Original Article

Research Synthesis Methods

Received 18 January 2014, Revised 2 March 2015, Accepted 27 March 2015 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1153

A Bayesian missing data framework for generalized multiple outcome mixed treatment comparisons

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

Bayesian statistical approaches to mixed treatment comparisons (MTCs) are becoming more popular because of their flexibility and interpretability. Many randomized clinical trials report multiple outcomes with possible inherent correlations. Moreover, MTC data are typically sparse (although richer than standard meta-analysis, comparing only two treatments), and researchers often choose study arms based upon which treatments emerge as superior in previous trials. In this paper, we summarize existing hierarchical Bayesian methods for MTCs with a single outcome and introduce novel Bayesian approaches for multiple outcomes simultaneously, rather than in separate MTC analyses. We do this by incorporating partially observed data and its correlation structure between outcomes through contrast-based and arm-based parameterizations that consider any unobserved treatment arms as missing data to be imputed. We also extend the model to apply to all types of generalized linear model outcomes, such as count or continuous responses. We offer a simulation study under various missingness mechanisms (e.g., missing completely at random, missing at random, and missing not at random) providing evidence that our models outperform existing models in terms of bias, mean squared error, and coverage probability then illustrate our methods with a real MTC dataset. We close with a discussion of our results, several contentious issues in MTC analysis, and a few avenues for future methodological development. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Bayesian hierarchical model; Markov chain Monte Carlo; missingness mechanism; network meta-analysis

1. Introduction

Mixed treatment comparisons (MTCs), also called network meta-analysis, are meta-analytic statistical techniques that extend traditional meta-analysis of two treatments (DerSimonian and Laird, 1986) to simultaneously incorporate the findings from several studies on multiple treatments, where in most cases none of the studies compared *all* the treatments at one time, to address the comparative effectiveness and safety of interventions accounting for all sources of data (Hoaglin *et al*., 2011; Jansen *et al*., 2011; Ades *et al*., 2012). MTCs are extremely useful in the real-world scenario of having to decide between many different competing interventions for the same population. In the MTC data framework, we can estimate a relative effect of two treatments rarely investigated in the same trial by using all available and relevant evidence (i.e., both indirect and direct comparisons) in a coherent way, strengthening inference (Lumley, 2002; Gartlehner and Moore, 2008). Bayesian hierarchical statistical meta-analysis for MTCs with a single binary outcome has been investigated actively (see e.g., Smith *et al*. (1995), Lu and Ades (2004, 2006), and Nixon *et al*. (2007)). However, we can easily generalize the method to non-binary settings (e.g., continuous or count outcomes) by utilizing appropriate link functions (Jansen *et al*., 2008; Welton *et al*., 2008; Dias *et al*., 2011a, 2013b).

[a]*Department of Mental Health, Johns Hopkins University, Baltimore, MD 21205, USA*
[b]*Division of Biostatistics, University of Minnesota, Minneapolis, MN 55405, USA*
[c]*Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, USA*
**Correspondence to: Hwanhee Hong, Department of Mental Health, Johns Hopkins University, Hampton House 810, 624 N. Broadway, Baltimore, MD 21205, USA.*
E-mail: hhong@jhu.edu

Two key issues in MTCs are heterogeneity on the treatment effects across studies and inconsistency, commonly defined as between-trial variability and apparent discrepancy between direct and indirect comparisons, respectively (Lumley, 2002; Lu and Ades, 2006; Higgins *et al*., 2012). We can model heterogeneity by assigning a distribution (usually an exchangeable normal) to the treatment effects across studies. For three treatments, a consistency equation can be defined as $m_{BC} = m_{AC} - m_{AB}$, where m_{AB} is the relative effect between treatments *A* and *B*, and inconsistency arises when this equality does not hold (Lumley, 2002; Lu and Ades, 2006; Cooper *et al*., 2009; Dias *et al*., 2010; Lu *et al*., 2011; Piepho *et al*., 2012). Dias *et al*. (2011b) and Jansen and Naci (2013) explain that inconsistency is another form of conflict among the means of the treatment effects, which should not exist unless the trials differ with respect to effect modifiers, such as experimental settings or patients' characteristics. Inclusion of multi-arm trials complicates the identification and interpretation of inconsistency, and the sparsity of the data can lead to large uncertainty (i.e., wide 95% credible sets) regarding inconsistency, making it difficult to distinguish from heterogeneity.

As many studies report multiple outcomes measured on the same subjects, correlations between both outcomes and treatments should be incorporated in an MTC model. For example, similar types of drugs or medical devices may tend to behave similarly, causing correlated results, and multiple outcomes can also induce correlations (e.g., negative correlation between efficacy and safety outcomes). Lu and Ades (2009) and Franchini *et al*. (2012) discuss modeling the correlation across treatments with a single outcome. Riley *et al*. (2007) and Jackson *et al*. (2011) discuss when one can estimate within-study correlations between outcomes in bivariate random effect meta-analysis, and thus produce estimates with smaller standard errors than in the independence setting.

In MTC data with multiple continuous outcomes, we usually observe only the sample mean and sample variance in each arm. However, sample correlations between outcomes are rarely reported for aggregated MTC data. Recently published papers have addressed this issue of unknown within-study correlations, and some recommended Bayesian approaches with informative priors. Efthimiou *et al*. (2014) develop a method to elicit expert opinion about the within-study correlation between two dichotomous outcomes. However, they only estimate a within-study correlation for a study reporting both outcomes, or else they impute a missing outcome for each study by assuming zero within-study correlation, instead of borrowing information from the estimated correlations among other studies. Bujkiewicz *et al*. (2013) use external individual patient-level data to construct an informative prior for the within-study correlation between mixed outcomes, one continuous and one binary. This method is obviously limited to the case where strong external evidence is available. Wei and Higgins (2013) investigate a pairwise meta-analysis model for two outcomes, but applying their method to an MTC setting appears too complicated. Schmid *et al*. (2014) investigate an MTC model for unordered categorical data with partially missing event counts for multiple outcomes.

As most randomized controlled trials compare only two or three treatment arms, the result is often extremely sparse MTC data from the perspective of missing data analysis. Although a standard MTC approach (e.g., Lu and Ades (2006)) models the observed data, we can gain additional information from the incomplete records. Suppose we calculate the missingness rate as the ratio of the summation of the number of missing arms and the total number of treatments times the number of studies times the number of outcomes. Then, when we compare five treatments for a single outcome, the missingness rate is typically 40% to 60% and could increase above 70% if 10 treatments are considered. We can impute such missing components based on a Bayesian hierarchical model that accounts for between-treatment and between-outcome correlations using Markov chain Monte Carlo (MCMC) algorithms instead of using observed data only. Especially when the missingness does not occur 'completely at random', but depends on some observed or unobserved information, ignoring such missing data can lead to biased estimators (Little and Rubin, 2002). For example, some treatments are discarded over time as accumulating evidence suggests they are inferior. Alternatively, study arms might be chosen to satisfy regulatory or reimbursement requirements, or even to further a company's marketing strategy.

In this paper (extending earlier work of Hong *et al*. (2013a)), we propose a new MTC modeling approach incorporating a missing data framework and compare this with the broadly used Lu and Ades (LA)-style approach. Both approaches can be applied to a generalized linear model framework, but our own models can more easily and flexibly incorporate correlations between treatments and outcomes. We model the correlation structure at the random effect level, instead of imposing such correlation structure into our likelihood. By doing so, our random effects can easily borrow information across outcomes and studies. In addition, our methods do not require any external data (although informative priors for the covariance matrices remain welcome) and impute missing data using full posterior inference, rather than by assigning unknown correlations arbitrarily. In addition, we present our methods with two parameterizations, contrast-based parameterization and arm-based parameterization, and discuss assumptions, advantages, and limitations of each parameterization.

The remainder of this paper is structured as follows. First, Section 2 describes our motivating data considering the bivariate continuous outcomes case. Section 3 provides details of our Bayesian missing data hierarchical modeling framework for MTCs under various assumptions to accommodate missing data and multiple outcomes. Section 4 reports the results of simulation studies validating our approaches, while Section 5 delivers the results of our analysis of the real data. Finally, Section 6 discusses our work, its limitations, several controversial topics in MTCs, and unmet methodological challenges.

2. Motivating data: knee pain osteoarthritis data

Figure 1(a) and (b) exhibits the trial network among physical therapy interventions for community-dwelling adults with knee pain secondary to osteoarthritis (OA) in terms of pain and disability outcomes, respectively (Shamliyan *et al*., 2012). A total of 54 randomized controlled trials are included to compare 11 therapies: no treatment, education, placebo, low-intensity diathermy, high-intensity diathermy, electrical stimulation, aerobic exercise, aquatic exercise, strength exercise, proprioception exercise, and ultrasound, coded 1 to 11 in that order. Each study reported sample means of standard scores to measure the level of pain only (28 studies), disability only (3 studies), or both outcomes (23 studies). The size of each node represents the number of studies investigating the drug, while the thickness of each edge denotes the total number of samples for the comparison. The numbers on the edges indicate the numbers of studies investigating the comparison (Chaimani *et al*., 2013). The network features are similar for both outcomes, but we have limited information on the disability outcome, with fewer connections between therapies and smaller total sample sizes overall than for the pain outcome.

The target population was pre-defined as adults with OA in outpatient settings who have had OA symptoms for at least 2 months. That is, the 54 studies were selected under several strict inclusion and exclusion criteria, so it is reasonable to assume their study populations are similar. Given literature concluding the various different OA-reporting systems are roughly equivalent to each other; we rescaled them to have the same, comparable range, 0 to 10 (e.g., for a score having range 0 to 100, we simply divided the sample mean and standard deviation by 10). Detailed description of the inclusion/exclusion criteria and outcomes with the full OA data are available in Hong *et al*. (2013a). The OA data themselves are also available in the Supporting Information of this article.

