A Bayesian Missing Data Framework for Generalized Multiple Outcome Mixed Treatment Comparisons

Hwanhee Hong*, Haitao Chu, Jing Zhang, and Bradley P. Carlin

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, 55405, USA

*email: hong0362@umn.edu

SUMMARY: Bayesian statistical approaches to mixed treatment comparisons (MTCs) are becoming more popular due to their flexibility and interpretability. Many randomized clinical trials report multiple outcomes with possible inherent correlations. Moreover, MTC data are typically sparse and researchers often choose study arms based on previous trials. In this paper, we summarize existing hierarchical Bayesian methods for MTCs with a single outcome, and we introduce novel Bayesian approaches for multiple outcomes simultaneously, rather than in separate MTC analyses. We do this by incorporating missing data and correlation structure between outcomes through contrast-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 develop a new measure of inconsistency under our missing data framework, having more straightforward interpretation and implementation than standard methods. We offer a simulation study under various missingness mechanisms (e.g., MCAR, MAR, and MNAR) providing evidence that our models outperform existing models in terms of bias and MSE, then illustrate our methods with two real MTC datasets. We close with a discussion of our results and a few avenues for future methodological development.

KEY WORDS: Bayesian hierarchical model; Markov chain Monte Carlo; missingness mechanism; network meta-analysis.
1 Introduction

Mixed treatment comparisons (MTCs) 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 (Jansen et al., 2011; Hoaglin et al., 2011; Ades et al., 2012). In the MTC data framework, since few head-to-head comparisons are available, we must combine indirect and direct comparisons, typically each investigated treatment against a control or a standard treatment (Lumley, 2002; Gartlehner and Moore, 2008). Bayesian hierarchical statistical meta-analysis for MTCs with a single binary outcome has been investigated actively since the 1980s (Smith et al., 1995; Nixon et al., 2007; Lu and Ades, 2004, 2006). 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).

The main issues in MTCs are heterogeneity 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; White 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 \( m_{23} = m_{13} - m_{12} \), where \( m_{AB} \) is the fixed 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) explain that inconsistency is another form of heterogeneity which should not exist unless the trials differ with respect to effect modifiers, such as experimental settings or patients’ characteristics. They also note that inclusion of multi-arm trials complicates the
identification and interpretation of the inconsistency, and the sparsity of the data can lead to large uncertainty (i.e., wide 95 percent credible sets) regarding inconsistency, making it difficult to distinguish from heterogeneity.

Recently, some studies consider the correlation across treatments with a single outcome in depth \cite{LuAdes2009, Franchini2012}, whereas correlation across outcomes has been less discussed \cite{Ades2010}. However, 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).

As most randomized controlled trials (RCTs) compare only two or three treatment arms, the result is often extremely sparse MTC data from the perspective of missing data analysis. Suppose we calculate the missingness rate as the ratio of the summation of the number of missing arms to the total number of treatments times the number of studies. Then, when we compare five treatments the missingness rate is typically 40 to 60 percent, and could increase above 70 percent if 10 treatments are considered. A standard MTC model (e.g., Lu and Ades, 2006) uses only the observed data. However, we can in a sense borrow strength from missing data by imputing them in Bayesian hierarchical modeling that accounts for between-treatment and between-outcome correlations using Markov chain Monte Carlo (MCMC) algorithms. Especially when the missingness does not occur “completely at random”, but depends on some observed or even unobserved information, simply ignoring such missing data can lead to biased estimators \cite[p.13]{LittleRubin2002}.

The remainder of the paper is structured as follows. First, Section 2 describes two data sets. The first considers a single count outcome, while the second considers the bivariate continuous outcomes case. Section 3 provides details of our Bayesian missing data hierarchical
modeling framework for MTC under various assumptions to incorporate missing data and multiple outcomes, and with it our new approach for measuring inconsistency. Section 4 reports the results of simulation studies validating our approaches, while Section 5 delivers the results of our analysis of the two data sets. Finally, Section 6 discusses our work, its limitations, and needed future methodological developments.

2 Motivating Data

2.1 NSAIDs data

These data (Trelle et al., 2011) originally consisted of 5 different outcomes connected with the safety of a collection of non-steroidal anti-inflammatory drugs (NSAIDs) which are frequently prescribed for pain relief. Here for illustrative purposes we select one outcome, myocardial infarction (MI). Figure 1(a) displays the network of 8 drugs (placebo, celecoxib, diclofenac, naproxen, ibuprofen, etoricoxib, rofecoxib, and lumiracoxib) from a total of 30 studies reporting the MI counts and associated patient years of follow-up for each arm in each study. Each drug is coded as 1 to 8 in a clockwise direction from placebo to lumiracoxib. The size of each node represents the number of studies investigating the drug, while the thickness of each edge denotes the total patient years for the relation. The numbers on the edges indicate the numbers of studies investigating the relation (see http://www.mtm.uoi.gr/ for how to draw the network graph). In these data, only 4 NSAIDs (celecoxib, naproxen, rofecoxib, and lumiracoxib) are compared to placebo, and 18 studies do not include a placebo as their control arm; that is, about two thirds of the trials have active drug-to-drug comparisons only.

