Statistics
in Medicine

Research Article

Received 27 May 2010 Accepted 7 June 2011 Published online 25 August 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4326

An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes

Eleanor M. Pullenayegum[a,b*†]

It is well known that when a Bayesian meta-analysis includes a small number of studies, inference can be sensitive to the choice of prior for the between-study variance. Choosing a vague prior does not solve the problem, as inferences can be substantially different depending on the degree of vagueness. Moreover, because the data provide little information on between-study heterogeneity, posterior inferences for the between-study variance based on vague priors will tend to be unrealistic. It is thus preferable to adopt a reasonable, informed prior for the between-study variance. However, relatively little is known about what constitutes a realistic distribution. On the basis of data from the Cochrane Database of Systematic Reviews, this paper describes the distribution of between-study variance in published meta-analyses, and proposes some realistic, informed priors for use in meta-analyses of binary outcomes. It is hoped that these priors will improve the calibration of inferences from Bayesian meta-analyses. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: meta-analysis; bayesian; random effects; prior

1. Introduction

Meta-analysis is a popular and valuable tool in evidence-based medicine, and indeed a meta-analysis of large randomized controlled trials is considered to be at the top of the hierarchy of evidence [1]. Meta-analysis can be by either fixed effects or random effects, and there are important motivations for using each [2].

Random effects meta-analysis will often need an estimate of the between-study variance. In frequentist analyses this is estimated from the data, but subsequently treated as known. For example, in a DerSimonian and Laird inverse-variance weighted approach [3], the weights are the reciprocal of the sum of the between and within-study variances, but nowhere is the uncertainty in the between-study variance accounted for. This is a particular issue when the number of studies included in the meta-analysis is small. Typically, meta-analyses include fewer than 10 studies, and variances estimated on so few degrees of freedom are likely to be estimated subject to considerable uncertainty.

Amongst their other benefits [4], Bayesian meta-analyses naturally account for the uncertainty in the between-study variance: the between-study variance is simply treated as another unknown parameter to be estimated, and the posterior distribution of the parameters of interest will automatically incorporate the uncertainty in between-study variance. Bayesian meta-analyses do, however, pose a new problem: how to specify a prior for the between-study variance.

Vague or noninformative priors are a popular choice, because, especially in the context of evidence-based medicine, subjectivity is not a desirable attribute in an analysis and hence it is attractive to allow the data to dominate. While it is easy to select a vague prior, it is much harder to select a prior that does not influence the results, particularly when the number of studies included in the meta-analysis is small. Lambert *et al*. [5] explored 13 possibilities for a vague prior and noted that while the posterior median

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada*
[b]*Biostatistics Unit, St Joseph's Healthcare Hamilton, 50 Charlton Ave E, Hamilton, ON L8N 4A6, Canada*
**Correspondence to: Eleanor M. Pullenayegum, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada.*
[†]*E-mail: pullena@mcmaster.ca*

for the pooled odds ratio is relatively consistent across priors, the width of the confidence interval is not. They recommend sensitivity analysis; however, in the case where results vary across analyses, this still poses the problem of which result to use. Senn [6] commented that one of the problems with a frequentist analysis of a small number of trials is that the estimate of between-trial variance is unlikely to be credible, and pointed out that this problem is not solved by using a vague prior distribution that incorporates implausible values for the between-study variability, as this will result in Bayesian results that are not well calibrated [7].

The obvious alternative to using a vague prior is to use an informed prior. This can be based on empirical evidence to avoid subjectivity. Smith [8] and Higgins and Whitehead [9] considered a meta-analysis of meta-analyses to construct a prior for the between-study heterogeneity. Using the between-study heterogeneity in meta-analyses of similar research questions, Smith constructed a prior, and Higgins and Whitehead, under the assumption of exchangeability, placed the heterogeneity in the meta-analysis of interest in the context of similar meta-analyses. A related problem is the estimation of the ICC in a cluster-randomized trial. Turner *et al*. [10] considered using informed priors for the ICC derived by using existing empirical data on ICCs for various health outcomes and using clusters of general practices, post code sectors, or towns. This can be done either by fitting a parametric distribution through maximum likelihood, or by specifying a parametric distribution and letting the parameters of that distribution be unknown parameters in the Bayesian model, information about which can be updated in the model through the use of the external empirical data.

A busy statistician may find the prospect of conducting a meta-analysis of meta-analyses too daunting to fit into an already overloaded schedule, and at times using a vague prior may seem appealing because it requires less work to specify. However, as the number of publications on the subject demonstrates (see, e.g. [6, 11–13]), specifying a vague prior requires a great deal of work. This paper describes an informed prior that might be suitable for general use, and thus avoid the sensitivity associated with vague priors without adding an additional workload. A review of the Cochrane database is conducted to describe the distribution of the between-study heterogeneity in Cochrane meta-analyses. The resulting distribution can be viewed as placing an upper bound on how vague we need to be about the between-study variance. It is hoped that this will provide an informed reference prior suitable for use in a wide range of meta-analyses.

The remainder of this paper is organised as follows. Section 2 gives a motivating dataset to illustrate the problem. Section 3 describes a review of the Cochrane Database of Systematic Reviews and Meta-analyses, and gives the methods used to describe the resulting distribution of between-study variances. The results are given in Section 4. A discussion and some recommendations follow in Section 5.

2. Motivating dataset

As our example dataset we use the same meta-analysis as Lambert and colleagues. The review investigates the use of antibiotics for acute otitis media in children and has since been updated [14]. We take as our comparison the effect of antibiotics versus placebo on the presence of pain at 24 h. The meta-analysis includes five studies, whose results are shown in Figure 1. Throughout this paper, τ^2 will denote the between-study variance in log odds ratios.

