12-30-2015

The Choice of Prior Distribution for A Covariance Matrix in Multivariate Meta-Analysis: A Simulation Study

Sandra M. Hurtado Rua
Cleveland State University, s.hurtadorua@csuohio.edu

Madhu Mazumdar
Icahn School of Medicine

Robert L. Strawderman
University of Rochester

The choice of prior distribution for a covariance matrix in multivariate meta-analysis: a simulation study

Sandra M. Hurtado Rúa, Madhu Mazumdar and Robert L. Strawderman

1. Introduction

Meta-analysis is used to systematically review and synthesize evidence for a possible association between an exposure and an outcome in several studies where the overall picture remains unclear about the significance of the effect size [1–3]. When jointly modeling p outcomes based on n studies, a multivariate meta-analysis formulation based on a random effects model [4] given by (1) is recommended [5–10],

$$\mathbf{Y}_i = \boldsymbol{\theta} + \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i;\ 1 \leqslant i \leqslant n, \tag{1}$$

where $\boldsymbol{Y}_i = \left(Y_{i1}, Y_{i2}, \ldots, Y_{ip}\right)^\top$ is the vector of study effect sizes, $\boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_p\right)^\top$ is the vector of population effect sizes, and $\boldsymbol{\mu}_i = \left(\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip}\right)^\top$ and $\boldsymbol{\varepsilon}_i = \left(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{ip}\right)^\top$ are the vectors of between-study random effects and within-study sampling errors for the ith study, respectively. It is often assumed that $\boldsymbol{\mu}_i$ and $\boldsymbol{\varepsilon}_i$ are independent, with $E[\boldsymbol{\mu}_i] = E[\boldsymbol{\varepsilon}_i] = \mathbf{0}$, $\mathrm{Var}\left(\boldsymbol{\mu}_i\right) = \boldsymbol{D}$ and $\mathrm{Var}\left(\boldsymbol{\varepsilon}_i\right) = \boldsymbol{\Sigma}_i$, where $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$ (assumed to be known) are the $p \times p$ between-study and within-study covariance matrices, respectively.

Several frequentist methods exist for estimating the parameters $\boldsymbol{\theta}$ and $\boldsymbol{D}$ in model (1). Maximum likelihood (ML) [11], restricted ML (REML) [12], the method of moments (MM) [13], the generalized least squares method (GLS) [14], and the nonparametric U-statistics method (UM) [15] are among the

frequentist options. These differ primarily in the estimation of the between-study covariance matrix, $\boldsymbol{D}$; hence, inferences about the between-study variances and correlations can be expected to vary from method to method.

A number of benefits and drawbacks to likelihood-based methods are discussed in the literature. Mavridis *et al*. [16] mention several benefits, including the fact that REML estimates of effect sizes are unbiased. Limitations of likelihood-based methods are discussed in References [16, 17]. For example, departures from the normality assumption of Y may seriously bias estimates of variance components. Reference [13] shows that MM performance is similar to ML and REML when p is small and there is moderate to large heterogeneity. The estimates based on UM have been shown to be robust to highly skewed data in a study that simulated non-normal random effects and sampling errors [15], while References [9, 10] and [15] have noted an upward bias in between-study variance estimates when using other frequentist methods. This bias is seen as a consequence of restricting $\boldsymbol{D}$ to be positive definite rather than an estimation problem.

There has been significant growth in the use of Bayesian methods over the past 25 years. A distinguishing characteristic of the Bayesian approach is that parameters are estimated from the posterior probability distribution, obtained via Bayes theorem upon specifying a likelihood and a prior distribution for all unknown parameters (e.g., [18–20]). The development of computer-based Markov chain Monte Carlo (MCMC) methods, combined with increases in computing power and the availability of free and commercial software for Bayesian computation like WinBUGS [21], R [22], STATA [23], SAS [24] and Mplus [25], are major contributors to the indicated growth.

Bayesian analysis brings a flexible framework to incorporate prior information when available. When non-informative priors are considered, inferences based on Bayesian and classical methods also provide results that are often very similar. Further advantages of Bayesian methods include the following: it is easier to make inferences for functions of parameters like tail probabilities; the interpretation of the inferences is straightforward, for example, a given parameter has a probability of 0.95 of falling in a 95% credible interval; and it provides a convenient setting for hierarchical models and missing data problems. Among the disadvantages are as follows: MCMC methods and computer power are often needed to facilitate Bayesian inferences; from a practical point of view, the prior selection process can be arbitrary and tedious; and when the likelihood is not that informative, posterior distributions can be highly sensitive to the prior distributions, often as a result of sample size, model specification, or both. Establishment of posterior robustness or sensitivity analysis to the choice of prior is highly recommended for any Bayesian analysis [18].

Applied to the hierarchical meta-analysis model (1), the use of Bayesian methods to estimate $\boldsymbol{\theta}$ and $\boldsymbol{D}$ is known as multivariate Bayesian meta-analysis (MBMA) and requires a prior specification for $(\boldsymbol{\theta}, \boldsymbol{D})$. In general, a challenging problem is to define a prior distribution for a covariance matrix (e.g., $\boldsymbol{D}$) such that the posterior estimate of the covariance matrix remains symmetric and positive definite. In a normal hierarchical model like (1), the most frequently used matrix prior is the conjugate inverse Wishart prior [20]. Subjective informative priors [26] can be potentially advantageous when prior information is elicitable. For example, Reference [27] used empirical data to obtain predictive distributions for the heterogeneity variances. However, in many applications, little or no prior information about the between-study covariance matrix (especially correlations) is available. In those settings, non-informative, weakly informative, least informative, objective, or reference priors are then preferred [28–31]. We will use the general term non-informative prior (NIP) to refer to any prior that does not incorporate subjective probability; objective and reference priors will be considered here because they have been rigorously studied for the multivariate normal model [30, 31]. A number of benefits and drawbacks in the use of NIPs are discussed in the literature. NIPs minimize subjectivity by constructing priors based only on the model and the observed data. However, they are often improper [18], which can sometimes lead to improper posterior distributions. Some NIPs are location-scale invariant distributions [18], but others are not invariant under re-parametrization. For example, a uniform prior on the variances may not be uniform on their logarithm. Finally, other types of priors, like hierarchical priors, have also been proposed for the estimation of $(\boldsymbol{\theta}, \boldsymbol{D})$ with desirable properties (shrinkable, interpretable, and invariant under re-parameterizations) [32–34].

The growth in the use of Bayesian methods cited earlier extends to medical research, including meta-analysis. We performed a literature review to assess the use of Bayesian meta-analysis in clinical papers published between 1995 and 2014 the types of priors used, use of sensitivity analysis, existence of multiple outcomes, and uses of MBMA. We found 102 clinical papers utilizing Bayesian meta-analysis with applications in a variety of medical fields. There were 7 papers published before 2000, 16 papers from 2000 to 2006, 36 papers from 2007 to 2010, and 43 papers published between 2011 and 2014. Of the

studies identified, 67 used non-informative priors and 19 did not report the prior choice. About half of the studies did not report any sensitivity analysis for the choice of prior. Fifty-nine studies analyzed multiple endpoints; however, 47 of those papers did not use a multivariate meta-analysis method. We also found that the majority of the clinical papers applying meta-analysis models do not fully list the data, making it difficult to reproduce the analysis. The clinical papers range from small size to large-size meta-analysis, with a mean of 28 studies analyzed per paper (standard deviation = 40.13) and a median of 16 (interquartile range = 20). Clinical papers using meta-analysis models with large number of studies are common on the following fields: cardiovascular disease (number of papers found: $n = 12$, median number of studies: $m = 19$; psychiatry; genetics ($n = 7$, $m = 17$); and vaccine therapy ($n = 2$, $m = 42$). Small-sized meta-analysis studies were found in fields such as oncology ($n = 7$, $m = 6$) and surgery ($n = 2$, $m = 7.5$).

Advantages of frequentist-based multivariate meta-analysis approaches have been established over the last decade [10]. Advantages of univariate Bayesian meta-analysis over frequentist methods have been discussed in [35]. Additionally, Reference [36] showed through a comprehensive simulation study, in the univariate meta-analysis case, that vague prior distributions are highly influential, particularly in small studies. With the rapid pace of adoption of Bayesian meta-analysis in the published literature and the frequent use of univariate meta-analysis in multivariate settings, guidelines for prior choice and implementation of multivariate meta-analysis in the Bayesian context would be valuable. In this paper, we investigate the impact of the prior specification for $\boldsymbol{D}$ on the overall inferences for $(\boldsymbol{\theta}, \boldsymbol{D})$ in order to help inform the guidelines about the use of priors and related sensitivity analyses. In Section 2, we review a Bayesian model framework for multivariate meta-analysis and summarize several suitable prior choices for the covariance matrix. A simulation study to examine the impact of the prior distribution in an MBMA setting is performed in Section 3. Two examples involving a small ($n = 5$) and medium ($n = 21$) meta-analysis from the periodontal and stroke fields are presented in Section 4. In Section 5, we discuss the findings. Additional plots, convergence analysis, code, and all simulation results are available in the Supporting Information.

2. Bayesian hierarchical models for multivariate meta-analysis

2.1. Notation and model formulation

Consider a systematic review of n studies where each one estimated the $p \times 1$ vector of effect sizes, $\boldsymbol{Y}_i$ along with their variance covariance matrices, $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, n$. We assume that each vector of effect sizes, $\boldsymbol{Y}_i$ (e.g., standardized mean differences, hazard ratios, or logarithm of the odds ratios) is normally distributed. The random effect model given by (1) can be rewritten as a two-level hierarchical model of the form: for $1 \leqslant i \leqslant n$,

$$\boldsymbol{Y}_i|\boldsymbol{\mu}_i \sim \mathrm{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \text{ and } \boldsymbol{\mu}_i|\boldsymbol{\theta}, \boldsymbol{D} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{D}) \tag{2}$$

The formulation of model (2) implies that multivariate meta-analysis models assume a known common set of endpoints across studies. When a set of endpoints is missing, other approaches like multiple imputation or network meta-analysis can be used. It is important to emphasize here that the $\boldsymbol{\Sigma}_i$ matrices are mostly assumed to be known for multivariate meta-analysis models. In practice, many publications report within-study variances, but within-study correlations are often only partially reported or missing. We assumed throughout the paper that the $\boldsymbol{\Sigma}_i$ matrices are known. In general, multiple imputation techniques are often used when some data points are missing at random. Reference [37] introduces a series of asymptotic estimators of within-study covariance for continuous and dichotomous outcomes.

Traditionally, the primary goal in a meta-analysis has been the estimation and inference of the mean vector $\boldsymbol{\theta}$ and the study mean effects, $\boldsymbol{\mu}_i$; however, the study of the whole distribution of random effects may also be considered as important [19]. In a multivariate meta-analysis, we need to estimate the between-study covariance matrix, $\boldsymbol{D}$. We assume that $(\boldsymbol{\theta}, \boldsymbol{D})$ are random with a prior distribution of the form $\pi(\boldsymbol{\theta}, \boldsymbol{D}) = \pi(\boldsymbol{\theta})\pi(\boldsymbol{D})$. Inference about $(\boldsymbol{\theta}, \boldsymbol{D})$ is based on the posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{D})$, given the observed data $(\boldsymbol{y}_i, \boldsymbol{\Sigma}_i)$, which are proportional to

$$\left[\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[(\boldsymbol{y}_i - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) + (\boldsymbol{\mu}_i - \boldsymbol{\theta})'\boldsymbol{D}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\theta})\right]\right\}\right]\pi(\boldsymbol{\theta})\pi(\boldsymbol{D}) \tag{3}$$

Prior distributions for $\boldsymbol{\theta}$ under model (2) have been studied in the literature [30, 32]. These priors include an improper prior of the form $\pi(\boldsymbol{\theta}) \propto 1$, and a normal prior $\mathrm{N}_p(0, \boldsymbol{C}_0)$, where $\boldsymbol{C}_0$ is a known $p \times p$ matrix. Assuming, for example, a known variance covariance matrix $\boldsymbol{D}$ and $\boldsymbol{\theta} \sim \mathrm{N}_p(0, \boldsymbol{C}_0)$, the conditional posterior densities of $\boldsymbol{\mu}_i$ and $\boldsymbol{\theta}$ on $\mathfrak{D}_i = \{\boldsymbol{y}_i, \boldsymbol{\Sigma}_i, \boldsymbol{D}\}$ are:

$$\boldsymbol{\mu}_i|\boldsymbol{\theta}, \mathfrak{D}_i \sim \mathrm{N}_p\left(\boldsymbol{V}_i(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i + \boldsymbol{D}^{-1}\boldsymbol{\theta}), \boldsymbol{V}_i\right) \text{ and } \boldsymbol{\theta}|\boldsymbol{\mu}_i, \mathfrak{D}_i \sim \mathrm{N}_p\left(V_{\boldsymbol{\theta}}\boldsymbol{D}^{-1}\sum_{i=1}^{n}\boldsymbol{\mu}_i, \boldsymbol{V}_{\boldsymbol{\theta}}\right), \tag{4}$$

where $\boldsymbol{V}_i = (\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{D}^{-1})^{-1}$ and $\boldsymbol{V}_{\boldsymbol{\theta}} = (n\boldsymbol{D}^{-1} + \boldsymbol{C}_0^{-1})^{-1}$.