2.2 Knee pain osteoarthritis (OA) data

Figures 1(b) and (c) exhibit the trial network among physical therapy interventions for community-dwelling adults with knee pain secondary to osteoarthritis in terms of pain
and disability outcomes [Shamliyan et al., 2012]. A total of 54 RCTs 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 (38 studies), disability only (2 studies), or both outcomes (23 studies). We rescaled different standard scores to make them have the same, comparable range (0 to 10). These two network graphs are interpreted similarly to panel (a) except that the thickness of each edge now indicates the total number of samples for that relation. 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.

[Figure 1 about here.]

3 Methods

3.1 Likelihood

In MTC, 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 to other active treatments. In the NSAIDs data, placebo is the reference treatment though only 12 studies included placebo as their arms. For the 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 (e.g., education and placebo in the OA data).

Suppose we are comparing $K$ treatments from $I$ studies in terms of $L$ outcomes. For any types of aggregated-level (i.e., summarizing over individuals) MTC data, we assume that the data for a specific outcome from each study follow an exponential family distributional model. That is,

$$Y_{ikl} \sim f_Y(y_{ikl} | \Delta_{ikl}, \xi_{ikl}), \quad i = 1, \ldots, I, \quad k = 1, \ldots, K, \quad l = 1, \ldots, L,$$

where $Y_{ikl}$ is the observed aggregated outcome, $f_Y(\cdot)$ is a density function, $\Delta_{ikl}$ is an unknown true population location parameter, and $\xi_{ikl}$ is a known nuisance (typically variance) parameter in the $k$th treatment arm from the $i$th study with respect to the $l$th outcome. For example, when the measurement type is continuous, $Y_{ikl}$ often represents the sample mean following a normal density $f_Y(\cdot)$ with unknown true mean $\Delta_{ikl}$ and known sample variance, $\xi_{ikl}$. We consider $k = 1$ to be the reference treatment.

### 3.2 Generalized Lu and Ades Model

In this paper, we only consider random effects models, although a fixed effects model can easily be implemented (Whitehead, 2002, pp.5 and 58). Extending Lu and Ades (2004, 2006) to the generalized linear model case, the mean structure can be modeled as

$$g(\Delta_{ikl}) = \alpha_{iB \ell} + \delta_{iBkl},$$

where $\delta_{iBkl} = 0$ if $k = B$ and $g(\cdot)$ is a known link function, such as the identity, logit, or log for a continuous, binary, or count response, respectively (McCulloch et al, 2008, p.136). Here, $B$ indicates the baseline treatment in each study $i$, $\alpha_{iB \ell}$ are the baseline treatment effects, and $\delta_{iBkl}$ are the random effects of contrasts between treatment $k$ and the baseline treatment for outcome $\ell$ in study $i$. We define $d_{jk \ell}$ as the fixed mean of contrasts between treatments $k$ and $j$ for outcome $\ell$, with $d_{jk \ell} \equiv 0$ when $j = k$. We infer the treatment effects
in terms of $d_{1\ell}$, always comparing the $k^{th}$ treatment to the reference treatment, so that we need only to assign a prior to the $d_{1\ell}$ to obtain $d_{j\ell} = d_{1\ell} - d_{1\ell}$ under the consistency assumption [Lu and Ades 2006].

We often assume homogeneous variance across random effects for all arms, i.e.,

$$\delta_{iB\ell} \sim N(d_{1\ell} - d_{1B\ell}, \tau^2_\ell).$$  \hspace{1cm} (3)

Here, $\delta_{iB\ell}$ is 0 when $k = B$, and $\tau_\ell$ is the standard deviation of the random effects for outcome $\ell$. We denote this model as the Lu and Ades (LA)-style homogeneous random effects model (LAREhom). For multi-arm trials, the $\delta_{iB\ell}$ in (3) are replaced by a vector that follows a multivariate normal distribution with dimension equal to the number of arms in study $i$ minus one, for each outcome $\ell$. Here, the between-arm-contrast correlation is 0.5 as a consequence of the homogeneous variance and consistency assumptions [Lu and Ades 2006]. LA-style models always assume that the baseline effects $\alpha_{iB\ell}$ are fixed and independent across studies. However, we can also assume them to be random effects, to shrink them towards their grand mean and possibly reduce model complexity; see Section 3.5.

3.3 Modeling for Missing Data and Correlation 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 [Dias et al. 2011a; Salanti et al. 2008]. Lu and Ades-style models use a CB approach. Note that the mean effect of random contrasts between treatment $k$ and reference treatment ($d_{1\ell}$) is the parameter of interest in CB models, whereas for AB models it is the mean of the random effects for treatment $k$ ($\mu_{k\ell}$), both for each outcome $\ell$.

In MTC, it is common that the number of treatments compared in the $i^{th}$ study is far less than the complete collection of $K$ treatments. Under the LA model framework, since each study contributes to the likelihood for a different set of treatments, using the observed
measurements only can complicate estimating the covariance matrix of the random effects $\delta_{iBk\ell}$, leading to difficulties in prior assignment and parameter inference for multi-arm trials. In addition, researchers may select study arms based on the trials conducted previously, resulting in missingness that may or may not be “at random.” Finally, it is difficult to interpret $\alpha_{iB\ell}$ in (2) when the baseline treatment is not always the same across studies.

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 (though possibly missing) baseline treatment, $B = 1$ in (2).

