Published in final edited form as:

Stat Biopharm Res. 2016 ; 8(1): 49–59. doi:10.1080/19466315.2015.1096823.

Bayesian Estimation and Testing in Random Effects Meta-analysis of Rare Binary Adverse Events

Ou Bai

Department of Statistical Science, Southern Methodist University

Min Chen

Department of Mathematical Sciences, University of Texas at Dallas

Dept. of Clinical Sciences, Univ. of Texas Southwestern Medical Center

Xinlei Wang*

Department of Statistical Science, Southern Methodist University

Abstract

Meta-analysis has been widely applied to rare adverse event data because it is very difficult to reliably detect the effect of a treatment on such events in an individual clinical study. However, it is known that standard meta-analysis methods are often biased, especially when the background incidence rate is very low. A recent work by Bhaumik et al. (2012) proposed new moment-based approaches under a natural random effects model, to improve estimation and testing of the treatment effect and the between-study heterogeneity parameter. It has been demonstrated that for rare binary events, their methods have superior performance to commonly-used meta-analysis methods. However, their comparison does not include any Bayesian methods, although Bayesian approaches are a natural and attractive choice under the random-effects model. In this paper, we study a Bayesian hierarchical approach to estimation and testing in meta-analysis of rare binary events using the random effects model in Bhaumik et al. (2012). We develop Bayesian estimators of the treatment effect and the heterogeneity parameter, as well as hypothesis testing methods based on Bayesian model selection procedures. We compare them with the existing methods through simulation. A data example is provided to illustrate the Bayesian approach as well.

Keywords

Bayesian hierarchical modeling; Bayesian model selection; deviance information criterion; fixed effect; heterogeneity; generalized linear mixed model; sparse data

1 Introduction

In clinical studies, it is important to assess adverse drug reactions such as drug toxicity because their negative impacts can be substantial. Many dangerous events associated with drug administration, such as cardiovascular diseases, cancers and chronic intoxication, have binary outcomes with very low background incidence rates. As a result, it is very difficult to

*All correspondence should be addressed to Xinlei Wang (swang@smu.edu)..

reliably detect adverse effects of a drug (or treatment) in an individual clinical study. Meta-analysis, which synthesizes data from multiple relevant clinical studies using formal statistical procedures, has played critical roles in the inference of such adverse effects to make informed decisions. In this research, we consider the problem of meta-analysis of rare binary adverse events.

Fix-effect meta-analysis (FEM) is widely used in practice and it assumes a homogenous treatment effect in all studies. Popular FEM methods include the Mantel-Haenszel method (MH, Mantel and Haenszel, 1959), the empirical logit method based on an inverse-variance weighting approach (EL, DerSimonian and Laird 1986) and Peto's method (Yusuf et al. 1985). However, the homogeneity assumption of FEM is sometimes restrictive due to the diversity in the studies included in the meta-analysis. Heterogeneity arises naturally when clinical studies have different protocols (e.g., doses, control medications, follow-up times) or contain patients of diverse characteristics. To account for the inter-study variability, random effects meta-analysis (REM) has been proposed to allow the treatment effects to vary from study to study (e.g., DerSimonian and Laird 1986; Bhaumik et al. 2012).

A special concern often arises when method-of-moments estimators, based on either FEM or REM, are used to estimate the treatment effect for rare event data (Sweeting et al. 2004, Bradburn et al. 2007, Bhaumik et al. 2012). That is, original moment-based estimators for the (log) odds ratio or (log) relative risk are often undefined in studies with zero adverse events, which occur frequently when the background incidence rate is too low and/or the sample sizes are not sufficiently large. A common way to handle this problem is adding a continuity correction factor. However, bias may be introduced by using the correction factor with the commonly used FEM methods, as pointed out by Bradburn et al. (2007). Bhaumik et al. (2012) formally extended the use of the correction factor to REM of sparse data and proposed a simple average (SA) estimator. This new SA estimator seeks a continuity correction to make the moment-based estimate of the treatment effect asymptotically unbiased for a single study, and then take the simple (unweighted) average of treatment effect estimates over multiple studies (see more detail in Section 3.1). It has been shown via simulation that the SA estimator is the least biased when compared with popular existing estimators. However, as the event rate gets lower and lower so that the number of zero-event studies gets large, bias still tends to increase for all the moment-based methods above. Besides, in the estimation of the heterogeneity parameter that is used to model the between-study variability of the treatment effect in REM models, standard approaches (e.g., Paule and Mandel 1982, DerSimonian and Laird 1986, DerSimonian and Kacker 2007) have been reported to consistently underestimate the heterogeneity (Bhaumik et al. 2012).

Bayesian methods may provide a competitive alternative to the frequentist methods mentioned above (Warren et al. 2012). However, compared to the abundant frequentist work, much less Bayesian work can be found in the literature of meta analysis of binary data (e.g., Carlin 1992, Smith et al. 1995, Warn et al. 2002, Higgins et al. 2009). When narrowed to *rare* binary data, existing Bayesian work is rather scant; and the only work that we are aware is Cai et al. (2010), in which Poisson random effects models have been proposed to combine multiple 2×2 tables for inference about the relative risk. Bhaumik et al. (2012) conducted a comprehensive simulation study, with focus on rare binary data, to compare their proposed

methods with other existing (frequentist) methods; and they have showed that their methods are the least biased. However, their study excluded comparison with Bayesian methods, even though it would be very appealing to consider a Bayesian approach under the REM. We attempt to address the question that is left unanswered in Bhaumik et al. (2012); that is, for rare binary data, can Bayesian methods offer some advantages over the moment-based methods in meta-analysis, and if yes, when? In Section 2, we will implement a Bayesian framework based on the random effects model in Bhaumik et al. (2012); and in Section 3, we will provide simulation-based comparisons between Bayesian estimators and their strong frequentist competitors in estimating the treatment effect and heterogeneity parameter.