The results with some of Lambert's vague priors ($\tau^2 \sim \mathrm{U}(0.001, 1000)$, $\tau^2 \sim \mathrm{U}(0.001, 4)$, $\tau \sim \mathrm{U}(0, 100)$, $\tau \sim \mathrm{U}(0, 2)$) are shown in the top four rows of Figure 2. As can be seen, the width of the

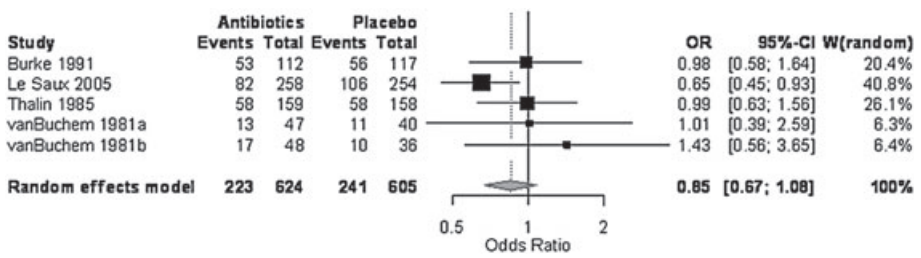

Figure 1. Cochrane Review by Glasziou *et al*. of the effects of antibiotics in pain relief at 24hours amongst children with acute otitis media. In the frequentist random effects analysis, $\tau^2 = 0.0020$, $I^2 = 2.5\%$ (95% CI 0% to 79.7%), $Q = 4.1$ on 4 df ($p = 0.39$).

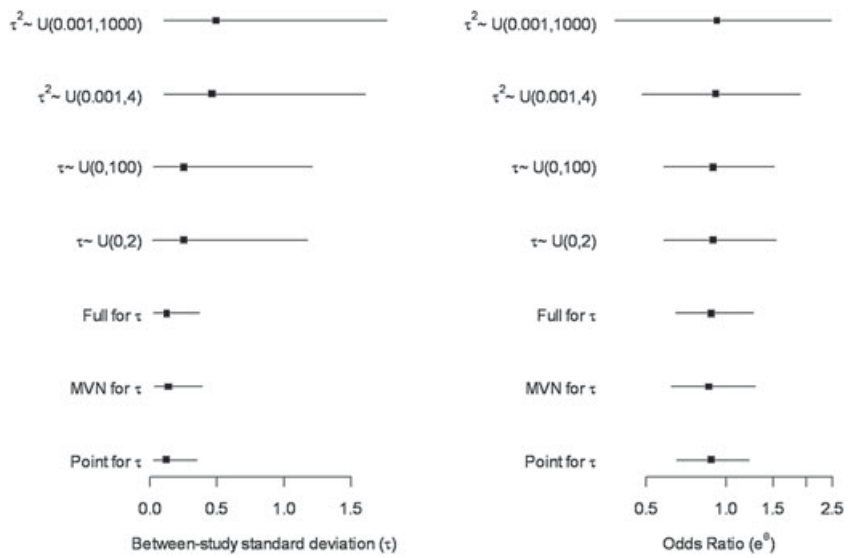

Figure 2. Posterior distributions for τ and θ from the meta-analysis of Glasziou *et al*., using the informed priors and a selection of Lambert's vague priors.

Bayesian 95% CrI for the pooled log odds ratio varies widely with the choice of prior, as do both the posterior medians and the posterior credible intervals for τ. Note that if the prior distribution for τ does not represent reasonable beliefs, there is no reason to believe the posterior distribution.

Spiegelhalter *et al.* [15] and Smith *et al.* [16] used theoretical reasoning to determine which values of τ are reasonable in practice. Smith *et al.* based their reasoning on the distribution of study-level odds ratios (ORs), showing that the ratio of the 97.5th percentile of the study odds ratios to the 2.5th percentile is $e^{3.92\tau}$, and arguing that usually the range of study ORs would be within an order of magnitude, that is, $e^{3.92\tau} < 10$ and so $\tau < 0.59$. Spiegelhalter *et al.* based their reasoning on the median of the ratio of two randomly drawn study-level odds ratios. The absolute difference between two randomly drawn log odds ratios follows a half-normal distribution $HN(2\tau^2)$, and so the median ratio of the smaller odds ratio to the larger is $e^{1.09\tau}$. Thus, for example, $\tau = 0.64$ leads to a median ratio of 2. On the basis of this type of reasoning, values of τ that are less than 0.1 would lead to a median ratio of 1.1 and might be considered small, and values that are greater than 1 represent a median ratio of 3 and thus represent extreme heterogeneity. Using this characterisation, the posterior credible intervals for τ in the otitis media example include values that are small, and values that represent extreme heterogeneity. For example, from the first row of Figure 2, the posterior credible interval includes 1.5, that is, the median ratio of two randomly drawn odds ratios would be over 5, representing very extreme heterogeneity. which we would not have deemed to be reasonable *a priori*. This illustrates the difficulty with the use of vague priors; because they do not represent reasonable *a priori* beliefs, the resulting posterior distributions will be similarly unreasonable when the number of included studies is small.

While theoretical reasoning can provide us with guidelines as to what is and is not reasonable, this does not tell us about the distribution of values in practice. The following review of the Cochrane database is intended to provide this information.

3. Methods

All the reviews in the Cochrane Database of Systematic Reviews published between January 2008 and 31 July 2009 were identified. Reviews were excluded from analysis if any of the following held: there were no pooled results; the first reported pooled result was not for a binary outcome; there were no events in any of the included studies. For each included review, the first forest plot was used. If this contained pooled results for several subgroups followed by a combined result across subgroups, the combined analysis was used; if results were not combined across subgroups, the first pooled result was used.

For the reviews contributing to the analysis, study-level results were retrieved (number of events, number of nonevents in each arm). We also recorded review-level characteristics: inclusion criteria for study type (exclusively RCTs, RCTs or observational studies, observational studies only), type of intervention

(pharmacological or other), outcome assessor (healthcare professional versus patient or patient proxy), and type of outcome (objective, defined as an outcome with clearly defined and objectively measured criteria, e.g., mortality; semi-objective, defined as requiring some subjective judgement but captured on a scale, e.g., the Rankin scale; subjective, defined as depending purely on personal opinion, e.g. pain; and dropout, which is used in many psychiatric and substance abuse studies). Regardless of whether the study reported an odds ratio or a relative risk, pooled results and measures of between-study variance were calculated on the odds ratio scale.