3.3.1 Contrast-based Approach

Under our missing data framework, the random effects model has the same form as in (2) with $B$ always equal to 1, and the distribution for the random effects $\delta_{iBk\ell}$ in (3) can be replaced with a matrix form as follows:

$$\delta_{i\ell}^{Trt} \sim MVN(d_{i\ell}^{Trt}, \Sigma_{i\ell}^{Trt}), \quad (4)$$

where $\delta_{i\ell}^{Trt} = (\delta_{i12\ell}, \ldots, \delta_{iK\ell})^T$, $d_{i\ell}^{Trt} = (d_{12\ell}, \ldots, d_{1K\ell})^T$, and $\Sigma_{i\ell}^{Trt}$ is a $(K-1) \times (K-1)$ unstructured covariance matrix for $\ell = 1, \ldots, L$. Note that since $\delta_{11\ell}$ and $d_{11\ell}$ are always 0, they are not included in $\delta_{i\ell}$ and $d_{i\ell}$. Here, $\Sigma_{i\ell}^{Trt}$ captures all random contrasts’ relations among treatments in each outcome $\ell$. We refer to this model as a contrast-based random effects model assuming independence between outcomes (CBRE1).

To allow correlation among outcomes, the distribution of $\delta_{iBk\ell}$ in (3) can be respecified to

$$\delta_{ik}^{Out} \sim MVN(d_{ik}^{Out}, \Sigma_{ik}^{Out}), \quad (5)$$

where $\delta_{ik}^{Out} = (\delta_{i1k1}, \ldots, \delta_{i1kL})^T$, $d_{ik}^{Out} = (d_{1k1}, \ldots, d_{1kL})^T$, and $\Sigma_{ik}^{Out}$ 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 $\Sigma_k^{Out}$. We call this model CBRE2. Alternatively, we can also use the same $\Sigma_k^{Out}$ for all $k$, if such an assumption is sensible.

We can also partition the random effects into two independent sources, and incorporate both into a model. Then (2) would be rewritten as

$$g(\Delta_{ik\ell}) = \alpha_{i1\ell} + d_{1k\ell} + \nu_{ik} + \omega_{i\ell},$$

where $\nu_i = (\nu_{i2}, \ldots, \nu_{iK})^T \sim MVN(0, R^{Trt})$, $\omega_i = (\omega_{i1}, \ldots, \omega_{iL})^T \sim MVN(0, R^{Out})$, and $\nu_{ik}$ and $\omega_{i\ell}$ are independent. Here, $R^{Trt}$ and $R^{Out}$ are $(K-1) \times (K-1)$ and $L \times L$ unstructured covariance matrices implying correlation between treatments and outcomes, respectively. We denote this model as CBRE3.

In this approach, we can always have $\delta_{ik}, \delta_{il}^{Out}, \nu_i$, and $\omega_i$ vectors of the same length for every study $i$, and incorporate all sources of uncertainty by considering unobserved arms as missing data to be imputed by our MCMC algorithm. For example, suppose Study 1 compares treatments 1, 2, and 3, giving information about two contrasts, $\delta_{i12\ell}$ and $\delta_{i13\ell}$, whereas Study 2 compares only treatments 1 and 2, and Study 3 includes only treatments 1 and 3. We can impute the missing contrasts $\delta_{i13\ell}$ and $\delta_{i12\ell}$ in Studies 2 and 3 respectively by using the information related to these contrasts observed in Study 1. Also, we can borrow information from the relation between outcomes $\ell$. Finally, in our CB approach, $\alpha_{iB\ell}$ becomes meaningful because the baseline treatment is the same across all studies.

3.3.2 Arm-based Approach

The CB method estimates the treatment contrasts, but the approach’s singular focus on relative treatment effects ultimately leads to some limitations. First, we cannot directly calculate the correlation between treatment effects via the correlation between effect contrasts. That is, the interpretation of correlations in terms of the relative effect scale is not straightforward. Furthermore, our CB model restricts the variance of a baseline effect to always be smaller
than that of other treatments. For example, when the outcome is continuous, the variance of the population mean of the baseline treatment, $\Delta_{iB\ell}$, is $\text{Var}(\alpha_{iB\ell})$, whereas for other treatments we have $\text{Var}(\alpha_{iB\ell}) + \text{Var}(\delta_{iB\ell})$, which is never smaller than $\text{Var}(\alpha_{iB\ell})$.

The AB model can be written by respecifying model (2) as

$$g(\Delta_{ik\ell}) = \mu_{k\ell} + \eta_{ik\ell}, \quad (7)$$

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

If we begin by assuming independent random effects between outcomes, then the random effect $\eta_{ik\ell}$ in (7) can be structured as $(\eta_{1k\ell}, \ldots, \eta_{Kk\ell})^T \sim \text{MVN}(0, \Lambda_{Trt}^k)$ with $\Lambda_{Trt}^k$ a $K \times K$ unstructured covariance matrix capturing relations of random effects between treatments, 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 $(\eta_{1k1}, \ldots, \eta_{LK})^T \sim \text{MVN}(0, \Lambda_{Out}^k)$ where $\Lambda_{Out}^k$ 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 can use the same $\Lambda_{Out}^k$ for all $k$ when it is reasonable to do so. Finally, we can allow both correlation sources by rewriting the mean structure (7) as $g(\Delta_{ik\ell}) = \mu_{k\ell} + \nu_{ik} + \omega_{i\ell}$, where $(\nu_{i1}, \ldots, \nu_{iK})^T \sim \text{MVN}(0, D_{Trt}^k)$, $(\omega_{i1}, \ldots, \omega_{iL})^T \sim \text{MVN}(0, D_{Out}^k)$, and $\nu_{ik}$ and $\omega_{i\ell}$ are independent. Here, $D_{Trt}^k$ and $D_{Out}^k$ are $K \times K$ and $L \times L$ unstructured covariance matrices.

