Solving Complex Statistical Problems in Network Meta-Analysis: Discussion

David C. Hoaglin[1]
[1]Independent consultant, 73 Hickory Road, Sudbury, MA 01776
Department of Quantitative Health Sciences, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605
dchoaglin@gmail.com

I thank the speakers for their stimulating presentations. I am glad to have the opportunity to discuss the three papers. For the paper by Schnitzer et al. and the paper by Yang et al., I based my prepared discussion on draft manuscripts, so some of my comments may not match today's presentations.

Schnitzer et al.

The authors have made an imaginative and thought-provoking application of potential outcomes and related methods.

Concrete examples, based on actual network meta-analyses, would help to clarify various issues. For example,

- What is the reason for using "superpopulation" and not simply "population"?
- For randomized controlled trials, what advantage does their approach have over current approaches?

The approach involves a number of strong assumptions:

- Randomized controlled trials sample subjects randomly. Ordinarily, from the patients who are being treated at a study center, a trial attempts to enroll those who belong to the population defined by the inclusion and exclusion criteria.
- At the design stage, the superpopulation for each randomized controlled trial is randomly drawn from a "metapopulation." I would like to see evidence of trials that actually operate in this way.
- The characteristics and data of randomized controlled trials come i.i.d. from the metapopulation. It is common to regard the studies in a meta-analysis as a sample from a population of studies, but this assumption is unrealistic. In practice, once the inclusion and exclusion criteria for the meta-analysis have been established, the literature search aims to deliver a census. Omission of an eligible study is likely to bring criticism.
- Subjects' (continuous) data come from normal distributions. In practice, real data are never normal.

How does the generalized propensity score, $P(a \in A_i | W_i)$, fit into the authors' approach? I am familiar with propensity scores at the subject level, but not at the level of the trial.

In the simulation study

- Full detail would be helpful.
- I would like to see examples of actual population average characteristics.
- In sampling 2 treatments for each study (from the 4 treatments), does each replication yield a different network?
- How frequently did the authors get networks that were not connected?

For further work, I suggest that the authors adapt and extend their approach to handle the challenges of incorporating data from both randomized controlled trials and observational studies. For any meta-analysis, it is helpful to ask how it contributes to the aim of synthesizing the totality of the evidence.

Yang et al.

Disclosure: The authors and I are coauthors on a previous paper.

Confidence distributions have potential for a range of applications. To become more widely known, they need substantial exposition and accessible software. (I was glad to learn, during JSM, that a book is now available: Tore Schweder and Nils Lid Hjort, *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*, Cambridge University Press, 2016.)

The draft paper is rich in content and fairly technical.

I applaud the substantial focus on robustness.

I am not reassured, however, by reliance on asymptotic normality:

- Meta-analyses often have few studies.
- Studies' sample sizes are often not "large" (in the authors' "Meta-analysis of a Set of Large Studies" the sample size per arm ranged from 39 to 154).

The usual approach to meta-analysis with log-odds-ratio as the measure of effect is problematic. To describe the difficulties, I focus on the log-odds scale:

- Denote the true log-odds for Treatment *j* in Study *i* by $\theta_{ij} = \log\left(p_{ij} / (1 - p_{ij})\right)$.

- The corresponding sample estimate is usually calculated as $y_{ij} = \log\left(r_{ij} \big/ (n_{ij} - r_{ij})\right)$.

- And its variance is estimated by $\hat{\sigma}_{ij}^2 = 1 \big/ r_{ij} + 1 \big/ (n_{ij} - r_{ij})$.

- y_{ij} is a biased estimate of θ_{ij}. (If the process generating r_{ij} puts positive probability on $r_{ij} = 0$ or $r_{ij} = n_{ij}$, the expectation of y_{ij} is undefined. The usual remedy adds 0.5 to all four cells of a 2×2 table that contains a zero cell.)

- Replacing $r_{ij} = 0$ by $r_{ij} = 0.5$ and $r_{ij} = n_{ij}$ by $r_{ij} = n_{ij} - 0.5$ does not remove enough of the bias. Instead, one should add 0.5 to all r_{ij} and $n_{ij} - r_{ij}$ (the "corrected logit").

- The expression for $\hat{\sigma}_{ij}^2$ is only valid asymptotically. (It is the variance of the limiting normal distribution of the log-odds.) Under the usual binomial model y_{ij} does not have finite variance, and even after adjusting zero cells its variance in finite samples has not been studied adequately.

- Reasonable estimates are available for the variance of the corrected logit.

- One can avoid the biases and approximations by basing the analysis on a binomial likelihood for r_{ij}.

For completeness it would have been helpful if the examples involving two treatments had also been analyzed by standard inverse-variance-weighted methods, especially the DerSimonian-Laird method for random-effects meta-analysis, which is very widely used but has well-documented shortcomings (for example, it can produce biased estimates with falsely high precision). For log-odds-ratio in particular, the inverse-variance-weighted methods do not take into account the association between the study-level estimates and their weights.

Multivariate meta-analysis aims to gain strength by borrowing information from related parameters. In the three examples the effect measures (log-odds, risk) are independent. What explains the apparent gain (in the width of the confidence interval) for log-odds-ratio and risk difference?

Zhang et al.

The authors are addressing an important problem. Single-arm studies are part of the totality of the evidence.

The diverse set of Bayesian models aids exploration of potential solutions.

For most of the parameters and effect measures in the two sets of data, estimates from the six models differed less than I expected.

Some of the models can be excluded a priori. For example, it is not acceptable to “break the randomization” in the randomized controlled trials. In one way or another the analysis must combine within-trials comparisons. Thus, I would exclude the Nondiscriminatory model. For BGLMM (I) I ask whether correlation of the arms’ random effects suffices to avoid breaking the randomization. The same question applies to BGLMM (II), BGLMM (III), HPP, and HCP.

If the single-arm studies serve as the basis for priors, one needs to manage their contribution. HPP and HCP do this.

BGLMM (I) and other models assume that the other arms of the single-arm studies are missing at random (MAR). This assumption needs careful discussion when it is applied to entire arms, rather than to data of individual subjects.

It is important to be alert to the possibility of heterogeneity. BGLMM (II) and other models have separate random-effect variances for the two types of single-arm studies. In analyses of actual data, one should start with a graphical display, such as a forest plot.