For descriptive purposes, the method-of-moments estimator of the between-study variance was used. If a meta-analysis contains k studies, and we observe study-level log odds ratios $\hat{\theta}_i$ with associated standard errors σ_i, we have $\hat{\theta}_i \sim N(\theta_i, \sigma_i^2); \theta_i \sim N(\theta, \tau^2)$, and the method-of-moments estimator is given by

$$\hat{\tau}^2 = (Q-(k-1)) \Bigg/ \left(\sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \right) \text{ if } Q \geq k-1$$
$$0 \text{ if } Q < k-1$$

where Q is Cochran's Q-statistic: $Q = \sum_{i=1}^{k} w_i (\hat{\theta}_i - \hat{\theta}^F)^2$

$\hat{\theta}^F$ is the frequentist fixed-effects estimate: $\hat{\theta}^F = \sum_{i=1}^{k} w_i \hat{\theta}_i \Big/ \sum_{i=1}^{k} w_i$

w_i are the study weights $w_i = 1/\sigma_i^2$

and $\hat{\theta}^R$, the frequentist random-effects estimate, is given by $\hat{\theta}^R = \sum_{i=1}^{k} w_i^* \hat{\theta}_i \Big/ \sum_{i=1}^{k} w_i^*, w_i^* = 1/(\sigma_i^2 + \tau^2)$.

For a formal analysis of the distribution of the between-study variance τ it is desirable to account for the uncertainty in the estimate of τ, and hence a Bayesian approach was used. The distribution of estimates of between-study variance suggests a gamma, inverse-gamma or log-normal distribution. A previous work by Higgins and Whitehead [9] found the inverse gamma distribution to be a good fit. We initially considered all three distributions; however, the MCMC sampler encountered problems with the gamma distribution, and in what follows, we discuss just the inverse gamma and log-normal distributions.

In the models below, $\log OR_{ij}$ denotes the observed log odds ratio for the i^{th} study in the j^{th} review, θ_{ij} is the true log OR for the i^{th} study in the j^{th} review, θ_j^{pooled} is the mean of the assumed normal distribution from which the true log odds ratios for the j^{th} review are drawn, and τ_j is the standard deviation of that distribution.

The inverse-gamma model was

$$\begin{aligned} \log OR_{ij} &\sim N(\theta_{ij}, var_{ij}) \\ \theta_{ij} &\sim N(\theta_j^{\text{pooled}}, \tau_j^2) \\ \tau_j^2 &\sim IG(a, b) \end{aligned}$$

and we used vague priors on the remaining parameters, that is, $\theta_j^{\text{pooled}} \sim N(0, 10)$, $a \sim \Gamma(0.001, 0.001)$, $b \sim \Gamma(0.001, 0.001)$.

The log-normal model was initially fit without covariates and then expanded to include review-level covariates

$$\begin{aligned} \log OR_{ij} &\sim N(\theta_{ij}, var_{ij}) \\ \theta_{ij} &\sim N(\theta_j^{\text{pooled}}, \tau_j^2) \end{aligned}$$

Without covariates:

$$\begin{aligned} &\log \tau_j^2 \sim N(\mu_\tau, \sigma_\tau^2); \theta_j^{\text{pooled}} \sim N(0, 10) \\ &\mu_\tau \sim N(0, 10); \sigma_\tau \sim U(0, 10) \end{aligned}$$

With covariates:

$$\log \tau_j^2 \sim N(\mathrm{X}_{\tau j}\alpha, \sigma_\tau^2); \theta_j^{\text{pooled}} \sim N(\mathrm{X}_{\theta j}\beta, \sigma_\theta^2)$$
$$\alpha_0 \sim N(0, 1000); \alpha_k \sim N(0, 100); \beta_0 \sim N(0, 10); \beta_2, \ldots, \beta_k \sim N(0, 0.25)$$
$$\sigma_\tau \sim U(0, 100); \sigma_\theta \sim U(0, 100)$$

where $\mathrm{X}_{\tau j}$ is a vector of review-level covariates for τ (initially including just the intercept to compare to the inverse-gamma model), and $\mathrm{X}_{\theta j}$ is a vector of review-level covariates for θ. The priors for $\mu, \sigma, \alpha, \beta$ are chosen to be vague. While $\tau^2 = 3$ would represent extreme heterogeneity, τ^2 could be close to zero and hence the mean of $\log(\tau^2)$ could be large in absolute value. This motivates choosing a prior for α_0 that is very diffuse. Priors for $\alpha_1, \ldots, \alpha_k$ were also chosen to be diffuse, but less so than α_0 because the impact of a covariate on between-study heterogeneity was likely to be limited. Choosing a $N(0,10)$ prior for β_0 represents a prior belief that 95% of odds ratios will be between $\exp(-6) = 0.002$ and $\exp(6) = 400$, that is, a very vague prior. Similarly, $N(0, 0.25)$ priors for the β_k result in a prior belief that the ratio of odds ratios associated with the covariate lie between $\exp(-1) = 0.37$ and $\exp(1) = 2.7$, which represents a reasonable, but still vague belief.

Because the sign of θ depends on whether the binary outcome is expressed as a favourable or an unfavourable outcome, and that choice is arbitrary, regression models for θ were fit after flipping the sign for θ for those reviews in which an outcome event represented a favourable response (i.e. for those reviews in which an odds ratio that was greater than 1 favoured treatment rather than control). When including logarithmic transforms of θ in the covariates X_τ for $\log \tau^2$, an offset is needed to avoid singularities at $\theta = 0$; and in an attempt to reduce collinearity between the offset, the regression coefficient for the logarithmic term and the intercept α_0, we replaced the covariate $\log(\text{offset} + |\theta|)$ with $\log((\text{offset} + |\theta|)/(\text{offset} + 0.5))$.