The matrix $\boldsymbol{D}$ is the covariance matrix on the second level of the hierarchical meta-analysis model given by (2) and is called the between-study covariance. It contributes to the marginal variability of $\mathbf{Y}_i$ because $Var(\mathbf{Y}_i) = \boldsymbol{D} + \boldsymbol{\Sigma}_i$. The diagonal elements of $\boldsymbol{D}$ represent the level of heterogeneity between studies for each outcome, and its off-diagonal elements are the outcomes between-study correlations. Multiple endpoints are likely to be correlated within a study, but the role of the between-study correlation is less obvious. The between-study correlations indicate how the population 'true' effect sizes are correlated across studies, and these can be affected by differences across studies, that is, study-design characteristics, study-level population characteristics, study-level variable definition, among others. One of the advantages of using multivariate meta-analysis is the estimation of effect sizes with higher degree of precision by borrowing of strength through the between-study and within-study correlations [5, 8, 15].

It is generally unreasonable to consider the between-study variance matrix, $\boldsymbol{D}$, as known. Therefore, a prior distribution for $\boldsymbol{D}$ needs to be selected. Modeling a covariance matrix is crucial in any multivariate setup and becomes more difficult as the dimension increases because of the quadratic growth in the number of parameters and the need to force the matrix to remain non-negative definite. Prior distributions for covariance matrices in hierarchical models are frequently chosen casually; for example, the inverse Wishart distribution is a common choice in MBMA clinical applications because of its conjugacy, although other prior distributions for matrices are available. Another modeling issue is the use of informative versus non-informative priors distributions for $\boldsymbol{D}$. Non-informative priors distributions are often improper. In the context of a hierarchical model such as (2), caution is advised because not all choices of improper priors for $\boldsymbol{D}$ will lead to a proper posterior distribution. There are methods to obtain informative distributions for covariance matrices [38–40]; however, obtaining informative priors for a second-level covariance matrix in a hierarchical model is more challenging because, for example, historical data are usually scarce. Reference [41] shows how expert knowledge and beliefs can formally be elicited. In particular, if a closed form prior distribution for $\boldsymbol{D}$ is chosen (i.e., inverse Wishart, $\mathrm{IW}_p(\nu, \boldsymbol{R})$), one could elicit information about the hyperparameters $\nu, \boldsymbol{R}$ and obtain an informative prior on either the covariance matrix $\boldsymbol{D}$ or the hyperparameters ν and $\boldsymbol{R}$. Some clinical applications have considered independent priors on each one of the covariance components instead of formulating a prior distribution for the whole matrix $\boldsymbol{R}$. This is in part because of the fact that the elicitation of priors for variances and correlations is easier than the specification of an informative prior for the covariance matrix. However, when considering independent priors, restrictions to ensure that $\boldsymbol{D}$ is positive definite need to be addressed before model implementation.

We briefly discuss several families of potentially suitable prior choices for $\boldsymbol{D}$ under model (2). Specifically, we consider the conjugate inverse Wishart prior parametrization under non-informative and informative set-ups; objective and reference priors for $\boldsymbol{D}$ that lead to a proper posterior distribution for the hierarchical model (2); independent priors on the covariance components; and, finally, the constrained and mixture Wishart prior distribution families. Where helpful or informative, we illustrate the parametrization of these priors assuming that $p = 2$ and parameterize $\boldsymbol{D}$ and $\boldsymbol{\Sigma}_i$ as follows:

$$\boldsymbol{D} = \begin{bmatrix} \tau_1^2 & \tau_{12} \\ \tau_{12} & \tau_2^2 \end{bmatrix}, \ \boldsymbol{\Sigma}_i = \begin{bmatrix} \sigma_{i1}^2 & \sigma_{i12} \\ \sigma_{i12} & \sigma_{i2}^2 \end{bmatrix}, \tag{5}$$
$$\tau_{12} = \tau_1\tau_2\rho_b, \ \sigma_{i12} = \sigma_{i1}\sigma_{i2}\rho_{iw},$$

where $\tau_1 > 0, \tau_2 > 0, -1 < \rho_b < 1, \sigma_{i1} > 0, \sigma_{i2} > 0$, and $-1 < \rho_{iw} < 1 \ \forall_i$. Here, τ_1^2, τ_2^2, and ρ_b describe the between-study variances and between-study correlation, whereas σ_{i1}^2, σ_{i2}^2, and ρ_{iw} capture the within-study variation and corresponding correlation. We also define $\boldsymbol{\Sigma}_0$ as $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Sigma}_i$ and denote by $\lambda_1 \geqslant \lambda_2$ the ordered eigenvalues of $\boldsymbol{D}$.

2.2. Conjugate prior

A $p \times p$ matrix $\boldsymbol{X}$ has a Wishart distribution, $\mathrm{W}_p(\nu, \boldsymbol{R})$, if its associated probability density function is proportional to $|\boldsymbol{X}|^{(\nu-p-1)/2} \exp\{tr(-\boldsymbol{X}\boldsymbol{R}^{-1})/2\}, \boldsymbol{X} > 0$, where $\nu > 0$ is the degrees of freedom and $\boldsymbol{R}$ is a $p \times p$ symmetric scale matrix.

An inverse Wishart prior distribution [18] for $\boldsymbol{D}$ is equivalent to assuming that $\boldsymbol{D}^{-1}$ has a Wishart distribution. Specifically, assuming that $\boldsymbol{D}^{-1} \sim \mathrm{W}_p(\nu, \boldsymbol{R}^{-1})$, the prior mean of the matrix $\boldsymbol{D}^{-1}$ is $\nu\boldsymbol{R}^{-1}$, which corresponds to a prior mean of $\boldsymbol{D}$ equal to $\boldsymbol{R}/(\nu - p - 1)$. The condition $\nu > p + 1$ implies a finite prior mean for $\boldsymbol{D}$, but the Wishart prior for $\boldsymbol{D}^{-1}$ is proper for any $\nu > p - 1$.

An inverse Wishart prior of the form $\mathrm{IW}_p(p, \boldsymbol{I}_p)$ is commonly used as a non-informative proper prior (see Section 1 of the Supporting Information for figures and discussion), and its implementation is straightforward (see the WinBUGS code provided in Section 4 of the Supporting Information). More generally, the selection of ν and $\boldsymbol{R}$ is not straightforward. The matrix $\boldsymbol{R}$ is often set as the ML estimator of $\boldsymbol{D}$ for non-hierarchical normal models. An empirical Bayes approach to estimate ν has been used by [42] in a non-hierarchical normal model. In general, the specification of $\boldsymbol{R}$ and ν can be influential. A simulation study evaluated several priors and showed that inverse Wishart prior distributions (hierarchical prior and fixed ν) for estimating the covariance matrix have suboptimal performance in terms of having high Bayes risk under the loss function $L(\tilde{\boldsymbol{D}}, \boldsymbol{D}) = tr(\tilde{\boldsymbol{D}}\boldsymbol{D}^{-1}) - \log|\tilde{\boldsymbol{D}}\boldsymbol{D}^{-1}| - p$.

2.3. Objective and reference priors

In general, NIPs may be useful and easier to specify when there is no expert-based or historical prior information. Among the NIPs, objective and reference priors have been rigorously studied in the context of multivariate normal models [30,31]. The two most popular objective prior distributions are the Laplace prior and Jeffreys's invariant prior. Under the parameterization given by (5), the Laplace prior is a constant prior that weighs all possibilities equally (i.e., $\pi_L(\tau_{ij}) \propto 1, \forall_{i,j} \in \{1, 2\}$); however, it is not invariant under reparameterization. The non-hierarchical Jeffreys's invariant prior $\pi_J(\boldsymbol{D}) = |\boldsymbol{D}|^{-(p-1)/2}$ yields an improper posterior in hierarchical settings. The hierarchical Jeffreys's invariant prior, $\pi_{HJ}(\boldsymbol{D}) = |\Sigma_0 + \boldsymbol{D}|^{-(p-1)/2}$, leads to a proper posterior, but it has been shown to behave poorly for $p > 1$ [33].

Under the parameterization (5), the non-hierarchical Jeffreys's invariant prior reduces to $\frac{1}{\tau_1\tau_2(1-\rho_b^2)^{1/2}}$. Reference [31] proposed other objective priors for non-hierarchical bivariate normal models. Among the recommended priors for $\boldsymbol{D}$ parameterized as in (5) are $\pi_H(\boldsymbol{D}) \propto \frac{1}{\tau_1\tau_2(1-\rho_b^2)^{3/2}}$, $\pi_{\boldsymbol{R}\rho_b}(\boldsymbol{D}) \propto \frac{1}{\tau_1\tau_2(1-\rho_b^2)}$ and $\pi_{R\sigma}(\boldsymbol{D}) \propto \frac{\sqrt{1+\rho_b^2}}{\tau_1\tau_2(1-\rho_b^2)}$. According to [30], priors $\pi_H, \pi_{R\rho_b}$, and $\pi_{R\sigma}$ could lead to an improper posterior distribution in the context of the hierarchical MBMA model (2).

Reference [30] proposed two reference priors appropriate for model (2). Recalling that λ_i denotes the *ith* eigenvalue of $\boldsymbol{D}$, these priors are $\pi_{HRPa}(\boldsymbol{D}) = \left[|\boldsymbol{D}|^{-(2p-1)/2p} \prod_{i<j}(\lambda_i - \lambda_j)\right]^{-1}$ and $\pi_{HRPb}(\boldsymbol{D}) = \left[|\Sigma_0 + \boldsymbol{D}| \prod_{i<j}(\lambda_i - \lambda_j)\right]^{-1}$. Section 1 in the Supporting Information provides a graphical comparison of the $\pi_{HRPa}(\boldsymbol{D})$ and $\pi_{HRPb}(\boldsymbol{D})$ distributions for $p = 2$. Reference priors $\pi_{HRPa}(\boldsymbol{D})$ and $\pi_{HRPb}(\boldsymbol{D})$ lead to a proper posterior distribution in multilevel hierarchical normal models and are easy to implement using a Gibbs sampler with Metropolis–Hastings steps; Section 4 in the Supporting Information contains a sample of the code implemented in R. In general, reference priors have been shown to outperform Jeffreys' priors in multivariate problems (i.e., when using both entropy and quadratic loss functions) [29,30,43]. Additionally, Jeffreys' prior fails to shrink the eigenvalues appropriately in multivariate situations [29].

2.4. Independent priors for the covariance components

There are several specifications of independent priors for the components of the covariance matrix $\boldsymbol{D}$. We considered a prior formulation assuming that all the elements of the covariance matrix are independent a priori. We discuss a prior formulation for $p = 2$ under (5), but the methods are easily extended to higher dimensions.

It is typical to assume that $1/\tau_j^2 \overset{\text{iid}}{\sim} Gamma(\epsilon, \epsilon), j = 1, 2$ for some $\epsilon > 0$. Other candidate priors for τ_1^2 and τ_2^2 include the log-uniform and the uniform distributions. Priors on transformation of τ_j^2, e.g., $\log(\tau_j^2)$, can also be considered. For modeling ρ_b, a uniform prior on $\frac{\rho_b}{1-\rho_b}$ was suggested in [10], a published discussion from a multivariate meta-analysis event held at the Royal Statistical Society;

however, this transformation is only defined for $\rho_b > 0$ and overwhelmingly favors boundary values. Reference [33] instead proposed a normal prior, $N(0, \sigma^2_{\rho_b})$, for Fisher-Z $(\rho_b) = \frac{1}{2}\log(\frac{1+\rho_b}{1-\rho_b})$. The normal prior on the Fisher-Z transformation has a shrinking effect toward zero (when $\sigma^2_{\rho_b}$ is small, this prior favors values of ρ_b close to 0). The normal distribution for the Fisher-Z transformation proposed earlier should be truncated over the relevant values of the correlations in order to guarantee a positive definite matrix $\boldsymbol{D}$; see the Supporting Information, namely Section 1 for plots and Section 4 for references about code implementation. Reference [44] considered a hyper-prior on $\sigma^2_{\rho_b}$, where $\pi(\sigma^2_{\rho_b}) \propto (c + \sigma^2_{\rho_b})^{-2}$ and c is a constant that represents a variance. For example, c can be set to be $\frac{1}{n-3}$, the variance of the Fisher-Z transformation.