The parameters in arm-based models permit more straightforward interpretation, especially in estimating a pure treatment effect. However, these models do require strong assumptions regarding the similarity and exchangeability of all populations, in order to permit meaningful clinical inference. Note that in AB models, all of our random effect covariance matrices are unstructured. That is, AB models are less constrained, but thus have a slightly larger number of parameters to be estimated than CB models.
3.4 Inconsistency

Lu and Ades (2006) and Dias et al. (2011b) explain how to measure and handle inconsistency in the LA approach. First, we must specify the number of inconsistency degrees of freedom (ICDF), defined as "the number of independent 'loops' of evidence" in the network [Dias et al. 2011b]. Here, we find an evidence loop wherever more than two treatments are connected as a cycle in the graph (e.g., the triangle generated by placebo (1), celecoxib (2), and naproxen (4) in Figure 1 (a)). Then, in the LA approach, we can estimate all the possible contrasts in each independent loop separately, without applying the consistency equation. In the example above, there are three possible contrasts, \( d_{12}, d_{14}, \) and \( d_{24}, \) and under the consistency assumption, we only assign priors to \( d_{12} \) and \( d_{14}, \) and obtain \( d_{24} \) as \( d_{14} - d_{12}. \) By contrast, inconsistency could occur if \( d_{24} \neq d_{14} - d_{12}, \) and this model allows all three contrasts to have independent prior distributions. Alternatively, we can follow Lu and Ades (2006) and measure the variability among inconsistencies by allowing for the so called 'w-factor' (defined as \( w_{124} = d_{24} - (d_{14} - d_{12}) \) in the above example) with a \( N(0, \sigma_w^2) \) prior on all nonzero w-factors. Dias et al. (2011b) do not recommend this method, since \( \hat{\sigma}_w \) could have a wide credible interval when ICDF is small. In any case, it is extremely difficult to define all the independent evidence loops when the data include multi-arm trials or uncommon baseline treatments across studies. Also, the model could be overparameterized when we compare too many treatments [Dias et al. 2011b Cipriani et al. 2011].

In our missing data framework, inconsistency will instead be defined as the discrepancy between observed and imputed treatment effects, in order to measure the data-driven magnitude of bias (DMB). For illustrative purposes, we consider the arm-based model with a single outcome, so for now we drop the \( \ell \) indices in (7). We define the set of studies observing the \( k^{th} \) treatment as \( S_{k}^{obs} \) and the set of remaining studies as \( S_{k}^{mis}. \) We can split (7) into two
parts as follows:

\[ g(\Delta_{ik}) = \begin{cases} 
\mu_k + \eta_{ik}^{obs} & \text{if } i \in S_{obs}^k, \\
\mu_k + \eta_{ik}^{mis} & \text{if } i \in S_{mis}^k,
\end{cases} \]  
(8)

where the superscripts ‘obs’ and ‘mis’ indicate which sources of information are used (observed and imputed, respectively), and the \( \eta_{ik} \) in (7) is either \( \eta_{ik}^{obs} \) or \( \eta_{ik}^{mis} \) depending on its origin. Then, we will define \( DMB_k \) as

\[ DMB_k = \Delta_k^{mis} - \Delta_k^{obs}, \]  
(9)

where \( \Delta_k^{mis} = \frac{\sum_{i \in S_{mis}^k} w_{i}^{mis} g^{-1}(\mu_k + \eta_{ik}^{mis})}{\sum_{i \in S_{mis}^k} w_{i}^{mis}} \) and \( \Delta_k^{obs} = \frac{\sum_{i \in S_{obs}^k} w_{i}^{obs} g^{-1}(\mu_k + \eta_{ik}^{obs})}{\sum_{i \in S_{obs}^k} w_{i}^{obs}} \). Here, the total sample size of the \( i \)-th study could be considered for weights \( w_{i}^{mis} \) and \( w_{i}^{obs} \), or we can simply select both to be 1 for the equal weight. Positive DMB values mean that the imputed treatment effect is larger than observed, and vice-versa for the negative DMBs. Also, larger absolute values of DMB could be a sign of significant inconsistency, and bias in the estimated treatment effects.