These analyses are based upon the summary statistics $\log OR_{ij}$ and the associated variance var_{ij}, rather than upon the raw data (counts of events and nonevents in each treatment group). While computationally easier, the use of summary statistics excludes studies with no events (or no nonevents) in both arms, and requires a continuity correction when there are no events or no nonevents in one arm of the study. The standard correction in which 0.5 is added to each cell of the study was used, because it represents the most commonly used method, despite the fact that it can be biased [17].

Convergence was monitored using the Geweke [18] and Heideberger and Welch [19] statistics as implemented in the R package BOA [20]. All MCMC simulations were run in WinBUGS [21] for 100 000 iterations in the first instance, and the first 50 000 iterations were discarded. Convergence diagnostics were run on the remaining 50 000 iterations, and further iterations were run if needed. Goodness-of-fit for these models was assessed using the Deviance Information Criterion (DIC) and through $q-q$ plots. Specifically, once convergence had been checked, a single iteration of the Gibbs sampler was run and the random draws from the posterior distributions of τ_j^2 and θ_j were extracted. These were then compared with their assumed sampling distributions using $q-q$ plots.

4. Results

The search retrieved 942 records. Of these, 314 provided data contributing to analysis. Of those excluded, 103 did not include any studies, 320 did not pool results for any outcome. For 198 the first forest plot did not relate to a binary outcome, in 4, the included studies either all had 0% event rates or all had 100% event rates. One review was excluded because it included only cross-over trials, and two were excluded because the numbers of events in each arm were not reported.

4.1. Descriptive statistics

Descriptive statistics of the included reviews are shown in Table I: 196 (62%) of the meta-analyses were of a pharmacological intervention, the remainder were of procedural or device interventions; 299 stipulated RCT as an inclusion criterion, 13 intended to include both RCTs and cohort studies, and two included cohort studies; 245 studied an objective outcome such as a diagnosis or an event, 28 studied a semi-objective outcome such as a scale, 26 studied a subjective outcome such as pain, and 15 had dropout as their primary outcome. The outcome was assessed by the patient or patient caregiver in 59 of the analyses and by a physician or other healthcare professional in the remaining 255.

The number of trials per review ranged from 2 to 111, with a median of 4, and quartiles 3 and 8. After recoding the outcome so that an OR less than one favoured the intervention, the median pooled

Table I. Descriptive statistics on included reviews.

Characteristic	$n/N(\%)$
Intervention - pharmacological versus device/procedure/service	196/314 (62%)
Study type	RCT: 299/314 (95%); RCT & cohort: 13/314 (4%); cohort: 2/314 (1%)
Outcome	Objective: 245/314 (78%); semi-objective: 28/314 (9%); subjective: 26/314 (8%); dropout: 15/314 (5%)
Assessment - healthcare professional versus patient or patient caregiver	255/314 (81%)

Examples of objective outcomes: mortality, diagnosis of cancer, febrile state after 6 days, live birth.
Examples of semi-objective outcomes: neurological improvement according to a validated scale, ACR20 (improvement according to a set of scales), need for analgesia, failure to respond to treatment amongst patients with depression, disease progression in MS, response to treatment using a depresseion scale (e.g. the Montgomery-Asberg Depression Rating Scale), clinical global impressions of anxiety.
Examples of subjective outcomes: increased cough, mood symptoms based on patient report only, pain relief, sore throat, substantial acute relief from migraine, hypersalivation, nausea.

OR was 0.72 (IQR 0.44, 1.02). Approximately half the reviews (153/314; 49%) showed no discernable heterogeneity using the method-of-moments estimate. Four reviews (1%) had heterogeneity between 0 and 0.1, 78 (25%) fell between 0.1 and 0.5, 58 (18%) between 0.5 and 1, and 21 (7%) exceeded 1. A histogram of the method-of-moments estimates is given in Figure 3.

4.2. Distributional form

Analysis first investigated the distributional form without accounting for covariates. Using a log-normal distribution for τ^2 led to a DIC of 4697, compared with 4711 for the inverse gamma distribution. The log-normal distribution thus seems to be the better choice, but $q-q$ plots based on a single iteration of the Gibbs sampler after convergence suggest that neither distribution is a good fit (see the top row of Figure 4). However, incorporating covariates may improve the fit and thus further analysis was based upon the log-normal distribution.

4.3. Review-level covariates associated with between-study heterogeneity

On including covariates, there was no important association between the number of studies included in a meta-analysis and the amount of heterogeneity. Because the number of included studies is an outcome

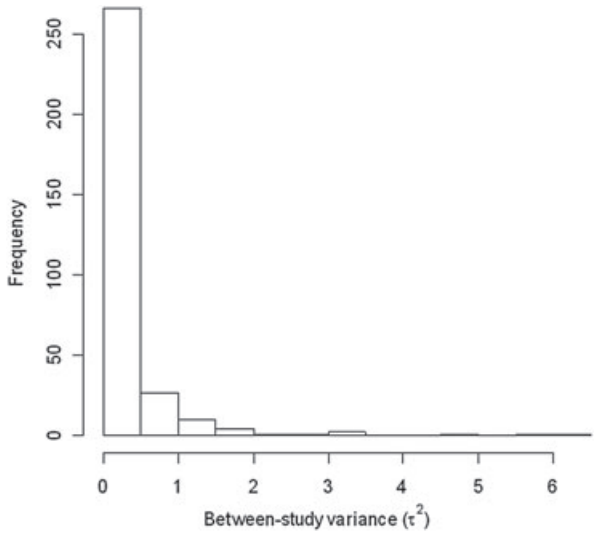

Figure 3. Histogram of the method-of-moments estimates of τ^2.

