Evidence synthesis assumes additivity on the scale of measurement: response to 'rank reversal in indirect comparisons' by Norton et al. [2012]

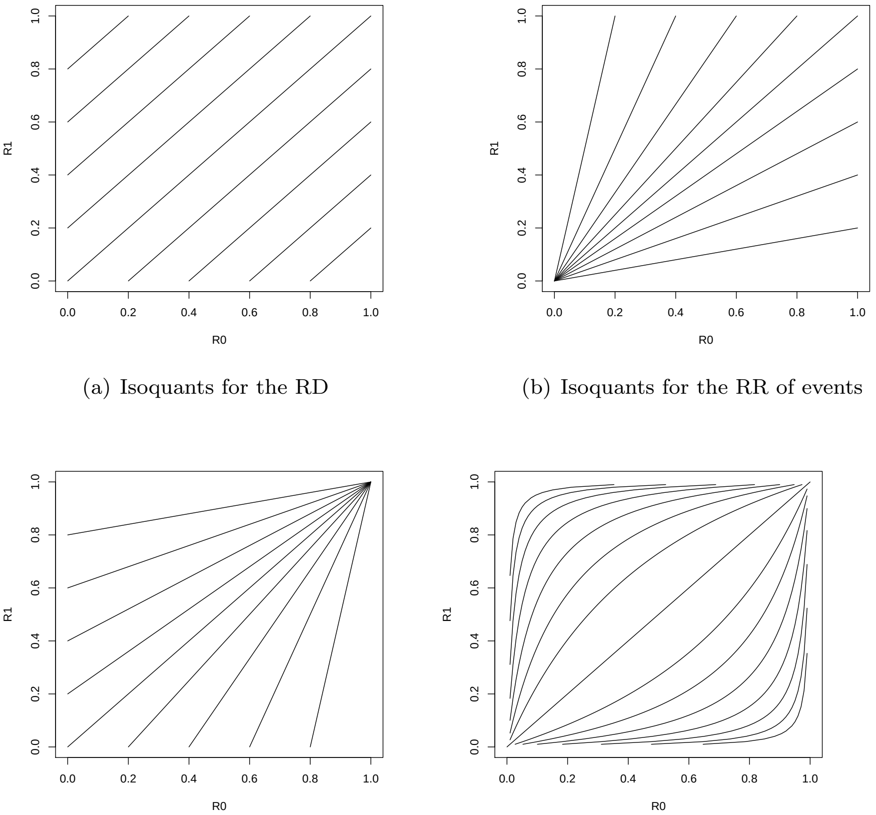
Gert van Valkenhoef A. E. Ades

Norton et al. [2012] point out that the risk ratio (RR), the risk difference (RD) and the odds ratio (OR) may lead to different rankings of treatment alternatives when combining trials with different baseline risks in indirect comparisons. We do not dispute this conclusion, but disagree with the way in which the authors attribute this problem to indirect comparisons. The issue can be clarified by careful consideration of the objectives of both pair-wise meta-analysis and indirect comparisons, and the assumptions made in evidence synthesis.

1 Assumptions underlying indirect comparisons

Indirect comparison meta-analysis [Bucher et al., 1997, Song et al., 2003] is part of a family of evidence synthesis methods that includes direct pair-wise meta-analysis [Hedges and Olkin, 1985] and network meta-analysis [Caldwell et al., 2005, Lumley, 2002, Lu and Ades, 2004, 2006]. The latter allows the combination of both direct and indirect evidence. The key assumption underlying all of these models is the assumption that the trials are *exchangeable* which, put simply, means that they all measure the same underlying relative effects, or effects drawn from a common distribution [Lu and Ades, 2009]. Exchangeability implies that direct and indirect evidence are consistent, and thus that indirect comparisons are valid [Lu and Ades, 2009]. All meta-analytic estimates, whether 'fixed effect' or 'random effects' are weighted averages of the study-specific relative effects [DerSimonian and Laird, 1986]. This is also true of the estimates from network meta-analysis [Lu et al., 2011]. This, in turn, tells us that the exchangeability assumptions can only be correct if the treatment effects are given on a specific linear additive scale. Thus, choosing between the (log) OR, (log) RR, or RD as a scale of measurement is not a matter of selecting a 'summary statistic' on the basis of ease of interpretation or convenience, but one of choosing the most appropriate statistical *model* for the data at hand. Note that an analysis on the (log) RR can be carried out either for the number of subjects experiencing an event or for the number of subjects not experiencing an event, and that these two models make incompatible assumptions. If the appropriate measurement scale is not the one desired for interpretation, a transformation can be applied after evidence synthesis [Caldwell et al., 2012].

