of the model used for completeness. A random effects Bayesian model for a continuous outcome is used. The continuous outcome is the logit of the probability of disease i.e. the log of the odds of disease. The term the log of the odds of disease is commonly contracted to the 'log odds'. Let b denotes the baseline treatment of the whole network (usually placebo but in this example A is used), and let b_i denotes the trial-specific baseline treatment of trial i. It could be the case that $b \neq b_i$. Suppose there are L treatments in a network. Assume a normal distribution for the continuous measure of the treatment effects of arm k relative to the trial-specific baseline arm b_i in trial i, y_{ib_ik}, with variance V_{ib_ik}, such that

$$y_{ib_ik} \sim N\left(\theta_{ib_ik}, V_{ib_ik}\right),$$

and

$$\theta_{ib_ik} \sim \begin{cases} N(d_{b_ik}, \sigma^2_{b_ik}), & \text{for } b_i = b, \\ N(d_{bk} - d_{bb_i}, \sigma^2_{b_ik}), & \text{for } b_i \neq b, \end{cases}$$

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press

Table 2. The estimated log odds ratio from all possible pairwise comparisons in the network meta-analysis of five treatment groups.

A	2.101	1.888	1.934	2.579
(1.582_2.645)	B	−0.213	−0.167	0.477
(1.395_2.397)	(−0.855_0.425)	C	0.046	0.690
(0.418_3.462)	(−1.782_1.447)	(−1.560_1.632)	D	0.644
(1.970_3.202)	(−0.098_1.047)	(−0.008_1.400)	(−0.985_2.278)	E

The row treatment is the numerator and the column as the denominator. The treatments are listed alphabetically. The event is disease, therefore a positive logOR indicates that the risk of disease is higher in the numerator.

where d_{bk} is the treatment effects (log of the odds ratio) of k relative to the network baseline treatment b and where $\sigma^2_{b_ik}$ is the between-trial variance. The priors of d_{bk} and σ_{b_ik} are

$$d_{bk} \sim N(0, 10000),$$

and there is a homogeneous variance assumption that $\sigma^2_{b_ik} = \sigma^2$, where $\sigma \sim U(0, 5)$. Thus, for L treatments, we have $L-1$ priors to d_{bl}, $l \in \{1, \ldots, L\}$, $l \neq b$. For $l = b$, we have $d_{bb} = 0$.

All Bayesian analyses require a "run in" period which is discarded before the estimates are considered to converge. The number of simulations required for convergence differs by model and data. In this analysis, the results of 5000 simulations were used to create the posterior distributions of the parameters of interest. For some network meta-analyses it is necessary to include an adjustment for trials with more than two arms, however our example network does not have this feature (Higgins and Whitehead, 1996; Lu and Ades, 2004). As mentioned, this paper is about interpreting the outcome from a valid network meta-analysis, therefore we do not include a discussion about the model used, the choice of prior distributions, the assessment of the consistency assumption or the assessment of convergence for the Bayesian analysis which are all aspects of the data analysis.

Estimates from the model, direct, and indirect

From the Bayesian analysis, the primary output is the posterior distribution of the log of the odds of disease for the baseline treatment (A in our example) and the log of the odds ratio (logOR) of treatment (B, C, D, or E) compared to the baseline treatment (A). In our example, the baseline treatment is A and we have five treatments. From the network meta-analysis, we therefore obtain four relative estimates of the treatment effects which are usually referred to as basic parameters:

- the log of the odds ratio (logOR) of treatment B compared to treatment A,
- the log of the odds ratio (logOR) of C compared to A, and
- the log of the odds ratio (logOR) of D compared to A, and
- the log of the odds ratio (logOR) of E compared to A.

After estimation of the basic parameters, all possible pairwise comparisons are derived from the basic parameters. These derived comparisons are sometimes called the functional parameters, because they are 'a function of' the basic parameters. If the basic parameter is estimated as B compared to A, and C compared to A, then the logOR of B compared to C is obtained by the difference in the logOR of B compared to A minus the logOR of C compared to A. Because we are using a Bayesian framework for the analysis, the reported information for the basic and functional parameters for each treatment is based on the posterior distribution. Once we have these results we can present them to the reader.