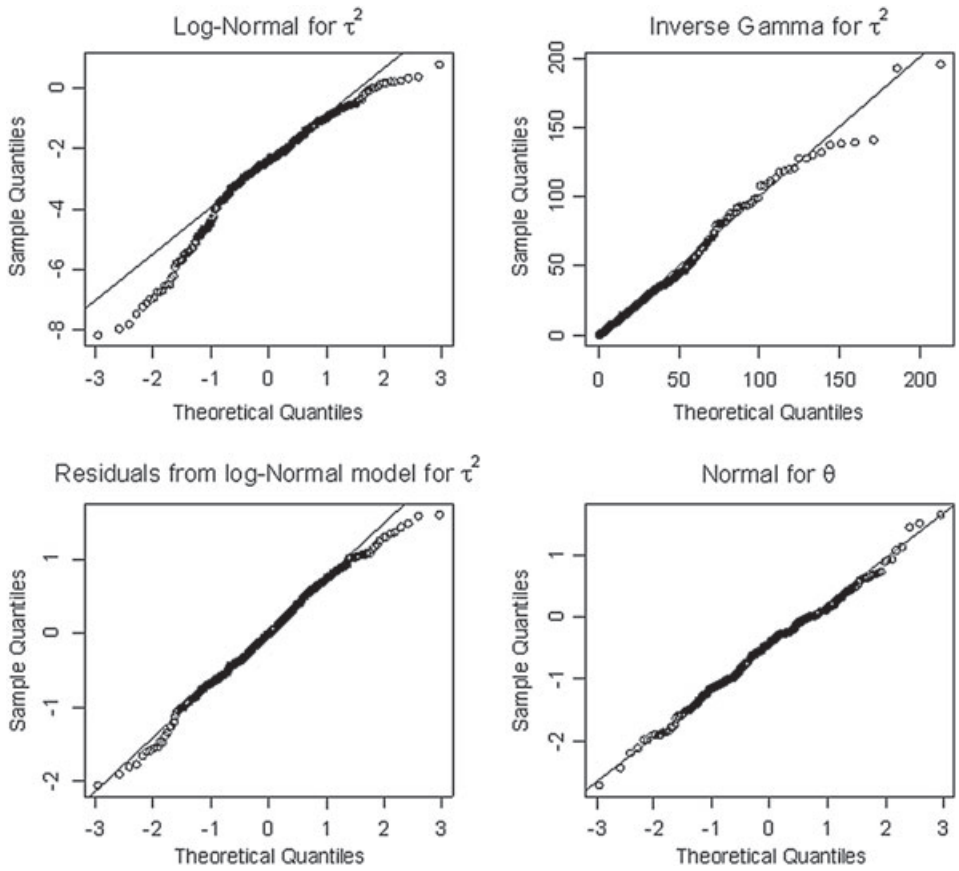

Figure 4. Quantile–quantile plots based on a single iteration of the Gibbs sampler after convergence. The top row evaluates goodness-of-fit for the models for τ^2 in the absence of covariates. The bottom row evaluates goodness of fit of the final models for both τ^2 and θ.

of the meta-analysis rather than an *a priori* decision, this was omitted from further analyses. There were too few meta-analyses including studies other than RCTs to allow assessment of the effects of study type (study type was collinear with the intercept). As can be seen in Table II, there was no improvement in fit upon including covariates for whether the intervention was pharmacological as opposed to a procedure or intervention, for whether the outcome was objective as opposed to subjective or semi-objective or whether the outcome was assessed by the patient or patient caregiver. There was a strong association between heterogeneity and the size of the pooled OR: for every unit increase in absolute pooled log OR, the distribution of the increase in mean log τ^2 had a posterior mean of 2.1 (standard deviation 0.28).

The shape of the association between the pooled effect and the heterogeneity deserves some attention. Figure 5 shows a loess fit of the relationship between the frequentist method-of-moments estimate of log τ^2 and the absolute pooled log effect size, amongst the subset of studies with $\tau^2 > 0$. This suggests a concave association. The DICs reported in Table II show that the fit of the model improves upon using square root or logarithmic transformations; the logarithmic transformation includes an offset to avoid singularities as $|\theta| \to 0$. When the offset is treated as stochastic with a Gamma(0.1,0.1) prior, its posterior distribution suggests that it is small, with a posterior mean of 0.0001. On fixing the offset at 0.0001, there is no loss of fit in the model, and thus $\log(0.0001 + |\theta|)/\log(0.0001 + 0.5)$ is used to model the shape of the association between the pooled effect size and the amount of between-study heterogeneity.

4.4. Exchangeability of reviews

Next, the association between the pooled effect size and covariates was investigated. Although this was not the primary concern of the analysis, it was deemed important because the model assumes exchangeability in θ amongst the reviews. There seemed to be little association between the magnitude of the pooled effect and whether or not the intervention was pharmacological; however, objective outcomes

Table II. Modelling details for the joint distribution of θ and τ. Numeric cell entries are posterior means and standard deviations, with the exception of the DIC.

