Commentary

Research
Synthesis Methods

Received 23 July 2015, Accepted 24 July 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1186

Rejoinder to the discussion of “a Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons,” by S. Dias and A.E. Ades

Hwanhee Hong,[a]* Haitao Chu,[b] Jing Zhang[c]
and Bradley P. Carlin[b]

First, we wish to thank Drs. Dias and Ades (henceforth DA) for their discussion of our work, as well as their thorough and passionate defense of the traditional contrast-based (CB) framework for meta-analysis. We are also very grateful to the editor, Dr. Christopher Schmid, for agreeing to publish our paper, the discussion by DA, and allowing us to provide this rejoinder. While CB methods have been and will likely remain the dominant school of thought in network meta-analysis (NMA), thanks to the proliferation of randomized clinical trial and observational datasets, hierarchical Bayesian modeling expertise, and associated computing power, arm-based (AB) methods are certainly in ascendancy (much to the chagrin of DA and others). This paper, its discussion, and this rejoinder have allowed all sides of the issue to be fully discussed and now offer practicing meta-analysts the chance to decide for themselves which model or models they will consider in their own work.

While we would like to avoid a tedious point-by-point response to DA, we will use their section headings in this rejoinder, in an attempt to organize our replies in an intelligible way.

The authors thank Prof. James Hodges for the helpful discussions that significantly influenced this rejoinder.

1. “Classic” contrast-based, contrast-based plus baseline, and arm-based models

DA begin by introducing the notion of a “classic” CB model that uses relative effect measures (perhaps even the “shrunken” trial-specific estimates) as the raw data, precluding estimation of absolute effects. (As a brief aside, in DA’s equation (5), $\sigma^2_{t_k}$ should really be replaced with a covariance matrix capturing correlations between arms.) They then go on to extend this model to one that uses an “arm-based likelihood”, and also assumes exchangeability of the trial-specific control arms. However, at this point, the “nonconstant baseline problem” forces them to place a noninformative prior on the random effects, leaving them “arbitrary and unrelated.” Only when this prior is instead placed on the absolute trial-specific effects of reference Treatment 1 (what DA now calls a “CB plus baseline” model) does the model become a true hierarchical model. The missing data framework described in our main paper makes this modeling more feasible from an MCMC–Bayes point of view but is already troublesome to DA, who worry that this model may make “the relative effect estimates vulnerable to misspecification of the absolute effects.” No wonder then that the AB model they describe next only heightens these worries, because the AB approach aims at an even loftier goal: a model that estimates absolute effects, from which any relative effects (on any scale) can then be inferred.

[a]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD 21205, USA
[b]Division of Biostatistics, University of Minnesota, 420 Delaware St. S.E., Minneapolis, MN 55455, USA
[c]Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, MD 20742, USA
*Correspondence to: Hong, Hwanhee, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, 624 N. Broadway, Baltimore, MD 21205, USA
E-mail: hhong@jhu.edu

But are the assumptions made by AB really “more dangerous” or “less justified” than those made by CB + baseline, or even classic CB? The CB + baseline assumption of exchangeability among Treatment 1 absolute effects is an AB assumption. Why is this assumption acceptable for this one arbitrarily selected treatment (Treatment 1), but not the others? Even the apparently less controversial assumption of exchangeability across relative effects requires the proper selection of scale, which as DA admit is very difficult and, in the end, empirically unverifiable. For example, the conventional scale on which to express a relative treatment effect for binary outcomes, the log odds ratio, is utterly arbitrary. Economists often use a different convention (the probit scale), and what’s exchangeable on one scale would not necessarily be so on another.

2. Comparing arm-based and contrast-based models

In this section of their discussion, D. A. compare AB models unfavorably to CB models, for a variety of reasons, which we address in turn:

(a) Fundamentals of meta-analysis: Overall, we find this subsection’s appeal to “the entire tradition of MA” uncompelling. Science inevitably moves ahead, with innovations welcomed and thoroughly investigated, not stifled merely for being different.