### 3.5 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 \( \alpha_{B\ell} \) and \( d_{1k\ell} \), and a vague Uniform distribution is assigned to \( \tau_{\ell} \) in the LAREHom model; \( Uniform(0.01, 5) \) and \( Uniform(0.01, 10) \) priors are selected for the NSAIDs and OA data analyses, respectively. Here, the upper bound 5 in the Uniform distribution for \( \tau_{\ell} \) in the NSAIDs data is considered to be large enough to be slack for any reasonable data on the log rate scale. In all CB models, we can assume a fixed baseline effect by assigning \( \alpha_{i1\ell} \sim N(0, 100^2) \), or a random baseline effect using a \( N(a_{\ell}, \tau_{\alpha,\ell}^2) \) here instead, where the hyperparameters \( a_{\ell} \) and \( \tau_{\alpha,\ell} \) follow noninformative normal and Uniform distributions, respectively (Gelman [2006]). Throughout all CB and AB models, the fixed effects (\( d_{1k\ell} \) and \( \mu_{k\ell} \), respectively) follow a \( N(0, 100^2) \) distribution, while
all inverse covariance matrices follow a $\text{Wishart}(\Omega, \gamma)$ having a mean of $\gamma \Omega^{-1}$ and degrees of freedom parameter $\gamma$ to be the matrix dimension, because this is the smallest value that will still yield a proper prior. For the NSAIDs data, we select $\Omega$ to be the identity matrix, which gives a 95 percent confidence interval of 0.45 to 31 for the prior mean standard deviation, in order to be fairly noninformative while ensuring MCMC convergence. For the OA data, we choose $\Omega$ to be $5\gamma$ times the identity matrix, yielding a 95 percent prior guess for the mean standard deviation of the random effects of 3.33 to 253 (Carlin and Louis, 2009, p.338).

We used the \texttt{BRugs} package (Gelman, 2007) in \texttt{R} to perform our simulation studies, where we call \texttt{OpenBUGS} (Lunn et al., 2009) 500 times from \texttt{R}, once for each simulated dataset. In each case, we obtain 20,000 samples, after a 30,000 sample burn-in, and collect medians of parameters across the 500 simulated datasets. For data analyses, \texttt{WinBUGS} (Lunn et al., 2000) is used 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.

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 to the LA method in terms of bias and MSE. We generate twenty studies (NS = 20) comparing two treatments, denoted 1 and 2, for a single continuous outcome and fit LAREhom and ABRE1 models. Then, we drop five, ten, and fifteen arms for Treatment 2 by applying three missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Thus there are ten scenarios: complete data (denoted Complete), MCAR with five, ten, and fifteen missing arms (denoted MCAR25, 50, and 75 indicating twenty-five,
fifty, and seventy-five percent of missing, respectively), and similarly for MAR and MNAR (denoted MAR25, 50, 75, and MNAR25, 50, 75, respectively). Similarly, we also consider a larger number of studies, NS = 100, and drop twenty-five, fifty, and seventy-five studies from the second treatment under the three missingness mechanisms. We generate 500 datasets under (7) without \( \ell \) indices, and we choose the identity link function with \( \mu_1 = 2, \mu_2 = 1, \) and \((\eta_{i1}, \eta_{i2})^T \sim MVN \left( 0, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \right)\) for the true parameters. For simplicity, we assume that every study has a sample size of 100 and a standard deviation of 2 for every arm.

To sample the partially missing data under MAR and MNAR mechanisms, we first calculate the probability of missing \( (p_{i,mis}) \) for study \( i \) by applying a logit model with the observed or missing data as covariates, respectively. We recall that \( y_{i2} \) is sometimes missing data in our design. As such, we use the following two logit models:

\[
\text{MAR : logit}(p_{i,mis}) = \beta_0^{MAR} - y_{i1} \quad (10)
\]

\[
\text{MNAR : logit}(p_{i,mis}) = \beta_0^{MNAR} - y_{i1} + y_{i2}, \quad (11)
\]

where coefficients and intercepts \((\beta_0^{MAR} \text{ and } \beta_0^{MNAR})\) are selected to result in a mean \( p_{i,mis} \) of about 25, 50, or 75 percent; \( \beta_0^{MAR} = 3, 2, \) or \( 1 \) for MAR25, 50, or 75, and \( \beta_0^{MNAR} = 2, 1, \) or \( -1 \) for MNAR25, 50, or 75, respectively. Given \( p_{i,mis} \), we generate the missingness indicator vector until twenty-five, fifty, or seventy-five percent of the studies are selected as missing.

### 4.2 Results

Figure 2 plots the bias and MSE of \( d_{12} \) across the 10 different settings with NS = 20, where we plot both the bias (solid symbols) and MSE (lines with open symbols). Both LAREhom and ABRE1 models yield almost zero bias under the Complete and MCAR settings, with similar MSE values. Under MAR, LAREhom gives a lot larger bias and MSE than ABRE1, and the pattern gets worse as the degree of missingness increases, whereas ABRE1 continues to produce unbiased estimates. Both methods give large bias and MSE under MNAR, again
with performance deteriorating as the degree of missingness increases, although ABRE1 performs slightly better than LAREhom. The trend for NS = 100 is similar.