Parameter	Covariates for $\log \tau^2$			Functional form for θ			Covariates for θ			
Functional form for θ	$\lvert\theta\rvert$	$\lvert\theta\rvert$	$\lvert\theta\rvert$	$\sqrt{\lvert\theta\rvert}$	$\log\left(\frac{\text{offset}+\lvert\theta\rvert}{\text{offset}+0.5}\right)$	$\log\left(\frac{0.0001+\lvert\theta\rvert}{0.0001+0.5}\right)$	$\log\left(\frac{0.0001+\lvert\theta\rvert}{0.0001+0.5}\right)$			
α_0	−3.00 (0.54)	−2.97 (0.52)	−4.12 (0.29)	−5.38 (0.41)	−2.50 (0.17)	−2.50 (0.17)	−2.52 (0.17)	−2.48 (0.18)	−2.45 (0.16)	−2.49 (0.18)
#studies	−000603 (0.0086)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
α_{drug}	−0.48 (0.32)	−0.48 (0.31)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
$\alpha_{objective}$	−0.42 (0.43)	−0.42 (0.42)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
$\alpha_{assessor}$	−0.60 (0.40)	−0.60 (0.38)	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
α_θ	1.93 (0.32)	1.90 (0.32)	2.10 (0.28)	3.67 (0.44)	1.36 (0.20)	1.36 (0.20)	1.36 (0.18)	1.34 (0.19)	1.33 (0.17)	1.33 (0.17)
offset	n/a	n/a	n/a	n/a	0.0001 ($3.9*10^{-20}$, 0.082)	n/a	n/a	n/a	n/a	n/a
β_0	−0.44 (0.043)	−0.44 (0.043)	−0.44 (0.043)	−0.44 (0.043)	−0.44 (0.043)	−0.44 (0.043)	−0.82 (0.11)	−0.89 (0.10)	−0.79 (0.090)	−0.73 (0.087)
β_{drug}	n/a	n/a	n/a	n/a	n/a	n/a	−0.089 (0.081)	n/a	n/a	n/a
$\beta_{objective}$	n/a	n/a	n/a	n/a	n/a	n/a	0.20 (0.11)	0.22 (0.10)	n/a	0.36 (0.096)
$\beta_{assessor}$	n/a	n/a	n/a	n/a	n/a	n/a	0.35 (0.11)	0.33 (0.11)	0.43 (0.098)	n/a
σ_θ	0.67 (0.036)	0.67 (0.037)	0.67 (0.037)	0.67 (0.037)	0.67 (0.037)	0.67 (0.037)	0.64 (0.036)	0.64 (0.036)	0.65 (0.036)	0.65 (0.035)
σ_{norm}	0.94 (0.26)	0.91 (0.22)	0.95 (0.19)	0.81 (0.21)	0.75 (0.19)	0.75 (0.19)	0.76 (0.21)	0.74 (0.20)	0.69 (0.20)	0.74 (0.20)
DIC	4662	4662	4660	4653	4653	4652	4641	4639	4645	4648

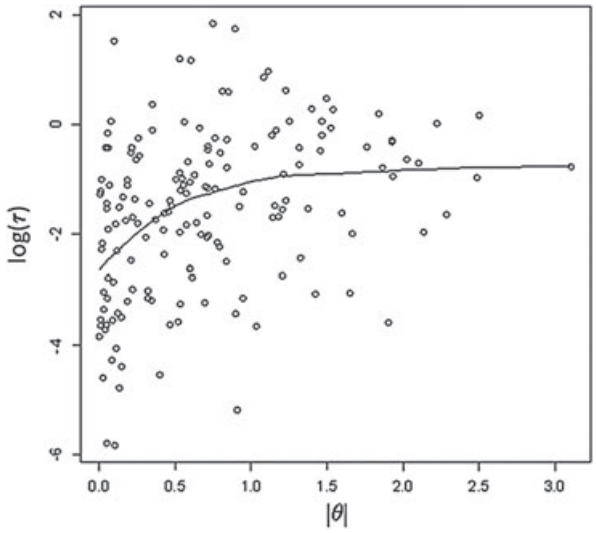

Figure 5. Scatterplot of between-study variance versus the absolute pooled log effect size, with loess curve.

Table III. Posterior summary statistics and fitted marginal distributions for the final joint model for θ and τ : $\log \tau^2 \sim N\left(\alpha_0 + \alpha_\theta * \log((0.0001 + |\theta|)/(0.0001 + 0.5)), \sigma_\tau^2\right)$, $\theta \sim N\left(\beta_0 + \beta_{objective} * \text{objective} + \beta_{assessor} * \text{assessor}, {\sigma_\theta}^2\right)$. For the marginal distribution, the posterior mean and standard deviation were used as parameters for the Normal distribution, after checking that the 95% intervals implied by the fitted distribution approximately matched the estimated posterior 95% CrIs.

Parameter	Mean	sd	Percentiles			Fitted marginal distribution
			2.5%	50%	97.5%	
for τ:						
α_0	−2.482	0.177	−2.861	−2.469	−2.167	$N(-2.48,0.18^2)$
α_θ	1.337	0.190	1.012	1.318	1.755	$N(1.34,0.19^2)$
σ_τ	0.738	0.195	0.401	0.723	1.158	$N(0.74,0.20^2)I(0,)$
for θ:						
β_0	−0.887	0.101	−1.086	−0.886	−0.687	$N(-0.89,0.10^2)$
$\beta_{objective}$	0.221	0.105	0.016	0.220	0.425	$N(0.22,0.11^2)$
$\beta_{assessor}$	0.334	0.110	0.119	0.334	0.551	$N(0.33,0.11^2)$
σ_θ	0.664	0.036	0.577	0.643	0.717	$N(0.64,0.036^2)I(0,)$

Variance-covariance matrix for α_0, α_θ, σ_τ, β_0, $\beta_{objective}$, $\beta_{assessor}$, σ_θ

	α_0	α_θ	σ_τ	β_0	$\beta_{objective}$	$\beta_{assessor}$	σ_θ
α_0	0.03610	−0.01485	−0.00045	0.00139	−0.00068	0.00764	−0.00055
α_θ	−0.01485	0.03240	0.00008	−0.00097	0.00094	−0.02139	−0.00008
σ_τ	−0.00045	0.00008	0.01000	−0.00473	−0.00414	−0.00073	0.00002
β_0	0.00139	−0.00097	−0.00473	0.01210	−0.00608	0.00151	0.00004
$\beta_{objective}$	−0.00068	0.00094	−0.00414	−0.00608	0.01000	−0.00072	−0.00015
$\beta_{assessor}$	0.00764	−0.02139	−0.00073	0.00151	−0.00072	0.04000	−0.00080
σ_θ	−0.00055	−0.00008	0.00002	0.00004	−0.00015	−0.00080	0.00130

and outcomes assessed by a healthcare professional were associated with larger (less favourable) overall effects. As can be seen from Table II, the fit of the model was best when both of these covariates were included.

Summary statistics for the parameters of the chosen model are shown in Table III. The bottom row of Figure 4 gives quantile–quantile plots for one random draw from the posterior distribution of θ, and of the residuals of the log-normal model for τ^2. As can be seen, the chosen distributions fit fairly well.

4.5. Approximating the prior

We then explored how this prior might be applied in practice. In practice analysts who wish to use our proposed priors for τ^2 would usually wish to use either a vague prior for θ or else an informed prior based on information external to their meta-analysis but still relevant to the question under consideration. Thus, although our analysis yields informative priors for both τ^2 and θ, we consider using the informative prior for τ^2 only in conjunction with a vague $N(0, 100)$ prior for θ.

