Research Article

Statistics
in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

A Bayesian approach to sequential meta-analysis

Graeme T. Spence[a], David Steinsaltz[b] and Thomas R. Fanshawe[a]*

As evidence accumulates within a meta-analysis, it is desirable to determine when the results could be considered conclusive to guide systematic review updates and future trial designs. Adapting sequential testing methodology from clinical trials for application to pooled meta-analytic effect size estimates appears well suited for this objective. In this paper we describe a Bayesian sequential meta-analysis method, in which an informative heterogeneity prior is employed and stopping rule criteria are applied directly to the posterior distribution for the treatment effect parameter. Using simulation studies, we examine how well this approach performs under different parameter combinations by monitoring the proportion of sequential meta-analyses that reach incorrect conclusions (to yield error rates), the number of studies required to reach conclusion, and the resulting parameter estimates. By adjusting the stopping rule thresholds, the overall error rates can be controlled within the target levels and are no higher than those of alternative frequentist and semi-Bayes methods for the majority of the simulation scenarios. To illustrate the potential application of this method, we consider two contrasting meta-analyses using data from the Cochrane Library and compare the results of employing different sequential methods, while examining the effect of the heterogeneity prior in the proposed Bayesian approach.

Keywords: meta-analysis; sequential methods; cumulative meta-analysis; Bayesian inference; Markov chain Monte Carlo methods

1. Introduction

Sequential methods of data analysis in clinical trials are well established, and are the subject of several book-length treatments (e.g. [1, 2]). In contrast, the full potential of sequential methods applied to meta-analysis has not yet been realised, despite recent promising methodological developments [3, 4, 5, 6]. There is an increasing need for statistical meta-analytic methods that have the potential to save time in the systematic review process, as highlighted by the theme of the 2015 Cochrane Colloquium ('Filtering the information overload for better decisions'). Of the more than 6,700 systematic reviews in the Cochrane Library at the end of 2015 [7], a large number of these reviews are for treatments on which many studies have now been published and whose results might now be considered definitive. The time required to update and initiate Cochrane reviews is considerable, and so it is becoming increasingly important to find ways to prioritise reviews whose results are inconclusive. [8] While the accumulation of evidence and the evolution of the pooled

[a] *Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK*
[b] *Department of Statistics, University of Oxford, Oxford, UK*
* *Correspondence to: Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK. E-mail: thomas.fanshawe@phc.ox.ac.uk*

effect size can be visualised using cumulative meta-analysis [9], appropriate statistical methods are required to formalise this procedure and reduce incorrect conclusions from multiple hypothesis testing.

Sequential methods appear particularly suited to this objective, as they allow analyses to be updated as new evidence is accrued. In conducting an analysis of systematic review using sequential methods, the analyst is faced with a variety of possible conclusions: an intervention may show a clear-cut benefit or harm, there may be sufficient evidence that the differences interventions being compared are too small to be of practical importance (termed 'futility'), or the conclusion may be uncertain, suggesting that more studies are required. The analogy with stopping rules in clinical trials that incorporate interim analyses is clear.

In addition to some philosophical differences relating to applying stopping rules to meta-analyses rather than within primary studies, most notably the extent to which researchers are able to recommend 'stopping' [5], there are several methodological factors that need to be considered. A key difference, which we explore in this paper, is that in a trial, the maximum anticipated sample size is typically set in advance as part of the design, whereas in a systematic review the analyst will usually not know in advance how many further studies are likely to become available in the future. Whilst frequentist methods are often highly reliant on this information, for example error-spending approaches which ration total error over the course of a fixed experiment, Bayesian methods could be considered more flexible in that they do not rely on similar requirements.

A further important issue that affects many meta-analyses is that of between-study variation, or heterogeneity. Estimates of the heterogeneity variance in the random-effects model perform poorly when the number of clusters (studies, in the context considered in the current paper) is small; this result applies both in general [10] and in the context of meta-analysis [11, 12, 13]. If an estimate of zero heterogeneity is obtained, the confidence interval of the summary effect size estimate may be too narrow. This has led some researchers to use the Bayesian paradigm in fitting the random effects model, which has the potential advantage of using prior expectations about the heterogeneity to prevent a zero estimate of heterogeneity being obtained [6, 14, 15]. For example, Turner *et al.* derived informative heterogeneity prior distributions from an empirical study of Cochrane meta-analyses [6]. They categorised these priors by outcome and comparison types and used them to undertake Bayesian meta-analysis via both MCMC (Markov chain Monte Carlo) and non-MCMC methods.

The aim of this paper is to build on the existing frequentist and semi-Bayes methods that have been proposed to examine a 'fully Bayesian' approach to sequential meta-analysis for general application. In Section 2 we review existing methods, explain how they can be extended to obtain posterior distributions of all relevant parameters, and show how these distributions might be used in a sequential meta-analysis scheme. The simulation studies described in Sections 3 and 4 demonstrate the performance of this approach in comparison to the existing methods, notably the semi-Bayes method proposed by Higgins *et al.* [16], which was previously found to display more favourable properties than non-Bayesian alternatives. In Section 5 we show how the fully Bayesian approach might be used in practice by applying it to two meta-analyses from the Cochrane Database of Systematic Reviews (CDSR) – one investigating the effect of erythropoiesis-stimulating agents on the need for red blood cell transfusion in patients with cancer, the other investigating mortality following antioxidant supplementation – and we end with a concluding discussion (Section 6).