In brief, to carry out a meta-analysis on any scale assumes that the observed effects are linearly additive and that the trials are exchangeable *on that scale*. Obviously, this condition cannot hold for all the different scales at the same



(a) Isoquants for the RD

(b) Isoquants for the RR of events

(c) Isoquants for the RR of non-events

(d) Isoquants for the OR

Figure 1: Isoquants for the risk difference (RD), the risk ratio (RR) of events, the RR of non-events, and the odds ratio (OR).

time! This is clearly illustrated by isoquant plots in Figure 1. That is not to say that rank reversal could not happen, but that if it did happen it would be because a dataset either fails to meet the exchangeability assumptions required for meaningful synthesis on any scale, or because the wrong scale of measurement has been chosen. It is well known that in pair-wise meta-analysis there is seldom sufficient data to determine the choice of appropriate scale on purely statistical grounds [Deeks, 2002], and it is interesting that network meta-analysis may enable such an empirical assessment [Caldwell et al., 2012].

2 Indirect comparisons for decision making

In practice, one could only observe a rank reversal on different scales, if the absolute response rates for each treatment vary from trial to trial. Rank reversal is not a discrete phenomenon, since the ranking of treatments based on the results of a meta-analysis is inherently uncertain. Thus, even if the point estimate shifts from one side of the 'no effect' line to another, this need not be relevant if the confidence interval around it is wide. The degree of between-trial variation in absolute response rates has to be quite extreme for relevant rank reversal to be observed. In fact, it has been suggested such variation in the absolute response rates is, in itself, potentially a sign that the exchangeability assumption is not being met [Dias et al., 2011]. Thus, in practice, the rank reversal phenomenon is unlikely to be observed unless the dataset violates the assumptions required for sensible synthesis, or unless the dataset is so sparse that rank reversals happen due to sampling error.

Setting this aside, consider the scale of measurement issue from a practical decision-making perspective. Suppose that in population 1 the response rates are 0.5% for treatment A, 10% for treatment B and 50% for treatment C. This corresponds to

$$\begin{array}{lll} \mathrm{OR}_{AB} = 22.1, & \mathrm{OR}_{BC} = 9, & \mathrm{OR}_{AC} = 199, \\ \mathrm{RD}_{AB} = 0.095, & \mathrm{RD}_{BC} = 0.40, & \mathrm{RD}_{AC} = 0.495. \end{array}$$

Suppose now that the scale on which we obtain linearity is the log(OR), and that in population 2 the absolute response rate is 35% for treatment A. This would mean that we would expect to see 92.3% response rate on treatment B, and 99.1% on treatment C in population 2. Note that whether we choose a population with response rate 0.5% or 35% on treatment A we do not see any rank reversal. The problem arises if we choose the wrong scale, such as the RD and then compare A and B in population 2 (a 57.3% difference), but compare A and C in population 1 (a 49.5% difference). Use of the wrong scale then leads to the false conclusion that B is better than C.

The example illustrates the importance of choosing the scale on which effects are additive. However, we see the issue as applying to any attempt to draw conclusions from one set of trials and generalise them to other situations, and not especially to indirect comparisons. For example, consider the position of an investigator who had observed just the effect of A and B in population 1, and who had assumed additivity on the RD scale. If this investigator then had to predict the response on B in patients with a response rate of 35% on A, he or she would predict 44.5%, seriously under-estimating the real response rate on B of 92.3%. Especially in decisions involving trade-offs between multiple outcomes, such as benefit-risk or cost-effectiveness, this difference could be as important as a rank reversal.

Of course, before embarking on such a hazardous prediction, one would hope that investigators would first ask themselves whether it is reasonable to expect the same effect size (on whatever scale) in two populations that are clearly completely different! There must, after all, be *some* limits on the possibility of generalising treatment effects, which are at the same time the limits of between-study exchangeability and valid synthesis. What is strange is that there appears to be very little in the process of literature identification and systematic review that is specifically designed to deliver the assumptions on which generalisability depends.

3 Conclusion

Indirect comparisons and network meta-analysis are being increasingly used to make coherent decisions when there are more than two alternatives [Jansen et al., 2011, Hoaglin et al., 2011]. Norton et al.'s paper is to be welcomed: in common with a number of recent publications [Song et al., 2011, Li et al., 2011, Mills et al., 2012] it creates a greater awareness and understanding of the what these methods can and cannot do, and draws attention to the assumptions underlying the statistical models that are being fitted.