2.5. A prior formulation based on the constrained Wishart distribution

Reference [45] proposed a constrained Wishart prior in the context of the two-level hierarchical normal model described by (2). A $p \times p$ matrix $\boldsymbol{X}$ has a constrained Wishart distribution denoted by $CW_p(\nu, \boldsymbol{R}; \boldsymbol{Q})$ if the density is proportional to $|\boldsymbol{X}|^{(\nu-p-1)/2} \exp\{tr(-\boldsymbol{X}\boldsymbol{R}^{-1})/2\}, \boldsymbol{X} > 0, \boldsymbol{Q} - \boldsymbol{X} \geqslant 0$, where $\nu > 0$ is the degrees of freedom, $\boldsymbol{R}$ is a $p \times p$ symmetric scale matrix, and $\boldsymbol{Q}$ is a diagonal constraint matrix.

A constrained Wishart prior on the transformation $\boldsymbol{B}_0 = \boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Sigma}_0 + \boldsymbol{D})^{-1}\boldsymbol{\Sigma}_0^{1/2}$ of the form $CW_p(\nu, \boldsymbol{R}; \boldsymbol{I}_p)$, where $\boldsymbol{R} \geqslant 0$ and $\nu > 1-n$, is the conjugate prior for the transformation $\boldsymbol{B}_0$ in model (2) when $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_0 \, \forall_i$. In the unequal covariance case, a constrained Wishart distribution of the form $CW_p(\nu^*, \boldsymbol{R}^*; \boldsymbol{I}_p)$ envelops the posterior distribution of $\boldsymbol{B_0}|\boldsymbol{Y}$ for any $\boldsymbol{R}^*$, provided $\nu^* \leqslant n - 1 + \nu$, where ν is the degrees of freedom of the constrained Wishart prior for $\boldsymbol{B}_0$. The fact that the CW distribution envelops the actual posterior makes it easy to implement the model using rejection sampling or importance sampling algorithms; see Section 4 of the Supporting Information for references about implementation using R. References [45] and [46] show that models using constrained Wishart priors do not require MCMC for implementation; hence, there are no MCMC convergence issues to worry about. Reference [45] also shows that their model has good frequentist performance (including coverage) when the prior on their transformation matrix is uniformly distributed.

This class of priors contains several interesting special cases. Setting the smallest eigenvalue of $\mathbf{Q}$ to ∞ gives the Wishart prior for $\mathbf{B}_0$; $\boldsymbol{R} = 0$ and $\nu = -(p+1)$ is equivalent to a uniform prior on $\boldsymbol{D}$ and $\boldsymbol{R} = 0$ and $\nu = 0$ is a Jeffrey's prior on $\mathbf{B}_0$; $\boldsymbol{R} = 0$ and $\nu = p + 1$ corresponds to the uniform shrinkage prior on $\mathbf{B}_0$. Section 1 in the Supporting Information contains a graphical comparison of constrained Wishart distributions of the form $\mathrm{CW}_2(\nu, \boldsymbol{I}_2; \boldsymbol{I}_2)$ using marginal histograms and scatter plots for several values of ν.

2.6. A prior formulation based on a mixture Wishart distribution

Reference [47] proposed a mixture Wishart distribution in the context of a random effect regression model. One of the advantages of this family is its conditional conjugancy, yet it is possible to select hyperparameters such that the distributions on the variances and correlations are not noninformative priors.

A $p \times p$ positive definite matrix $\boldsymbol{X}$ has a mixture Wishart distribution denoted by $MW_p(\nu, A_1, \ldots, A_p)$ if $\boldsymbol{X}|a_1, \ldots, a_p \sim IW_p(\nu + p - 1, 2\nu \, \mathrm{diag}(1/a_1, \ldots, 1/a_p))$, and $a_k \overset{\text{ind.}}{\sim} Inv - Gamma(1/2, 1/A_k^2); k = 1, \ldots, p$, where $\mathrm{diag}(1/a_1, \ldots, 1/a_p)$ denotes the diagonal matrix with diagonal elements $1/a_1, \ldots, 1/a_p$, $\overset{\text{ind.}}{\sim}$ stands for independently distributed, ν and $A_1, \ldots, A_p$ are positive scalars. The closed form density of $\boldsymbol{X}$ is proportional to $|\boldsymbol{X}|^{-(\nu+2p)/2} \prod_{k=1}^{p} \{\nu(X^{-1})_{kk} + 1/A_k^2\}^{-(\nu+p)/2}$, where $(C)_{kk}$ denotes the k, k element of C, but the scale mixture representation leads to a simpler implementation.

The assumption that $\boldsymbol{D} \sim MW_p(\nu, A_1, \ldots, A_p)$ implies that a priori $\tau_k \sim Half - t(\nu, A_k)$; hence, large values of A_k lead to weakly informative priors for τ_k. It can also be shown that the distribution of ρ_b is proportional to $(1 - \rho_b^2)^{\nu/2-1}$ so that $\nu = 2$ induces a uniform prior on ρ_b. Assuming a mixture Wishart prior $MW_p(\nu, A_1, \ldots, A_p)$ for $\boldsymbol{D}$ and a normal $N_p(0, \sigma^2\boldsymbol{I}_p)$ prior for θ in model (2) leads to conjugate conditional distributions for all the parameters including $a_k; k = 1, \ldots, p$ and $\boldsymbol{D}$. Section 1 in the Supporting Information contains plots and further discussion about the $MW_p(\nu, A_1, \ldots, A_p)$; Section 4 contains code.

3. Simulation study

3.1. Data generation and prior specifications

Each simulated meta-analysis is a set of hypothetical clinical trials comparing a treatment arm with a control. Data sets representing small ($n = 10$), medium ($n = 30$) and large meta-analyses ($n = 50$) were generated following [15]. Equation 2 is used to generate each meta-analysis data set, assuming that $\boldsymbol{Y}_i$ follows a bivariate normal distribution with given marginal mean $\boldsymbol{\theta}^*$ and given variance $\boldsymbol{D}^* + \boldsymbol{\Sigma}_i$. The variance matrices $\boldsymbol{D}^*$ and $\boldsymbol{\Sigma}_i$ are assumed to satisfy (5) such that (a) $\rho_{iw} = \rho_w$ for each i and (b) $\rho_w = \rho_b$. The parameters $(\boldsymbol{\theta}^*, \boldsymbol{D}^*)$ and variance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n$ used to generate each simulated dataset are further determined according to the following specifications:

Specification 1. The vector $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*)^\top$ is fixed at $(10, 10)^\top$.

Specification 2. The within-study variance matrices, $\boldsymbol{\Sigma}_i$, are obtained by fixing the within-study correlation ρ_w and then independently generating variance parameters σ_{i1}^2 and σ_{i2}^2 from an exponential distribution. Without loss of generality, the rate of the exponential distribution is fixed to 0.50 and all within-study variances were truncated to the range of $[0.50, 10]$, resulting in σ_i^2 with a mean of 2.50, median of 1.80, and first and third quartiles of 1.10 and 3.20, respectively. A new set of n matrices is generated for each simulated meta-analysis data set.

Specification 3. The between-study variances τ_j^{*2} (i.e., the diagonal elements of $\boldsymbol{D}^*$) are generated using the formulas

$$\tau_j^{*2} = \left(\frac{I_j^2}{1 - I_j^2} \right) s_j^2, \text{ where } s_j^2 = \lim_{n \to \infty} \frac{(n-1) \sum_{i=1}^n \sigma_{ij}^{-2}}{\left(\sum_{i=1}^n \sigma_{ij}^{-2} \right)^2 - \sum_{i=1}^n \sigma_{ij}^{-4}}; \; j = 1, 2, \qquad (6)$$

where I_j^2 captures the percentage of variability in the treatment estimates that can be attributed to heterogeneity between studies rather than to sampling error [6]. Three values of I_j^2 are considered, respectively reflecting low $\left(I_j^2 = 0.25\right)$, moderate $\left(I_j^2 = 0.50\right)$, and high $\left(I_j^2 = 0.85\right)$ between-study heterogeneity among estimates for each outcome. In combination with Specification 2, we obtain $\tau_j^{*2} = 0.49$ when $I_j^2 = 0.25$, $\tau_j^{*2} = 1.48$ when $I_j^2 = 0.50$, and $\tau_j^{*2} = 8.40$ when $I_j^2 = 0.85$.

With the aforementioned specifications for generating data, ten simulation scenarios are then considered for each of the three sample sizes. These ten simulation scenarios correspond to ten different choices of $\boldsymbol{D}^*$ (see (5)). Specifically we consider mild ($\rho_b = \rho_w = 0.3$) and strong ($\rho_b = \rho_w = 0.8$) correlations in combination with the following choices of (I_1^2, I_2^2): $(0.25, 0.25)$; $(0.50, 0.50)$; $(0.85, 0.85)$; $(0.25, 0.50)$; $(0.50, 0.85)$. For each of the ten scenarios for each sample size, 1000 data sets are generated; the MBMA model as described by (2) is then fit using each of the following prior specifications:

Prior (1) Six specifications of Wishart conjugate priors (Section 2.2) of the form $W_2(\nu, \boldsymbol{R})$ using $\nu = 2, 3, 4$, and $\boldsymbol{R} = c\boldsymbol{I}_2$ for $c \in \{0.1, 1\}$, where $\boldsymbol{I}_2$ denotes the 2×2 identity matrix.

Prior (2) Two reference priors (Section 2.3) of the form:

2.1 $\pi_{HRPa}(\boldsymbol{D}) = \left[|\boldsymbol{D}|^{-3/4} (\lambda_1 - \lambda_2) \right]^{-1}$, and
2.2 $\pi_{HRPb}(\boldsymbol{D}) = \left[|\boldsymbol{\Sigma}_0 + \boldsymbol{D}| (\lambda_1 - \lambda_2) \right]^{-1}$.

Prior (3) Twelve different specifications of independent priors (Section 2.4), where $\epsilon \in \{0.10, 1\}$,

3.1 $1/\tau_1^2 \sim \text{Gamma}(\epsilon, \epsilon)$, $1/\tau_2^2 \sim \text{Gamma}(\epsilon, \epsilon)$; and $\rho_b \sim \text{Uniform}[-1, 1]$,
3.2 $1/\tau_1^2 \sim \text{Gamma}(\epsilon, \epsilon)$, $1/\tau_2^2 \sim \text{Gamma}(\epsilon, \epsilon)$; and ρ_b is fixed at 0 (independent case),
3.3 $1/\tau_1^2 \sim \text{Gamma}(\epsilon, \epsilon)$, $1/\tau_2^2 \sim \text{Gamma}(\epsilon, \epsilon)$; and ρ_b is fixed at its true value,
3.4 $1/\tau_1^2 \sim \text{Gamma}(\epsilon, \epsilon)$, $1/\tau_2^2 \sim \text{Gamma}(\epsilon, \epsilon)$; and Fisher-Z$(\rho_b) = \frac{1}{2} \log(\frac{1+\rho_b}{1-\rho_b}) \sim N(0, r^{-1})$, for $r = 0.20, 0.40, 0.80$.

Prior (4) Twelve specifications of constrained Wishart prior distributions for $\mathbf{B}_0$ using $CW_2(\nu, \mathbf{0}; \mathbf{I}_2)$ (Section 2.5) with $\nu = 1 - n, 4 - n, -3, 0, 2, 3, 6, 9, 12, 15, 18, 21$, where n is the sample size.

Theoretically, any value of $\nu > 1 - n$ leads to a proper posterior distribution for $\mathbf{B}_0$; some exploratory analysis shows substantial variability in the distribution of matrix components corresponding to distributions with those degrees of freedom.

Prior (5) Six different specifications of the mixture of Wishart priors (Section 2.6) of the form $MW_p(\nu, A_1, \ldots, A_p)$ with $p = 2$, $\nu = 2, 3, 4$ and $A_1 = A_2 = c$ where $c \in \{10^2, 10^5\}$.

3.2. Model implementation

Models implemented with Priors (1) and (3) were fit using a Gibbs sampling algorithm (WinBUGS [21]). Models using Prior (2) were implemented by a Gibbs sampler with metropolis Hastings steps (R [22]), models using Prior (5) were implemented using Gibbs sampling (R [22]), and finally, models that considered Prior (4) were fit using a rejection sampling algorithm in R [22]. We used 5000 burn-in iterations and 10 000 additional sampling iterations when fitting models by any MCMC algorithm.