Bhaumik et al. (2012) also considered the problem of hypothesis testing involving the parameters of the treatment effect and the heterogeneity under the REM. To test the existence of the treatment effect, they constructed a large-sample test based on their newly developed SA estimator. For testing the existence of inter-study heterogeneity, they proposed two asymptotic tests based on the logarithm of Cochran's Q statistic (Cochran 1950) and the SA estimator, respectively. By contrast, we are not aware of any existing work that has applied Bayesian testing procedures in meta-analysis of rare binary events, even though they are applicable in general. Thus, the implementation detail and how well they perform under this specific context remain unclear. One advantage of Bayesian hypothesis testing is, unlike most frequentist approaches, it does not depend on asymptotic theories and thus it may be well suited for applications of small sample sizes. In Section 4 we will adopt a Bayesian model selection framework to address the hypothesis testing problem, where deviance information criteria based on different likelihood functions are developed to select models. We further compare the Bayesian procedures to competitive frequentist testing procedures through simulation.

A data example is presented to illustrate the Bayesian approach in Section 5. In Section 6, a brief discussion concludes the paper.

2 A Bayesian hierarchical approach

Below we present a Bayesian hierarchical method for meta-analysis of multiple studies of rare binary events. Let I be the number of studies and n_{it} (n_{ic}) be the number of subjects assigned to the treatment (control) group in the i th study. Let x_{it} (x_{ic}) be the number of subjects that experienced the adverse event in the treatment (control) group. Let p_{it} (p_{ic}) denote the probability of the event in the treatment (control) group. Let μ_i be the log-odds of the event in the control group and θ_i be the treatment effect in the logit scale. Following Bhaumik et al. (2012), we express the binomial-normal hierarchical model by

$$\begin{aligned} x_{ic} &\sim \text{Binomial}(n_{ic}, p_{ic}), & x_{it} &\sim \text{Binomial}(n_{it}, p_{it}), \\ \text{logit}(p_{ic}) &= \mu_i, & \text{logit}(p_{it}) &= \mu_i + \theta_i, \\ \mu_i &\sim \mathcal{N}(\mu_0, \sigma^2), & \theta_i &\sim \mathcal{N}(\theta_0, \tau^2), \end{aligned} \quad (1)$$

where the hyper-parameters μ_0 and σ^2 (or, θ_0 and τ^2) represent the mean and variance of μ_i 's (or θ_i 's) across all studies. Obviously, τ^2 measures the inter-study heterogeneity of the treatment effects.

To further develop our Bayesian model, priors for the hyper-parameters need to be specified. We consider non-informative prior distributions for both μ_0 and θ_0 ,

$$\mu_0 \text{Uniform}(L_\mu, U_\mu), \theta_0 \text{Uniform}(L_\theta, U_\theta).$$

The lower and upper bounds for each uniform distribution must be assigned to allow for all realistic values of μ_0 and θ_0 in a specific meta-analysis. A conservative choice for these bounds, for example, for θ_0 can be obtained in the following way,

$$\begin{aligned} L_\theta &= \min \{ \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_I \} - \alpha \\ U_\theta &= \max \{ \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_I \} + \beta, \end{aligned}$$

where $\hat{\theta}_i$'s are the estimates for the treatment effects in individual studies that can be easily obtained using any moment-based method with capability to handle the case when the odds ratio is undefined (e.g., DSL, MH, EL and SA with continuity correction). Based on preliminary simulation results, α and β can be conservatively chosen as 5. Similarly, we can specify L_μ and U_μ for the prior distribution of μ_0 . Note that choices of the lower and upper bounds are not critical because μ_i and θ_i will still cover all real numbers, even when the intervals of the uniform distributions are not wide. We choose conditional conjugate priors for the variance components σ^2 and τ^2 , that is

$$\sigma^2, \tau^2 \text{InverseGamma}(a, b).$$

Here, a and b are typically chosen to be very close to zero (e.g., $a = b = 0.01$), to make the priors very vague, reflecting the common situation of no meaningful knowledge available about σ^2 and τ^2 .

Next, we derive the joint posterior probability density and full conditionals of all the parameters based on the above specification. Let Θ denote the set of the (hyper)parameters involved in our Bayesian model:

$$\Theta = (\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\psi}), \quad (2)$$

where $\boldsymbol{\mu} = (\mu_i)_{i=1}^I$, $\boldsymbol{\theta} = (\theta_i)_{i=1}^I$, $\boldsymbol{\psi} = (\mu_0, \theta_0, \sigma^2, \tau^2)$. Under the independence assumption of all prior or hyperprior distributions, the full probability model is given by

$$P(X, \Theta) = \prod_{i=1}^I P(x_{ic}, x_{it} | \mu_i, \theta_i) \cdot P(\mu_i, \theta_i | \mu_0, \theta_0, \sigma^2, \tau^2) \cdot P(\mu_0, \theta_0, \sigma^2, \tau^2),$$