As discussed by Shuster *et al*. (2012), there are two types of assumptions in meta-analysis. The first, often referred to as “studies at random” (SR), assumes that the studies are independently chosen from a conceptual urn containing a large number of possible studies. The second, often called “effects at random” (ER), assumes that the relative effects in each study are randomly drawn from a conceptual urn, while the studies are fixed. The AB model adopts the SR assumption, while the CB model assumes ER. Arguably, ER makes assumptions over and above SR, namely that the distribution of relative effects is independent of study design.This subsection also brings up the familiar refrain that AB models “break randomization,” because their assumption of exchangeable absolute effects across studies cannot be guaranteed unless all trial arms can be thought of as a sample from a single, reasonably homogeneous superpopulation. While relative effects expressed on a “suitable” (but arbitrary?) relative scale may indeed “have been seen as relatively stable” by past meta-analysts using L’Abbé plots, we see no theoretical reason why they should always vary less between trials than absolute treatment effects.Moreover, taking this thinking to its logical conclusion, it suggests clinical trial’s eligibility criteria do not matter, because clinical trials measure relative treatment effects that “stay constant” even across different populations and by assumption, they are exchangeable across trials. So, the AB assumptions do not strike us as more “extreme” than those of CB + baseline at all; indeed, both models allow the control rate in one study to influence estimation of the treatment effect in another. In addition, if we know something about how bias may arise (say, because of changes in the absolute effect of the control preparation over time), the AB framework encourages us to model this.

(b) Evidence synthesis for decision-making: DA are so concerned about the potential for bias in relative effects wrought by misspecification of the absolute effects model; they go on to suggest that absolute effects should be estimated independently from (preferably totally) separate data sources, such as “cohort studies, a carefully selected subset of the trials included in the meta-analysis, or expert opinion.” They go on to defend this approach largely on the basis that this is the way this problem has always been approached, referring to previous NICE work and “standard texts.” Again, such backward-looking justifications of CB leave us cold, but what is more troubling to us is the fact that this suggestion strikes us as inconsistent with fundamental principles of hierarchical Bayesian modeling that DA allegedly support. To begin with, if the absolute effects use a subset of the same trials used to estimate the relative effects, then it is clear these estimates are not independent. Instead, one would need to model the correlation among the two sets of estimates, which is in effect what AB models do. Results using cohort studies (which are subject to their own sorts of biases) and expert opinion are not likely to be better for estimating absolute effects. The latter option is particularly fraught with danger, as experts are well known to be overconfident in their assessments (part and parcel of being an “expert,” it seems), and appropriate correction of such overconfidence is tricky at best (e.g., O’Hagan *et al*., 2006). In addition, we think that DA are making an even stronger assumption here, namely that estimates obtained in an NMA are generalizable to studies not belonging to the NMA.

We agree that trials in an NMA dataset should represent the target population well to have a practical (and certainly not “arbitrary” and “unusable”) interpretation of absolute effects (although we do think the same condition should be assumed for CB models). As most systematic reviews adopting meta-analytic methods have thorough inclusion and exclusion criteria reflecting their potential target populations, the absolute effect estimates are practically meaningful in many NMAs with well-collected studies. In addition, another publication by Hong *et al*. (2015) shows an example where AB models provide more useful and straightforward interpretations for diabetes patients needing to select the best treatment given their characteristics and disease severity levels.As researchers steeped not in the work of the Cochrane Collaboration but instead in hierarchical Bayesian theory, methods, and computing, we naturally prefer to model all correlations we think might be present in all data sources we use. The potential for shrinkage of estimates from like groups toward each other is after all what supports the use of Bayesian methods in NMA: this can produce an improved ensemble of estimates, where here “improved” actually refers not merely to Bayesian properties but to long-run frequency properties as well; this line of research goes all the way back to Stein (1955). DA dislike that AB modelers are “obliged” to use the same data to estimate absolute and relative effects, but because these effects are almost certain to be correlated *a posteriori*, this obligation is no different than the one any statistician faces when modeling any extremely large, complex dataset. If you acknowledge and try to model the correlations you know are there, you will do better in the long run.

(c) The empirical question: First, we disagree that posterior precision of the mean relative effects in the AB models was “severely degraded” compared with the CB models. Our Fig. 6 compares the LAREhom (classic CB), two CBRE (CB + baseline), and two ABRE (AB) models, where all but the first incorporate the missing data framework. Compared with the classic CB model, the CB + baseline and AB models do introduce larger posterior variability, but this is entirely due to the uncertainty correctly acknowledged by the missing data framework. However, an “apples to apples” (CBRE2 to ABRE2, and CBRE3 to ABRE3) comparison of the widths of the 95% Bayesian credible intervals for the mean relative effect estimates reveals them to be quite similar. So, Fig. 6 does not make the case that AB models allow “huge variation in absolute effects to affect the estimates of relative effects.”