When implementing Bayesian models, it is important to consider identifiability and its impact on MCMC convergence. We discuss identifiability issues for model (2) from two perspectives: Bayesian identifiability (proper posterior distributions) and classical identifiability (estimable models from the perspective of the likelihood). Bayesian identifiability refers to the property of the posterior distribution; in this sense, a model is identifiable (from the Bayesian perspective) if its posterior distribution is proper. In general, a model with proper priors always has a proper posterior distribution; however, improper priors may lead to improper posteriors. All the models proposed here have proper posterior distributions. We implement model (2) using non-informative and weakly informative proper prior distributions with exception of the family of reference priors. Reference [30] shows that the reference priors $\pi_{HRPa}(\boldsymbol{D})$ and $\pi_{HRPb}(\boldsymbol{D})$ lead to proper posterior distributions in multilevel hierarchical normal models.

Classically, models with non-identifiable parameters are those in which all possible sets of observations have identical probabilities for two different sets of parameters [48]. In this sense, identifiability is a property of the likelihood (model specification) and is closely related to overparameterization. Bayesian inference is possible for non-identifiable models (from the perspective of the likelihood) when informative priors are used. However, non-identifiable models typically yield highly correlated parameters in the posterior density, creating problems with autocorrelated parameters and slow convergence in MCMC; in practice, it is also possible for identifiable models to exhibit these problems.

The impact of lack of identifiability on MCMC convergence for Bayesian hierarchical models incorporating random effects has been studied in the literature [49, 50]. The main issues in this regard are the convergence of the MCMC method (simulated values are in fact draws from the posterior distribution) and autocorrelation in the sequence of Markov chain iterations. According to Reference [49], estimable functions of mean parameters of multilevel normal models should exhibit lagged autocorrelations tending to zero even when non-informative priors are used. These authors also show that iterative sampling results in convergence of estimable parameters even when priors have large variances.

We monitor convergence by measuring the mixing of independent chains on a fraction of the simulated data sets by checking that each chain seems to have reached a stationary distribution. In principle, convergence diagnostics should be assessed for each model that is fit; however, it is impossible to assess convergence for all of the 1.14×10^6 ($38 \times 10 \times 3 \times 1000$) models in our simulation. Therefore, we checked for convergence in a random subgroup of simulated data sets and the examples presented in Section 4. According to Reference [48], more than one criterion should be considered when monitoring convergence. In particular, we used trace plots, autocorrelation plots, Gelman and Rubin diagnostics [51], and the Geweke test [52] to monitor convergence. Figures 7 to 13 in Section 2 in the Supporting Information show an example of the plots and statistics we used to analyze convergence for a random subgroup of simulated data sets and for the examples.

Overall, we did not detect strong evidence of lack of convergence. In general, the trace plots show good mixing, the sample autocorrelation function plots indicate a low degree of correlation between the draws (i.e., acceptable mixing), and plots of the kernel density estimates of posterior density functions show unimodal distributions. The Geweke tests or Gelman and Rubin diagnostics were above the desired cutoff for a few parameters (18 cases out of $2 \times 60 \times 5$ combinations of sample sizes/models/parameters we considered), but in no case did we see Geweke tests and Gelman and Rubin diagnostics to be unsatisfactory for the same parameter. A few Gelman and Rubin diagnostics (G.R) were above 1 for some of the τs. Models with conjugate family of priors were the only ones reporting Geweke test statistics (G) for some θs above the cutoff of 3, but this never occurred for all three chains.

3.3. Simulation results

The simulation results are summarized using frequentist criteria with the rationale that (i) Bayesian analysis under non-informative or weakly informative priors often leads to estimates that are similar to those derived from frequentist methods and (ii) the data are generated using specific choices of several key parameters. The relative bias of the averaged posterior parameters are reported over the 1000 data set replicates, where R-Bias$(z) = (\hat{z} - z) \times 100/z$ with $z \in \{\theta_1^*, \theta_2^*, \tau_1^{*2}, \tau_2^{*2}, \rho_b\}$, $\hat{z}_j$ is the corresponding posterior mean estimate of z obtained at the j^{th} simulation run and $\hat{z} = \frac{1}{1000}\sum \hat{z}_j$. The mean-squared error MSE $(\hat{z})$ is also estimated using $\sum_{j=1}^{1000}(\hat{z}_j - z)^2/1000$. Frequentist coverage, computed for the marginal 95% credible intervals (Cr. I), is also reported because Bayesian intervals constructed using certain classes of prior tend to exhibit comparable frequentist behavior. From this point forward, we do not make a distinction between the random variables $\boldsymbol{\theta}$ and $\boldsymbol{D}$ and the fixed values of these parameters that were used to generate the data, it being implicitly understood that the posterior distributions corresponding to the former should increasingly concentrate around the latter as n becomes larger.

This section discusses Figures 1–5; Section 3 in the Supporting Information (Tables S1–18) presents and discusses simulation results concerning all prior specifications and all sample sizes. Figures 1–5 summarize simulation results of 9 of the 38 prior specifications reported on this simulation for a sample size of $n = 10$ and concentrates on posterior means for θ_1, τ_1^2, and ρ_b, but conclusions can be extended to estimates for θ_2 and τ_2^2. Relative to what is reported here, the trends across priors are generally similar for $n = 30$ and 50, but as expected, the performance under all priors also improves and becomes more similar across priors. Figures 1 to 5 compare the best priors for each of the five families of distributions described earlier in terms of the trend of the relative bias of $\hat{\theta}_1$, the mean squared error of $\hat{\theta}_1$, the coverage probability of $\hat{\theta}_1$, and the relative biases of $\hat{\tau}_1^2$ and $\hat{\rho}_b$. In general, point estimates of effect sizes (θ_1, θ_2) were not observed to be particularly sensitive to the choice of prior for $\boldsymbol{D}$, whereas those for covariance estimates $(\tau_1^2, \tau_2^2, \rho_b)$ were observed to be much more sensitive to prior choice.

The relative bias of $\hat{\theta}_1$ was less than 0.88% (in absolute value) for all priors with the exception of the reference prior π_{HRPb} (Figure 1). Prior π_{HRPb} displays higher absolute relative bias of $\hat{\theta}_1$ for all scenarios with a maximum relative bias of 2.50% when $I_1^2 = I_2^2 = 0.85, \rho = 0.80$. In general, the absolute relative bias of $\hat{\theta}_1$ tends to increase slightly as I_1^2 increases; this bias does not seem to depend on I_2^2 or ρ_b.

The MSE of $\hat{\theta}_1$ behaves similarly among all priors: it is observed to increase moderately as the associated heterogeneity index (I_1^2) moves from 0.25 to 0.5, then rises more substantially when $I_1^2 = 0.85$ (Figure 2). The MSE of $\hat{\theta}_1$ seems to behave similarly for scenarios with $I_1^2 \neq I_2^2$ and scenarios with $I_1^2 = I_2^2$ for the vast majority of priors. As seen in Figure 2, the mixture of Wishart priors $MW(4, 10^5)$ displays a higher MSE of $\hat{\theta}_1$ for almost all scenarios. Generally speaking, the MSE of $\hat{\theta}_1$ is not observed to depend on ρ_b. In view of the fact that the biases of the posterior mean estimators tend to be low, this suggests that the variability of the posterior mean estimate primarily depends on the underlying heterogeneity for that component.

Figure 3 displays the coverage probability of $\hat{\theta}_1$. The reference prior π_{HRPb} consistently overcovered θ_1 (maximum coverage probability of 0.99), the coverage level depending little on the heterogeneity level or correlation. The mixture of Wishart priors was observed to have coverage reasonably close to nominal, typically overcovering when correlation was high. The remaining priors had close to or somewhat higher than nominal coverage when the heterogeneity levels $\left(I_1^2 = I_2^2 = 0.25\right)$ were low and exhibited a tendency to undercover θ_1 in the presence of higher heterogeneity, with $CW(15, 0; \boldsymbol{I}_2)$ exhibiting the worst performance and greatest sensitivity to the heterogeneity level. Within each prior, the coverage probability of $\hat{\theta}_1$ varies with I^2, but there is no uniformly worst or best prior among all the scenarios.

In summary, for a sample size of $n = 10$, all estimators of $\boldsymbol{\theta}$ have reasonable relative bias, although the prior with the worse relative bias performance for all scenarios is the reference prior π_{HRPb}. According to the MSE criteria, the worst performing prior for all scenarios is the $MW(4, 10^5)$. The picture is not as clear when evaluating coverage probabilities: the same prior can have close-to-nominal probabilities for some scenarios and be farther from nominal coverage for others. In general, the Wishart priors on $\boldsymbol{D}^{-1}$ ($W(4, \boldsymbol{I}_2)$ and $W(2, 0.1\boldsymbol{I}_2)$), the reference prior π_{HRPb}, and $CW(15, 0; \boldsymbol{I}_2)$ have the least desirable performance when estimating θ_1. The most reasonable trade-off between small relative bias, low MSEs and close-to-nominal coverage probabilities for $\hat{\theta}_1$ seems to be obtained using the following priors: $1/\tau_j^2 \overset{\text{iid}}{\sim} Gamma(\epsilon, \epsilon), j = 1, 2$ and Fisher-Z $\left(\rho_b\right) \sim \text{N}(0, r)$, for $\epsilon = 1$ and $r \in \{0.20, 0.40\}$ and $CW_2(3, \boldsymbol{0}; \boldsymbol{I}_2)$ and $CW_2(6, \boldsymbol{0}; \boldsymbol{I}_2)$.

In general, the absolute relative bias of $\hat{\tau}_1^2$ is observed to be highly sensitive to the choice of prior, especially for scenarios where $I_1^2 = 0.25$. Over all priors, the absolute relative bias of $\hat{\tau}_1^2$ typically decreased as

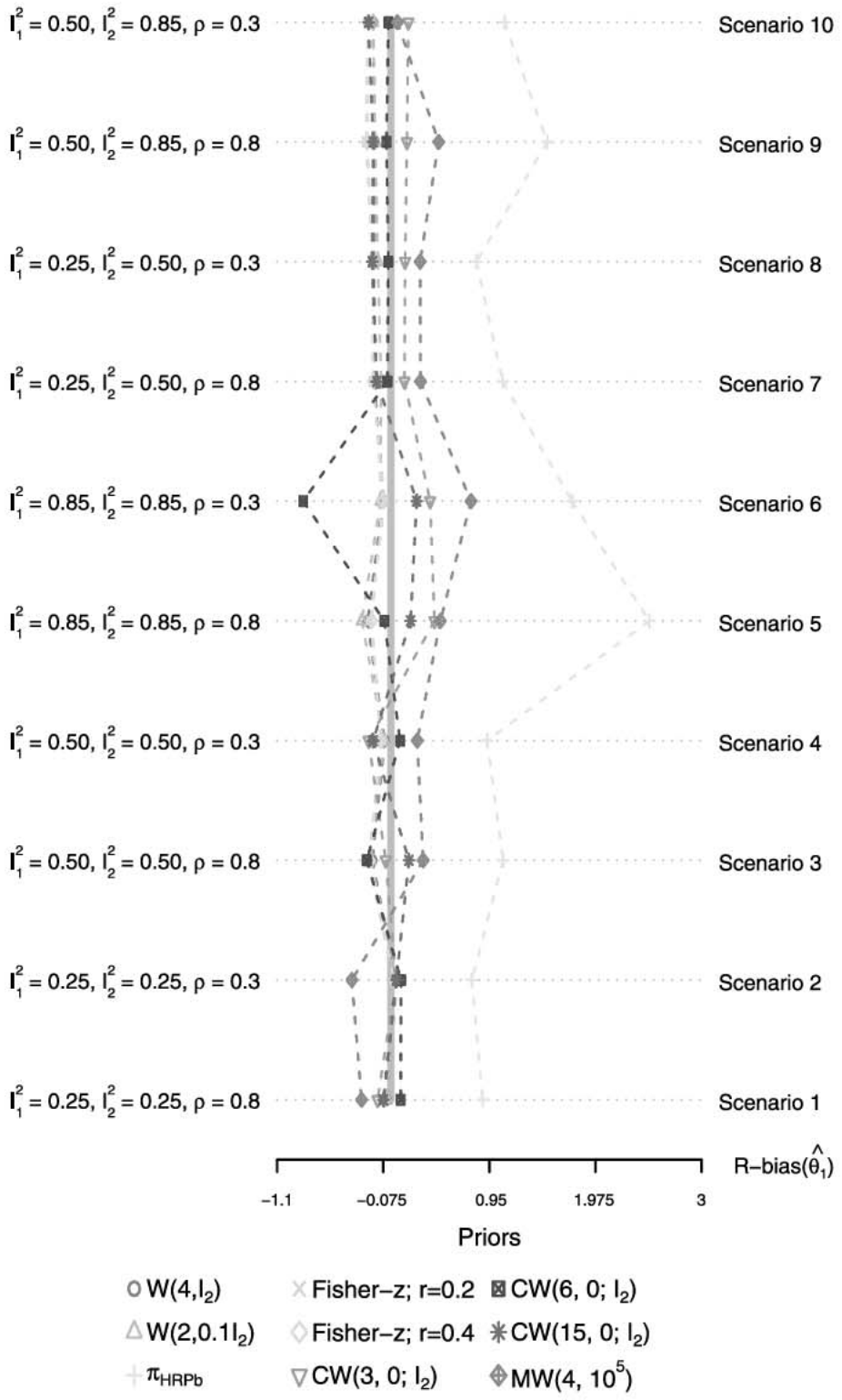

Figure 1. Summary of R-bias ($\hat{\theta}_1$) results for selected priors when sample size n=10.