3. Methods

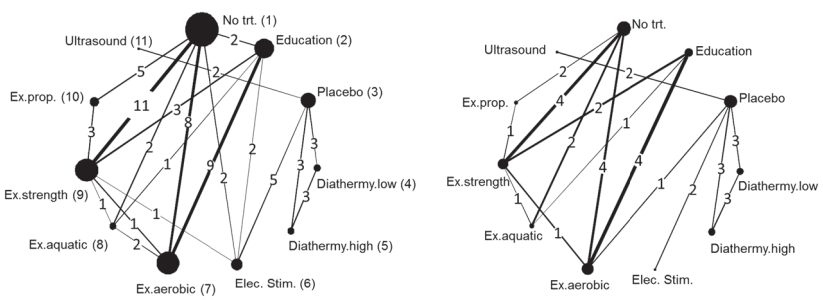
3.1. Likelihood

In MTCs, we must carefully distinguish between the terms *treatment* and *arm*. The former refers to a drug or device being tested, while the latter to the data on patients randomized to a particular drug or device in a *single* study. We must also distinguish between *reference* and *baseline* treatments. The reference treatment is a standard control treatment (often placebo, or simply no treatment), which can be compared with other active treatments. For our OA data, we take no treatment as the reference treatment among three possible choices (no treatment, education, and placebo). The baseline treatment is defined as the treatment assigned as the control arm *in each study*. That is, each study has its own baseline treatment, which is often the same as the reference treatment, but could differ.

Suppose we are comparing K treatments from I studies in terms of L outcomes. For any type of aggregated-level (i.e., summarizing over individuals) MTC data, we assume that the observations for a specific outcome from each study arise conditionally independently from a parametric statistical distributional model. Ignoring covariates for the moment, we write the distribution as

$$Y_{ik\ell} \sim f_{Y_\ell}(y_{ik\ell}|\Delta_{ik\ell}, \xi_{ik\ell}), \quad i = 1, \ldots, I,\ k = 1, \ldots, K,\ l = 1, \ldots, L, \tag{1}$$

where $Y_{ik\ell}$ is the observed aggregated outcome, $f_{Y_\ell}(\cdot)$ is a density function having an unknown parameter $\Delta_{ik\ell}$ (true population location parameter in the context of MTCs) and an assumed-to-be-known nuisance (typically variance) parameter $\xi_{ik\ell}$ in the k^{th} treatment arm from the i^{th} study with respect to the ℓ^{th} outcome. For example, when the measurements $Y_{ik\ell}$ are continuous, $Y_{ik\ell}$ often follows a normal density $f_{Y_\ell}(\cdot)$ with unknown true mean $\Delta_{ik\ell}$ and known standard deviation $\xi_{ik\ell}$. We let $k = 1$ index the reference treatment.



(a) pain (b) disability

Figure 1. Network graphs of osteoarthritis data: (a) pain outcome and (b) disability outcome. The size of each node represents the number of studies investigating the drug, the thickness of each edge denotes the total number of samples for the comparison, and the numbers on the edges indicate the numbers of studies investigating the comparison.

3.2. Generalized Lu and Ades model

In this paper, we only consider random treatment effects models, although a fixed treatment effects model can easily be implemented. For the Lu and Ades (2004, 2006) model framework, we observe a total of M data points (i.e., sum of the number of observed arms in every study), and we estimate random effects by using only those observed data. Extending Lu and Ades (2004, 2006) to the GLM case (Dias *et al*., 2013b), we enhance (1) with a linear predictor and link function for the unknown parameter $\Delta_{ik\ell}$, namely

$$g_\ell(\Delta_{ik\ell}) = \theta_{ik\ell} = \alpha_{iB\ell} + \delta_{iBk\,\ell}, \tag{2}$$

where $g_\ell(\cdot)$ is the known link function, such as the identity, logit, or log link for a continuous, binary, or count response, respectively, and $\theta_{ik\ell}$ is the linear predictor (McCulloch *et al*., 2008; again we assume no covariate at the moment). Given that B indicates the baseline treatment in each study i, $\alpha_{iB\ell}$ are the baseline treatment effects, and $\delta_{iBk\ell}$ are the random effects of contrasts (estimated only by observed data) between treatment k and the baseline treatment B for outcome ℓ in study i, with $\delta_{iBk\ell}\equiv 0$ if $k = B$. We define $d_{jk\ell}$ as the fixed mean parameter of contrasts between treatments k and j for outcome ℓ, with $d_{jk\ell}\equiv 0$ when $j = k$. We infer the treatment effects in terms of $d_{1k\ell}$, always comparing the k^{th} treatment with the reference treatment, so that we need to only assign a prior to the $d_{1k\ell}$ to obtain $d_{jk\ell} = d_{1k\ell} - d_{1j\ell}$ under the consistency assumption (Lu and Ades, 2006).

We often assume homogeneous variance across random effects for all two-arm studies, that is,

$$\delta_{iBk\,\ell} \sim N\left(d_{1k\ell} - d_{1B\ell}, \tau_\ell^2\right), \quad \text{for } k \neq B. \tag{3}$$

Again, $\delta_{iBk\ell}$ is 0 when $k = B$, and τ_ℓ is the standard deviation of the random contrast effects for outcome ℓ. We denote this model as the LA-style homogeneous random effects model (LAREhom). For multi-arm trials, the $\delta_{iBk\ell}$ in (3) are replaced by a vector that follows a multivariate normal distribution with dimension equals to the number of arms in study i minus one, for each outcome ℓ. Here, the between-arm-contrast correlation is 0.5 as a consequence of the homogeneous variance and consistency assumptions (Higgins and Whitehead, 1996; Lu and Ades, 2006).

Lu and Ades-style models always assume fixed study-specific baseline effects $\alpha_{iB\ell}$, generally considering them to be nuisance parameters. That is, the estimation and interpretation of baseline effects are not of interest in LA models. However, they may well be of interest in many situations and may often be sensibly assumed to be exchangeable across studies (Hong *et al*., 2013b). Moreover, a hierarchical Bayes model requires all parameters to have valid interpretations, with nuisance parameters integrated out from joint posterior distributions, in order for the resulting marginal posteriors of the remaining parameters of interest to be valid (Carlin and Louis, 2009).

3.3. Modeling for missing data and correlations between treatments and outcomes

We denote a model that parameterizes *relative* effect (e.g., mean difference or odds ratio) as a *contrast-based* (CB) model and *absolute* effect (e.g., mean effect or odds) as an *arm-based* (AB) model. LA-style models use a CB approach. Note that the mean effect of random contrasts between treatment k and reference treatment is the canonical parameter in CB models, whereas for AB models, it is the mean of the random treatment effects for treatment k, for each outcome ℓ. Although CB models are widely applied in MTC settings, AB models offer advantages in ease of interpretation, prior specification, and model fitting. Moreover, absolute measures of effect will often be of genuine interest, for example, the absolute amount of reduction in blood glucose produced by a given diabetes treatment. CB models assume exchangeability across relative effects, a weaker assumption than the AB models' assumption of exchangeability among absolute effects. Some researchers think that random baselines and arm-based modeling 'break the randomization' (because they assume treatment arms are exchangeable across studies) and are unlikely to hold in practice (because trial populations will often not be exchangeable). Note that the terms 'contrast-based' and 'arm-based' are used differently by other authors (Dias *et al*., 2011a; Salanti *et al*., 2008; Franchini *et al*., 2012), for example, to indicate different types of *data entry* for meta-analysis: a trial could report only the relative measurements between arms (contrast-based) or measurements for each arm (arm-based).

However, LA-style models have some limitations as well. It is common that the number of treatments compared in the i^{th} study is far less than the complete collection of K treatments in the given MTC dataset. Under the LA model framework, because each study contributes to the likelihood for a different set of treatments, using only the observed measurements can complicate specifying the *unstructured* covariance matrix of the random effects $\delta_{iBk\ell}$ (i.e., assuming uncommon random effect variability across treatments) leading to difficulties in prior assignment and parameter inference for multi-arm trials. That is, different treatments are compared in different studies, and not all studies investigate all arms. For example, suppose two trials investigate three treatments but each trial has a different set of compared treatments. The length of $\delta_{iBk\ell}$ vector is two, the same for both studies, and each vector follows a bivariate normal distribution with a 2×2 covariance matrix. However, it might be difficult to assign a prior distribution (e.g., a conjugate inverse Wishart specification) on the covariance matrix because the two covariance matrices could contain partially or fully different information, even though their dimensions are the same. Lu and Ades (2009) discuss modeling heterogeneous random effect variability, but their

recommendations seem too complicated to extend to multiple outcomes. Note that this is not an issue when we assume common random variability across treatments.

In addition, $\alpha_{iB\ell}$ is considered to be a nuisance parameter under the LA model framework and is typically left uninterpreted by CB modelers. However, NMA baseline effects are getting more attention as a way of understanding the baseline natural history or baseline risk (Dias *et al*., 2013a; Achana *et al*., 2013). Dias *et al*. (2013a) caution against assuming exchangeable baseline effects, although Achana *et al*. (2013) found that their results were not too influenced by this assumption in their examples. Finally, researchers may select study arms based on the trials conducted previously, resulting in missingness that may be missing at random when those preceding trials are included as a part of the data, or missing not at random when they excluded from the data.

To remedy this, we assume that all studies can in principle contain arms for every treatment, but in practice, much of this information is missing for various reasons. Thus, we will impute such unobserved arms by considering them as unknown parameters to be imputed along with the other model unknowns. Under this assumption, all studies can always share a common (although possibly missing) baseline treatment, $B = 1$ in (2).