Measures of association from the models: direct and indirect

A common approach to presenting such data is in a table with the treatments on the diagonal. The row treatment is usually the numerator and the column as the denominator. By convention we have listed the treatments alphabetically. Here because the event is disease, we therefore a positive logOR, an OR greater than one and a risk ratio greater than one, indicates that the risk of disease is higher in the numerator. The posterior mean is a key characteristic of the posterior distribution. The means of the distribution of the pairwise comparisons of the log odds ratios are given in the upper right hand side of Table 2. When the posterior distribution is not symmetrical, the posterior median can also be reported together with the mean to characterize the distribution. The 95% credible intervals of the pairwise comparisons of the log odds ratios appear in the lower left hand side of the table. The estimate of the log OR of A compared to D is 2.101 and the 95% credible interval is in the range of 1.582 to 2.645. Because there is only one study available for the comparison of D to A, the Bayesian estimate of the effect of D used only data from that study. The estimate of the baseline treatment A used data from multiple studies. Examining the raw study data for the single comparison between A and D in Table 1 we can see that the log odds ratio is 1.925 which is given by:

$$\log\left(\frac{204/96}{71/229}\right) = 1.925$$

The difference in the estimates of D to A obtained from the raw data and the network meta-analysis relates to the Bayesian estimation procedure because the Bayesian method synthesizes the data information as well as the prior information. For the other basic parameters, there are multiple estimates of the treatment effect, both direct and indirect. These data contribute to the estimation of the treatment effect. For example, from a frequentist pairwise meta-analysis of A to B using a random effects model we would obtain an estimate of the logOR of 2.230. The network meta-analysis estimate of the logOR of A to B differs from this frequentist pairwise meta-analysis estimate for two reasons: (1) the data from other indirect comparisons are used to determine the effect of B and (2) Bayesian estimation method.

For example, in Table 2, the estimate of B compared to C is −0.213, is obtained from 1.888–2.101 = −0.213. Similarly, the estimate of C compared to E is 0.690, which is obtained from 2.579–1.888 (with rounding error). In our example, the functional parameters are assumed to be valid because we have informed the

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press

Table 3. The estimated OR from all possible pairwise comparisons in the network meta-analysis of five treatment groups.

A	8.161	6.587	6.91	13.166
(4.864_14.085)	B	0.809	0.850	1.611
(4.033_10.986)	(0.425_1.530)	C	1.049	1.993
(1.520_31.871)	(0.168_4.252)	(0.210_5.115)	D	1.899
(7.174_24.576)	(0.906_2.849)	(0.992_4.055)	(0.373_9.759)	E

The row treatment is the numerator and the column as the denominator. The treatments are listed alphabetically. The event is disease, therefore an OR greater than one indicates that the risk of disease is higher in the numerator.

Table 4. The estimated risk ratio from all possible pairwise comparisons in the network meta-analysis of five treatment groups with the summary of baseline risk to be mean = 0.713, median = 0.728, 2.5% limit = 0.45, 97.5% limit = 0.899.

A	2.929	2.508	2.557	4.266
(1.616_5.942)	B	0.864	0.894	1.438
(1.488_4.814)	(0.527_1.355)	C	1.033	1.671
(1.114_10.832)	(0.330_3.322)	(0.396_3.801)	D	1.604
(2.009_9.828)	(0.929_2.359)	(0.995_3.068)	(0.431_4.717)	E

The row treatment is the numerator and the column as the denominator. The treatments are listed alphabetically. The event is disease, therefore an risk ratio greater than one indicates that the risk of disease is higher in the numerator.

reader that the consistency assumption has been previously checked. There are approaches to assessing the consistency assumption and authors of network meta-analyses should conduct and report this assessment (Dias *et al*., 2010; Dias *et al*., 2014). We can look also at the raw data for the comparison of C to E, and we obtain the estimate of the logOR of 0.630 using eqn 1. The mean of the functional parameter for the comparison of C to E, which uses both direct and indirect evidence, can be found in Table 2 (0.690) and does seem consistent with the direct estimate.

$$\log\left(\frac{111/412}{66/460}\right) = 0.630 \tag{1}$$

Most authors do not report the logOR because it is difficult to interpret (O'Connor, 2013; O'Connor, 2014). Instead most authors of network meta-analyses transform the basic parameters and the functional parameters to another scale such as the OR or the risk ratio.