where

$$\begin{aligned}
P(x_{ic}, x_{it} | \mu_i, \theta_i) &= \binom{n_{ic}}{x_{ic}} \binom{n_{it}}{x_{it}} e^{(x_{ic}+x_{it})\mu_i + x_{it}\theta_i} \left(\frac{1}{1+e^{\mu_i}}\right)^{n_{ic}} \left(\frac{1}{1+e^{\mu_i+\theta_i}}\right)^{n_{it}} \\
P(\mu_i, \theta_i | \mu_0, \theta_0, \sigma^2, \tau^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\mu_i-\mu_0)^2}{2\sigma^2}\right] \frac{1}{\sqrt{2\pi}\tau} \exp\left[-\frac{(\theta_i-\theta_0)^2}{2\tau^2}\right], \\
P(\mu_0, \theta_0, \sigma^2, \tau^2) &= \frac{I_{(L_\theta, U_\theta)}(\theta_0)}{U_\theta - L_\theta} \frac{I_{(L_\mu, U_\mu)}(\mu_0)}{U_\mu - L_\mu} \left(\frac{b^a}{\Gamma(a)}\right)^2 (\sigma^2)^{-a-1} e^{\left(\frac{b}{-\sigma^2}\right)} (\tau^2)^{-a-1} e^{\left(\frac{b}{-\tau^2}\right)}.
\end{aligned}$$

Thus, the posterior distribution can be specified as

$$\begin{aligned}
P(\Theta | X) &\propto \prod_{i=1}^I \left\{ e^{(x_{ic}+x_{it})\mu_i + x_{it}\theta_i} \left(\frac{1}{1+e^{\mu_i}}\right)^{n_{ic}} \left(\frac{1}{1+e^{\mu_i+\theta_i}}\right)^{n_{it}} \right\} \\
&\times \prod_{i=1}^I \left\{ \frac{1}{\sigma} \exp\left[-\frac{(\mu_i-\mu_0)^2}{2\sigma^2}\right] \frac{1}{\tau} \exp\left[-\frac{(\theta_i-\theta_0)^2}{2\tau^2}\right] \right\} \\
&\times I_{(L_\theta, U_\theta)}(\theta_0) I_{(L_\mu, U_\mu)}(\mu_0) (\sigma^2)^{-a-1} e^{\left(\frac{b}{-\sigma^2}\right)} (\tau^2)^{-a-1} e^{\left(\frac{b}{-\tau^2}\right)}.
\end{aligned}$$

To obtain posterior samples, we use a Gibbs sampler, one of the most popular Markov Chain Monte Carlo (MCMC) methods. All the full conditionals can be derived from the joint posterior distribution as follows:

$$\begin{aligned}
P(\mu_1 | X, \Theta / \mu_i) &\propto e^{(x_{ic}+x_{it})\mu_i} (1+e^{\mu_i})^{-n_{ic}} (1+e^{\mu_i+\theta_i})^{-n_{it}} \exp\left[-\frac{(\mu_i-\mu_0)^2}{2\sigma^2}\right], \quad i=1, 2, \dots, I; \\
P(\theta_i | X, \Theta / \theta_i) &\propto e^{x_{it}\theta_i} (1+e^{\mu_i+\theta_i})^{-n_{it}} \exp\left[-\frac{(\theta_i-\theta_0)^2}{2\tau^2}\right], \quad i=1, 2, \dots, I; \\
P(\mu_0 | X, \Theta / \mu_0) &\propto N\left(\bar{\mu}, \frac{\sigma^2}{I}\right) I_{(L_\mu, U_\mu)}(\mu_0), \quad \text{where } \bar{\mu} = \frac{\sum_{i=1}^I \mu_i}{I}, \\
P(\theta_0 | X, \Theta / \theta_0) &\propto N\left(\bar{\theta}, \frac{\tau^2}{I}\right) I_{(L_\theta, U_\theta)}(\theta_0), \quad \text{where } \bar{\theta} = \frac{\sum_{i=1}^I \theta_i}{I}, \\
P(\sigma^2 | X, \Theta / \sigma^2) &\sim \text{InverseGamma}\left(a + \frac{I}{2}, b + \frac{\sum_{i=1}^I (\mu_i - \mu_0)^2}{2}\right), \\
P(\tau^2 | X, \Theta / \tau^2) &\sim \text{InverseGamma}\left(a + \frac{I}{2}, b + \frac{\sum_{i=1}^I (\theta_i - \theta_0)^2}{2}\right),
\end{aligned}$$

It can be seen that except for μ_i 's and θ_i 's, all the full conditionals are known distributions, from which samples can be directly drawn. For example, $P(\mu_0 | X, \Theta / \mu_0)$ and $P(\theta_0 | X, \Theta / \theta_0)$ are truncated normal distributions; and $P(\sigma^2 | X, \Theta / \sigma^2)$ and $P(\tau^2 | X, \Theta / \tau^2)$ are inverse Gamma distributions.

We consider rejection sampling methods to draw posterior samples of μ_i 's and θ_i 's. To do so, we specify $N(\mu_0, \sigma^2)$ as the proposal distribution of μ_i with pdf $\phi(\mu_i; \mu_0, \sigma^2)$, and define

$$\begin{aligned}
g(\mu_i | \theta_i, \mu_0, \sigma^2) &= \frac{P(\mu_i | X, \Theta / \mu_i)}{\phi(\mu_i; \mu_0, \sigma^2)} = \sqrt{2\pi}\sigma e^{(x_{ic}+x_{it})\mu_i} (1+e^{\mu_i})^{-n_{ic}} (1+e^{\mu_i+\theta_i})^{-n_{it}}, \\
M &= \sup_{\mu_i \in (-\infty, \infty)} \{g(\mu_i | \theta_i, \mu_0, \sigma^2)\}.
\end{aligned}$$

Then the rejection sampling algorithm for drawing posterior samples of each μ_i is outlined below.