3.3.1. Contrast-based approach. Under our missing data framework, now $\Delta_{ik\ell}$ in (1) contains all true means from both *M* observed and *IKL-M* unobserved data points. We propose three contrast-based random effects models, denoted by CBRE1, 2, and 3, under three different assumptions, and they can be written as

$$\text{CBRE1, 2}: \; g_\ell(\Delta_{ik\ell}) = \theta_{ik\ell} = \alpha_{i1\ell} + d_{1k\ell} + \eta_{i1k\ell}, \tag{4}$$

$$\text{CBRE3}: \; g_\ell(\Delta_{ik\ell}) = \theta_{ik\ell} = \alpha_{i1\ell} + d_{1k\ell} + v_{ik} + \omega_{i\ell}. \tag{5}$$

Note that in (4), $\eta_{i1k\ell}$ are the random effects of contrasts, distinct from $\delta_{iBk\ell}$ in (2), $\eta_{i1k\ell}$ contains both observed and unobserved information, and $d_{1k\ell}$, $\eta_{i1k\ell}$, v_{ik}, and $\omega_{i\ell}$ set to be zero when $k = 1$ for all ℓ. The CBRE1 model assumes independence between outcomes, and $\eta_{i\ell} \sim N_{K-1}(0, \Sigma_\ell)$, where $\eta_{i\ell} = (\eta_{i12\ell}, \ldots, \eta_{i1K\ell})^T$ and Σ_ℓ is a $(K-1)\times(K-1)$ unstructured covariance matrix for $\ell = 1, \ldots, L$. Here, we cannot adopt a common Σ^* instead of Σ_ℓ because the scale or range of the data could differ across outcomes. Alternatively, for CBRE2, we can allow correlation among outcomes by respecifying $\eta_{ik} \sim N_L(0, \Lambda_k)$, where $\eta_{ik} = (\eta_{i1k1}, \ldots, \eta_{i1kL})^T$ and Λ_k is a $L \times L$ unstructured covariance matrix for $k = 2, \ldots, K$. In this model, we assume independent random contrasts between treatments but incorporate the correlation structure of those contrasts between outcomes through Λ_k. Unlike CBRE1, we can replace the Λ_k with a single Λ^*, which gives homogeneous variance across treatments, but heterogeneous variance across outcomes. Here, the number of random effects in the CBRE1 and 2 models is *I(K-1)L*, pretty large compared with the size of data. However, these random effects share common covariance matrices, so they can borrow information from each other and remain Bayesianly estimable.

To reduce the number of random effects and incorporate both treatment-wise and outcome-wise correlations, we partition random effects into two independent sources in (5) and assume $v_i = (v_{i2}, \ldots, v_{iK})^T \sim N_{K-1}(0, \Sigma)$ and $\omega_i = (\omega_{i1}, \ldots, \omega_{iL})^T \sim N_L(0, \Lambda)$. Here, Σ and Λ are $(K-1)\times(K-1)$ and $L \times L$ unstructured matrices capturing correlation between treatments and outcomes, respectively, and only $I(K-1) + IL$ random effects are estimated. In addition, the total number of parameters in Σ and Λ is smaller than that in Σ_ℓ under CBRE1 when *K* is large and *L* is small, which is commonly observed in a MTC dataset. We denote this parsimonious random effects model as CBRE3. Note that because the unstructured covariance matrices are positive definite (thanks to the Wishart prior, which will be discussed in Section 3.4), all CBRE models respect the second-order consistency condition defined by Lu and Ades (2009), see the Supporting Information. Recently, White *et al*. (2012) suggested a similar approach that is based on imputation (but only for the reference treatment) and implemented their frequentist (not Bayesian) models in Stata. A similar imputation strategy for the baseline adjustment is also suggested by Dias *et al*. (2013a) and Achana *et al*. (2013).

In our proposed CB methods, because we always have $\eta_{i\ell}$, η_{ik}, v_i, and ω_i, vectors of the same length for every study *i*, common covariance matrices can be used for every study with a simple prior assignment. In addition, our modeling allows us to incorporate all sources of uncertainty by considering unobserved arms as missing data to be imputed by our MCMC algorithm. Suppose Study 1 compares treatments 1, 2, and 3, with respect to a single outcome, so that it provides information about two contrasts, η_{112} and η_{113}, under the notation of (4) and dropping the ℓ subscript. Suppose further that Study 2 compares only treatments 1 and 2, and Study 3 includes only treatments 1 and 3. Here, we assume that $(\eta_{i12}, \eta_{i13})^T \sim N(0, \Sigma)$. In this simple example, we can impute the missing contrasts η_{213} and η_{312} using Gibbs sampling with the following steps:

Step 0. Specify initial values $\eta^{(i)}, \Sigma^{(i)}$, and $\lambda^{(i)}$, where η is the set of random effect parameters and λ is the set of other unknown parameters. Set $i = 0$.

Step 1. Draw a value $\eta_{213}^{(i+1)}$ from the full conditional posterior distribution $p\left(\eta_{213} | \eta_{312}^{(i)}, \Sigma^{(i)}, \lambda^{(i)}, \eta_{112}^{(i)}, \eta_{113}^{(i)}, \eta_{212}^{(i)}, \eta_{313}^{(i)}, \mathbf{y}\right)$, where y is the observed data

Step 2. Draw a value $\eta_{312}^{(i+1)}$ from $p\left(\eta_{312} | \eta_{213}^{(i+1)}, \Sigma^{(i)}, \lambda^{(i)}, \eta_{112}^{(i)}, \eta_{113}^{(i)}, \eta_{212}^{(i)}, \eta_{313}^{(i)}, \mathbf{y}\right)$

Step 3. Draw a value $\Sigma^{(i+1)}$ from $p\left(\Sigma | \eta_{213}^{(i+1)}, \eta_{312}^{(i+1)}, \lambda^{(i)}, \eta_{112}^{(i)}, \eta_{113}^{(i)}, \eta_{212}^{(i)}, \eta_{313}^{(i)}, \mathbf{y}\right)$

Step 4. Draw a value $\theta^{(i+1)}$ from $p\left(\theta|\eta_{213}^{(i+1)},\eta_{312}^{(i+1)},\lambda^{(i+1)},\eta_{112}^{(i)},\eta_{113}^{(i)},\eta_{212}^{(i)},\eta_{313}^{(i)},\mathbf{y}\right)$

Step 5. Similarly, draw values $\eta_{112}^{(i+1)},\eta_{113}^{(i+1)},\eta_{212}^{(i+1)}$, and $\eta_{313}^{(i+1)}$

Step 6. Set $i=i+1$ and repeat Steps 1 to 5 until all MCMC chains converge.

Note that the order of the draw does not matter. Similarly, we can apply the same algorithm with multiple outcomes borrowing information from the relationships between outcomes. Finally, in our CB approach, $\alpha_{i1\ell}$ becomes meaningful because the baseline treatment is the same across all studies.

3.3.2. Arm-based approach. Random effects models with the AB parameterization analogous to those for three CBRE models can be written by respecifying models (4) and (5) as

$$\text{ABRE1, 2}: g_\ell(\Delta_{ik\ell}) = \theta_{ik\ell} = \mu_{k\ell} + \eta_{ik\ell}, \tag{6}$$

$$\text{ABRE3}: g_\ell(\Delta_{ik\ell}) = \theta_{ik\ell} = \mu_{k\ell} + v_{ik} + \omega_{i\ell}, \tag{7}$$

where $\mu_{k\ell}$ is the fixed mean effect of treatment k associated with the link function $g_\ell(\cdot)$ with respect to outcome ℓ and $\eta_{ik\ell}$ is the study-specific random effect.

If we begin by assuming independent random effects between outcomes, the random effects $\eta_{ik\ell}$ in (6) will have the following distribution: $\eta_{i\ell}\sim N_K(0,\Sigma_\ell)$, where $\eta_{i\ell}=(\eta_{i1\ell},\ldots,\eta_{iK\ell})^T$ and Σ_ℓ is a $K\times K$ unstructured covariance matrix for $\ell=1,\ldots,L$. We denote this model as ABRE1. Similarly to CBRE2, we can instead allow dependence of random effects between outcomes but independence between treatments by defining $\eta_{ik}=(\eta_{ik1},\ldots,\eta_{ikL})^T\sim N_L(0,\Lambda_k)$ where Λ_k is a $L\times L$ unstructured covariance matrix capturing relations between outcomes, for $k=1,\ldots,K$. We refer to this model as ABRE2. Again, we would not likely adopt a common Σ^* instead of Σ_ℓ in ABRE1, whereas we might well use Λ^* instead of Λ_k in ABRE2.

As in the CBRE3 model, we partition the $\eta_{ik\ell}$ into two independent sources as in (7), denoted by ABRE3, and assign the following distributions: $v_i=(v_{i1},\ldots,v_{iK})^T\sim N_K(0,\Sigma)$ and $\omega_i=(\omega_{i1},\ldots,\omega_{iL})^T\sim N_L(0,\Lambda)$. Here, Σ and Λ are $K\times K$ and $L\times L$ unstructured covariance matrices. Again, ABRE3 has not only fewer parameters to be estimated but also fewer constraints because relationships between treatments and outcomes are modeled simultaneously, yet independently. In the Supporting Information, we derive the full correlation matrix under the ABRE3 model.

The parameters in arm-based models permit more straightforward interpretation, especially in estimating an absolute treatment effect. However, these models do require fairly strong assumptions regarding similarity and exchangeability between arms across all populations, in order to permit meaningful clinical inference. In fact, CB models also require exchangeability, but among treatment contrasts, not arms, and this weaker assumption is more readily accepted by the meta-analysis community. Thus, this could be an advantage of contrast-based models in settings where we doubt cross-trial similarity of like-named arms. However, a possible advantage of AB over CB models is that AB computation is relatively straightforward because it avoids the additional baseline effect parameters ($\alpha_{i1\ell}$) and the $d_{11\ell},\eta_{i11\ell}\equiv 0$ constraints. Note also that in AB models, all of our random effect covariance matrices are unstructured. That is, AB models not only are less constrained but also have a slightly larger number of parameters to estimate than CB models.

3.4. Prior distributions

We assume as noninformative a prior as possible, in order to let the data dominate the posterior calculation. Specifically, a $N(0,100^2)$ is used for $\alpha_{iB\ell}$ and $d_{1k\ell}$, and a vague uniform distribution, namely, *Uniform*$(0.01,10)$, is assigned to τ_ℓ in the LAREhom model (Gelman, 2006). In our proposed CB models, we may choose to assume *exchangeable* baseline effects by assigning $\alpha_{i1\ell}\sim N\left(a_\ell,\tau_{\alpha,\ell}^2\right)$, where the hyperparameters a_ℓ and $\tau_{\alpha,\ell}$ follow noninformative normal and uniform distributions, respectively. Throughout all CB and AB models, the canonical parameters ($d_{1k\ell}$ and $\mu_{k\ell}$, respectively) follow a $N(0,100^2)$ prior distribution, while all inverse covariance matrices follow a *Wishart*(Ω,γ) having a mean of $\gamma\Omega^{-1}$ and degrees of freedom parameter γ to be the matrix dimension, because this is the smallest value that will still yield a proper prior. For our OA data, we choose Ω to be 5γ times the identity matrix, yielding a 95% prior credible interval from 3.33 to 253 for the mean standard deviation of the random effects, in order to be fairly noninformative while ensuring MCMC convergence (Carlin and Louis, 2009).