We certainly do agree that the relative merits of AB and CB models can and should be judged empirically but reject the idea that missing data-acknowledging AB or CB + baseline estimates are somehow less valid because their relative effect estimates may have larger posterior variances. Certainly, every analyst wants narrow confidence intervals for treatment effects, but to deliberately choose a statistical model to “insulate” such relative effects from genuine uncertainty smacks of choosing a model designed to obtain the answer you want. After all, another way to do this would be to use a highly informative prior, something that would likely be anathema to most NMA practitioners (and objective Bayesians). Indeed, the “insulation” that DA recommend, when combined with the idea of estimating absolute effects separately from expert opinion, amounts to the most informative (and least objective) prior we have seen in a while! So, any improved precision in classic CB relative treatment effect estimates is illusory to the extent that it derives from ignoring uncertainty elsewhere in the hierarchical model.

3. Missingness of treatment arms in network meta-analysis

DA mention that the missing at random (MAR) assumption is made for traditional NMA, which differs from the missingness assumption we made; DA assume that the missingness is defined by the relative effects, while our missingness is defined by the absolute effects. We believe that our MCAR scenarios are closely related (if not equivalent) to the MAR assumption on relative effects made in the classic CB model. In addition, our paper only suggests it is beneficial to incorporate the missing data framework in NMA and does not differentiate between how missingness is modeled in CB versus AB settings. As mentioned in the preceding text, it is not the CB + baseline versus AB choice that greatly affects the inference; it is the decision whether or not to incorporate a model for the missing data (as classic CB does not). Furthermore, it is worth mentioning that, in an NMA with K treatments in I studies, the total number of relative effects is $I \times K \times (K-1)/2$. This is commonly much bigger than the total number of absolute effects, $I \times K$, when $K > 3$. As most randomized clinical trials include only two arms, the number of observed relative effects is often much smaller than the number of observed absolute effects, leading to a much higher proportion of missing data for relative effects. Consequently, it is much more difficult to justify the MAR assumption on relative effects and to conduct sensitivity analyses.

DA also introduce a meta-regression that takes into account the “relationships between relative treatment effects and baseline severity” in the classic CB model and wonder about the counterpart under the AB models. The AB approach models all correlations between arms, so we do not need to consider baseline and relative treatment effects separately.

4. Simulation studies

DA point out that the classic CB model cannot generate an absolute baseline effect by averaging the $\mu_{i,1}$. In addition, they confess that they have made a “tactical blunder” in previous publications in an attempt to illustrate that “the absolute effects a_k can be composed from the relative effect d_{1k} and an estimate of the absolute effect of the reference treatment *that does not originate from the NMA itself*.”

However, we agree that comparing absolute effect estimates obtained from the classic CB and AB models is not fair, and this is why our simulations compare relative effects estimated from the classic CB and CB + baseline models (for which results are presented in panels (d) and (e) of Figs. 3–5 in our paper).

Note that, in our simulations, we first simulate a complete dataset, where all trials have three arms, then drop arms randomly, with the result that missingness is independent of the absolute effects and the relative effects remain exchangeable across studies. If we interpret our simulation results favorable to DA's claim, both classic CB and CB + baseline models perform well under MAR in terms of relative effects (which is denoted by MCAR in our paper). However, the classic CB model is more vulnerable to the assumption of MAR in terms of absolute effects than our CB + baseline model. That is, incorporating the missing data framework into the CB models helps to obtain less biased relative effect estimates. In the case of the MNAR scenario, we admit that the interpretation is not clear; our "oddly unbiased" remark was intended to connote our surprise at the strong performance of classic CB here. But this is not always the case: in the additional single-outcome simulation study reported in Fig. 2 of the paper's Supporting Information, the classic CB method does not outperform CB + baseline in terms of bias in the MNAR case, as it does in Fig. 3(d) and (e) in our paper. It is expected that our models do not perform well in the MNAR case because we do not model the missingness explicitly. Recent work by Zhang *et al*. (2015) studies the AB approach in the MNAR case using selection models.