1. Sample μ_i from $N(\mu_0, \sigma^2)$ and u from $\text{Uniform}(0, 1)$.

2. If $u < g(\mu_i | \theta_i, \mu_0, \sigma^2) / M$, then accept μ_i as a realization of $P(\mu_i | X, \Theta / \mu_i)$ and stop.
If $u \geq g(\mu_i | \theta_i, \mu_0, \sigma^2) / M$, go to step 1.

Note that μ_0 and σ^2 are specified as current values in the iterations of the Gibbs sampler; and M can be evaluated numerically in R. The steps to draw θ_i 's can be outlined similarly with the proposal distribution $N(\theta_0, \tau^2)$.

3 Parameter estimation

A natural Bayesian estimator for the treatment effect θ_0 is the average of θ_0 's sampled using MCMC after a burn-in period. For the heterogeneity parameter τ^2 , the posterior mean is not a good estimator because the density plot shows that the posterior distribution of τ^2 is generally skewed. In fact, results of our simulation studies indicate that the posterior mean estimator of τ^2 consistently overestimated τ^2 . Thus, we use the posterior median as an estimator of τ^2 instead.

3.1 Performance in estimating the treatment effect θ_0

Here, we examine the performance of the Bayesian estimator of θ_0 (BAYES) and compare it with the commonly-used existing estimators, including SA by Bhaumik et al. (2012), DSL, EL, MH, and the estimator based on generalized linear mixed models (GLMM; Agresti 2013, Breslow and Clayton 1993). As in Bhaumik et al. (2012), whether zero events were observed or not, a 0.5 continuity correction was added for SA, DSL, EL and MH in all studies while for BAYES and GLMM, there is no need to do so. As mentioned in the introduction, Bhaumik et al. (2012) showed that SA is the least biased compared with the other moment-based estimators DSL, MH, and EL. Since SA is relatively new, it would be beneficial to describe this method below. Let $\hat{\theta}_{ia}$ be an estimate of the random treatment effect in the i th study:

$$\hat{\theta}_{ia} = \ln \left[\frac{(x_{it} + a) / (n_{it} - x_{it} + a)}{(x_{ic} + a) / (n_{ic} - x_{ic} + a)} \right],$$

where a is a positive number added for continuity correction. The SA estimator of θ is then given by taking the average of $\hat{\theta}_{i\frac{1}{2}}$'s over all the I studies, where a is fixed at 1/2 so that each $\hat{\theta}_{i\frac{1}{2}}$'s is unbiased up to the order of n_i^{-1} where $n_i = n_{ic} + n_{it}$.

Typically, REM requires the number of studies I not to be small for the purpose of estimating the heterogeneity parameter τ^2 . We set I to be 10 and 20, representing REM with a relatively small and large number of studies in practice, respectively. For each I value, we simulate rare binary event data in the same way as Bhaumik et al. (2012). We set the values of μ_0 to be -2.5 and -5 , corresponding to a moderate background incidence rate of 8% and a very low rate of 0.7%, respectively. The settings of the other parameters are as follows: $\theta_0 = \{-1, -0.9, \dots, 1\}$; $\sigma^2 = 0.5$; and $\tau^2 = 0.8$. With all the above parameters, p_{ic} 's and p_{it} 's can be simulated as $p_{ic} = e^{\mu_i} / (1 + e^{\mu_i})$ and $p_{it} = e^{\mu_i + \theta_i} / (1 + e^{\mu_i + \theta_i})$, respectively, for $i = 1, 2, \dots, I$. For each study, we draw a random number from Uniform(50, 1000), round it to the nearest

integer, and then set it as the number of subjects in the control group (n_{ic}). Next, a random number R is generated from the set $\{1, 1.5, 2\}$ with equal probability; and the number of subjects in the treatment group (n_{it}) is set to $R \cdot n_{ic}$ and rounded to the nearest integer. Last, x_{ic} 's and x_{it} 's are simulated from $\text{Binomial}(n_{ic}, p_{ic})$ and $\text{Binomial}(n_{it}, p_{it})$, respectively. Figure 1 reports how the different estimators perform when θ_0 changes from -1 to 1 for rare events (i.e., $\mu_0 = -5$) using 1000 simulated data sets for each setting.

It can be seen in Figure 1 that, the Bayesian and GLMM estimators are essentially unbiased regardless of the value of θ_0 since they almost overlap with the straight line $y = x$. All the other methods seem to overestimate the overall treatment effect θ_0 , especially when θ_0 is negative, meaning that the mean incidence rate in the treatment arm is lower than that in the control arm. Among them, SA outperforms the three moment-based methods MH, DSL, and EL as expected; and it is nearly unbiased when θ_0 is positive, but overestimates the treatment effect when θ_0 is negative. When $\mu_0 = -2.5$ (the simulation results are not presented), there is almost no difference among SA, GLMM and BAYES, which are nearly unbiased while the other methods are not.

Next, we fix θ_0 at zero, and let μ_0 vary from -5 to 0 with step size 0.5 ; and we keep the other parameters the same as before. Figure 2 shows that the standard estimators DSL, EL and MH are biased for all μ_0 s except $\mu_0 = 0$. For $\mu_0 < 0$, they consistently overestimate the treatment effect. Their biases follow the order $\text{MH} > \text{EL} > \text{DSL}$; and they are getting greater as the mean background incidence rate μ_0 is getting lower. GLMM, BAYES and SA are nearly unbiased when $\mu_0 > -3$. When $\mu_0 < -3$, SA, GLMM and BAYES maintain their performance quite well. The biases of GLMM, BAYES and SA are much smaller, compared to those of the other estimators; and they follow the order $\text{SA} > \text{GLMM} \approx \text{Bayes}$.