The first option is to use a hierarchical specification, in which $\log \tau^2 \sim N(\alpha_0 + \alpha_\theta^* \log((0.0001 + |\theta|)/(0.0001 + 0.5)), \sigma_\tau^2)$, $\theta \sim N(0, 100)$ and the full joint posterior distribution of the parameters α_0, α_θ, and σ_τ^2 from the above analyses is used in specifying the prior for $\log \tau^2$ in a new meta-analysis. Doing this would require the analyst to run the code given in the Appendix, and load in the data from the 314 studies included in this review. For ease of use, two approximations to this distribution are considered. First, because the marginal distributions of each of the parameters seem to be approximately normal, a multivariate normal (MVN) seems a natural approximation to joint distribution of the parameters α_0, α_θ, and σ_τ^2 (hereafter referred to as the MVN prior). The means and variance–covariance matrix for this MVN distribution are given in Table III; although Table III gives the joint distribution for α_0, α_θ, σ_τ^2, β_0, β_{assessor}, $\beta_{\text{objective}}$ and β_θ, the joint distribution for just the parameters of interest can easily be inferred. The second approximation (hereafter referred to as the point prior) holds α_0, α_θ, σ_τ fixed at their posterior means, thus ignoring the associated uncertainty. Clearly, the MVN and point priors are easier to use in practice than the full prior, and so the question of interest is whether they characterize the distribution sufficiently to give similar results to the full prior.

4.6. Simulation study: evaluating the MVN and point approximations

Data was simulated from the full joint distribution, then analysed using the MVN and point priors to assess how closely the MVN and point priors approximate the full joint prior. If these two priors are good approximations to the full prior, the posterior 95% credible intervals should cover the true value of τ 95% of the time.

The simulation set-up was as follows. Firstly, 1000 values of θ and τ were simulated from the full joint distribution by simulating 1000 values of α_0, α_θ, β_0, $\beta_{\text{objective}}$, β_{assessor}, σ_τ and σ_θ from their full joint posterior distribution, taking a lag of 100 between values so that their autocorrelation was close to zero. One thousand values of $\log(\tau^2)$ and θ were generated on the basis of their assumed normal distributions given the simulated parameters. For each of these (τ, θ) pairs, a single meta-analysis was simulated, consisting of two studies, one with 200 patients per arm and one with 100 patients per arm; both studies had a 50% event rate in the control arm. Each of the resulting simulated reviews was analysed using the full prior, the MVN prior, and the point prior, and a vague uniform(0,100) prior for τ for comparison. The coverage rates (i.e. the percentage of time that the 95% CrI contained the true value) are given in Table IV, and were close to 95% for all the informed priors, indicating that the MVN and point priors both adequately approximate the full prior in this respect. As would be expected, the uniform prior had poor coverage, with excessive coverage for θ and coverage for τ that was too small.

4.7. Applying the priors to the otitis media example

The priors were then applied to the otitis media example, and the results contrasted with the vague priors considered earlier. Compared with the results with vague priors, the results of all the informed priors were consistent with one another, both in terms of the posterior medians and the credible intervals, for both τ and θ (see Figure 2). Using an informed prior for τ led to smaller posterior medians for τ and θ than using a vague prior for τ. The credible intervals for both parameters were narrower with the informed priors than with the vague priors.

Table IV. Simulation study: coverage probabilities for the full, MVN, and point priors, with the uniform(0,100) prior for comparison.

Parameter	Full	MVN	Point	Uniform
τ	94.7	94.9	95.0	83.5
θ	95.3	95.4	95.2	100

5. Discussion and conclusions

It has previously been noted that it is difficult to specify a prior for the between-study variance in a meta-analysis that is vague in the sense of allowing the data to dominate, especially when the number of studies included in the review is small. Moreover, vague priors tend to lead to posterior distributions for the between-study variance that are not believable. Consequently, it is advisable to use an informed prior for the variance parameters in a hierarchical analysis. This paper has described the distribution of the between-study variance amongst Cochrane reviews published between 2008 and 2009, and investigating a binary outcome. A log-normal distribution incorporating the association between the between-study variance and the pooled effect size gave the best fit.

This work does have its limitations. Some studies may have been included in more than one Cochrane review, making independence questionable; however, the extent to which this happened, and the consequent impact, is likely minimal. Work was restricted to Cochrane reviews, as the standardised reporting makes data extraction more straightforward. However, Cochrane reviews may differ from other meta-analyses in important ways. For example, Cochrane reviews usually focus on randomized trials. The extent to which the proposed distribution would apply to meta-analyses in general is unclear.

Both the I^2 and τ have been used as measures of heterogeneity in meta-analyses [22]. While the I^2 has been more popular because of its ease of interpretation, it is dependent on the size of the studies included in the meta-analysis and so should be evaluated alongside the between-study variance τ^2 [23]. Because of this dependence, we chose to characterise the distribution of τ rather than I^2. While the distribution of I^2 could be characterised as a function of study size, because it is τ itself that is needed in the analysis, this seems the more logical choice to describe. Moreover, as noted by Higgins *et al.*, although in the past τ has often been viewed as a nuisance parameter, it is in fact an important output of a random effects meta-analysis, because it helps to describe the distribution of treatment effects in the population [2].

The Bayesian hierarchical models assume that the Cochrane reviews may be considered exchangeable, that is that there is no *a priori* reason to believe that any given review should have more or less heterogeneity than another, or a larger or smaller effect size than another. The extent to which this assumption holds was investigated by exploring review-level covariates in the model for τ and θ. After accounting for the pooled effect θ, we found there was no need to incorporate review-level covariates for τ, adding credibility to the assumption of exchangeability. In the model for θ, objectivity of the outcome and whether the outcome was assessed by a physician or other healthcare professional were associated with values of θ that were less favourable towards the intervention than outcomes that were not objective or that were patient-assessed. There were however some covariates that might be considered important, namely blinding, that we were not able to include, as these are in many cases study-level rather than review-level covariates.