In terms of the DMB inconsistency factors, we set \( w_i^{\text{obs}} = w_i^{\text{mis}} = 1 \) for all \( i \). We first calculate the true mean DMB from 500 simulated datasets with NS = 20 and compare them with their estimated values, \( \hat{\text{DMB}} \). True DMBs are -0.03, 0.73, and -0.75 with \( \hat{\text{DMB}}_2 = -0.03, 0.74, \) and 0.41 under MCAR50, MAR50, and MNAR50, respectively (DMB1 is not obtained because every study includes Treatment 1). Our ABRE1 model estimates the DMB well including its direction under MCAR and MAR, but produces a poor \( \hat{\text{DMB}} \) under MNAR. This is because our model can correctly impute ignorable missing data based on the observed data and model, which occurs under MCAR and MAR, but not MNAR. However, our model does not specify any sort of MNAR missingness (e.g., a selection model as in Little and Rubin, 2002, p.313), so such inconsistencies remain unmodeled. Although we cannot distinguish MAR from MNAR based on \( \hat{\text{DMB}} \), large \( \hat{\text{DMB}} \) values could be the indication of non-random missingness with possible inconsistency. We cannot define and calculate any inconsistencies in the LAREhom model because no evidence loop exists when only two treatments are compared in the MTC data, though inconsistency clearly exists under MAR and MNAR. Again, the trend for NS = 100 is similar.

We also test whether the DMB factors significantly differ from zero or not, and calculate the probability of significance (similar to power in a frequentist method) by finding the empirical proportion of DMB factors having 95 percent credible sets that exclude zero. For NS = 20, these probabilities are 0.008, 0.104, and 0.060 under MCAR50, MAR50, and MNAR50, respectively. However, under the scenario with NS = 100, these probabilities become 0.002, 0.604, and 0.404. That is, given a large enough set of studies, we can test the DMB factor’s difference from zero, with significance suggesting either MAR or MNAR missingness.

[Figure 2 about here.]
5 Data Analysis

5.1 NSAIDs data

Since these data have a count outcome, we assume the data from each study follow a Poisson distribution, namely $Y_{ik} \sim \text{Poisson}(\lambda_{ik} \times PY_{ik})$, where $Y_{ik}$ is the MI count, $\lambda_{ik}$ is the unknown rate of occurrence, and $PY_{ik}$ is the known number of observed person-years of treatment exposure for treatment $k$ from study $i$. Models (2) and (7) can be specified as

$$\log(\lambda_{ik}) = \alpha_i B + \delta_i B_k + \log(PY_{ik}/1000)$$

$$\log(\lambda_{ik}) = \mu_k + \eta_{ik} + \log(PY_{ik}/1000),$$

respectively, where $\log(PY_{ik}/1000)$ is the offset adjusting the different total follow-up times across studies. Note that, with a single outcome, we can only incorporate between-treatment correlation, thus we fit only CBRE1 and ABRE1 models among all our CB and AB models. For inference, we can obtain the rate ratio (RR) between treatment $k$ and the placebo per 1000 person-years as $RR_k = \exp(d_{1k})$ and $\exp(\mu_k - \mu_1)$ for the CBRE1 and ABRE1 models, respectively. Here, a RR smaller than 1 means that the drug is safer than placebo with respect to the MI event.

We fitted 5 models: LAREhom with consistency and inconsistency assumptions, CBRE1 with fixed and random baseline effects, and ABRE1, where the LA-style models of course do not employ the missingness framework. Table 1 displays the model choice statistics DIC (Spiegelhalter et al., 2002) and posterior medians of $RR_k$, with the latter with standard deviations in the parentheses. Regarding DIC, the two LAREhom models and CBRE1 with the fixed baseline effect give similar DIC, whereas CBRE1 with the random baseline effect and ABRE1 produce similar and improved DIC. Although CBRE1 with the fixed baseline effect borrows strength from the missing data, it behaves similarly to the LAREhom consistency model, whereas CBRE1 with the random baseline effect produces a $p_D$ 10 units smaller,
resulting from the shrinkage of the baseline effects. Overall, ABRE1 gives the smallest DIC, closely followed by random baseline CBRE1.

The estimated standard deviation, $\tau$, in the LAREhom consistency model is 0.28 with 95 percent credible interval from 0.03 to 0.75 (the LAREhom inconsistency model produces a similar estimate). The NSAIDs data provide six independent evidence loops, ICDF=6 (see Trelle et al., 2011, Web appendix 2, p.6), and the estimated w-factors are between -0.6 and 0.4 with $\hat{\sigma}_w = 1$ (95 percent CI 0.07 to 4.17) (results not shown in Table 1). The inclusion of the w-factors yields somewhat different $RR_k$ estimates; the direction of $RR_3$, $RR_4$, and $RR_6$ are changed, and standard deviations of $RR_3$ and $RR_4$ become larger in the LAREhom inconsistency model. Under ABRE1, the estimated variability on the standard deviation scale is between 0.4 and 0.8, with associated 95 percent credible interval mainly from 0.3 to 1.7. The absolute values of estimated equally weighted $DMB_k$ (not shown) are between 0.01 to 0.06 with none significantly different from zero, providing no statistical evidence of inconsistency. Compared to the LAREhom inconsistency model, ABRE1 gives smaller standard deviations for $RR_k$ except that for the fifth treatment.

Here, a w-factor is only associated with the specific loop which can be defined based on a few studies, not all studies. For example, the loop producing -0.6 (95 percent CI -1.98 to 0.23) w-factor is based on the triangle constructed by placebo, naproxen, and celecoxib in Figure 1(a), and only five studies are involved. The interpretation of the w-factor is limited to the relation between naproxen and celecoxib, and this result says nothing about the consistency of any other relationship or studies except those five. In other words, we cannot estimate inconsistency of a relation if its loop is not defined in the given data (e.g., the relation between diclofenac and etoricoxib in Figure 1(a)). On the other hand, our DMB-factor is specific to a particular treatment (not evidence loop), and is estimated for treatments having partially missing data from every study. That is, we can easily define and
measure the inconsistency, and interpret the results based on all trials in the given MTC data under our missing data framework.