We use WinBUGS to generate two parallel chains of 50,000 MCMC samples after a 50,000-iteration burn-in. To check MCMC convergence, we used standard diagnostics, including trace plots and lag 1 sample autocorrelations. Independence and normality were checked using plots of standardized residuals, which emerged as scattered randomly around 0 between -2 and 2 under all models (Lunn *et al.*, 2012). The WinBUGS programs for the LAREhom, CBRE3, and ABRE3 models are available in the Supporting Information.

4. Simulation study

4.1. Setting

In this simulation study, we will investigate how well our missing data approaches perform under different missingness mechanisms compared with the LA method in terms of bias, mean squared error (MSE), and coverage probability. The outline of our simulation study is

1. Generate data under the ABRE3 model framework, fit the LAREhom, CBRE3, and ABRE3 models, and compare bias, MSE, and coverage probability of absolute treatment effect estimates
2. Generate data under the LAREhom model framework, fit the LAREhom, CBRE3, and ABRE3 models, and compare bias, MSE, and coverage probability of relative treatment effect estimates

A simulated dataset (y_{ik1}, y_{ik2}) has 30 studies ($i = 1, \ldots, 30$) comparing three treatments, denoted 1, 2, and 3, for two continuous outcomes, and the number of iterations is 500. We consider four different data structures illustrated in Figure 2.

We start from simulating complete data (denoted complete) as in panel (a) of Figure 2. Under the ABRE3 framework in (7), we choose the identity link function and set $(\mu_{11}, \mu_{21}, \mu_{31}) = (2, 0, 1)$, $(\mu_{12}, u_{22}, \mu_{32}) = (-1, 2, 1)$, $(v_{i1}, v_{i2}, v_{i3})^T \sim N_3(0, \Sigma)$, where Σ has variances 1 and a common between-treatment correlation 0.5, and $(\omega_{i1}, \omega_{i2})^T \sim N_2(0, \Lambda)$,where Λ has variances 1 and a between-outcome correlation ρ for the true parameters. We consider ρ to be 0.2 and 0.8. Next, under the LAREhom framework in (2), we again choose the identity link function and let the true fixed study-specific baseline effect $\alpha_{iB\ell}$ be sampled as *Uniform*$(0, 4)$ and *Uniform*$(-3, 1)$ for $\ell = 1$ and 2, respectively. So, we assign somewhat arbitrary numbers to $\alpha_{iB\ell}$, and the ranges of the uniform distributions are selected to have means to be the same as μ_{11} and μ_{12} and variances around 1.3. Because LAREhom does not model correlation between outcomes, we borrow the CBRE3 model parameterization for random effects in (5) and set $(d_{111}, d_{121}, d_{131}) = (0, -2, -1)$, $(d_{112}, d_{122}, d_{132}) = (0, 3, 2)$, $(v_{i2}, v_{i3})^T \sim N_2(0, \Sigma)$, where Σ has variances 1 and a common between-treatment correlation 0.5, and $(a_{i1}, \omega_{i2})^T \sim N_2(0, \Lambda)$, where Λ has variances 1 and a between-outcome correlation ρ for the true parameters. Again, we set $\rho = 0.2$ and 0.8. For unequivocal investigation of the impacts of various missingness patterns, we set other features of our simulated data as simple as possible: namely, we assume that every study has a sample size of 200, and a standard deviation of 2 in every arm.

We create three more partially missing data structures by having two separate sets of five, ten, or fifteen studies drop one of their arms from the complete data, where one set compares Treatments 1 and 2 and the other compares Treatments 2 and 3, shown in panels (b), (c) and, (d) of Figure 2, as Missing1, 2, and 3, respectively. Note that the Missing 3 structure contains no multi-arm trials at all. We investigate three missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR with Missing1, 2, and 3 are denoted by MCAR1, 2, and 3, respectively, and similarly for MAR and MNAR (denoted MAR1, 2, 3 and MNAR1, 2, 3).

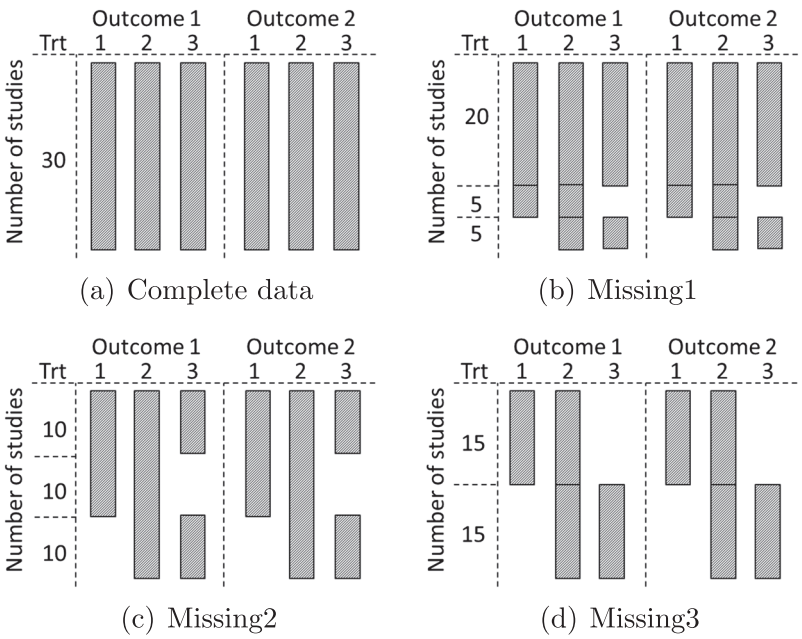


Figure 2. Data structures for simulation study: (a) complete data; (b) Missing1; (c) Missing2; and (d) Missing3.

We use vague prior distributions for all model parameters as described in Section 3.4. We also consider weakly informative priors for covariance matrices of random effects in ABRE3 and CBRE3 models; we set the mean of Wishart prior distribution equal to the true inverse covariance matrix and degrees of freedom γ to be 30. We used the BRugs package in R to perform our simulation studies, where we call OpenBUGS (Lunn *et al*., 2009) 500 times from R, once for each simulated dataset. In each case, we obtain 20,000 samples, after 5000-sample and 30,000-sample burn-in for LA-style models and our models, respectively. LA-style models needed a shorter burn-in period because they converge relatively quickly.

4.2. Results

Although we fit three models to the two generated data sets as described earlier, we only report partial results here because the results under CBRE3 and ABRE3 are similar regardless of the data generation. First, we consider the LAREhom and ABRE3 models with the data generated under the ABRE3 framework and compare estimates of the absolute treatment effect $\mu_{k\ell}$. Under the LAREhom model, $\mu_{1\ell}$, $\mu_{2\ell}$, and $\mu_{3\ell}$ are estimated by $E(\alpha_{i1\ell})$, $E(\alpha_{i1\ell})+d_{12\ell}$, and $E(\alpha_{i1\ell})+d_{13\ell}$, respectively. Second, we report results under LAREhom and CBRE3 models when the data are generated by the LAREhom framework and compare estimates of the relative treatment effect $d_{1k\ell}$, which is estimated by $\mu_{k\ell}-\mu_{1\ell}$ in the ABRE3 model. Note that LAREhom does not model correlation between outcomes. As such, although we simulate a dataset under the basic LAREhom framework, this model is still slightly misspecified for $\rho>0$.

Figures 3 and 4 plot the bias and MSE of μ_{k1} and d_{1k1} when the data are generated under ABRE3 [panels (a), (b), and (c)] and LAREhom [panels (d) and (e)], respectively, with $\rho=0.8$ and vague priors, across the 10 different missingness settings. We only report those parameters for the first outcome ($\ell=1$) because parameters associated with the second outcome produce results very similar to those for the first outcome. Note that under LAREhom models, $E(\alpha_{i1\ell})$ is calculated using studies including Treatment 1. In panels (a), (b), and (c) of Figures 3 and 4, both LAREhom and ABRE3 models yield almost zero bias under the Complete and MCAR settings, with slightly larger MSE under LAREhom. However, under MAR, LAREhom leads to significantly larger bias and MSE than ABRE3, and the pattern gets worse as the degree of missingness increases, whereas ABRE3 continues to produce unbiased estimates with small variabilities. Under MNAR, LAREhom performs worse than ABRE3 in terms of both bias and MSE. Compared with the MAR setting, ABRE3 produces somewhat larger bias and MSE than those under MNAR. Note that although every study observes the Treatment 2 arm, LAREhom fails to estimate μ_{21} correctly, while ABRE3 delivers performance equivalent to that seen with complete data.

In panels (d) and (e) of Figures 3 and 4, the results are similar under the Complete, MCAR, and MAR settings as in panels (a), (b), and (c). However, under MNAR, LAREhom gives a lot smaller bias and MSE for d_{121} and d_{131} than CBRE3. A similar trend is observed when we apply different logit models of missingness for MNAR. In particular, we tried three more sets of logit models: (1) $\left(\text{logit}\left(p_{i,mis}^{Trt1}\right)=\beta_{MNAR}^{Trt1}+2y_{i11}+y_{i12},\text{logit}\left(p_{i,mis}^{Trt3}\right)=\beta_{MNAR}^{Trt3}+y_{i31}-y_{i32}\right)$; (2) $\left(\text{logit}\left(p_{i,mis}^{Trt1}\right)=\beta_{MNAR}^{Trt1}+y_{i11}+y_{i31}+y_{i12}+y_{i32},\text{logit}\left(p_{i,mis}^{Trt3}\right)=\beta_{MNAR}^{Trt3}+y_{i11}+y_{i31}-y_{i12}-y_{i32}\right)$; and (3) $\left(\text{logit}\left(p_{i,mis}^{Trt1}\right)=\beta^{\ MNARTrt1}+y_{i11}+y_{i21}+y_{i31}+y_{i12}-y_{i22}-y_{i32},\text{logit}\left(p_{i,mis}^{Trt3}\right)=\beta_{MNAR}^{Trt3}+y_{i11}+y_{i21}+y_{i31}-y_{i12}+y_{i22}+y_{i32}\right)$. In addition, bias and MSE of d_{121} and d_{131} in LAREhom when the data are generated by ABRE3 under MNAR have the same patterns as seen in panels (d) and (e).