its associated heterogeneity index I_1^2 increased. Except for the Wishart priors, the level of bias observed did not depend strongly on whether the heterogeneity indices are equal; there is clear evidence of a negative impact for the Wishart priors when heterogeneity levels are unequal. The most consistent estimators of τ_1^2 among all scenarios are given by models using the $MW(4, 10^5)$ prior distribution. The $CW_2(3, \mathbf{0}; \boldsymbol{I}_2)$ and $CW_2(6, \mathbf{0}; \boldsymbol{I}_2)$ priors had the highest relative biases when the heterogeneity level was 0.25, the former

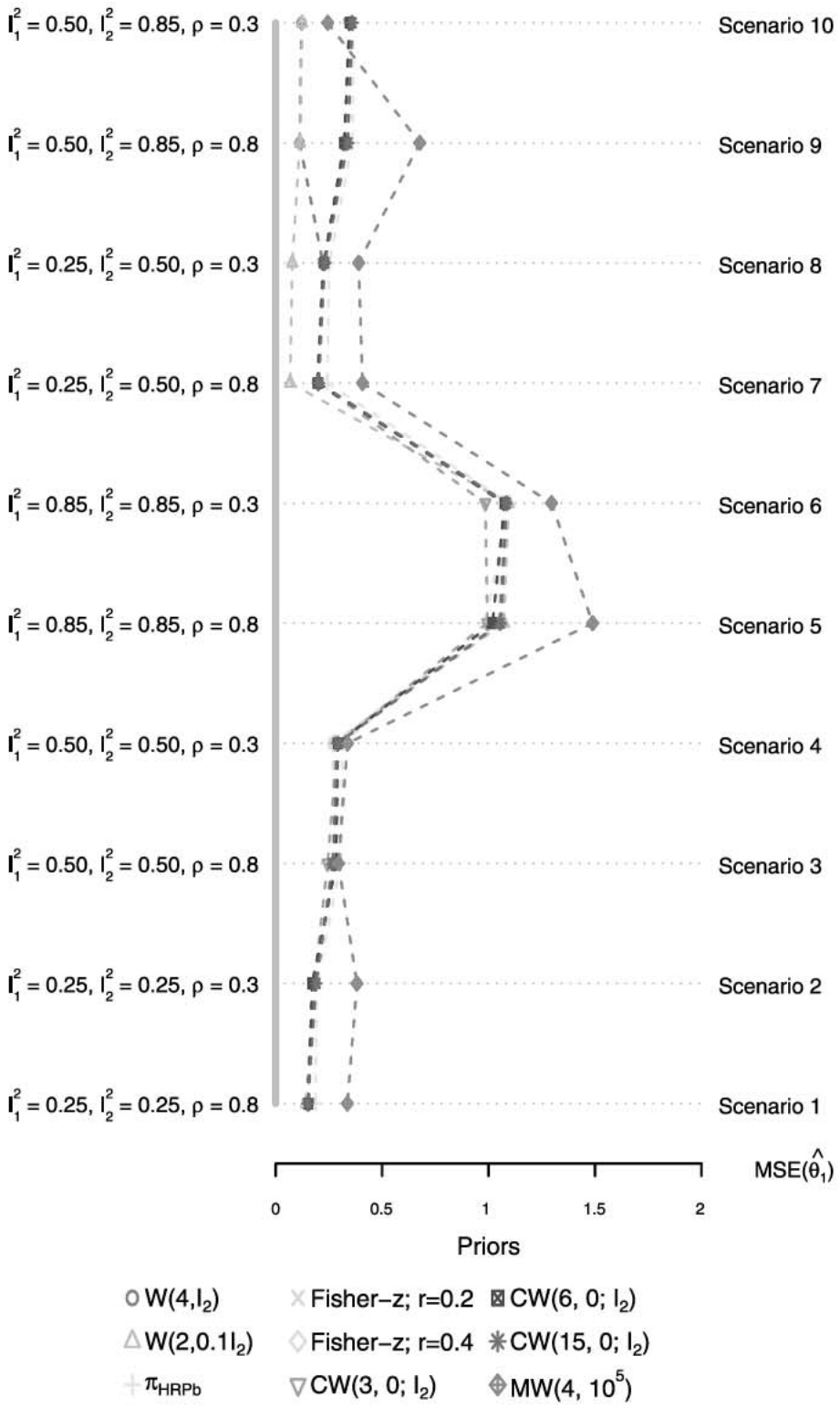

Figure 2. Summary of mean-squared error ($\hat{\theta}_1$) results for selected priors when sample size n=10.

also being a bit worse than the latter; the level of bias decreased dramatically with increased heterogeneity levels. Interestingly, the $CW_2(15, \mathbf{0}; \boldsymbol{I}_2)$ prior tended to exhibit complementary behavior, providing among the best performance when heterogeneity is low and exhibiting larger biases when heterogeneity is high. A small absolute relative bias of $\widehat{\tau}_1^2$ ($< 5\%$) was usually associated with coverage probabilities closer to the nominal 95%, but it did not influence the trend of the MSE of $\hat{\theta}_1$. The estimate of ρ_b was

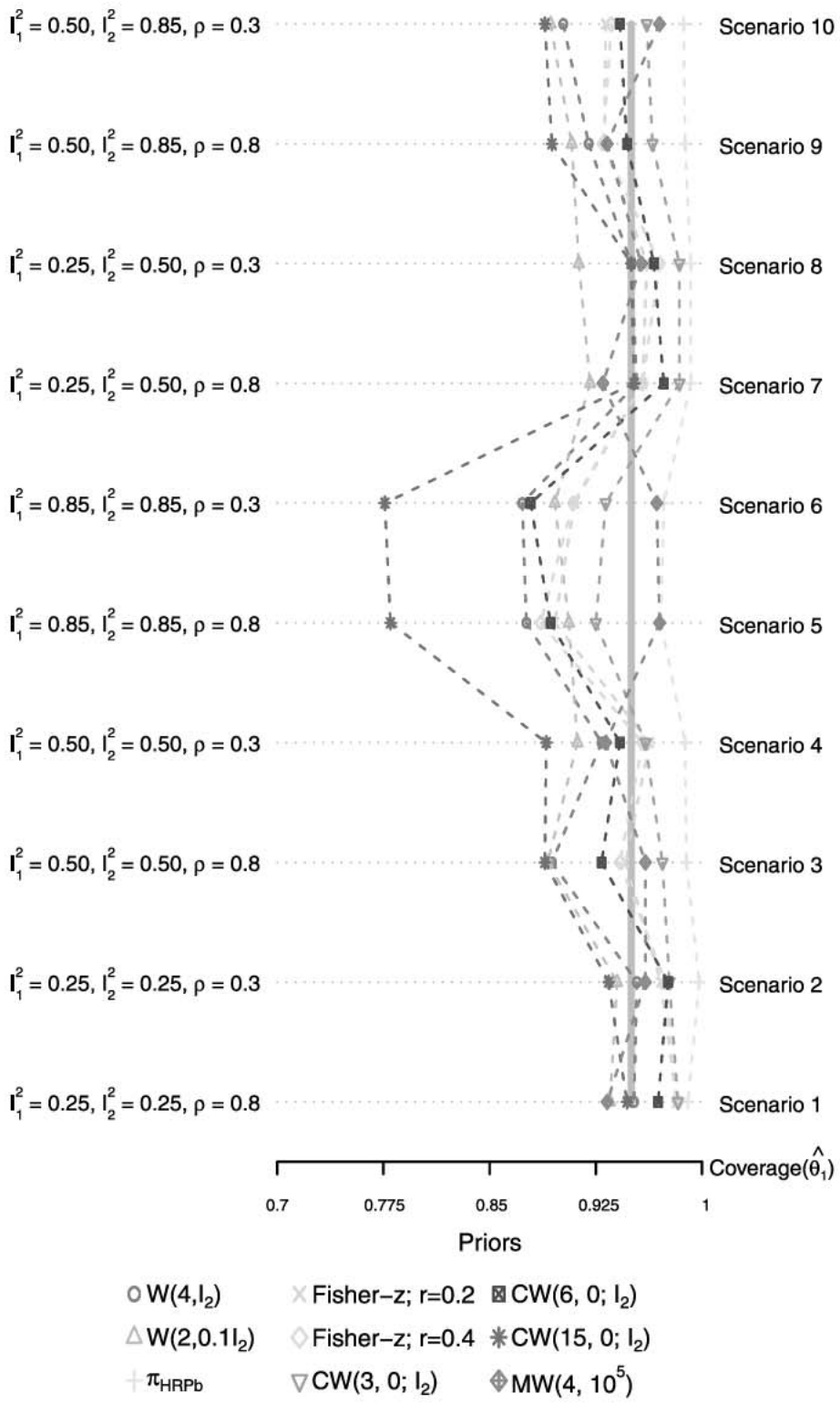

Figure 3. Summary of coverage ($\hat{\theta}_1$) results for selected priors when sample size n=10.

observed to be unusually sensitive to the choice of prior (Figure 5). The absolute value of the relative bias of $\hat{\rho}_b$ is usually higher than 20% for all priors, with the exception of the constrained Wishart priors and the reference priors. The Wishart and independence priors always underestimated ρ_b; the mixture of Wishart priors always overestimated ρ_b, the bias being observed to be relatively insensitive to the $\boldsymbol{D}$ matrix. The constrained Wishart priors and the reference priors resulted in the smallest relative biases for $\hat{\rho}_b$

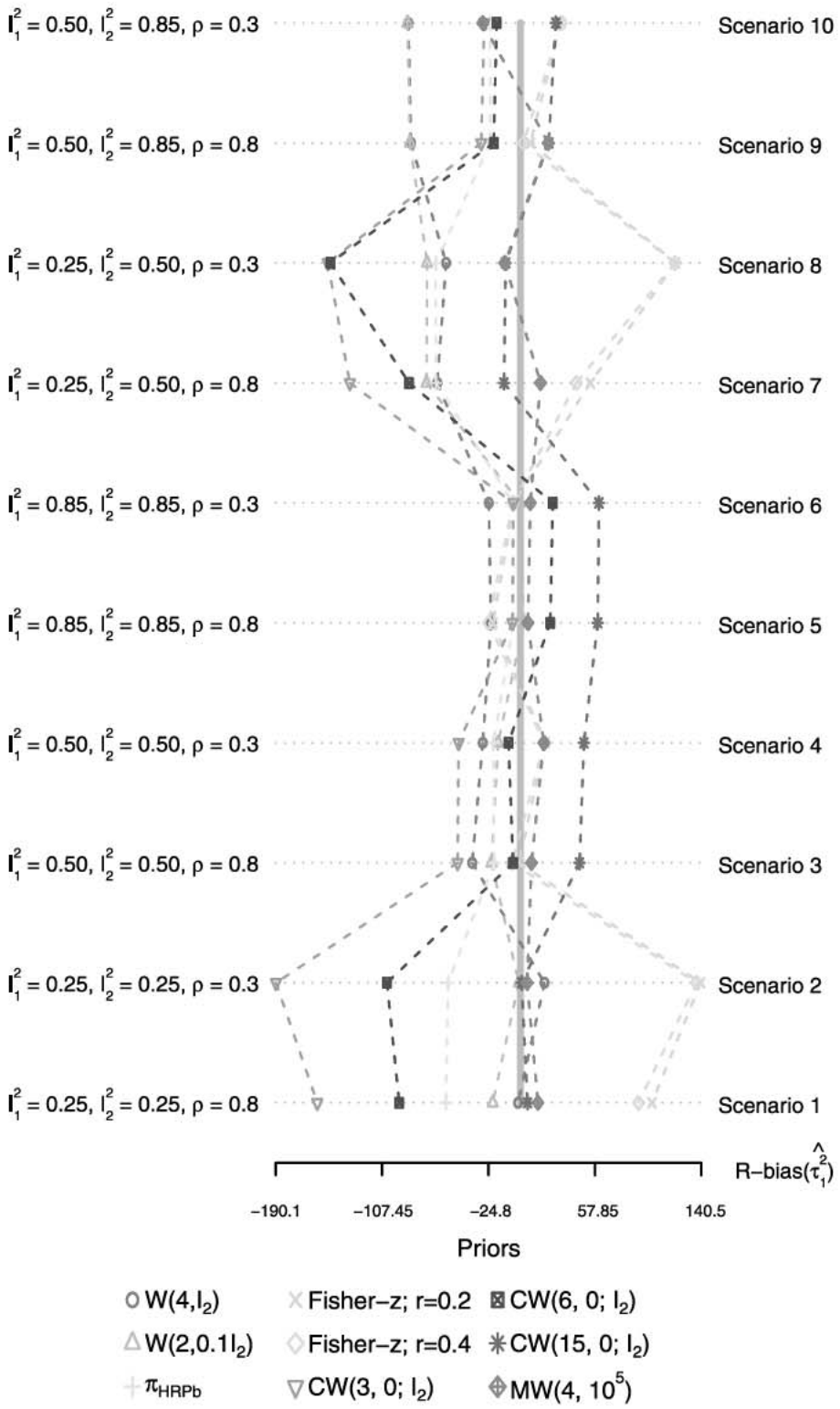

Figure 4. Summary of R-bias ($\hat{\tau}_1$) results for selected priors when sample size n =10.