We begin by discussing the OR estimates and 95% credible intervals for these. These estimates shown in Table 3 are obtained by exponentiation of the logOR for each chain. Some readers will notice that the exponent of the point estimates in Table 2 does not equal the point estimates of the OR reported in Table 3. This is because the exponent of the point estimates in Table 2 is the exponent of the mean of the log odds ratios, whereas the point estimate in Table 3 is the mean of the exponent of the log odds ratios. Here we have reported the mean of the posterior distribution of the OR.

The issue of concern about reporting the OR scale is that this measure of association is also difficult for researchers and clinicians to interpret correctly (although more interpretable than the log odds ratio) and is non-collapsible (Sainani *et al*., 2009; O'Connor, 2013; Grant, 2014; Mansournia and Greenland, 2015). Non-collapsibility of the OR is the phenomenon that when estimating the exposure outcome association with the OR, collapsing over the other covariate(s), the conditional OR does not necessarily equal the marginal OR even in the absence of confounding and effect modification. Further, when the outcome is common, the OR is misinterpreted as the risk ratio, it can lead to an overestimation of the magnitude of effect (Grant, 2014). To illustrate, consider an example where the risk of disease is 80% (160/200) in the exposed group and 20% (40/200) in the unexposed group. The point estimate of the OR is 15.7 while the risk ratio is 4. Clearly, it is incorrect to discuss the risk as increasing 16 fold with exposure.

As a consequence of this potential for misinterpretation, there is preference for reporting risk-based measures of association (Grant, 2014). Therefore, network meta-analyses often report the basic and functional parameters on the risk ratio scale as in Table 4. To obtain these risk ratio estimates, the risk ratio is calculated for each simulation from the logit. As mentioned the continuous outcome is the logit. We can use the expit to calculate the risk of disease in each treatment group using the same baseline in each simulation. The risk ratio is then obtained by dividing these two risks for each simulation. The mean risk ratios in Table 4 are therefore not the direct exponent of the mean OR reported in Table 3. Some readers might try to back calculate the mean risk ratio using the mean baseline risk and the mean odd ratios reported in Table 3 because they are familiar with eqn 2 and find an imperfect match. When consumers of network meta-analyses use this approach to matching the mean OR to the mean risk ratio and the match is not exact, this can seem like an error. However, such an approach is not valid because the risk ratio calculation from a Bayesian framework samples from a distribution of baseline risks in each simulation when calculating the relative risks.

$$\mathrm{RR} = \frac{\mathrm{OR}}{1 - p_0 + p_0 \times \mathrm{OR}} \tag{2}$$

Therefore the data in Table 4 are the mean and distribution characteristics for the formula above arising from the Markov Chain Monte Carlo (MCMC) process. However, if the reader

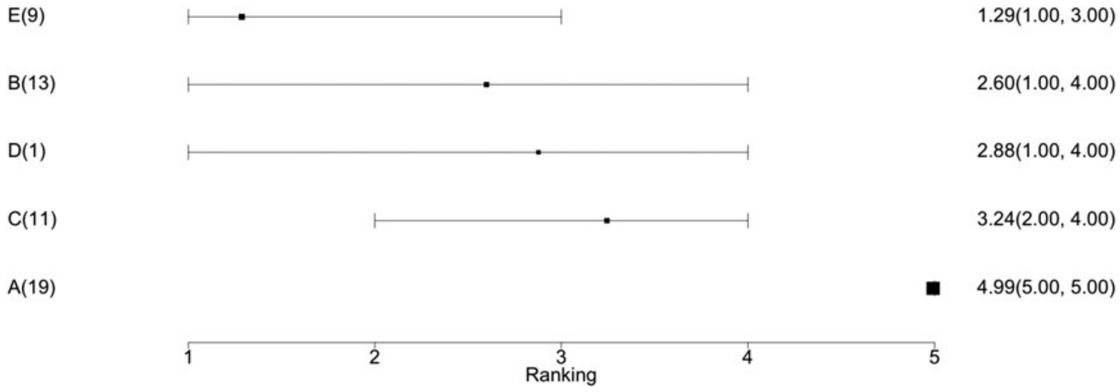

Fig. 2. The ranking plot of five treatments included in the meta-analysis. Lower rankings have lower incidence of the disease.