Recall that $d_{1k\ell}$ can be defined as $\mu_{k\ell}-\mu_{1\ell}$. After trying various logit models of missingness for MNAR, we find that the estimated $\mu_{k\ell}$ under LAREhom tends to be biased to the same degree, producing oddly unbiased $d_{1k\ell}$ estimates under MNAR. This is hard to interpret because the missingness mechanism behind MNAR is complicated and not intuitive compared with MAR. However, these results do not imply that LAREhom performs better than CBRE3 under MNAR because LAREhom does not correctly estimate $\mu_{k\ell}$ in panels (a), (b), and (c). Although CBRE3 imputes missing data given observed information, it does not model the missingness specifically and gives poorer estimates of both $\mu_{k\ell}$ and $d_{1k\ell}$ under MNAR than MAR.

Figure 5 displays coverage probabilities of 95% equal-tail credible intervals for parameters corresponding to those in Figures 3 and 4. In panel (a), the coverage probability of μ_{11} under LAREhom is below 0.2 across all simulation settings even when it is not biased. Recall μ_{11} is calculated by $E(\alpha_{i1\ell})$ under LAREhom. Because $E(\alpha_{i1\ell})$ is not a parameter but just the average of estimated $\alpha_{i1\ell}$, the 95% credible intervals for $E(\alpha_{i1\ell})$ under LAREhom are too narrow, resulting in the consistently low coverage probability. The coverage probability for μ_{11} under ABRE3 is around 0.95, the nominal coverage probability, except under MNAR. Panels (b) and (c) reveal that coverage probability gets lower as the bias gets larger, and estimates under LAREhom have low coverage probabilities under MAR and MNAR. Here, the estimates of d_{121} and d_{131} under LAREhom cover the narrow credible intervals for $E(\alpha_{i1\ell})$, and the LAREhom model yield coverage probabilities for μ_{21} and μ_{31} greater than 0.8 under Complete and MCAR, although they still cannot reach 0.95. In panels (d) and (e), coverage probabilities under CBRE3 are always around 0.95, except under MNAR, while LAREhom performs poorly only under MAR. Here, again, d_{121} and d_{131} estimates in LAREhom models under MNAR give coverage probabilities close to 0.95, resulting from the aforementioned oddly unbiased estimates.

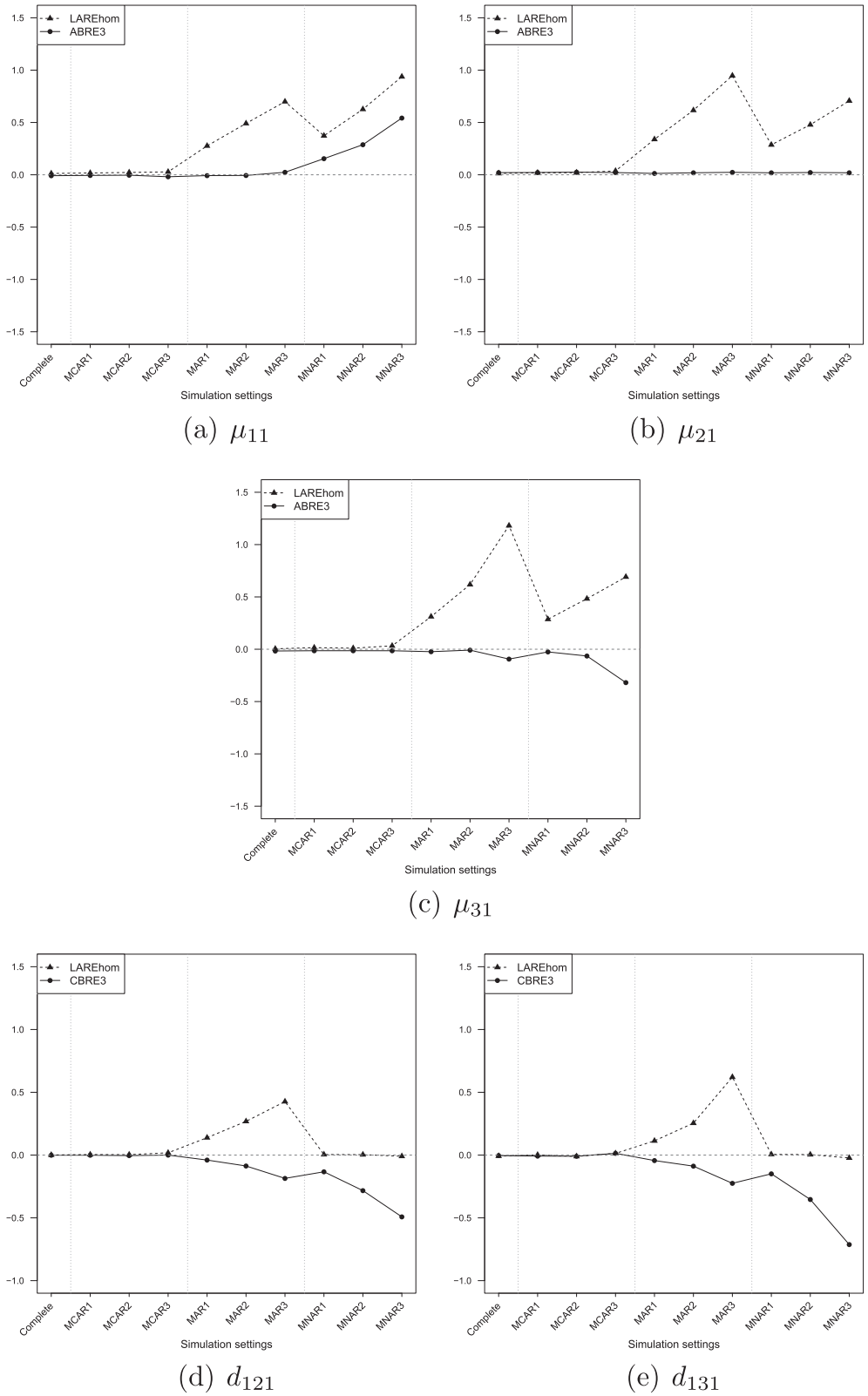


Figure 3. Bias of model parameters from the simulation study. Data are generated under the ABRE3 model framework for panels (a), (b), and (c) and under the LAREhom model framework for panels (d) and (e): (a) μ_{11}; (b) μ_{21}; (c) μ_{31}; (d) d_{121}; and (e) d_{131}.

With vague priors, the between-outcome correlation ρ is well estimated, while the between-treatment correlation is somewhat underestimated (estimates are around 0.2). Both correlations are correctly estimated when we apply weakly informative priors, but this does not directly affect the estimation of treatment effects

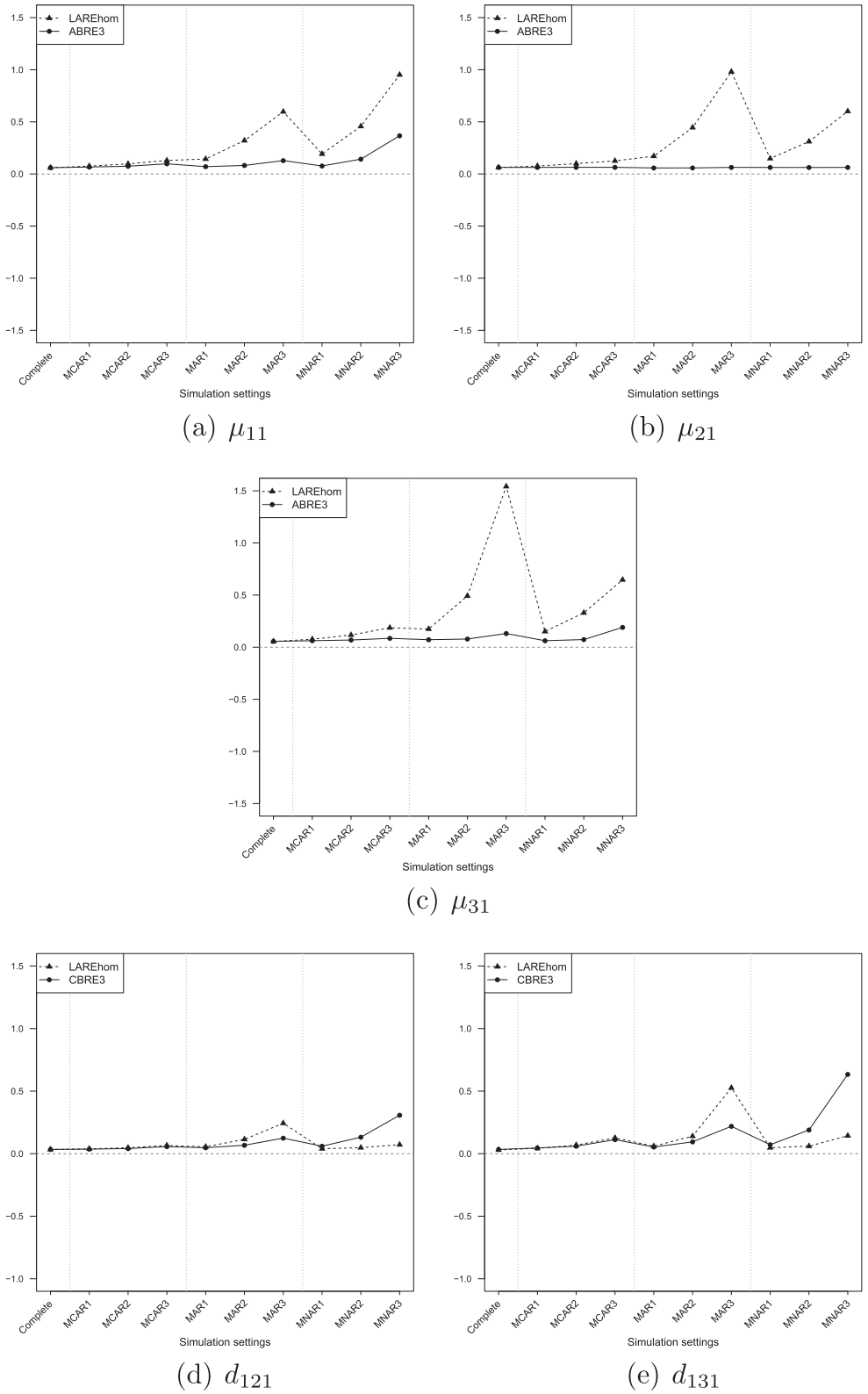


Figure 4. MSE of model parameters from the simulation study. Data are generated under the ABRE3 model framework for panels (a), (b), and (c) and under the LAREhom model framework for panels (d) and (e): (a) μ_{11}; (b) μ_{21}; (c) μ_{31}; (d) d_{121}; and (e) d_{131}.

and yield equivalent results to those under vague priors overall. Results are roughly the same when we generate 500 datasets under the setting $\rho = 0.2$.