By comparing Panel (a) with Panel (b) in both Figures 1 and 2, we find that all the results are not sensitive to the choice of I and all our findings remain virtually the same for both large and small I .

3.2 Performance in estimating the heterogeneity parameter τ^2

We proceed to examine the performance in estimating τ^2 for rare binary event data. We compare the Bayesian method (BAYES) with three popular moment-based methods including DerSimonian and Laird (DSL), DerSimonian and Kacher (DSK), Paule and Mandel (PM), plus the improved Paule and Mandel (IPM) proposed by Bhaumik et al. (2012). We include the GLMM estimator of τ^2 for comparison as well. We set $\theta_0 = 0$, $\mu_0 = \{-2.5, -5\}$, and $\tau^2 = \{0, 0.1, \dots, 1\}$. The settings of the other parameters are the same as before. Figure 3 shows that when there exists no heterogeneity (i.e., $\tau^2 = 0$), all the methods appear to overestimate τ^2 in some degree, especially for the small I . When τ^2 moves away from zero, all the frequentist methods become to underestimate τ^2 , and their biases appear to increase when the heterogeneity gets larger; on the other hand, Bayes keeps overestimating τ^2 (slightly). Overall, the bias follows the following order approximately: $\text{BAYES} < \text{GLMM} < \text{IPM} < \text{DSL} < \text{PM} < \text{DSK}$ for $\tau^2 > 0$. For $\mu_0 = -2.5$ (results not reported), observations are almost the same as above.

In Figure 4, we fix τ^2 at 0.8 and vary μ_0 from -5 to 0 . It can be seen that when μ_0 is close to zero, DSK, PM and IPM seem to be a bit better than the other three methods. As μ_0 moves away from 0 , the performance for DSL, DSK, PM and IPM deteriorates. Bayes and GLMM is much less sensitive to the change of μ_0 . When μ_0 becomes close to -5 , Bayes and GLMM outperform the other methods that all drastically underestimate τ^2 ; and overall, BAYES is less biased than GLMM, especially for small I .

4 Bayesian hypothesis testing and model selection

4.1 Deviance information criteria

In Bayesian meta-analysis, the hypothesis testing problem can be viewed as a Bayesian model selection problem. When both the overall treatment effect θ_0 and the heterogeneity parameter τ^2 are considered, there are four possible models denoted by $M_1 - M_4$ that correspond to the following hypotheses:

$$\begin{aligned} H_1(M_1) &: \theta_0=0, \quad \tau^2=0; \\ H_2(M_2) &: \theta_0 \neq 0, \quad \tau^2=0; \\ H_3(M_3) &: \theta_0=0, \quad \tau^2 \neq 0; \\ H_4(M_4) &: \theta_0 \neq 0, \quad \tau^2 \neq 0. \end{aligned}$$

Here, M_4 is referred to as the full model while the other three are called reduced models (say $M_1 - M_3$). For instance, under H_1 , the reduced Bayesian model M_1 can be specified as follows:

$$\begin{aligned} x_{ic} &\text{Binomial}(n_{ic}, p_{ic}), \quad x_{it} \text{Binomial}(n_{it}, P_{it}); \\ \text{logit}(p_{ic}) &= \mu_i, \quad \text{logit}(P_{it}) = \mu_i; \\ \mu_i &\sim N(\mu_0, \sigma^2), \\ \mu_0 &\sim \text{Uniform}(L_\mu, U_\mu), \\ \sigma^2 &\sim \text{InverseGamma}(a, b). \end{aligned}$$

From a Bayesian perspective, the problem of hypothesis testing can be solved via model selection among the above four models. One commonly accepted criterion for model selection is to compare the Bayes factors (Kass and Raftery 1995). However, in the meta-analysis, evaluating Bayes factors requires high-dimensional integrations, which is hard to evaluate reliably, especially for rare event data. We instead consider the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002) that selects the best model from a point view of prediction given the data currently observed. Under a specific model M , let the Bayesian deviance be

$$D(\beta) = -2\log\{P(\text{data}|\beta, M)\} + 2\log\{f(\text{data})\},$$

where $P(\text{data}|\beta, M)$ is the likelihood function of β under model M , given the observed data; and $f(\text{data})$ is a fully standardizing term that only depends on the observed data. Define

$$\begin{aligned}\bar{D} &= E_{\beta}[D(\beta)] \\ p_D &= \bar{D} - D(\beta),\end{aligned}$$

where $\bar{\beta} = E(\beta|data)$. Then,

$$DIC = p_D + \bar{D}.$$

The model with the smallest DIC is deemed as the best model in terms of the ability of prediction. An advantage of using DIC is that it can be easily calculated from posterior samples. An estimate of $\bar{D}$ can be obtained by taking the sample average of the simulated values of $D(\beta)$; and $\bar{\beta}$ is estimated by the posterior average of β . To evaluate the DIC under each competitive model (hypothesis), the corresponding likelihood needs to be evaluated. Note that though we refer $P(data|\beta, M)$ as the likelihood function, for a model (e.g., $M_1 - M_4$) with hierarchical parameter structures, the likelihood function cannot be defined uniquely (Spiegelhalter et al. 2002). For example, with the notations in (2), we have the marginal distribution of the data X under the full model:

$$P(X|M_4) = \int P(X|\mu, \theta, M_4) \cdot P(\mu, \theta|M_4) d\mu d\theta = \int P(X|\psi, M_4) \cdot P(\psi|M_4) d\psi,$$

where

$$\begin{aligned}P(\mu, \theta|M_4) &= \int P(\mu, \theta|\psi, M_4) P(\psi|M_4) d\psi, \\ P(X|\psi, M_4) &= \int P(X|\mu, \theta, M_4) P(\mu, \theta|\psi, M_4) d\mu d\theta.\end{aligned}$$