It is also important to note that in the case where there are few studies in the meta-analysis, the prior will have an important effect on the results, and thus while the informed priors should lead to results that have good Bayesian properties (i.e. good calibration), they may have poor frequentist properties. For example, the confidence interval coverage probabilities will likely be off from the nominal levels.

There are three ways the proposed log-normal prior for τ might be implemented. The full joint distribution representing the uncertainty in the mean and precision of the normal distribution can be used, and for this purpose the raw data and the code used to estimate the parameters of interest has been provided in the Appendix. The second option is to use a multivariate normal approximation to the joint posterior distribution of the unknown parameters, and the final option is simply to use the posterior medians for the parameters. Simulated meta-analyses with just two studies per review suggest that we can simply use the posterior means for these parameters (the point prior), that is, take $\log(\tau^2) \sim N(-2.48 + 1.34^* \log((0.0001 + |\theta|)/(0.0001 + 0.5)), \text{sd} = 0.74)$.

Typically analysts use vague priors either because they want to let the data dominate, or because they know little about the parameter in question (or both). When there are few trials in a meta-analysis, it may be impossible to let the data dominate. In this case informed priors for variance components are helpful. This review of the Cochrane database provides empirical data on what might be considered a plausible prior. It is hoped that this provides a reference prior that is suitable for general use and, at the least, places an upper bound on how vague the prior needs to be.

APPENDIX A. WinBUGS code for the full joint posterior

```
for(i in 1:n){
# logor.all and std.dev are the study-specific odds ratios and std deviations, and form the data in the model
logor.all[i] ~ dnorm(mu.all[i],prec.all[i])
prec.all[i] <- 1/pow(std.dev[i],2)
# The true underlying study-level logORs cluster around a review-level pooled OR
# id gives the review id, i.e. it identifies which studies were from the sample review
# prec.tau.all is the between-study precision; tausq.all is the between-study variance
mu.all[i] ~ dnorm(theta.all[id[i]],prec.tau.all[id[i]])
}

for(i in 1:314){
prec.tau.all[i] <- 1/tausq.all[i]
tausq.all[i] <- exp(log.tausq.all[i])

# Between-study variances form a log-Normal distribution that is correlated with the pooled effect
log.tausq.all[i] ~ dnorm(mu.norm[i],prec.norm)
mu.norm[i] <- alpha0 + alpha.theta*log((offset + abs(theta.all[i]))/(offset + 0.5)) + alpha.drug*drug[i]
+ alpha.phys*physician[i] + alpha.objective*objective[i]

# The pooled effect sizes follow a Normal distribution theta.signed[i] ~ dnorm(mu.theta[i],prec.theta)
mu.theta[i] <- beta0 + beta.drug*drug[i] + beta.phys*physician[i] + beta.objective*objective[i]
theta.all[i] <- theta.signed[i]*pow(-1,1-bad.outcome[i])
# The sign of the logOR is meaningless unless we code so that the outcome represents something unfavourable. Once this is done, a logOR< 0 represents evidence in favour of the intervention
}

offset <- 0.0001

# priors for the hyper-parameters
prec.norm <- 1/(sd.norm*sd.norm)
sd.norm ~ dunif(0,100)
prec.theta <- 1/(sd.theta*sd.theta)
sd.theta ~ dunif(0,100)
alpha0 ~ dnorm(0,0.001)
alpha.theta ~ dnorm(0,0.01)
beta0 ~ dnorm(0,0.1)
beta.objective ~ dnorm(0,16)
beta.phys ~ dnorm(0,16)

alpha.drug <- 0
alpha.phys <- 0
alpha.objective <- 0
beta.drug <- 0
}
```

Acknowledgements

This work was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council.

References

1. Phillips B, Ball C, Sackett DL, Badenoch D, Straus S, Haynes RB, Dawes M. *Oxford Centre for Evidence-based Medicine-Levels of Evidence*. Oxford University Press: Oxford, 2009.
2. Higgins JP, 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.

3. DerSimonian R, Laird NM. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
4. Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 2001; **10**(4):277–303.
5. 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.
6. Senn S. Trying to be precise about vagueness. *Statistics in Medicine* 2007; **26**:1417–1430.
7. Dawid AP. The well-calibrated Bayesian. *Journal of the American Statistical Association* 1982; **77**:605–610.
8. Smith TC. *Interpreting evidence from multiple randomised and non-randomised studies*. University of Cambridge: Cambridge, 1995.
9. Higgins JP, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 1996; **15**(24):2733–2749.
10. Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials* 2005; **2**(2):108–118.
11. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; **1**:515–533.
12. Daniels MJ. A prior for the variance in hierarchical linear models. *Canadian Journal of Statistics* 1999; **27**:567–578.
13. Gustafson P, Hossain S, MacNab YC. Conservative prior distributions for variance parameters in hierarchical linear models. *Canadian Journal of Statistics* 2006; **34**:377–390.
14. Glasziou PP, Del Mar CB, Sanders SL, Hayem M. Antibiotics for acute otitis media in children. *Cochrane Database of Systematic Reviews* 2004; **0**(1):CD000219.
15. Spiegelhalter DJ. Bayesian methods for cluster-randomized trials with continuous responses. *Statistics in Medicine* 2001; **20**:435–452.
16. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* 1995; **14**(24):2685–2699.
17. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? use and avoidance of continuity corrections in meta-analysis of sparse data. *Statistics in Medicine* 2004; **23**(9):1351–1375.
18. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, Bernardo JM, Berger J, Dawid AP, Smith AFM (eds), 4th ed. Oxford University Press: Oxford, U.K., 1992.
19. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. *Operations Research* 1983; **31**:1109–1144.
20. Smith BJ. Bayesian Output Analysis Program (BOA) v. 1.1.5, March 23 2005.
21. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS–a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
22. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**(11):1539–1558.
23. Rücker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC Medical Research Methodology* 2008; **8**:79.