However, while the growing use of these multi-treatment comparison methods is quite appropriately raising interest and awareness in the fundamental assumptions being made, there has perhaps not been a corresponding awareness that these exact same assumptions apply to pair-wise meta-analysis, and indeed to any process for making valid generalisations based on evidence.

References

H. C. Bucher, G. H. Guyatt, L. E. Griffith, and S. D. Walter. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*, 50(6):683–691, 1997. doi: 10.1016/S0895-4356(97)00049-8.

D. M. Caldwell, N. J. Welton, S. Dias, and A. E. Ades. Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis. *Res Synth Methods*, 3(2):126–141, 2012. doi: 10.1002/jrsm.1040.

D. M. Caldwell, A. E. Ades, and J. P. T. Higgins. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ*, 331 (7521):897–900, 2005. doi: 10.1136/bmj.331.7521.897.

J. J. Deeks. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med*, 21(11):1575–1600, 2002. doi: 10.1002/sim.1188.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clin Trials*, 7(3):177–188, 1986. doi: 10.1016/0197-2456(86)90046-2.

S. Dias, N. J. Welton, A. J. Sutton, and A. E. Ades. NICE DSU technical support document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. Technical report, 2011. URL `http://www.nicedsu.org.uk/`.

L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, 1985. ISBN 9780123363817.

D. C. Hoaglin, N. Hawkins, J. P. Jansen, D. A. Scott, R. Itzler, J. C. Cappelleri, C. Boersma, D. Thompson, K. M. Larholt, M. Diaz, and A. Barrett. 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 Health*, 14(4):429–437, 2011. doi: 10.1016/j.jval.2011.01.011.

J. P. Jansen, R. Fleurence, B. Devine, R. Itzler, A. Barrett, N. Hawkins, K. Lee, C. Boersma, L. Annemans, and J. C. Cappelleri. 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*, 14(4):417–428, 2011. doi: 10.1016/j.jval.2011.04.002.

T. Li, M. A. Puhan, S. S. Vedula, S. Singh, K. Dickersin, C. Cameron, K. Dickersin, S. N. Goodman, T. Li, E. Mills, D. Musch, M. A. Puhan, G. ter Riet, K. Robinson, C. Schmid, S. Singh, F. Song, K. Thorlund, T. Trikalinos, and S. S. Vedula. Network meta-analysis-highly attractive but more methodological research is needed. *BMC Med*, 9:79, 2011. doi: 10.1186/1741-7015-9-79.

G. Lu and A. E. Ades. Combination of direct and indirect evidence in mixed treatment comparisons. *Stat Med*, 23(20):3105–3124, 2004. doi: 10.1002/sim.1875.

G. Lu and A. E. Ades. Assessing evidence inconsistency in mixed treatment comparisons. *J Am Stat Assoc*, 101(474):447–459, 2006. doi: 10.1198/016214505000001302.

G. Lu and A. E. Ades. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics*, 10(4):792–805, 2009. doi: 10.1093/biostatistics/kxp032.

G. Lu, N. J. Welton, J. P. T. Higgins, I. R. White, and A. E. Ades. Linear inference for mixed treatment comparison meta-analysis: A two-stage approach. *Res Synth Methods*, 2(1):43–60, 2011. doi: 10.1002/jrsm.34.

T. Lumley. Network meta-analysis for indirect treatment comparisons. *Stat Med*, 21(16):2313–2324, 2002. doi: 10.1002/sim.1201.

E. J. Mills, J. P. Ioannidis, K. Thorlund, H. J. Schunemann, M. A. Puhan, and G. H. Guyatt. How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA*, 308(12):1246–1253, Sep 2012. doi: 10.1001/2012.jama.11228.

E. C. Norton, M. M. Miller, J. J. Wang, K. Coyne, and L. C. Kleinman. Rank reversal in indirect comparisons. *Value Health*, (in press), 2012. doi: 10.1016/j.jval.2012.06.001.

F. Song, D. G. Altman, A.-M. Glenny, and J. J. Deeks. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *BMJ*, 326(7387):472–476, 2003. doi: 10.1136/bmj.326.7387.472.

F. Song, T. Xiong, S. Parekh-Bhurke, Y. K. Loke, A. J. Sutton, A. J. Eastwood, R. Holland, Y. F. Chen, A. M. Glenny, J. J. Deeks, and D. G. Altman. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. *BMJ*, 343, 8 2011. doi: 10.1136/bmj.d4909.