Proponents of LA-style models might think that our simulation setup is favorable to our proposed models, because it employs random baseline effects. Although LA models do not adopt such effects, we used the same

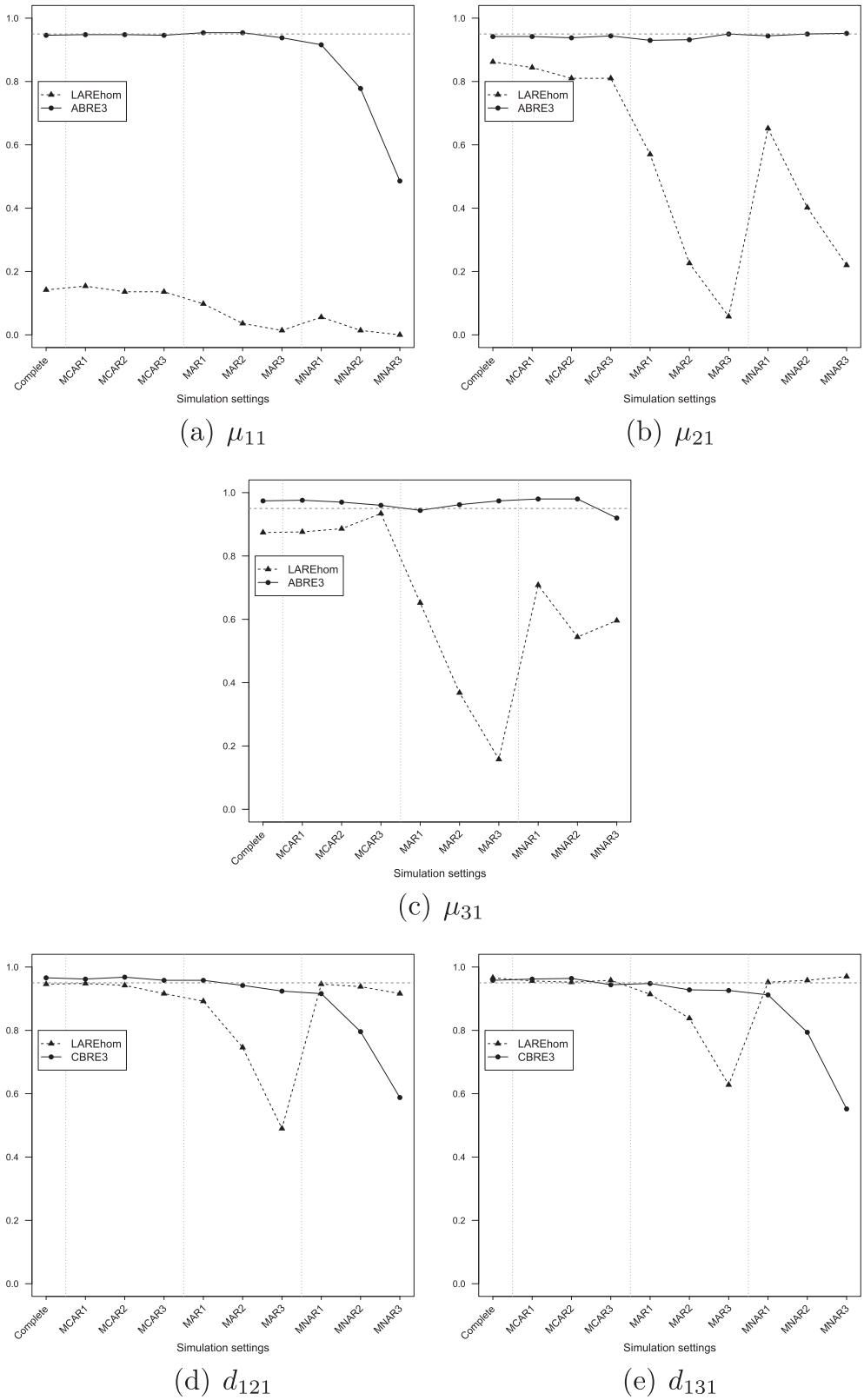


Figure 5. Coverage probabilities of model parameters from the simulation study. Data are generated under the ABRE3 model framework for panels (a), (b), and (c) and under the Lu and Ades-style homogeneous random effects model (LAREhom) model framework for panels (d) and (e): (a) μ_{11}; (b) μ_{21}; (c) μ_{31}; (d) d_{121}; and (e) d_{131}.

uniform distribution across all studies to sample these effects, a natural way to simulate MTC data. In addition, when we simulated data under LAREhom, we set the covariance matrix to have a homogeneous variance with 0.5 correlation – exactly what LAREhom fits, suggesting this simulation may be slightly favorable to LAREhom.

We acknowledge that LAREhom does not allow the between-outcome correlation while our proposed models do. However, we suspect this is not the reason that LAREhom performs poorly in our simulation study; rather, it is because LAREhom does not impute any missing arms into the model, inflating bias and MSE of the estimates under MCAR and MAR. In fact, we found the same patterns of bias and MSE when we only considered a single outcome with fixed baseline effects not sampled from any distribution. This alternate simulation setting and its results are added in the Supporting Information.

5. Osteoarthritis data analysis

We now use our models to analyze the OA data introduced in Section 2. Our OA data have two continuous outcomes, pain and disability ($\ell = 1$ and 2, respectively), and we can assume a normal likelihood for the data and an identity link function. That is, the likelihood (1) is $y_{ik\ell} \sim N\left(\Delta_{ik\ell}, \xi^2_{ik\ell}/n_{ik}\right)$, where $y_{ik\ell}$ is the observed sample mean of the pain and disability score change, $\Delta_{ik\ell}$ is the unknown true mean score change, $\xi^2_{ik\ell}$ is the known sample variance, and $n_{ik\ell}$ is the number of subjects in the k^{th} treatment arm from the i^{th} study with respect to the ℓ^{th} outcome. Under the identity link, we can replace $g_\ell(\Delta_{ik\ell})$ in all equations in Section 3 with $\Delta_{ik\ell}$. We fit seven models: LAREhom, CBRE1, CBRE2, CBRE3, ABRE1, ABRE2, and ABRE3, where the LAREhom model does not employ the missingness framework. Note that the LAREhom model assumes study-specific baseline effects and the three CBRE models assume exchangeable baseline effects. We apply homogeneous covariance matrices for CBRE2 and ABRE2; that is, a common Λ^* is applied, instead of a separate Λ_k for each k. All CB and AB models incorporate missingness; CBRE1 and ABRE1 allow correlation structure between treatments, CBRE2 and ABRE2 allow correlation structure between outcomes, and CBRE3 and ABRE3 permit correlations between outcomes and treatments simultaneously.

Table 1 compares the fit of the seven models with our OA data (Spiegelhalter *et al.* (2002)). CBRE3 fits the data best (smallest $\overline{D}$) while ABRE3 fits poorly (largest $\overline{D}$) because ABRE3 has not only fewer assumptions but also fewer parameters to be estimated than CBRE3, and this is shown as higher p_D in CBRE3. The remaining models give similar $\overline{D}$ values. All AB models except ABRE3 give very slightly higher p_D than the corresponding CB models because again, they are less constrained and more parameters need to be estimated. However, across three models under each CB or AB parameterization, CBRE3 and ABRE3 have the smallest p_D values because these models are slightly more parsimonious. Because our data are sparse and we lack sufficient data to accurately estimate all the treatment-by-treatment covariances, the heterogeneous variance assumption, a feature of CBRE1 and ABRE1, is not a good choice here. Considering both goodness of fit and complexity, CBRE3 or ABRE3 are the best overall, although the deviance information criterion (DIC) differences between either model and LAREhom are not of practical importance (less than five units).

Figure 6 exhibits posterior medians of $d_{1k\ell}$, the mean difference between therapy k and no treatment, with 95% Bayesian credible intervals (BCIs) for each outcome across five of the models (all but CBRE1 and ABRE1). We define $d_{1k\ell}$ in AB-style models as $\mu_{k\ell} - \mu_{1\ell}$. Smaller scores mean better conditions for both pain and disability. We indicate the best treatment with respect to each outcome in each model with a triangle character, and the worst treatment with a square, based on $\hat{d}_{1k\ell}$ posterior medians. Although the DICs of LAREhom, CBRE2, and ABRE2 models are similar, the estimates differ for some therapies; for the pain outcome, our CB and AB models agree that low and high intensity diathermy and ultrasound perform worse than no treatment, while LAREhom oddly reverses this conclusion (although the difference is not significant). As we expected, the 95% BCIs from LAREhom, CBRE2, and ABRE2 are narrower than those for CBRE3 and ABRE3 models. The estimated standard deviations of random effects (τ_ℓ in LAREhom; square roots of diagonal elements of Λ^* in CBRE2 and ABRE2) are always between 1 and 1.5 for these three models, with associated 95% BCI widths around 0.4. By contrast, CBRE3 and ABRE3 give a bit larger standard deviations, between 2 and 3.5 with 95% BCI widths mainly from 1.7 to 6. Under LAREhom, CBRE2, and ABRE2, proprioception exercise and strength exercise reduce pain significantly better than no active treatment, while aerobic exercise reduces pain significantly better than no treatment under LAREhom and CBRE2, but reduces disability significantly better than no treatment only under LAREhom. Compared with the pain outcome, the 95% BCIs for the disability outcome are wider because only about half as many studies reported this outcome, and the best therapy now varies across models.

Table 1. Model comparisons for osteoarthritis data analysis using DIC.