Since DIC selects a model from a perspective of best prediction without assuming there is actually a true model, both $P(X|\mu, \theta, M_4)$ and $P(X|\psi, M_4)$ can be chosen as the likelihood function for DIC. That is, if we choose $\beta \equiv (\mu, \theta)$, then the likelihood function can be defined by

$$P(X|\mu, \theta, M_4) = \prod_{i=1}^I \binom{n_{ic}}{x_{ic}} \binom{n_{it}}{x_{it}} e^{(x_{ic}+x_{it})\mu_i + x_{it}\theta_i} \left(\frac{1}{1+e^{\mu_i}}\right)^{n_{ic}} \left(\frac{1}{1+e^{\mu_i+\theta_i}}\right)^{n_{it}}.$$

If we choose $\beta \equiv \psi$, then the likelihood function becomes

$$\begin{aligned}P(X|\psi, M_4) &= \prod_{i=1}^I \left\{ \int \int \binom{n_{ic}}{x_{ic}} \binom{n_{it}}{x_{it}} e^{(e_{ic}+x_{it})\mu_i + x_{it}\theta_i} \left(\frac{1}{1+e^{\mu_i}}\right)^{n_{ic}} \left(\frac{1}{1+e^{\mu_i+\theta_i}}\right)^{n_{it}} \right. \\ &\quad \times \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(\mu_i-\mu_0)^2}{2\sigma^2}\right] \frac{1}{\sqrt{2\pi}\tau} \exp\left[-\frac{(\theta_i-\theta_0)^2}{2\tau^2}\right] d\theta_i d\mu_i \Big\}.\end{aligned}$$

In the software of WinBUGS, the first one with $\beta \equiv (\mu, \theta)$ is used for all models, while in this study, DIC based on both forms will be studied. This is because the effects of individual studies are not the main focus of the meta-analysis; instead, we focus on inference about the

overall effect θ_0 (and τ^2), which is a part of ψ . So the second form with $\beta \equiv \psi$ seems to be a reasonable choice. We denote the DIC using the first one as DIC^1 and the second as DIC^2 . Note that two-dimensional integrations needed to be evaluated to compute $P(X|\psi, M)$. Though none of them has a closed form, they can be numerically evaluated.

4.2 Performance in testing both θ_0 and τ^2

We compare the performance of hypothesis testing using classical methods and model selection methods including Akaike information criterion (AIC, Akaike 1974) based on GLMM, Bayesian information criterion (BIC, Schwarz 1978) based on GLMM, and the DIC's (DIC^1 and DIC^2) based on the the Bayesian hierarchical model in Section 2. We consider testing both θ_0 and τ^2 which is equivalent to model selection from $M_1 - M_4$. Using classical tests, the procedure contains two steps including testing $H_{0(\theta_0)} : \theta_0 = 0$ and testing $H_{0(\tau^2)} : \tau^2 = 0$. For instance, if $H_{0(\theta_0)}$ is not rejected but $H_{0(\tau^2)}$ is rejected, then M_2 ($\theta_0 = 0, \tau^2 = 0$) is selected. To test $H_{0(\theta_0)}$, we consider four tests as in Bhaumik et al. (2012), which are based on different estimators of θ_0 including the simple average (SA), DerSimonian and Laird (DSL), empirical logit (EL), and Mantel and Haenszel (MH). To test $H_{0(\tau^2)}$, we use the procedure based on SA in Bhaumik et al. (2012), which has been shown to outperform the alternative test based on Cochran's Q for rare event data. All tests are performed at the significance level $\alpha = 0.05$. For our simulation study, the value of θ_0 is set to be $\{0, 0.25, 0.5, 0.75, 1.0\}$; the value of τ^2 is set to be $\{0, 0.2, 0.4, 0.6, 0.8\}$; the number of studies is fixed at $I = 20$; and $\mu_0 = -5$ so that we focus on rare events. The settings of other parameters are the same as in Section 3. For every combination of θ_0 and τ^2 , $n = 500$ data sets are simulated. Then the classical, AIC, BIC, and DIC methods are used to select the "best" model, respectively.

For comparison, we use the percentage of correct selection (POC) as the performance measure, defined by

$$\text{POC} = \frac{\# \text{ of correct selection of a model}}{n}.$$

Because there are four classical methods considered for testing $H_{0(\theta_0)}$ and one for testing $H_{0(\tau^2)}$, there are four in combination for the two-step testing of both $H_{0(\theta_0)}$ and $H_{0(\tau^2)}$. Our simulation shows that there is no universal winner among the four procedures under different settings. So we use the maximum value of their POC's (maxPOC) to measure the best performance achieved from the classical procedures. Figure 5 contains the contour plots of POC's for different methods in all simulation scenarios. We observe that DIC's appear to have the best overall performance, followed by AIC, then by the classical procedures and BIC. Further, Figure 5a and 5b show that both DIC^1 and DIC^2 perform relatively well when the overall treatment effect is not weak. It can also be seen that the performance is improved by using DIC^2 instead of DIC^1 . Thus, we prefer DIC^2 to DIC^1 .