in Scenarios 1–6 (same heterogeneity index for both outcomes), underestimating ρ_b; these priors overestimated ρ_b when heterogeneity indexes were not equal, but the bias tended to remain among the smallest.

Overall, for estimating the covariance matrix, the mixture of Wishart priors tended to exhibit the most stable, if not always the best, performance, being relatively insensitive to the choice of $\boldsymbol{D}$. The reference

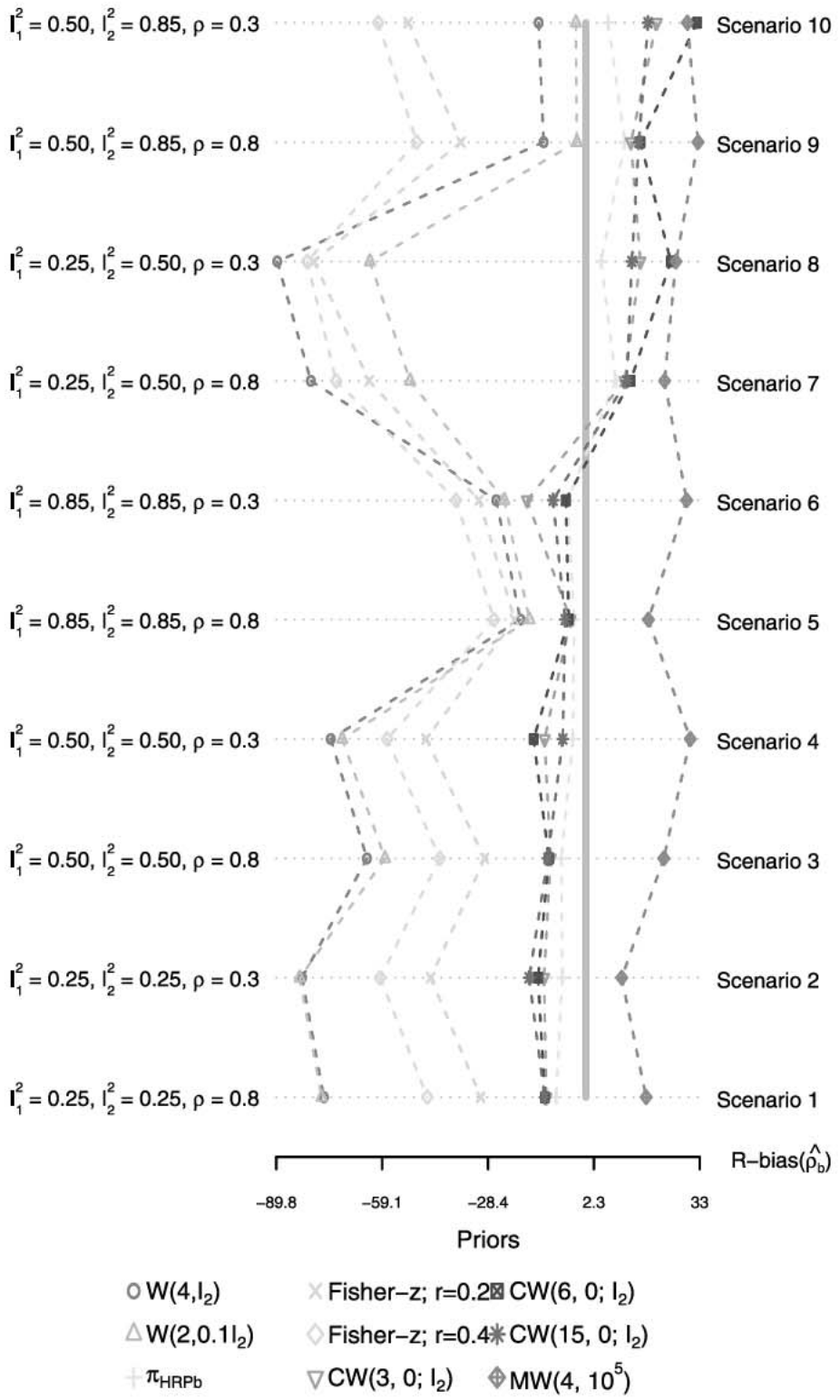

Figure 5. Summary of R-bias($\hat{\rho}_b$) results for selected priors when sample size n =10.

prior and the constrained Wishart priors appear to be better choices (i.e., with respect to the criteria used here), provided that the heterogeneity level for each outcome is not too low; here, the $CW_2(15, \mathbf{0}; \boldsymbol{I}_2)$ prior exhibited the most stable performance among the set of choices considered for the constrained Wishart prior.

3.4. Computational efficiency

A model with a constrained Wishart prior had the highest computational efficiency. Models using constrained Wishart priors are fit based on a rejection sampling algorithm [45], being more efficient than MCMC sampling methods. For a simulated data set of size 10 (under the assumptions of scenario 1), it took a model with a constrained Wishart prior (see Supporting Information for code) on average 0.97 s to produce the results. In contrast, a model with a reference prior was the slowest to run (average of 86.76 s). A model with a mixture of Wishart distributions took 68.50 s to produce a result. The other two MCMC-based algorithms (models with Wishart and independent priors) were written in Winbugs; it took 18.72 and 25.19 s on average to finish the computations.

4. Examples

We apply the proposed method to two published meta-analyses. The first one is a meta-analysis of 21 placebo-controlled trials of vasoactive drugs for acute stroke and the second one is a small-sized meta-analysis of five randomized controlled trials that compared a surgical versus a non-surgical procedure for treating periodontal disease.

4.1. A meta-analysis of 21 trials comparing vasoactive drugs for acute stroke

We consider a meta-analysis of placebo-controlled trials of vasoactive drugs for acute stroke [53] with two endpoints: systolic blood pressure (SBP, in mHg) and diastolic blood pressure (DBP, in mHg). The data are available in the published literature [37, 53] and summarized in terms of mean differences of each endpoint between the treatment and the control group, $\left(Y_{it}, Y_{ic}\right)$; the corresponding within-study variances, $(\sigma_{it}^2, \sigma_{ic}^2)$; and the within-study correlations (ρ_{iw}). We follow techniques and assumptions in Reference [37] to approximate the within-study covariances. A positive Y_{ij} indicates that the treatment group has lower outcome (SBP or DBP) than the control group.

The vasoactive drugs for acute stroke data were analyzed using the model given by (2). This model was fit using MBMA techniques with each of the nine prior choices for $\boldsymbol{D}$ deemed to be the best, based on the simulation results. The heterogeneity indices are $45\%\left(I_1^2\right)$ and $70\%\left(I_2^2\right)$, meaning that $45\%\left(I_1^2\right)$ of the total variation in SBP was from the between-study variation and 70% of the total variation in DBP was from the between-study variation.

As shown in Table I, all priors produced similar point estimates for θ. The negative estimate implies that SBP and DBP were lower in the treatment group than in the control group; however, the difference is not more than 2.70 mHg for both outcomes. The 95% marginal credible intervals for θ_1 and θ_2 are quite different among the fit models. For example, models with $W(4, \boldsymbol{I}_2)$ and π_{HRP_b} priors produce 95% credible intervals that contain zero for both θ_1 and θ_2, while all the other credible intervals suggest a significantly different from zero effect for both outcomes (with the exception of the model with a $W(2, 0.1\boldsymbol{I}_2)$ prior).

The model with the $W(4, \boldsymbol{I}_2)$ prior has the lowest point estimate for τ_1 and τ_2, while the model with π_{HRP_b} has the highest point estimate for τ_1 and τ_2. The constrained Wishart priors give much tighter credible intervals for τ_1 and τ_2 than the other methods.

The models considered here do not agree about point estimates and credible intervals for ρ_b. Among all the 95% credible intervals for ρ_b, the ones that did not include zero are those obtained using $W(2, 0.1\boldsymbol{I}_2)$, π_{HRP_b}, constrained Wishart, and mixture of Wishart prior distributions. It is also interesting to note that the frequentist point estimates reported by Reference [53] for all parameters are more consistent with those obtained by the model with constrained Wishart and mixture of Wishart prior distributions.

4.2. A meta-analysis of five trials comparing a surgical versus non-surgical procedure for treating periodontal disease

Riley *et al*. [8] illustrated the use of frequentist multivariate meta-analysis methods with a study of five randomized controlled trials that compared a surgical versus a non-surgical procedure for treating periodontal disease with two endpoints: probing depth (mm) and attachment level (mm) 1 year after treatment. The data are available in the published literature [15, 54] and summarized in terms of mean differences of each endpoint between the two treatments (surgical minus non-surgical), $\left(Y_{i1}, Y_{i2}\right)$; the corresponding within-study variances, $(\sigma_{i1}^2, \sigma_{i2}^2)$; and the within-study correlations (ρ_{iw}). A positive Y_{ij} indicates that the surgical procedure produces a better patient outcome than the non-surgical one.

Table I. Bivariate meta-analysis results of Example 1—vasoactive drugs for acute stroke.

	SBP ($I_1^2 = 0.45$)				DBP ($I_2^2 = 0.70$)				Correlation	
Prior	$\hat{\theta}_1$	95% Cr.I	$\hat{\tau}_1$	95% Cr.I	$\hat{\theta}_2$	95% Cr.I	$\hat{\tau}_2$	95% Cr.I	$\hat{\rho}_b$	95% Cr.I
$W(4, I_2)$	−2.223	(−7.969, 4.879)	0.400	(0.063, 1.883)	−2.598	(−7.845, 5.323)	0.379	(0.071, 2.087)	0.024	(−0.997, 0.958)
$W(2, 0.1I_2)$	−2.409	(−4.881, 0.134)	2.976	(0.125, 5.433)	−2.387	(−4.289,−0.479)	3.083	(1.378, 4.754)	0.994	(0.948, 0.999)
π_{HRPb}	−2.663	(−6.682, 1.683)	7.818	(5.248,10.610)	−2.296	(−4.649, 0.156)	4.470	(2.983, 6.095)	0.702	(0.433, 0.902)
$\frac{1}{\tau_j} \sim \gamma(1, 1)$:Fisher-z ($\rho_b$);r=.2	−2.503	(−4.768,−0.098)	1.962	(0.132, 4.556)	−2.307	(−3.985,−0.612)	2.436	(1.086,3.789)	0.207	(−0.227, 0.558)
$\frac{1}{\tau_j} \sim \gamma(1, 1)$:Fisher-z ($\rho_b$);r=.4	−2.575	(−4.960,−0.215)	2.395	(0.139, 4.883)	−2.380	(−4.070,−0.703)	2.497	(1.054, 3.886)	0.339	(−0.177, 0.791)
$CW(3, 0; I_2)$	−2.670	(−2.739,−2.600)	4.953	(4.804, 5.112)	−2.417	(−2.466,−2.369)	3.880	(3.763, 4.004)	0.885	(0.875, 0.894)
$CW(6, 0; I_2)$	−2.676	(−2.743,−2.610)	4.590	(4.452, 4.737)	−2.432	(−2.478,−2.386)	3.588	(3.480, 3.703)	0.881	(0.870, 0.890)
$CW(15, 0; I_2)$	−2.666	(−2.727,−2.606)	3.874	(3.758, 3.998)	−2.454	(−2.496, −2.412)	3.093	(2.989, 3.192)	0.893	(0.883, 0.901)
$MW(4, 10^5, 10^5)$	−2.662	(−5.534,−0.123)	3.713	(0.346, 6.114)	−2.401	(−4.177,−0.389)	3.060	(1.521, 4.639)	0.624	(0.060, 0.942)

I_j^2 denotes the I^2 statistic for the j^{th} outcome. $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are the estimates of the population effect sizes, $\widehat{\tau}_1^2$ and $\widehat{\tau}_2^2$ are the estimates of the between-study variances, and $\widehat{\rho}_b$ denotes the between-study correlation. 95% Cr.I corresponds to the 95% credible interval.