CBRE1 with the fixed baseline effect has similar RR estimates as those under the LAREhom consistency model (albeit with slightly larger standard deviations for $\hat{RR}_5$ and $\hat{RR}_8$), but rather different estimates from those under CBRE1 with the random baseline effect. RR estimates under CBRE1 with the random baseline effect and ABRE1 are quite similar, meaning these methods order the drugs in the same way. However, this safety ordering does not always agree with those of the LAREhom methods, which flip several of the RR values to the other side of the null value, 1.0; indeed all of the estimated RRs exceed this value under the LAREhom inconsistency model. Under random baseline CBRE1 and ABRE1, naproxen (4) and rofecoxib (7) are the best and worst drugs with respect to the occurrence of MI, respectively, although they are not significantly different from placebo.

[Table 1 about here.]

5.2 OA data

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 with the identity link function. That is, we can replace $g(\Delta_{ik\ell})$ with $\Delta_{ik\ell}$, the unknown true mean score of therapy $k$ from study $i$ with respect to outcome $\ell$, in all the equations. Table 2 compares the fit of seven models to our OA data. We apply homogeneous covariance matrices for CBRE2 and ABRE2; that is, $\Sigma^\text{Out}_k$ and $\Lambda^\text{Out}_k$ are the same for all $k$, respectively. 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. ABRE2 fits the data best (smallest $D$), but there is no significant difference in fit across random effects models. All AB models give very slightly higher $p_D$ than the corresponding CB models because they are
less constrained and more parameters need to be estimated. Since our data are sparse, 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, though the DIC differences between either model and LARE are not of practical importance (less than five units).

Figure 3 exhibits posterior medians of $d_{1k\ell}$, the mean difference between therapy $k$ and no treatment, with 95 percent credible intervals for each outcome across five of the models (all but CBRE1 and ABRE1). We indicate the best treatment with respect to each outcome in each model with a triangle character, and the worst treatment with a square. We find that the credible intervals from LARE, CBRE2, and ABRE2 are narrower than the other models. This is because the estimated variability on the standard deviation scale is always between 1 and 1.5 for these models, with associated 95 percent credible interval widths around 0.4. By contrast, CBRE3 and ABRE3 give a bit larger standard deviations, between 2 and 3.5 with intervals mainly from 1.7 to 6. For the pain outcome, our CB and AB models agree that low and high intensity diathermy and ultrasound perform worse than no treatment, while LARE oddly reverses this conclusion (though again, the difference is not significant). Strength exercise reduces pain significantly better than no active treatment across all models, although proprioception exercise has a slightly better (lower) mean across all models. Compared to the pain outcome, the 95 percent credible sets for disability are wider because only about half as many studies reported this outcome, and the best therapy now varies across models. The posterior median correlations between the two outcomes are 0.38 (0.06 to 0.61) and 0.04 (-0.41 to 0.48) for the ABRE2 and ABRE3, respectively. ABRE3 gives a relatively small between-outcome correlation estimate since between-treatment correlation is also implied in this model, and these two correlations might not be distinguished well by the data.

The ABRE3 model produces absolute equally weighted DMB values from 0.02 to 0.38 and
0.04 to 1.04 for the pain and disability outcomes, respectively, again no DMB factors emerging as significant. For the disability outcome, education, electrical stimulation, aerobic exercise, and aquatic exercise give the largest DMBs, and we see correspondingly big differences in the estimated $d_{ik2}$ for those treatments except aquatic exercise between ABRE3 and LARE in Figure 3. We note that ABRE2 gives a lot smaller absolute DMB values (all smaller than 0.2) than ABRE3, again because ABRE3 captures more uncertainty by attempting to model two sources of correlation.

[Table 2 about here.]

[Figure 3 about here.]

6 Discussion

In this paper, we have proposed Bayesian MTC approaches for multiple outcomes under a novel missing data framework, developed a corresponding new concept of evidence inconsistency, and compared our results to previous hierarchical modeling methods. Various types of outcomes can be handled in our framework using an appropriate distributional model and link function. We considered unobserved arms to be missing data, and imputed them by borrowing information from the observed relationships. 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 arm-based model can outperform existing Lu and Ades-style models in terms of bias and MSE of estimators under various missingness mechanisms. Finally, we illustrated our methods using real data.

Several articles assume that missing arms occur at random (Caldwell et al., 2005; Giovane et al., 2012), but such missingness is not considered in the standard Lu-and-Ades approach, implying that they technically assume missingness 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 incorporated into a model. Although we do not specify the missingness pattern into our model and the choices of coefficients in missingness-generating logit functions certainly matter, our simulation study shows that our imputation method based on the observed data performs better than the LA method in terms of bias and MSE, even under MNAR. Modeling such nonignorable missingness into the Bayesian MTC hierarchical approach is an ongoing research area.