uses the point estimates from Table 3 and the mean of the posterior of the baseline from formula on eqn 2 then the calculation is equal to 3

$$\mathrm{RR} = \frac{\mathrm{meanOR}}{1 - \mathrm{mean}p_0 + \mathrm{mean}p_0 \times \mathrm{meanOR}} \tag{3}$$

Additional outputs: ranking plots

The major advantage of a Bayesian analysis is that additional outputs can be obtained that can enhance the information obtained from the MCMC simulations. One of the most commonly used outputs is the ranking plot. The risk of disease for each treatment from each simulation is used for the ranking plot. This risk of disease was discussed above in the calculation of the risk ratio. For each simulation, the absolute risk of the outcome is ranked from the lowest to highest, and then converted to a number from 1 to the number of treatments. In our example, the ranks are from 1 to 5. What is ranked high or low depends upon the review question. For this example, the goal is to have the lowest risk of disease, therefore, we would want a low risk of disease to have a lower rank. In our simulation, a rank of one is similar to coming first. After the simulation is complete, the mean rank for each treatment over all simulations is calculated. The rank of the upper 97.5% of the rankings across all simulations and the rank that marks the lowest 2.5% of the rankings across all simulations are also reported. Together these two points are referred to as the 95% credible interval in Bayesian analyses. For our example, the ranking plot is shown in Fig. 2. The rankings show that A is consistently ranked lowest, i.e. it has the highest treatment risk (4.99, upper and lower intervals of 5). Note that these results are not incorrect because the limits are based on the observed data at the 2.5, 50, and 97.5% points of the distribution. For example, suppose we have 100 rankings of A, and only 1 ranking is 4, and all other rankings are 5. The mean ranking is not five, however, the rankings at the 2.5, 50, and 97.5% quantiles of the distribution are 5. E has the highest mean rank and therefore the 'best' treatment with a mean rank of 1.28, and the 95% credible interval of 1 to 3. The ranking plot describes on average (and jointly) which treatment ranks 'best'. There are several interesting points about the ranking plot that aid interpretation. First, it is inherent in the concept of ranking in that all treatments are ordered. Therefore, the clinical relevance of differences in ranks might be meaningless and the evidence to support differences in ranks should also include evaluation of the mean and the distribution of ranks. In our example, the difference in the average rank between the 1st and 2nd ranked treatments is much larger than the difference between the average rank of the 2nd and 3rd ranked treatments, however the difference in ranks is the same. Therefore, being 'one' rank apart does not have the same interpretation across the treatments

Despite combining all the ranks onto one plot, the ranking plot should be carefully interpreted when thinking of the comparison of pairs of treatments. A more meaningful interpretation is to focus on the range of ranks a single treatment can have. A wide 95% credible interval suggests that a single treatment has a wide variety of rankings over all the simulations i.e. the posterior distribution of the rank has large variation. As the rank is a relative measure about the joint rankings, a posterior distribution of the rank with large variation can be a result of either the posterior distribution of the disease risk for a treatment having large variation or the posterior distribution of the disease risk of other treatments having large variation. Notice here that we are discussing the posterior distribution of two outputs, the disease risk and the rankings.