Table II. Bivariate meta-analysis results of Example 2—surgical versus non-surgical procedure for treating periodontal disease.

	Probing depth ($I_1^2 = 0.69$)				Attachment level ($I_2^2 = 0.96$)				Correlation	
Prior	$\hat{\theta}_1$	95% Cr.I	$\hat{\tau}_1$	95% Cr.I	$\hat{\theta}_2$	95% Cr.I	$\hat{\tau}_2$	95% Cr.I	$\hat{\rho}_b$	95% Cr.I
$W(4, \boldsymbol{I}_2)$	0.373	(−0.051, 0.782)	0.174	(0.042, 0.499)	−0.353	(−0.778, 0.082)	0.180	(0.046, 0.539)	0.051	(−0.616, 0.715)
$W(2, 0.1\boldsymbol{I}_2)$	0.367	(0.151, 0.605)	0.040	(0.006, 0.165)	−0.348	(−0.600,−0.093)	0.054	(0.011, 0.214)	0.225	(−0.592, 0.909)
π_{HRPb}	0.367	(0.115, 0.634)	0.039	(0.005, 0.264)	−0.345	(−0.627,−0.038)	0.051	(0.006, 0.358)	0.436	(−0.591, 0.990)
$\frac{1}{\tau_j} \sim \gamma(1,1)$:Fisher-z ($\rho_b$);r=.2	0.373	(0.045, 0.710)	0.081	(0.015, 0.390)	−0.347	(−0.700,−0.006)	0.098	(0.015, 0.431)	0.030	(−0.397, 0.427)
$\frac{1}{\tau_j} \sim \gamma(1,1)$:Fisher-z ($\rho_b$);r=.4	0.374	(0.053, 0.702)	0.083	(0.014, 0.389)	−0.346	(−0.703,−0.001)	0.097	(0.016, 0.448)	0.056	(−0.521, 0.604)
$CW(3, 0; \boldsymbol{I}_2)$	0.348	(0.345, 0.351)	0.014	(0.013, 0.015)	−0.340	(−0.344,−0.335)	0.040	(0.037, 0.043)	0.583	(0.554, 0.612)
$CW(6, 0; \boldsymbol{I}_2)$	0.343	(0.340, 0.345)	0.007	(0.008, 0.009)	−0.340	(−0.343,−0.336)	0.025	(0.024, 0.027)	0.609	(0.580, 0.635)
$CW(15, 0; \boldsymbol{I}_2)$	0.335	(0.333, 0.337)	0.003	(0.002, 0.004)	−0.341	(−0.344,−0.339)	0.012	(0.011, 0.013)	0.677	(0.653, 0.700)
$MW(4, 10^5, 10^5)$	0.357	(0.131, 0.592)	0.020	(0.001, 0.198)	−0.354	(−0.655,−0.026)	0.058	(0.008, 0.421)	0.145	(−0.528, 0.843)

I_j^2 denotes the I^2 statistic for the j^{th} outcome. $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are the estimates of the population effect sizes, $\widehat{\tau}_1^2$ and $\widehat{\tau}_2^2$ are the estimates of the between-study variances, and $\widehat{\rho}_b$ denotes the between-study correlation. 95% Cr.I corresponds to the 95% credible interval.

The periodontal disease data were analyzed using the model given by (2). This model was fit using MBMA techniques with each of the nine prior choices for $\boldsymbol{D}$ deemed to be the best, based on the simulation results. The heterogeneity indices are $69\%\left(I_1^2\right)$ and $96\%\left(I_2^2\right)$, meaning that $69\%\left(I_1^2\right)$ of the total variation in probing depth was from the between-study variation and 96% of the total variation in attachment level was from the between-study variation.

As shown in Table II, all priors produced significant estimates for θ, indicating that the probing depth improved for the surgical group and that the attachment level improved for the non-surgical procedure. However, the 95% marginal credible intervals for θ_1 and θ_2 are quite different. For example, a model with a $W(4, \boldsymbol{I}_2)$ prior produces 95% credible intervals that contain zero for both θ_1 and θ_2, while all the other credible intervals suggest a significantly different from zero effect for both outcomes. Point estimates and credible intervals for τ_1 and τ_2 are also different among the fit models. The model with the $W(4, \boldsymbol{I}_2)$ prior produces higher point estimates and wider credible intervals than the other models. The models considered here do not agree about point estimates and credible intervals for ρ_b. Among all the 95% credible intervals for ρ_b, the only ones that did not include zero are those obtained using constrained Wishart prior distributions. It is also interesting to note that the frequentist point estimates for ρ_b reported in [15] are pretty similar to those obtained by the model with constrained Wishart prior distributions.

5. Discussion

We have carried out an extensive simulation study designed to illustrate the impact of the choice of prior distribution for the covariance matrix ($\boldsymbol{D}$) on estimating in a MBMA model with normal random effects (see (2)). The data were generated as a set of hypothetical clinical trials comparing a treatment arm with a control.

Five families of priors for $\boldsymbol{D}$, for a total of 38 different prior distributions, were studied under ten different simulation scenarios for three samples sizes representing small ($n = 10$), medium ($n = 30$), and large meta-analyses ($n = 50$). The simulation results are summarized using the frequentist criteria of relative bias, MSE, and coverage probability with the rationale that Bayesian analysis with non-informative or weakly informative priors can lead to the same estimates derived from the frequentist methods.

The simulation study shows that point estimates of effect sizes (θ_1, θ_2) are not particularly sensitive to the choice of prior for $\boldsymbol{D}$; in contrast, and not unexpected, the results for the covariance parameters $(\tau_1^2, \tau_2^2, \rho_b)$ varied more widely under different prior specifications, especially so when the number of studies was small. Large biases in the posterior mean estimators of covariance matrix parameters were observed to be associated with poor frequentist performance of Bayesian credible intervals for the effect sizes, suggesting that the use of different priors for the covariance matrix $\boldsymbol{D}$ can lead to different inferences for (θ_1, θ_2).

The results we obtained do not point to a uniformly best choice of prior family, but do provide insights on the differences in inferences produced by different priors. The conjugate Wishart prior and the independent prior family of distributions did not perform well in terms of the MSE and coverage probability criteria for θ_1, θ_2, nor in terms of relative bias for τ_1, τ_2, ρ_b. In general, the reference prior family provided good summary measures, except for large relative biases of ρ. At least one of the constrained Wishart priors and the mixture of Wishart prior typically had equal or better performance for all the parameters, but there was no consistent winner. In the case where heterogeneity was higher, however, the uniform shrinkage prior $CW(3, \mathbf{0}; \boldsymbol{I}_2)$ arguably provided the best overall balance in frequentist performance. For these reasons, we recommend the use of either the class of constrained Wishart or mixture of Wishart prior distributions, especially when the number of studies is less than 30.

Models using constrained Wishart priors do not require MCMC for implementation. Reference [45] also shows that their model has good frequentist performance (including coverage) when the prior on their transformation matrix is uniformly distributed; however, they do not provide theoretical results or discussion that helps to explain why the constrained Wishart priors perform well. Reference [47] shows through a simulation that the mixture of Wishart distributions family has better performance over the classical Wishart prior. In our study, the performance of the constrained Wishart priors for a given heterogeneity level was observed to depend on the specified degree of freedom parameter (ν). As shown in (6), the heterogeneity index I_j^2 is a statistic that solely depends on the value of τ_j^{*2} and s_j^2, not on the value of ρ_b. Considering a fixed value of s_j^2, it is easy to see that τ_j^{*2} increases as I_j^2 increases. Hence, with higher heterogeneity levels, the matrix $\boldsymbol{D}^*$ used to generate each meta-analysis data set had larger values on the diagonal compared with that when heterogeneity was smaller. A closer look at $\boldsymbol{B}$ matrices

generated from constrained Wishart distributions, $\boldsymbol{B} \sim CW(\nu, \boldsymbol{R}; \boldsymbol{I}_2)$, shows that the diagonal elements of $\boldsymbol{B}$ tend to increase with ν for a fixed precision matrix $\boldsymbol{R}$. The opposite happens to the transformed matrix $\boldsymbol{D} = \Sigma_0^{1/2}(\boldsymbol{B}_0^{-1} - \boldsymbol{I})\boldsymbol{\Sigma}_0^{1/2}$, $\boldsymbol{B}_0 \sim CW_2(\nu, \boldsymbol{I_2}; \boldsymbol{I_2})$: as ν increases, the diagonal elements of $\boldsymbol{D}$ tend to get smaller on average. Additionally, the distribution of the values of ρ_b becomes increasingly concentrated about 1 when the correlation value corresponding to $\boldsymbol{\Sigma}_0$ is large ($\rho_0 = 0.8$), but it does not change much in qualitative terms as ν increases. These observations help to explain the performance of the constrained Wishart distributions in our simulation: the $CW(\nu, \boldsymbol{R}; \boldsymbol{I}_2)$ prior distribution with small degrees of freedom ($\nu = 3$) leads to random $\boldsymbol{D}$ matrices that tend to have larger diagonal elements when compared with that of a constrained Wishart distribution with $\nu > 3$, thus being more consistent with how the data were actually generated. One useful take-away message here is that a constrained Wishart prior distribution with fewer degrees of freedom may be preferred for MBMA when greater levels of heterogeneity are anticipated.

The constrained Wishart prior does not exhibit strong dependence between variance and correlation, probably because of the constraints it implicitly places on the eigenvalues. However, the support of the distribution of the correlation parameter changes with the degrees of freedom ν. Samples of the 50% equiprobability ellipse (the contour plot in which 50% of the bivariate normal density lies) for $\boldsymbol{\mu}_i|\boldsymbol{\theta} = \boldsymbol{0}, \boldsymbol{D}$ where $\boldsymbol{D} \sim CW_2(\nu, \boldsymbol{0}; \boldsymbol{I}_2)$ look more like spheres for values of $\nu = 3, 6, 15$, which is expected because the squared relative lengths of the principal axes are given by the corresponding eigenvalues ($\boldsymbol{I}_2$ is the constraint matrix). Another comparatively strong performer for estimating the covariance matrix in the class of priors considered was the mixture of Wishart prior distributions. Like the constrained Wishart priors, these priors do not exhibit strong dependence between variance and correlation. Hierarchical normal models using a mixture of Wishart distributions have conditional conjugacy; compared with other priors requiring MCMC, this class of priors has the computational advantage of allowing for exact Gibbs sampling from the posterior distributions. Samples of the 50% equiprobability ellipses (the contour plot in which 50% of the bivariate normal density lies) for $\boldsymbol{\mu}_i|\boldsymbol{\theta} = \boldsymbol{0}, \boldsymbol{D}$ where $\boldsymbol{D} \sim MW_2(\nu, \boldsymbol{A}_k)$ show a set of ellipses with several orientations and axes lengths that do not strongly depend on the values of the hyperparameters.

As expected, independent prior distributions also do not exhibit much dependence between variance and correlation. However, samples of the 50% equiprobability ellipse (the contour plot in which 50% of the bivariate normal density lies) for $\boldsymbol{\mu}_i|\boldsymbol{\theta} = \boldsymbol{0}, \boldsymbol{D}$, where $\boldsymbol{D}$ is generated by one of the independent prior distributions, are significantly different from the ones obtained using any other of the prior distributions. The orientation of the points along both axes is quite different: horizontal or vertical ellipses only. The directions of the principal axes of the ellipsoids are given by the eigenvectors of the covariance matrix $\boldsymbol{D}$, which implies that the eigenvectors of $\boldsymbol{D}$ are constrained when using independent priors. These restrictive orientation patterns in the equiprobability ellipses suggest that sampling takes place in a limited sample space, potentially impacting performance.