Regarding evidence inconsistency, our DMB factors are more easily calculated and readily interpreted than the $w$-factors in the LA approach. It is difficult to define ICDF when the data have multi-arm trials and various different baseline treatments across studies; it may even be difficult to identify all the evidence loops in the graph. 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 the interaction in the two-way linear mixed model. However, again, the degrees of freedom for the interaction (similar to ICDF) often gets larger than the total number of treatments as the data get complicated. By contrast, our missing data framework directly measures the difference of observed and imputed treatment effects, providing a more direct and treatment-specific measure of inconsistency.

In the NSAIDs data analysis, random baseline CBRE1 gives a smaller DIC, resulting from less model complexity than with fixed baseline CBRE1. We prefer the random baseline effect especially when we lack a common baseline treatment, because we can borrow information more easily. Our OA data offer similar results in terms of parameter estimates and DIC between CBRE and ABRE models when they have similar correlation structures, though they parameterize differently. The AB model seems preferable because it is less constrained, prior specification is more straightforward, and its parameters are easily interpreted. For the OA data, although CBRE3 and ABRE3 models enjoy the smallest DIC scores, these methods also produce wide 95 percent treatment effect credible intervals. For decision making, we
only compare the estimated treatment effects for each outcome, but we could also obtain the probability of being the best treatment by utilizing a weighted score with multiple outcomes. In both data analyses, we cannot find any significant DMB factors, and this might due to the insufficient number of studies.

Of course, our methods have some limitations. First, our models often result in slow MCMC convergence because we work with the full imported missing data-parameter space. Our Wishart priors must be carefully selected to ensure MCMC convergence and correct estimation of variability; when informative, they require careful sensitivity analysis. We also assume that the within-study correlations are zero in the likelihood (i.e., the data between arms in one study are independent). However, Riley et al. (2007) discuss when we can estimate within-study correlations in bivariate random effect meta-analysis and thus produce estimates with smaller standard errors than in the independent setting. Finally, our estimates do not always have narrower 95 percent credible intervals than those from other models.

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 aggregated 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.

References


ISPOR task force on indirect treatment comparisons good research practices: part 1. 


Piepho, H.P., Williams, E.R. and Madden, L.V. (2012). The Use of Two-Way Linear Mixed
Models in Multitreatment Meta-Analysis. To appear *Biometrics*.


Figure 1: Network graphs: (a) MI outcome from NSAIDs data; (b) pain outcome from OA data; (c) disability outcome from OA data.
Figure 2: Bias and MSE of \( d_{12} \) from the simulation study are plotted.
Mean difference between therapy and no treatment

Figure 3: Posterior median of $d_{1k\ell}$ with 95 percent credible set in each outcome for OA data.
Table 1: Analysis results for the NSAIDs data: DIC and posterior medians of RRₖ with standard deviations in the parentheses are reported.

<table>
<thead>
<tr>
<th></th>
<th>LAREhom consistency</th>
<th>LAREhom inconsistency</th>
<th>CBRE1 fixed baseline</th>
<th>CBRE1 random baseline</th>
<th>ABRE1</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>66.4</td>
<td>65.8</td>
<td>61.6</td>
<td>63.4</td>
<td>60.9</td>
</tr>
<tr>
<td>pD</td>
<td>40.6</td>
<td>41.0</td>
<td>45.6</td>
<td>34.2</td>
<td>35.3</td>
</tr>
<tr>
<td>DIC</td>
<td>107.0</td>
<td>106.8</td>
<td>107.2</td>
<td>97.6</td>
<td>96.2</td>
</tr>
</tbody>
</table>

RR₂ 1.39 (1.47) 1.34 (0.44) 1.59 (0.89) 1.38 (0.38) 1.43 (0.52)
RR₃ 0.81 (0.93) 1.34 (2.02) 0.81 (0.72) 1.07 (0.41) 1.13 (0.49)
RR₄ 0.86 (0.91) 1.15 (0.50) 0.73 (0.48) 0.60 (0.30) 0.68 (0.36)
RR₅ 1.74 (2.20) 1.59 (1.47) 2.10 (5.26) 0.89 (2.65) 0.89 (2.31)
RR₆ 0.74 (0.91) 1.23 (2.22) 0.72 (1.25) 1.18 (0.81) 1.24 (0.73)
RR₇ 2.08 (2.14) 2.01 (0.51) 1.99 (0.78) 1.88 (0.52) 1.97 (0.71)
RR₈ 2.11 (2.57) 1.95 (2.14) 2.73 (4.41) 0.81 (0.46) 0.79 (0.48)

Table 2: Model comparisons for OA data.

<table>
<thead>
<tr>
<th></th>
<th>LAREhom consistency</th>
<th>CBRE1</th>
<th>CBRE2</th>
<th>CBRE3</th>
<th>ABRE1</th>
<th>ABRE2</th>
<th>ABRE3</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>169.4</td>
<td>168.4</td>
<td>169.0</td>
<td>174.8</td>
<td>169.5</td>
<td>167.8</td>
<td>174.9</td>
</tr>
<tr>
<td>pD</td>
<td>154.5</td>
<td>162.9</td>
<td>153.9</td>
<td>144.7</td>
<td>164.4</td>
<td>158.0</td>
<td>145.9</td>
</tr>
<tr>
<td>DIC</td>
<td>323.9</td>
<td>331.3</td>
<td>322.9</td>
<td>319.5</td>
<td>330.3</td>
<td>325.8</td>
<td>320.8</td>
</tr>
</tbody>
</table>