	LAREhom	CBRE1	CBRE2	CBRE3	ABRE1	ABRE2	ABRE3
$\overline{D}$	62.6	61.6	62.2	60.8	59.1	61.1	68.2
p_D	154.5	162.9	153.9	152.1	164.4	158.0	145.9
DIC	217.1	224.6	216.1	212.9	223.5	219.1	214.1

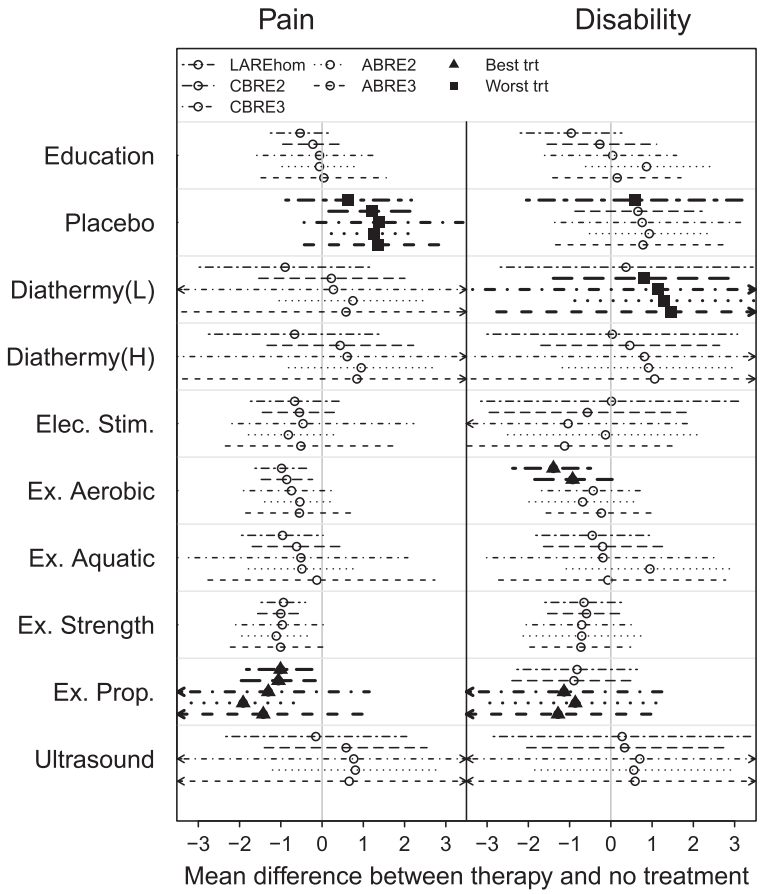


Figure 6. Posterior medians of $d_{1k\ell}$ with 95% Bayesian posterior credible sets for each outcome, five hierarchical models for the osteoarthritis data.

The posterior median correlations between the two outcomes are 0.49 (95% BCI: 0.18, 0.71) and 0.38 (0.06, 0.61) for the CBRE2 and ABRE2 models, respectively, suggesting mild positive linear association between reported pain and disability scores in both contrast-based and arm-based parameterizations. CBRE3 and ABRE3 give relatively small between-outcome correlation estimates, 0.30 (−0.22, 0.68) and 0.04 (−0.41, 0.48), respectively, because between-treatment correlations are also implied in these models (again, see the Appendix for the full calculation of the correlation matrix).

Using random baselines can make a material difference in the treatment effect estimates, although the degree of difference will change for different data sets. To investigate this, we fit the LAREhom, CBRE1 with random effects, and CBRE1 with fixed baseline effects (i.e., assuming $\alpha_{i1\ell} \sim N(0, 100^2)$) models to the OA data, considering the pain outcome only, and compare estimates of relative treatment effects. Comparing the two CBRE1 models, CBRE1 with random baseline effects flips the point estimates for low diathermy, high diathermy, and ultrasound to the other side of the null value and provides smaller standard errors for all estimates. LAREhom and CBRE1 with fixed baseline effects agree on the directions of treatment effects except for ultrasound (although the corresponding standard error estimates under both models are very large), but LAREhom yields smaller standard errors because the model does not acknowledge the missing data.

As a sensitivity analysis, we repeated our work under weakly informative Wishart priors for ABRE2 and 3 based on the estimates obtained under vague priors. For ABRE2, we assume 0.5 correlation between outcomes with mean variances 1.5 with 30 degrees of freedom; for ABRE3, we assume zero between-outcome correlation and mean variances 1, also with 30 degrees of freedom. DIC values decrease about a half unit, and standard deviations of posterior estimates tend to decrease a bit, but the estimated treatment effect parameters do not change much. This agrees with our simulation study results.

6. Discussion

In this paper, we have proposed Bayesian MTC approaches for multiple outcomes under a novel missing data framework and compared our results to those using standard LA hierarchical modeling methods. Our framework

handled varying outcome types using appropriate distributional models and link functions. We considered unobserved trial arms to be missing data and imputed them by borrowing information from the observed relationships in other trials. We also incorporated multiple outcomes into contrast-based and arm-based models through random effects with a variety of correlation structures. We used simulation to show that our imputation models can outperform standard models in terms of bias, MSE, and coverage probability of estimators under various missingness mechanisms. Finally, we illustrated our methods using real data. Note that this paper is a successor to that of Hong *et al*. (2013a), who focused on treatment ranking methods for multiple continuous outcomes and investigated the impact of the ignorance of outcome-wise correlation on hypothesis testing via simulation. By contrast, our paper focused on extending the Gaussian linear model to the GLM setting; also, its simulation studies sought to examine the effect of missingness pattern on bias and MSE of parameters of interest.

Several articles assume that missing arms occur at random (Caldwell *et al*., 2005; Giovane *et al*., 2013), but such missingness is not considered in the standard LA approach, implying that, technically, LA assumes missingness is *completely* at random. However, when missing data have any known patterns or relations to the observed (or even unobserved) data, the missingness mechanism should be taken into consideration. Although we do not explicitly specify the missingness pattern in our models, our imputation approach implicitly assumes MAR because we use only the observed relations to impute missing information. When the data are MNAR, this might be a sign of inconsistency. Explicit modeling of noningnorable missingness mechanisms in the Bayesian MTC modeling is an ongoing area of research.

One of the contentious issues in MTCs is the long meta-analytic tradition of mistrust for models with exchangeable baseline effects, a topic that has been discussed extensively for pairwise meta-analysis (van Houwelingen and Senn, 1999; Senn, 2010). Senn (2010) summarizes van Houwelingen's work and concludes that using random baseline effects could raise an issue of bias in theory, but the approach has some advantages, and the bias is likely to be small in practice. Achana *et al*. (2013) present work along the lines of our random baseline approach to model the association between baseline risk and effectiveness. They impute non-zero data for the unobserved control arms by assigning them a shared distribution. They acknowledge that the exchangeablility of baseline effects is a strong assumption and compromises the randomization. Similarly, Dias *et al*. (2013a) suggest modeling the baseline and relative effects separately and assume exchangeable baselines as one possible way of constructing a baseline history model. Dewilde and Hawkins (2012) treat the baseline effect as a nuisance parameter to ensure that the treatment effect estimates from MTCs are influenced only by the within-trial relative treatment effects, not by differences in absolute response across trials, and believe that this is the way to preserve unbiased estimates, a benefit of randomization. Dias *et al*. (2013b) caution that using random baselines might deliver biased relative treatment effect estimates unless the model is correct.

The advantage of randomized controlled trials as MTC inputs is that they control for the heterogeneity of arms across studies, thus delivering unbiased evidence on relative treatment effects. In meta-analysis, we first collect and screen randomized studies with several inclusion and exclusion criteria, which often concern characteristics of the target populations. Then, we conduct a meta-analysis (usually contrast-based) under the fundamental assumption that relative treatment effects are exchangeable across studies, and this approach is widely accepted by the meta-analysis community. However, specifically in the LA modeling, the baseline treatment *changes across trials*. This means the CB model's assumption of exchangeability of effects relative to an arbitrary baseline remains a lot to assume, and its failure need not logically preclude exchangeability of effects of a *common* (albeit sometimes imputed) baseline, as our AB models typically assume. This latter assumption will often be fair given that the MTC data arise from similar study populations, although we acknowledge the assumption under arm-based models is significantly stronger than that under contrast-based models. Unfortunately, both relative and absolute effect exchangeability assumptions would be flawed if the collected studies were heterogeneous and thus did not represent a common target population. For example, suppose that a drug performs much better with sicker people than with healthier people. In this case, the exchangeability of relative effects may not hold. In any case, while AB methods may risk more bias, the control of which is typically most important in NMA, the inherent statistical tradeoff between bias and variance (Carlin and Louis, 2009, Ch.1) means that AB methods may still outperform CB in terms of mean-squared error. Of course, the homogeneity of studies can be empirically checked if additional covariates are available, although sadly our OA data do not provide any such covariates.

Our methods have some limitations. First, exchangeability assumptions are typically not empirically verifiable. MTCs are used to pool data from a systematic review of the literature designed to answer a specific question, but the starting point assumes inclusion of all studies relevant for the target population of interest. Although our OA data example has a clear pre-defined target population, in general, AB inference should be made with caution because we lack information on any trials that were *not* selected for inclusion in the MTC. That is, individual studies typically enjoy internal validity, but possibly limited external validity across studies, especially when baseline patient population characteristics are poorly reported in publications, or there is sparse evidence on the population of interest. Plausibility of all NMA assumptions (including exchangeability) needs to be considered on a case-by-case basis. This issue, and the associated risk of selection bias, was not considered in this paper.

Second, our models sometimes result in slow MCMC convergence because we work with the full imputed missing data-parameter space. When data are sparse, some covariances in unstructured covariance matrices of random effects are estimated as prior means. Although all MCMC chains in our simulation study and data analysis

converge well under vague Wishart priors, these priors must be carefully selected to ensure MCMC convergence and correct estimation of variability, and alternatively structured covariance matrices should perhaps also can be considered. Lambert *et al*. (2005) and Gelman (2006) caution that Wishart and inverse-gamma priors on variance–covariance matrices can sometimes perform poorly and have computational problems. It is also difficult to ensure non-informativeness, because large variances are often associated with large correlations. The separation strategy of Barnard *et al*. (2000) may offer a reasonable alternative here.

In addition, for decision making, we compared only the estimated treatment effects for each outcome, but we could also have obtained the probability of being the best treatment by utilizing a weighted score across our multiple outcomes (Hong *et al*., 2013b).