The reference prior distribution $\left(\pi_{HRPb}(\boldsymbol{D}) = \left[|\boldsymbol{\Sigma}_0 + \boldsymbol{D}| \prod_{i<j}(\lambda_i - \lambda_j)\right]^{-1}\right)$ is skewed to the left for ρ_b when $\boldsymbol{\Sigma}_0$ has a high positive correlation ($\rho_0 > 0.70$) and skewed to the right for ρ_b when $\boldsymbol{\Sigma}_0$ has a high negative correlation ($\rho_0 < -0.70$). There is no strong dependence between variance and correlation when the correlation values are sampled in the neighborhood of ρ_0; however, variances are highly correlated (concentrated around the same value) when correlation values are sampled far away from ρ_0. This in turn suggests that MCMC may converge fairly quickly in some cases or need extra runs in others. Samples of the 50% equiprobability ellipses (the contour plot in which 50% of the bivariate normal density lies) for $\boldsymbol{\mu}_i|\boldsymbol{\theta} = \boldsymbol{0}, \boldsymbol{D}$, where $\boldsymbol{D}$ is generated by one of the reference prior distributions, show a set of ellipses with several orientations and axes lengths.

The usual conjugate choice, the inverse Wishart prior distribution, has problems that appear to stem from the strong dependence between variance and correlation: high variance implies high correlation (in absolute value), and low variance implies low-to-moderate correlation (in absolute value). This is a problem for inference, because it means that correlation will tend to be exaggerated with higher variances. This may help to explain the subpar performance of the inverse Wishart priors for estimating variance in the case of unequal heterogeneity levels.

The simulation results have a number of implications on the use of MBMA in medical clinical research. Even least informative prior distributions have an impact on the overall inferences for both the effect sizes and the covariance parameters. Therefore, sensitivity analysis to prior family, as well as to choice of prior within family, is important; not surprisingly, this is especially true when the sample size is small. Of the priors considered, the constrained Wishart and mixture of Wishart priors have clear computational advantages, and these families were typically observed to behave (comparatively) well across the

frequentist performance measures used here. An advantage of the mixture of Wishart distributions is that its performance was only observed to depend weakly on the specification of its hyperparameters. We could not identify a single best choice for the optimal degrees of freedom for $CW_2(\nu, \mathbf{0}; \boldsymbol{I}_2)$; however, as demonstrated earlier and discussed previously, improved performance results when the degrees of freedom vary inversely with the heterogeneity index. The availability of historical information on heterogeneity indices could be used to help inform the selection of an appropriate choice of ν; alternatively, an empirical Bayes approach to prior specification can be taken, where the heterogeneity indices are computed from the observed data (e.g., [6]; see also the Supporting Information for code). Because the relationship between the degrees of freedom and the heterogeneity index cannot be determined mathematically, sensitivity analysis should still be used.

All models fit in this paper utilized R or a combination of R and WinBUGS; the code is available in the Supporting Information, and these platforms were selected because of the relatively high penetration of these softwares in applied research. Software implementation of MCMC algorithms, including model and prior parameterizations, may have an impact on the results. The available code fits the same model using the five families of priors for $\boldsymbol{D}$ considered in this paper, and the results are displayed in a similar format to those in Table I, allowing an immediate comparison of results among families. Intra-prior family sensitivity analysis can also be performed using the code available and varying the input parametric values that characterize the distribution. As noted earlier, the code also computes the heterogeneity index for each effect size, as proposed in [6], and additionally produces convergence diagnostic plots and tables for each model.

Although the simulation presented here can be extended in many ways, the implementation of a simulation study such as this one requires large computational and time resources. The normality assumption in (2) is considered reasonably precise for large studies and is an attractive choice more generally because of its relative computational simplicity and the ease of interpretation of its model parameters. Extensions to non-normal models, such as the Dirichlet process model [55] and the Polya tree mixture model [56], may also be worthwhile. For future research, the impact of different fixed effect sizes (specification 1), additional heterogeneity indices (specification 3), and different within-correlation and between-correlation coefficients should be considered, as should missing data and robustness to model assumptions. The assumption of independent diagonal elements for each within-study variance matrix (i.e., for each i, $\sigma^2_{ij}, j = 1, 2$ are independently distributed as exponential with a rate of 0.50, with all within-study variances truncated to the range of [0.50, 10]) is designed to create some heterogeneity in the within-study-level variances (i.e., heterogeneity in the diagonal elements of the Σ_is). We do not expect that the manner in which we have created this heterogeneity to have a material impact on our results. To provide some rationale for this argument, note that we may write $\sigma^2_{ij} = v_{ij}/n_j$ for each i, j. Because the matrices Σ_i are fixed throughout the analysis, we are in effect conditioning on the set of observed variances, hence study sample sizes $n_1 \ldots n_n$. As a result, it should not matter much whether unequal diagonals in the Σ_is arise from a setting in which we have equal variances and unequal study sample sizes; unequal variances and equal study sample sizes; or unequal variances and unequal study sample sizes. To test this intuition, and specifically whether variation in the study sample sizes might influence our results, we ran an additional mini-simulation with 1000 replicates using the constrained Wishart priors with 10 studies under the specification of independent error variances (*i.e.*, (v_{1j}, v_{2j}) are independent exponentials with rate equal to 0.05) and random study sample sizes generated from a Poisson distribution with rate 10. Here, the study sample sizes vary, with σ^2_{i1} and σ^2_{i2} being dependent for each i but independent across i. We did not find substantial differences between the results for the dependent and independent within-study-level variances (results available upon request).

Acknowledgements

Madhu Mazumdar and Sandra Hurtado Rúa were professor and post-doctoral associate at the Division of Biostatistics and Epidemiology, Department of Public Health, Weill Cornell Medical College, at the time this manuscript was initially drafted. Their current appointments are at Icahn School of Medicine at Mount Sinai and Cleveland State University, respectively. They were partially funded by NIH CTSA funding to Weill Medical College of Cornell University (2UL1 TR000457-06). The authors thank Dr. Xia Wang and Dr. Yan Ma, the editor and two anonymous referees for their constructive critiques that helped to improve this manuscript. Ms. Lyn Liu and Mr. Jonathan Eskreis-Winkler assisted with editing many of the supporting information tables and figures.

References

1. Peto R. Why do we need systematic overviews of randomized trials? *Statistics in Medicine* 1987; **6**:233–240.
2. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
3. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. *Methods for Meta-analysis in Medical Research*. Wiley: Chichester, UK, 2000.
4. Demidenko E. *Mixed Models: Theory and Applications*. Wiley: New York, 2004.
5. Riley RD, Abrams KR, Lambert PC, Sutton AJ, Thompson JR. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Statistics in Medicine* 2007; **26**:78–97.
6. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**:1539–1558.
7. White IR. Multivariate random effects meta-analysis. *The Stata Journal* 2008; **9**:40–56.
8. Riley RD. Multivariate meta-analysis: the effect of ignoring within-study correlation. *Journal of the Royal Statistical Society: Series A* 2009; **172**:789–811.
9. Riley RD, Abrams KR, Sutton AJ, Lambert PC, Thompson JR. Bivariate random effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology* 2007; **7**:3.
10. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. *Statistics in Medicine* 2011; **30**(20): 2481–2498.
11. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**: 619–629.
12. van Houwelingen H C, Arends LR, Stijnen T. Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine* 2002; **21**:589–624.
13. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analysis. *Statistics in Medicine* 2010; **29**:1282–1297.
14. Berkey CS, Hoaglin DC, Antczak-Bouckoms A, Mosteller F, Colditz GA. Meta-analysis of multiple outcomes by regression with random effects. *Statistics in Medicine* 1998; **17**:2537–2550.
15. Ma Y, Mazumdar M. Multivariate meta-analysis: a robust approach based on the theory of U-statistics. *Statistics in Medicine* 2011; **30**(24):2911–2929.
16. Mavridis D, Salanti G. A practical introduction to multivariate meta-analysis. *Statistical Methods Medical Research* 2013; **22**(2):133–158.
17. Mehrotra DV. Non-iterative robust estimators of variance components in within-subject designs. *Statistics in Medicine* 1997; **16**:1465–1479.
18. Berger JO. *Statistical Decision Theory and Bayesian Analysis*. Springer: New York, 1980.
19. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A* 2009; **172**(1):137–159.
20. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*, 2nd ed. CRC Press: Boca Raton, 2003.
21. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBugs: a Bayesian modeling framework: concepts, structure and extensibility. *Statics and Computing* 2011; **10**:325–337.
22. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2011; Available from: http://www.R-project.org.
23. StataCorp. *Stata Statistical Software: Release 12: College Station,* TX, 2011. StataCorp LP.
24. SAS Institute Inc. *SAS 9.1.3. Help and Documentation*. SAS Institute Inc.: Cary, NC, 2000–2004.
25. Muthen BO, Muthen L. *Mplus: The Comprehensive Modeling Program to Applied Researchers*. M-plus: Los Angeles, CA, 2000.
26. Raudenbush SW, Bryk AS. Empirical Bayes meta-analysis. *Journal of Educational Statistics* 1985; **10**(2):75–98.
27. Turner R, Davey J, Clarke M, Thompson S, Higgins J. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database on Systematic Reviews. *International Journal of Epideiology* 2012; **41**(3): 818–827.
28. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 1996; **91**(435):1343–1370.
29. Yang R, Berger JO. Estimation of a covariance matrix using the reference prior. *Annals of Statistics* 1994; **22**:1195–1211.
30. Berger JO, Strawderman W, Tang D. Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *The Annals of Statistics* 2005; **33**(2):606–646.
31. Berger JO, Sun D. Objective priors for the bivariate normal model. *Annals of Statistics* 2008; **36**:963–982.
32. Morris CN, Lysy M. Shrinkage estimation in multilevel normal models. *Statistical Science* 2012; **27**(1):115–134.
33. Daniels MJ, Kass RE. Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *Journal of the American Statistical Association* 1999; **94**(448):1254–1263.
34. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **3**:515–533.
35. Tweedie RL, Scott DJ, Biggerstaff BJ, Mengersen KL. Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer* 1996; **14**(Suppl 1):S171–94.
36. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* 2005; **24**:2401–2428.
37. Wei Y, Higgins JP. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. *Statistics in Medicine* 2013; **32**(7):1191–205.
38. Daniels M, Pourahmadi M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika* 2002; **89**(3):553–566.
39. Gaskins J, Daniels M. A nonparametric prior for simultaneous covariance estimation. *Biometrika* 2013; **100**(1):125–138.
40. Barnard J, McCulloch R, Meng X. Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica* 2000; **10**:1281–1311.

41. Garthwaite P, Kadane JB, O'Hagan A. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association* 2005; **100**(470):680–701.
42. Chen CF. Bayesian inference for a normal dispersion matrix and its application to stochastic multiple regression analysis. *Journal of the Royal Statistical Society: Series B* 1979; **41**:235–248.
43. Berger JO, Bernardo JM. On the development of reference priors (with discussion). *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. Clarendon Press: Oxford, 1992, 35–49.
44. Christiansen CL, Morris CN. Hierarchical Poisson regression modeling. *Journal of the American Statistical Association* 1997; **92**:618–632.
45. Everson PJ, Morris CN. Inference for multivariate normal hierarchical models. *Journal of the Royal Statistical Society: Series B* 2000; **62**:399–412.
46. Everson P, Morris C. Simulation from Wishart distributions with eigenvalue constraints. *Journal of Computational and Graphical Statistics* 2000; **9**(2):380–389.
47. Huang A, Wand MP. Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis* 2013; **8**(1):1–14.
48. Rannala B. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Systematic Biology* 2002; **51**: 754–760.
49. Gelfand AE, Carlin BP, Trevisani M. On computation using Gibbs sampling for multilevel models. *Statistica Sinica* 2001; **11**:981–1003.
50. Hobert JP, Casella G. The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 1996; **91**(436):1461–1473.
51. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–472.
52. Geweke J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics*, Bernardo JM, Berger JO, Dawiv AP, Smith AFM (eds), Vol. 4. Clarendon Press: Oxford, UK, 1992.
53. Geeganage C, Bath PM. Vasoactive drugs for acute stroke. *Cochrane Database of Systematic Reviews* 2010; **7:** CD002839.
54. Antczak-Bouckoms A, Joshipura K, Burdick E, Tulloch JFC. Meta-analysis of surgical versus non-surgical method of treatment for periodontal disease. *Journal of Clinical Periodontology* 1993; **20**:259–268.
55. Burr D, Ross H. A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association* 2005; **100**:242–251.
56. Branscum A, Hanson T. Bayesian nonparametric meta-analysis using Polya tree mixture models. *Biometrics* 2008; **64**: 825–833.

Supporting information

Additional supporting information may be found in the online version of this article at the publishers web site.