5 Data Example

Nissen and Wolski (2007) conducted a meta-analysis of 48 trials to evaluate potential adverse effects of rosiglitazone, a widely used drug to lower blood glucose levels in patients

with type 2 diabetes. The adverse endpoints are cardiovascular outcomes including myocardial infarction (MI) and cardiovascular death (CVD). In the 48 trials, 16,856 patients were assigned to the treatment arms with rosiglitazone and 12,962 were assigned to the control groups. There were 86 MI cases in the treatment groups and 72 MI in the control groups. For CVD, the numbers were 39 and 22 for the treatment and control arms, respectively. Here, we use this data set to illustrate the Bayesian approach. Figure 6 shows the posterior densities of the treatment effect θ_0 and the heterogeneity parameter τ^2 for MI and CVD data; and Table 1 presents a summary of Bayesian estimates of θ_0 , τ^2 , μ_0 and σ^2 including the posterior mean, standard error (SE), median, and a 95% credible interval (CI) using the 2.5th and 97.5th percentiles of posterior samples. In this example, the Bayesian estimate of μ_0 is -5.954 and -6.478 for MI and CVD data, respectively, indicating both events of interest are very rare.

We also applied the existing FEM (MH, EL) and REM (DSL, SA, GLMM) methods to estimate and test the overall treatment effect θ_0 , where continuity corrections were used for all the studies. The results are summarized in Table 2, along with the Bayesian estimates listed at the bottom. For MI data, all the methods provide positive estimates except for DSL that gives a slightly negative estimate. For CVD data, the SA method gives a slightly negative estimate while the other methods report positive estimates. The values of the estimates from the different methods are quite variable in both data. Among them, the MH method provides the largest estimates; and only MH reports a significant effect of the treatment in both data (at the significance level $\alpha = 0.05$). This might not be surprising – recall that in Section 3.1, we observe that for rare events, MH has the largest positive bias in estimating θ_0 . GLMM provides the second largest estimates for both data; but it only reports a significant treatment effect for MI data while its p-value for CVD data is at the borderline. All the other tests including SA and BAYES do not reject H_0 . As we can see from Table 1, the reported Bayesian CIs for θ_0 cover 0, indicating that for both MI and CVD data, there is no strong evidence to support the existence of the treatment effect. We mention that Bayes, SA and GLMM are the top three methods in estimating θ_0 , as shown by our simulation results in Section 3.1; and in this example, Bayes and SA reach the same conclusion for testing the existence of the treatment effects. As to estimation of the heterogeneity parameter τ^2 , all DSL, DSK, PM, IPM and GLMM give an estimate of zero for both MI and CVD data. It has been seen in Figure 3 that for a very low background incidence rate μ_0 (-5 or even smaller), all the methods tend to overestimate τ^2 when τ^2 is equal or very close to zero; as τ^2 moves away from zero, all the frequentist methods underestimate τ^2 , but BAYES is nearly unbiased. As shown by Figure 6b and 6d, the posterior densities of τ^2 are very skewed. So τ^2 is estimated by the posterior median, 0.044 and 0.055 for MI and CVD, respectively. Since the above estimates are so close to zero, we conjecture that $\tau^2 = 0$. In this case, BAYES only slightly overestimates τ^2 ; and BAYES is able to draw the same conclusion for testing the treatment effect (the fundamental interest of researchers) as the majority does.

As to model selection, three out of the four classical testing procedures discussed in Section 4.2 select model 1 ($\theta_0 = 0$, $\tau^2 = 0$) except that the one using the MH estimator selects model 2 ($\theta_0 \neq 0$, $\tau^2 = 0$); DIC² selects model 3 ($\theta_0 \neq 0$, $\tau^2 \neq 0$); AIC based on GLMM selects

model 2 ($\theta_0 = 0, \tau^2 = 0$); and BIC based on GLMM selects model 1 ($\theta_0 = 0, \tau^2 = 0$). Note that each of the methods picks up the same model for both MI and CVD data.

6 Discussion

In this article, we examine a Bayesian hierarchical approach to random-effects meta-analysis of binary adverse event data. From simulation studies we find that when the overall background rate μ_0 is very small and there exist moderate or large heterogeneity, the Bayesian approach can provide (much) less biased estimates of the overall treatment effect and/or the heterogeneity parameter than their frequentist competitors. Among the best methods (Bayes, SA/IPM, GLMM), Bayes is less biased than SA and is comparable to GLMM in estimating the treatment effect; and it is less biased than both IPM and GLMM in estimating the heterogeneity parameter. In terms of testing both μ_0 and τ^2 , the performance of the Bayesian procedure using DIC², which is based on the (marginalized) likelihood function of the parameters of main interest, is better in most cases than the classical methods for rare event data.

We note that this article is an attempt from a Bayesian perspective for analysis of rare adverse event data. It provides complementary and comparative results to Bhaumik et al. (2012) with a relatively easy Bayesian implementation of their REM model. More advanced Bayesian modeling techniques can be used to optimize the performance of Bayesian methods in future research.

Acknowledgements

This work was supported in part by the National Institutes of Health grants R15GM113157 and K25AR063761. The authors thank the reviewers for helpful and constructive comments.