Lastly, we acknowledged all recent publications about NMA for multiple outcomes and have summarized these works in the Introduction. All of these methods require additional assumptions or external information regarding the unknown within-study correlations, while our approach uses random effects to capture and estimate such correlations.

5. Conclusions

Arm-based methods are hardly new to network meta-analysis and have also been used in conventional pairwise meta-analysis (e.g., Chu *et al*., 2012). Throughout our research into this topic, our viewpoint has been that while we agree that AB models do require a different set of assumptions than CB models, it is not obvious that they are less reasonable, and the payoffs they can provide (significantly increased modeling flexibility, as well as greater ease of interpretation, prior specification, and model fitting) can be substantial. Because every CB model can be expressed as an AB model, but the reverse is not true, the higher assumption burden for AB is not surprising. Moreover, CB models' assumption of exchangeability of relative effects, typically on an arbitrary scale and using an arbitrarily baseline, is often heroic but rarely acknowledged as such. Furthermore, the MAR assumption on relative effects is often unrealistic and difficult to justify.

In summary, we find some of the language used by DA to describe AB models ("radical – even revolutionary"; "thoroughly misguided"; "a huge step back") to be badly overblown and verging on hysteria; AB modelers are not savages banging on the gates, looking to sack the CB castle. AB models are simply another more complete and fully Bayesian approach to modeling what is without doubt a very, very challenging type of "big data": a network of clinical studies, gleaned in an ad hoc manner from an often uneven published literature. AB modeling may well lead to increases in bias but except in cases where missingness is nonrandom, can often lead to improvements in mean squared error behavior. And as mentioned in the preceding text, the perceived gains resulting from CB modeling can often be illusory, the result of misplaced trust in tradition and convention. Even if one can select the proper scale and link function (a huge "if"), there is often little more reason to think that relative effects are exchangeable across studies than absolute effects would be. For example, as the standard of care improves, the relative benefits of treatments may well become smaller. Combining results from phases II, III, and IV trials (different as they are with respect to enrollment criteria) would also likely call the exchangeability of relative effects into question.

As a final thought, we think this entire debate may well be something of a tempest in a teapot, with biases resulting from the selection of the studies included in the NMA often swamping any differences resulting from the application of AB versus CB methods. Literature searches may exclude key studies not published in the meta-analyst's own language, published in obscure places, or most seriously, not published at all because of their failure to achieve statistical significance (the celebrated "file drawer problem" in meta-analysis; Iyengar and Greenhouse, 1988). Sources from which NMA data are abstracted are often incomplete, lacking specific information about study entry criteria, dosing information, and other key quantities needed for their proper incorporation. Subjects enrolling in clinical trials are fundamentally different from the much broader population of persons about which an NMA hopes to infer, and their characteristics change with trial phase and calendar year. While we continue to favor the fully Bayesian AB approach, we agree with DA that it is important to do the best we can to guard against all sources of bias while still using models that seek to account for all correlations in the data. We look forward to the inevitable appearance of further comparisons of CB and AB approaches that extend the investigation of their relative strengths and weaknesses.

References

Chu H, Nie L, Chen Y, Huang Y, Sun W. 2012. Bivariate random effects models for meta-analysis of comparative studies with binary outcomes: methods for the absolute risk difference and relative risk. *Statistical Methods in Medical Research* **21**(6):621–633.

Hong H, Fu H, Price KL, Carlin BP. 2015. Incorporation of individual-patient data in network meta-analysis for multiple continuous endpoints, with application to diabetes treatment. *Statistics in Medicine* **34**(20):2794–2819.

Iyengar S, Greenhouse JB. 1988. Selection models and the file drawer problem. *Statistical Science* **3**(1):109–117.

O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T. 2006. Uncertain Judgements: Eliciting Experts' Probabilities. New York: Wiley.

Shuster JJ, Guo JD, Skyler JS. 2012. Meta-analysis of safety for low event-rate binomial trials. *Research Synthesis Methods* **3**(1): 30–50.

Stein C. 1955. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proc. Third Berkeley Symp. on Math. Statist. and Prob.(pp. 197–206). Berkeley, CA: Univ. of California Press.

Zhang J, Chu H, Hong H, Virnig BA, Carlin BP. 2015. Bayesian hierarchical models for network meta-analysis incorporating nonignorable missingness. *Statistical Methods in Medical Research*. DOI: 10.1177/0962280215596185.