Lastly, although we devoted scant attention to measuring evidence inconsistency, there have been several suggestions in this regard. Lu and Ades (2006) add an extra set of terms called *w-factors* into the model based on the MTC network graph's 'evidence loops', but these are hard to identify in the presence of multi-arm trials and varying baseline treatments across studies. Dias *et al*. (2010) suggest a *node-splitting* method that allows one to split the information estimating a model parameter into two distinct components: direct and indirect. The resulting two posterior distributions can then be examined for inconsistency. Presanis *et al*. (2013) and Piepho *et al*. (2012) define inconsistency as the interaction between trial types and treatments and then test consistency by conducting a global Wald test for interaction in a two-way linear mixed model.

Future work looks to extending our methods to mixed types of outcomes (say, a binary safety outcome paired with a continuous efficacy outcome). Another important future model enhancement is to the case of differential borrowing of strength across non-exchangeable subgroups, say determined by similarities across trials or treatments. Furthermore, we hope to extend our models to incorporate both aggregate and individual patient-level data, potentially permitting the borrowing of strength from patient-level covariates to investigate how those personal clinical characteristics impact estimated treatment effects.

Acknowledgements

Hwanhe Hong was supported by a grant from the Eli Lilly and Company Research Award Program. Haitao Chu was supported in part by the NIAID AI103012, NIDCR R03DE024750, NCI P30CA077598, and NIMHD U54-MD008620. Jing Zhang was supported in part by NIAID AI103012. Bradley P. Carlin was supported in part by the US NCI 1R01-CA157458-01A1 and NIAID AI103012.

References

Achana FA, Cooper NJ, Dias S, Lu G, Rice SJC, Kendrick D, Sutton AJ. 2013. Extending methods for investigating the relationship between treatment effect and baseline risk from pairwise meta-analysis to network meta-analysis. *Statistics in Medicine* **32**: 752–771.

Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. 2012. NICE DSU Technical Support Document 7: evidence synthesis of treatment efficacy in decision making: a reviewers checklist. Available at: http://www.nicedsu.org.uk. [Accessed on May 2015].

Barnard J, McCulloch R, Meng X. 2000. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**: 1281–1312.

Bujkiewicz S, Thompson JR, Sutton AJ, Cooper NJ, Harrison MJ, Symmons DPM, Abrams KR. 2013. Multivariate meta-analysis of mixed outcomes: a Bayesian approach. *Statistics in Medicine* **32**: 3926–3943.

Caldwell DM, Ades AE, Higgins JPT. 2005. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *BMJ* **331**: 897–900.

Carlin BP, Louis TA. 2009. *Bayesian Methods for Data Analysis*. 3rd edn. Boca Raton, FL: Chapman & Hall/CRC.

Chaimani A, Higgins JPT, Mavridis D, Spyridonos P, Salanti G. 2013. Graphical tools for network meta-analysis in STATA. *PLoS ONE* **8**(10): e76654. doi:10.1371/journal.pone.007

Cooper NJ, Sutton AJ, Morris D, Ades AE, Welton NJ. 2009. Addressing between-study heterogeneity and inconsistency in mixed treatment comparisons: application to stroke prevention treatments in individuals with non-rheumatic atrial fibrillation. *Statistics in Medicine* **28**: 1861–1881.

DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**: 177–188.

Dewilde S, Hawkins N. 2012. Investigating incoherence gives insight: clopidogrel is equivalent to extended-release dipyridamole plus aspirin in secondary stroke prevention. *Journal of Clinical Epidemiology* **65**: 835–845.

Dias S, Welton NJ, Caldwell DM, Ades AE. 2010. Checking consistency in mixed treatment comparison meta-analysis. *Statistics in Medicine* **29**: 932–944.

Dias S, Welton NJ, Sutton AJ, Ades AE. 2011a. NICE DSU Technical Support Document 2: a generalised linear modelling framework for pairwise and network meta-analysis of randomised controlled trials. Last updated April 2012; Available at: http://www.nicedsu.org.uk.

Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. 2011b. NICE DSU Technical Support Document 4: inconsistency in networks of evidence based on randomised controlled trials. Last updated April 2012; Available at: http://www.nicedsu.org.uk.

Dias S, Welton NJ, Sutton AJ, Ades AE. 2013a. Evidence synthesis for decision making 5: the baseline natural history model. *Medical Decision Making* **33**: 657–670.

Dias S, Sutton AB, Ades AE, Welton NJ. 2013b. A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making* **33**: 607–617.

Efthimiou O, Mavridis D, Cipriani A, Leucht S, Bagos P, Salanti G. 2014. An approach for modelling multiple correlated outcomes in a network of interventions using odds ratios. *Statistics in Medicine* **33**: 2275–2287.

Franchini AJ, Dias S, Ades AE, Jansen JP, Welton NJ. 2012. Accounting for correlation in network meta-analysis with multi-arm trials. *Research Synthesis Methods* **3**: 142–160.

Gartlehner G, Moore CG. 2008. Direct *versus* indirect comparisons: a summary of the evidence. *International Journal of Technology Assessment in Health Care* **24**: 170–177.

Gelman A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**: 515–533.

Giovane CD, Vacchi L, Mavridis D, Filippini G, Salanti G. 2013. Network meta-analysis models to account for variability in treatment definitions: application to dose effects. *Statistics in Medicine* **32**: 25–39.

Higgins JPT, Whitehead A. 1996. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* **15**: 2733–2749.

Higgins JPT, Jackson D, Barrett JK, Lu G, Ades AE, White IR. 2012. Consistency and inconsistency in network meta-analysis: concepts and models for multi-arm studies. *Research Synthesis Methods* **3**: 98–110.

Hoaglin DC, Hawkins N, Jansen JP, Scott DA, Itzler R, Cappelleri JC, Boersma C, Thompson D, Larholt KM, Diaz M, 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.

Hong H, Carlin BP, Chu H, Shamliyan TA, Wang S, Kane RL. 2013a. A Bayesian missing data framework for multiple continuous outcome mixed treatment comparisons. Methods Research Report. (Prepared by the Minnesota Evidence-based Practice Center under Contract No. 290-2007-10064-I.) AHRQ Publication No.13-EHC004-EF. Rockville, MD: Agency for Healthcare Research and Quality; January 2013. Available at: http://www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Hong H, Carlin BP, Shamliyan TA, Wyman JF, Ramakrishnan R, Sainfort F, Kane RL. 2013b. Comparing Bayesian and frequentist approaches for multiple outcome mixed treatment comparisons. *Medical Decision Making* **18**: 702–714.

van Houwelingen H, Senn S. 1999. Investigating underlying risk as a source of heterogeneity in meta-analysis by S. G. Thompson, T. C. Smith and S. J. Sharp, Statistics in Medicine, 16, 27412758 (1997). *Statistics in Medicine* **18**: 110–115.

Jackson D, Riley R, White IR. 2011. Multivariate meta-analysis: potential and promise. *Statistics in Medicine* **30**: 2481–2498.

Jansen JP, Crawford B, Bergman G, Stam W. 2008. Bayesian meta-analysis of multiple treatment comparisons: an introduction to mixed treatment comparisons. *Value in Health* **11**: 956–964.

Jansen JP, Fleurence R, Devine B, Itzler R, Barrett A, Hawkins N, Lee K, Boersma C, Annemans L, 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.

Jansen J, Naci H. 2013. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. *BMC Medicine* **11**: 159.

Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. 2005. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* **24**: 2401–2428.

Little RJ, Rubin DB. 2002. *Statistical Analysis with Missing Data*. 2nd edn. New York: Wiley.

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

Lu G, Ades AE. 2006. Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association* **101**: 447–459.

Lu G, Ades AE. 2009. Modeling between-trial variance structure in mixed treatment comparisons. *Biostatistics* **10**: 792–805.

Lu G, Welton NJ, Higgins JPT, White IR, Ades AE. 2011. Linear inference for mixed treatment comparison meta-analysis: a two-stage approach. *Research Synthesis Methods* **2**: 43–60.

Lunn D, Jackson C, Spiegelhalter DJ, Best N, Thomas A. 2012. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. **98**. Boca Raton, FL: CRC Press.

Lunn D, Spiegelhalter D, Thomas A, Best N. 2009. The BUGS project: evolution, critique, and future directions. *Statistics in Medicine* **28**: 3049–3067.

Lumley T. 2002. Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine* **21**: 2313–2324.

McCulloch CE, Searle SR, Neuhaus JM. 2008. *Generalized, Linear, and Mixed Models*. 2nd edn. Wiley, New York.

Nixon RM, Bansback N, Brennan A. 2007. Using mixed treatment comparisons and meta-regression to perform indirect comparisons to estimate the efficacy of biologic treatments in rheumatoid arthritis. *Statistics in Medicine* **26**: 1237–1254.

Piepho HP, Williams ER, Madden LV. 2012. The use of two-way linear mixed models in multitreatment meta-analysis. *Biometrics* **68**: 1269–1277.

Presanis AM, Ohlssen D, Spiegelhalter DJ, De Angelis D. 2013. Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science* **28**: 376–397.

Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. 2007. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* **26**: 78–97.

Salanti G, Higgins JPT, Ades AE, Ioannidis JP. 2008. Evaluation of networks of randomized trials. *Statistical Methods in Medical Research* **17**: 279–301.

Schmid CH, Trikalinos TA, Olkin I. 2014. Bayesian network meta-analysis for unordered categorical outcomes with incomplete data. *Research Synthesis Methods* **5**: 162–185.

Senn S. 2010. Hans van Houwelingen and the art of summing up. *Biometrical Journal* **52**: 85–94.

Shamliyan TA, Wang S-Y, Olson-Kellogg B, Robert KL. 2012. Physical therapy for knee pain secondary to osteoarthritis. Comparative Effectiveness Review. Prepared by the Minnesota Evidence-based Practice Center under Contract No. 290-1007-10064 I. Agency for Healthcare Research and Quality. Rockville, MD: In press.

Smith TC, Spiegelhalter DJ, Thomas A. 1995. Bayesian approaches to random-effects meta-analysis: a comparative study. *Statistics in Medicine* **14**: 2685–2699.

Spiegelhalter DJ, Best NG, Carlin BP, Avd L. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **64**: 583–639.

Wei Y, Higgins JPT. 2013. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* **32**: 1191–1205.

Welton NJ, Cooper NJ, Ades AE, Lu G, Sutton AJ. 2008. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. *Statistics in Medicine* **27**: 5620–5639.

White IR, Barrett JK, Jackson D, Higgins JPT. 2012. Consistency and inconsistency in network meta-analysis: model estimation using multivariate meta-regression. *Research Synthesis Methods* **3**: 111–125.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.