For example, let's imagine a ranking plot with three treatments X, Y, and Z. Let each of the treatments Y and Z have a posterior distribution of the disease risk which has small variation. Also, let treatment X have a posterior distribution of the disease risk which has large variation. There are only three possible rankings 1, 2, and 3. When the rankings are calculated based on sampling from the posterior distributions of disease risk for X, Y, and Z, the variation that is observed in the posterior distribution of the rank of Y and Z will largely be driven by the large variation in the posterior distribution of disease risk for X in each simulation. Because the posterior distribution of disease risk of X is highly variable, one time it may be ranked 1st and another time 3rd. This means although Y and Z have quite consistent disease risk, the ranking they obtain is more variable than suggested by the posterior distribution of the disease risk. Of course, the example is overly simplified, and is more complicated when there are numerous treatments. In our example data, treatment B and D have the same credible intervals (1 to 4) but as B has a better average rank than D, we would conclude that B on average ranks better than D. If we further imagine that the treatments B and D retained the mean ranks reported in Fig. 2 and Table 5 but B has the same credible interval from 1 to 4 but D had a credible interval from 2 to 3.5. In this situation, we should still conclude that B has a better average rank than D but B is more variable. The rankings do not translate to a direct comparison of how

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press

Table 5. Summary of the distribution of the rankings for the five treatments

Treatment	Mean	SD	2.5%	50%	97.5%
A	4.99	0.09	5	5	5
C	3.25	0.74	2	3	4
D	2.87	1.21	1	3	4
B	2.61	0.74	1	3	4
E	1.28	0.53	1	1	3

Table 6. The probability that one treatment has a lower disease risk than another treatment.

A	0.000	0.000	0.008	0.000
1.000	B	0.755	0.588	0.050
1.000	0.245	C	0.475	0.026
0.992	0.412	0.525	D	0.205
1.000	0.950	0.974	0.795	E

The upper quadrant provides the probability that the row treatment has a lower disease risk than the column treatment. The treatments are listed alphabetically. The event is disease and therefore a lower disease risk is the preferred outcome.

often B is better than D and rankings plots should not be interrupted as pairwise comparisons. The inference about pairwise comparisons with particular treatments is discussed below. A wide range of ranks might be a function of variation in estimates from the studies or few data points to base the estimate upon. If the 95% credible intervals of treatments never overlap then it can be concluded that one treatment is better than the other at least 95% of the time. For example, we can conclude that at least 95% of the time that E, B, D, and C had lower disease risk than A. The analysis in the next sections provides more transparent metrics for these inferences.

Additional outputs: pairwise probability of being lower disease risk

Another possible output of interest is the probability that one treatment has a lower disease risk than another i.e. is better. Using the risk data from each simulation, it is possible to determine the probability that B has a lower risk of disease than D i.e. lower risk of disease is a better outcome. The risk of disease in each simulation is determined, and then all possible comparisons made. An indicator is made for the treatment with the lowest risk and summed over all simulations and the proportion calculated. For example, if the following are the risk of disease of three simulations

- 25% for A, 17% for B, 16% for C, 12% for D, and 10% for E
- 25% for A, 16% for B, 17% for C, 12% for D, and 10% for E,
- 23% for A, 18% for B, 17% for C, 12% for D, and 9% for E,

Then the probability that E is better than A (B, C or D) would be 1 because the disease risk is always lower in E. The probability that C is better than B would be 0.66, and the probability that B is better than C is 0.33 because the disease risk for C is lower than B in 2 of the 3 simulations. For our example, these pairwise probabilities are provided in Table 6. It can be seen that the credible intervals are not normally distributed about the mean rank.

Table 7. The probability of having the lowest disease risk (best) and the probability of being the highest disease risk (worst)

Treatment	Probability of lowest disease risk	Probability of highest disease risk
A	0.00	0.992
B	0.031	0.00
C	0.014	0.00
D	0.201	0.008
E	0.754	0.00

The treatments are listed alphabetically. The event is disease and therefore a lower risk is the preferred outcome.

Additional outputs: probability of having the lowest or highest disease risk

Another common output in network meta-analyses is the probability of being the 'best' or 'worst' treatment option which, more precisely, is the probability of having the lowest or highest absolute risk of disease. This is calculated in much the same way as the pairwise probability of being 'better'. For each simulation, the lowest disease risk is identified among the five treatments i.e. which treatment has the lowest disease risk and an indicator is created. For each simulation, the treatment with the lowest risk of disease receives the indicator is 1 and all other treatments receive the indicator 0. After all the simulations, the number of 1's for each treatment are summed and the proportion calculated. This gives the probability of having the lowest risk. The same approach is used to obtain the probability of having the highest disease risk. For our example, these proportions are reported in Table 7. Again, we see that treatment A is less efficacious than the other treatments. The probability that A is the best treatment is zero and the probability that it is the worst treatment is extremely high.