References

- Agresti, A. Categorical Data Analysis. 3rd. John Wiley & Sons; 2013.
- Akaike H. A new look at the statistical model identification. Automatic Control, IEEE Transactions on. 1974; 19(6):716–723.
- Bhaumik DK, Amatya A, Normand S-LT, Greenhouse J, Kaizar E, Neelon B, Gibbons RD. Meta-analysis of rare binary adverse event data. Journal of the American Statistical Association. 2012; 107(498):555–567. [PubMed: 23734068]
- Bradburn MJ, Deeks JJ, Berlin JA, Russell Localio A. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. Statistics in medicine. 2007; 26(1):53–77. [PubMed: 16596572]
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. 1993; 88:9–25.
- Cai T, Parast L, Ryan L. Meta-analysis for rare events. Statistics in medicine. 2010; 29(20):2078–2089. [PubMed: 20623822]
- Carlin JB. Meta-analysis for 2×2 tables: A Bayesian approach. Statistics in medicine. 1992; 11(2): 141–158. [PubMed: 1349763]
- Cochran WG. The comparison of percentages in matched samples. Biometrika. 1950; 37(3/4):256–266. [PubMed: 14801052]
- DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemporary clinical trials. 2007; 28(2):105–114. [PubMed: 16807131]

- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled clinical trials*. 1986; 7(3):177–188. [PubMed: 3802833]
- Higgins J, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 2009; 172(1):137–159.
- Kass RE, Raftery AE. Bayes factors. *Journal of the american statistical association*. 1995; 90(430):773–795.
- Mantel N, Haenszel W. Statistical aspects of the analysis of data from retro-spective studies of disease. *Journal of the National Cancer Institute*. 1959; 22(4):719. [PubMed: 13655060]
- Nissen SE, Wolski K. Effect of rosiglitazone on the risk of myocardial infarction and death from cardiovascular causes. *New England Journal of Medicine*. 2007; 356(24):2457–2471. [PubMed: 17517853]
- Paule RC, Mandel J. Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*. 1982; 87(5):377–385.
- Schwarz G. Estimating the dimension of a model. *The annals of statistics*. 1978; 6(2):461–464.
- Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in medicine*. 1995; 14(24):2685–2699. [PubMed: 8619108]
- Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2002; 64(4):583–639.
- Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Stat Med*. May; 2004 23(9):1351–75. [PubMed: 15116347]
- Warn D, Thompson S, Spiegelhalter D. Bayesian random effects meta-analysis of trials with binary outcomes: methods for the absolute risk difference and relative risk scales. *Statistics in medicine*. 2002; 21(11):1601–1623. [PubMed: 12111922]
- Warren FC, Abrams KR, Golder S, Sutton AJ. Systematic review of methods used in meta-analyses where a primary outcome is an adverse or unintended event. *BMC medical research methodology*. 2012; 12(1):64. [PubMed: 22553987]
- Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in cardiovascular diseases*. 1985; 27(5):335–371. [PubMed: 2858114]

Figure 1.

Comparison of estimates of θ_0 by simple average (SA), DerSimonian & Laird (DSL), empirical logit (EL), Mantel & Haenszel (MH), generalized linear mixed model (GLMM) and Bayesian (BAYES) methods for different values of θ_0 . Panel (a) is for $I=10$ and Panel (b) for $I=20$.

Figure 2.

Comparison of estimates of θ_0 by simple average (SA), DerSimonian & Laird (DSL), empirical logit (EL), Mantel & Haenszel (MH), generalized linear mixed model (GLMM) and Bayesian (BAYES) methods for different values of μ_0 . Panel (a) is for $I = 10$ and Panel (b) for $I = 20$.

Figure 3.

Comparison of estimates of τ^2 by DerSimonian & Laird (DSL), DerSimonian & Kacher (DSK), Paule & Mandel (PM) and improved Paule & Mandel (IPM), GLMM and Bayesian (BAYES) methods for different values of τ^2 . Panel (a) is for $I = 10$ and Panel (b) for $I = 20$.

Figure 4.

Comparison of estimates of τ^2 by DerSimonian & Laird (DSL), DerSimonian & Kacher (DSK), Paule & Mandel (PM) and improved Paule & Mandel (IPM) GLMM, and Bayesian (BAYES) methods for different values of μ_0 . Panel (a) is for $I = 10$ and Panel (b) for $I = 20$.

Figure 5.

Contour plots for comparison of testing both θ_0 and τ^2 : POC for DIC^1 (Panel a), POC for DIC^2 (Panel b), POC for AIC (Panel c), POC for BIC (Panel d) and the maximum POC for the two-step testing using four classical procedures (Panel e).

Figure 6.
Data example: posterior densities for θ_0 and τ^2 in MI and CVD data

Table 1

A summary of Bayesian estimates for MI and CVD data

	MI				CVD			
	Mean	SE	Median	95% C.I.	Mean	SE	Median	95% C.I.
θ_0	0.259	0.201	0.261	(-0.071,0.588)	0.142	0.229	0.154	(-0.243,0.492)
τ^2	0.080	0.100	0.044	(0.007,0.271)	0.148	0.242	0.055	(0.008,0.625)
μ_0	-5.954	0.234	-5.948	(-6.353,-5.578)	-6.478	0.137	-6.513	(-6.634,-6.209)
σ^2	0.769	0.311	0.721	(0.357,1.345)	0.563	0.393	0.480	(0.096,1.294)

Table 2

Results of five methods of analyzing the odds ratios

	MI		CVD	
	$\hat{\theta}_0$	p-value	$\hat{\theta}_0$	p-value
MH	0.356	0.011	0.529	0.006
EL	0.181	0.117	0.0877	0.335
SA	0.0213	0.463	-0.0365	0.444
DSL	-0.0047	0.488	0.176	0.176
GLMM	0.299	0.037	0.439	0.054
BAYES	0.259	*	0.142	*