Additional outputs: posterior distribution of probability of the event

Above we have described how rankings, pairwise probabilities, and best and worst ranks are obtained from the Bayesian network meta-analysis. As discussed, these outputs are created for each simulation using the risk of disease estimated. Each of these risk estimates can be plotted as a distribution. Figure 3 and Table 8 present the posterior distribution of the disease risk for all treatments. These plots provide a visual representation of the risk data over all the simulations. What can be seen is that for treatment A, the estimates of disease risk are skewed to higher disease risks. Similarly the other treatments are skewed to lower disease risks. It can also be seen that treatment E has the highest peak close to 0 which means a lower average risk of disease and a narrower credible interval compared to other treatments. If a reader wants to know exactly how often treatment E had a lower risk of disease than treatment A this information is reported in the pairwise probability. Similarly, this plot helps the reader to understand the rankings for treatments B and C, which are very similar. We see that treatment B has a higher peak closer to zero than treatment C, which would explain why it ranks better a small number of times i.e. its mean rankings are slightly better.

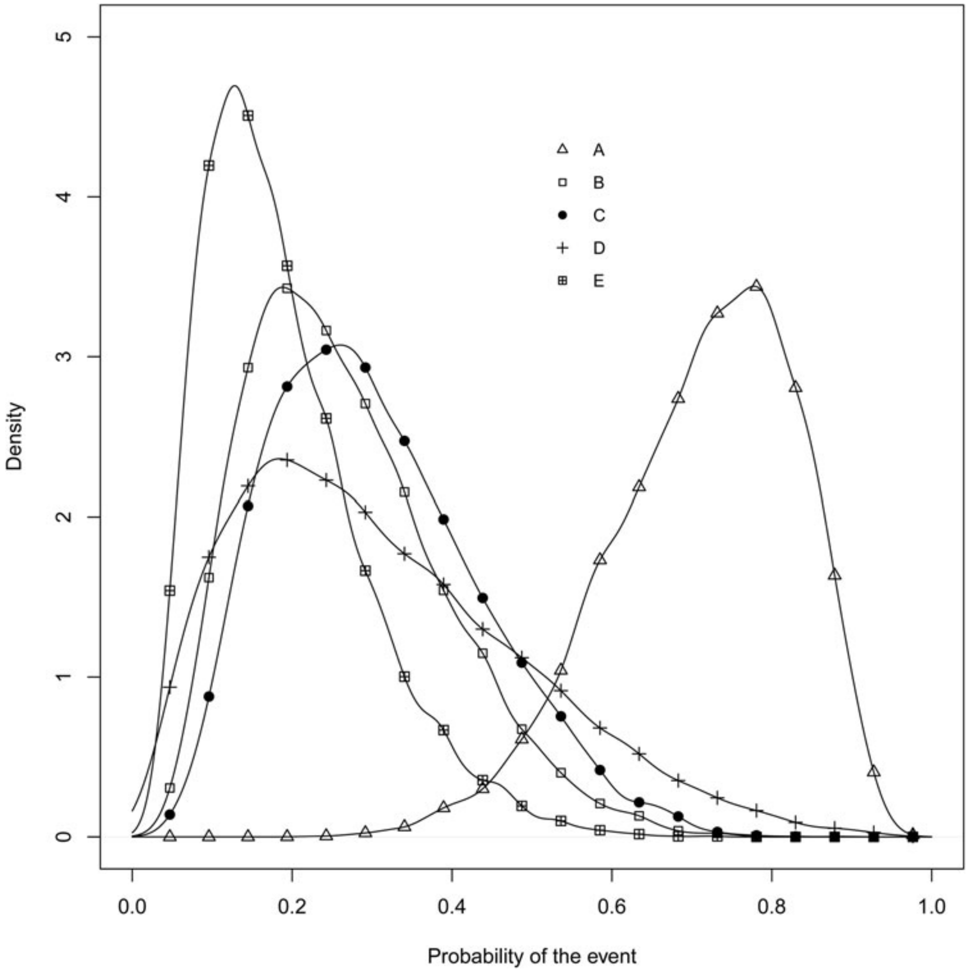

Fig. 3. The distribution of the probability of the event for each treatment. The event is disease and therefore a lower risk is the preferred outcome.

Table 8. Summary of the distribution of the probability of disease risk for five treatments

Treatment	Mean	SD	2.5%	50%	97.5%
A	0.713	0.117	0.45	0.728	0.899
B	0.265	0.122	0.081	0.247	0.549
C	0.305	0.13	0.1	0.289	0.595
D	0.313	0.18	0.053	0.282	0.721
E	0.189	0.102	0.05	0.169	0.44

The treatments are listed alphabetically. The event is disease and therefore a lower risk is the preferred outcome.

We can also see that the limits of the distributions of the posterior distributions of B and C are almost exactly the same i.e. 1 to 4. If the reader want to better understand the workings of the average ranking in the ranking plot this posterior distribution plot often clarifies that information.

Overall interpretation

The role of a network meta-analysis is not to provide recommendations but rather to synthesize the research in a manner that facilitates interpretation. Therefore, the conclusion drawn from these results would depend upon the end-users question and externalities such as cost and potential for adverse events and the validity of the data informing the network meta-analysis. The results of network meta-analyses are a decision-supporting tool rather than a decision-making tool. Still, here we illustrate examples of some conclusions that might be drawn if we assume that all externalities are equal (something that would rarely be true) and that the data included represent all current knowledge.

If the end-user were interested in a particular pairwise comparison, such as knowing if E is likely to be more effective than A, then results show there is a 100% probability that E is more effective than A (Table 6). If the reader is interested in how much more worse A is compared to E, this information is found in the magnitude of the mean OR (Table 3) and mean risk ratios (Table 4). If the reader wanted to know how much better E is compared to A, these OR and RR can be inverted. Further, the credible intervals of these posterior distributions suggest the reader can have reasonable confidence in those estimates. The number of studies available means that end-user can have a confidence in this conclusion i.e. it seems unlikely that another study would dramatically change this inference. Alternatively, if the end-user is interested in another pair of treatments such D versus B, the end-user would have difficulty reaching an interpretation, because we only have one study involving D. From Table 6 we

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press

can see that there is a 60% posterior probability that Treatment B has a lower risk of the adverse event than D. If the reader wanted to know more about the magnitude of effect, this is in Table 3 and Table 4. The average difference in relative risks is not large, indicated by the risk ratio is close to 1, i.e. 0.894 and the credible interval is very wide (0.330 to 3.322). This is reflective of the fact that we only have 1 study with results about D, therefore the posterior distribution of the risk is quite flat. The end-user should reach a cautious conclusion that B is more effective than D for these reasons, and should expect that as data about D becomes available, the inference could change.

Alternatively, perhaps the end-user is interested in all treatments relatively rather than pairwise comparisons. Treatment E has the highest mean rank 5 and the highest probability of having the lowest disease risk 7. The expected magnitude of the effect of E compared to all other treatments is reported in either Table 3 or Table 4. The credible intervals are narrow and a reasonable number of studies inform this conclusion. Interestingly, D has the second highest probability of having the lowest disease risk. This result might be unexpected if the end-user has interpreted the pairwise results for treatments B and D. Figure 3 helps the end-user to understand this result. We can see the distribution of D has some density to the left of B, and these points result in the conclusion that D has a higher probability of being ranked 1st than B. This differs from the pairwise probability data which looks at how often Treatment B had a lower risk of the event than treatment B. Because we only have one study for D, the flat prior used in the Bayesian model is influencing the distribution of D. Again, for treatment D, the end-user should reach a cautious conclusion about the rank of D relative to the other treatments, and should expect that as data about D becomes available, the inference could change.

Conclusions

The aim of this paper was to describe in lay terms the interpretation of outputs frequently reported with network meta-analyses. Other rarer outputs may be included as authors of reviews seek to provide the readers with information about the results.

Authorship

AOC developed the draft of the manuscript. DH created the data for the example, and conducted the data analysis and figures and worked with AOC to ensure the interpretation was correct. CW provided guidance on the conduct of the analyses and interpretation of the results, commented on the manuscript drafts, and approved the final manuscript version. CBW provided feedback on drafts to ensure the description of the examples was clear. JS provided feedback on drafts to ensure the description of the examples was clear

Publication declaration

The authors declare that this is a full and accurate description of the project and that no important information or analyses are omitted.

Funding source

Support for this project was not supported by any external agencies.

Conflict of interest

Dh, AOC, CW, CBW, and JS have no conflicts to declare.

References

Coleman CI, Phung OJ, Cappelleri JC, Baker WL, Kluger J, White CM and Sobieraj DM (2010) Use of mixed treatment comparisons in systematic reviews [Internet]. Agency for Healthcare Research and Quality. Available at https://www.ncbi.nlm.nih.gov/books/NBK107330/.

Dias S and Caldwell DM (2019) Network meta-analysis explained. *Archives of Disease in Childhood – Fetal and Neonatal Edition* **104**, F8–F12.

Dias S, Welton N, Caldwell D and Ades A (2010) Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* **29**, 932–944.

Dias S, Welton NJ, Sutton AJ and Ades A (2014) NICE DSU technical support document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials, National Institute for Health and Care Excellence (NICE). Available at www.ncbi.nlm.nih.gov/books/NBK310366/.

Grant RL (2014) Converting an odds ratio to a range of plausible relative risks for better communication of research findings. *BMJ* **348**, f7450.

Higgins JP and Whitehead A (1996) Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15**, 2733–2749.

Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, Boersma C, Thompson D, Larholt KM, Diaz M and Barrett A (2011) Conducting indirect-treatment-comparison and network-meta-analysis studies: report of the ispor task force on indirect treatment comparisons good research practices: part 2. *Value in Health* **14**, 429–437.

Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, Ioannidis JP, Straus S, Thorlund K, Jansen JP, Mulrow C, Catala-Lopez F, Gotzsche PC, Dickersin K, Boutron I, Altman DG and Moher D (2015) The prisma extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. *Annals of Internal Medicine* **162**, 777–784.

Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L and Cappelleri JC (2011) 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 in Health* **14**, 417–428.

Lu G and Ades A (2004) Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in Medicine* **23**, 3105–3124.

Lu G and Ades A (2009) Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics (Oxford, England)* **10**, 792–805.

Mansournia MA and Greenland S (2015) The relation of collapsibility and confounding to faithfulness and stability. *Epidemiology* **26**, 466–472.

O'Connor AM (2013) Interpretation of odds and risk ratios. *Journal of Veterinary Internal Medicine* **27**, 600–603.

O'Connor AM (2014) Letter to the editor. *Journal of Veterinary Internal Medicine* **28**, 1382–1383.

O'Connor AM, Coetzee JF, da Silva N and Wang C (2014) A mixed treatment comparison meta-analysis of antibiotic treatments for bovine respiratory disease. *Systematic Reviews* **3**, 99.

O'Connor AM, Yuan C, Cullen JN, Coetzee JF, da Silva N and Wang C (2016) A mixed treatment meta-analysis of antibiotic treatment options for bovine respiratory disease – an update. *Preventive Veterinary Medicine* **132**, 130–139.

Reken S, Sturtz S, Kiefer C, Bohler YB and Wieseler B (2016) Assumptions of mixed treatment comparisons in health technology assessments – challenges and possible steps for practical application. *PLoS One* **11**, e0160712.

Sainani KL, Schmajuk G and Liu V (2009) A caution on interpreting odds ratios. *Sleep* **32**, 976.

Salanti G, Kavvoura FK and Ioannidis JP (2008) Exploring the geometry of treatment networks. *Annals of Internal Medicine* **148**, 544–553.

Wellman NG and O'Connor AM (2007) Meta-analysis of treatment of cattle with bovine respiratory disease with tulathromycin. *Journal of Veterinary Pharmacology and Therapeutics* **30**, 234–241.

https://doi.org/10.1017/S1466252319000343 Published online by Cambridge University Press