2. Sequential Meta-Analysis Methods

2.1. Frequentist and Bayesian Random Effects Meta-Analysis

In random effects meta-analysis, we suppose that the incorporated studies exhibit slight differences in their sample populations or study designs, leading to greater differences between the study-specific effect sizes than would be expected under the fixed effects model. This between-study variation is termed heterogeneity [16]. We assume that the observed treatment effect, y_i, of the i-th trial can be described within a two-level hierarchical model along with its 'true' study

effect, θ_i, and the overall parameter of interest, θ:

$$\begin{aligned} y_i &\overset{\text{iid}}{\sim} \text{N}(\theta_i, \sigma_i^2) \\ \theta_i &\overset{\text{iid}}{\sim} \text{N}(\theta, \tau^2) \end{aligned} \tag{1}$$

The σ_i^2 term is the within-study variance, which is usually estimated using the variance v_i obtained from the study results, and τ^2 corresponds to the heterogeneity variance parameter.

This normal-normal hierarchical model can be applied both to continuous effect measures (e.g. mean difference) and to binary measures using a logarithmic transformation (e.g. log odds ratio), and as such will be used in this study. However for binary data, an alternative binomial-normal hierarchical model can also be employed [17, 18].

For pooling results using the inverse variance method for frequentist meta-analysis, the treatment effects are appropriately weighted using the weights

$$w_i = \left(\sigma_i^2 + \tau^2\right)^{-1}$$

and θ is estimated as

$$\hat{\theta} = \frac{\sum_i w_i y_i}{\sum_i w_i},$$

where estimates of unknown parameters are inserted as required. These results can be used to construct confidence intervals for θ and to calculate the standardised test statistic, $Z = \hat{\theta}/\sqrt{\text{Var}(\hat{\theta})}$, which can be used for significance testing. For τ^2, the method of moments (DerSimonian-Laird) estimator [11] is often used:

$$\hat{\tau}^2_{DL} = \max\left(0, \frac{Q - k + 1}{\sum_i w_i - \sum_i w_i^2 / \sum_i w_i}\right),$$

where k is the number of studies and $Q = \sum_i w_i (y_i - \hat{\theta})^2$.

In Bayesian meta-analysis, prior distributions are placed on parameters θ and τ^2. The appropriate likelihood expressions derived from the above hierarchical model are then applied to the observed study effect sizes (y_i) and the estimated variances ($\hat{\sigma}_i^2$), with the true study effect sizes (θ_i) as latent variables. The posterior distributions of θ and τ^2 can then be sampled using MCMC or non-MCMC methods.

2.2. Frequentist Sequential Methods

Pogue and Yusuf proposed the application of Lan-DeMets monitoring boundaries to fixed effects meta-analysis ($\tau^2 = 0$) by assuming that the pooled results can be considered to be from a single large trial [19]. An overall sample size can be then calculated using control and treatment effect estimates to obtain the desired size for a 'conclusive' meta-analysis. Two-sided Lan-DeMets monitoring boundaries are subsequently calculated and used for the cumulative test statistic Z. Specifically, if the test statistic crosses a monitoring boundary an effect (benefit or harm) is concluded; if it crosses neither boundary before the overall sample size is attained, the null hypothesis of no effect is accepted.

The *Trial Sequential Analysis* (TSA) method from researchers at the Copenhagen Trial Unit adapts Lan-DeMets boundaries for use in a random effects meta-analysis model [4, 20]. They also consider the meta-analysis as a large single trial, but define a heterogeneity adjustment scale factor in terms of a 'diversity' measure, D^2. The diversity expresses the relative variance reduction when changing from a random effects to a fixed effects model, and if D^2 is greater than zero, the desired overall size is increased accordingly. This inflated sample size can then be used to calculate both modified inner and outer Lan-DeMets boundaries.

In the TSA method, a uniform adjustment factor is most often calculated retrospectively at the time of the latest sequential analysis using heterogeneity estimates from all of the available trials. However, this can be problematic when only a small number of studies have been published, as the variability estimates (especially that of τ^2_{DL}) can be

highly unstable. In these situations, *a priori* estimates could instead be used for these parameters. The heterogeneity adjusted sample sizes used in TSA may be used to inform the design of new trials, indicating the additional number of participants needed for a meta-analysis to be 'conclusive'. However, unlike in fixed effects models, the number of future trials undertaken also has to be considered for random effects meta-analysis – one large trial would provide much less information on the effect size of interest and its associated heterogeneity than many smaller studies [21].

Whitehead boundaries have also been used in sequential meta-analysis [3, 5, 22], and are applied to the score statistic, S_k, rather than Z_k:

$$S_k = Z_k\sqrt{\mathcal{I}_k} \quad \sim \mathrm{N}(\theta\mathcal{I}_k, \mathcal{I}_k),$$

where the information $\mathcal{I}_k$ is the inverse variance of $\hat{\theta}$ at the k-th analysis.

The related 'restricted' versions of the Whitehead boundaries are equivalent to O'Brien-Fleming monitoring boundaries [1], with the corresponding upper and lower boundaries ($\pm d_k$) given by

$$d_k = H - 0.583\sqrt{\mathcal{I}_k - \mathcal{I}_{k-1}},$$

where the value of the horizontal boundary H is determined numerically. The resulting so-called 'Christmas tree' correction, based on the exact information steps observed at the interim analyses, is described in detail elsewhere [23].

Both the score statistic and the information are cumulative, and are given by

$$\mathcal{I}_k = \sum_{i=1}^{k} \frac{1}{\sigma_i^2 + \tau^2} = \sum_{i=1}^{k} w_i,$$

$$S_k = \sum_{i=1}^{k} \frac{y_i}{\sigma_i^2 + \tau^2} = \sum_{i=1}^{k} w_i y_i,$$

again with estimators substituted for unknown parameter values. At any interim analysis, the pooled effect size ($\hat{\theta}$) is given by $S_k/\mathcal{I}_k$. Although no explicit sample size calculations are undertaken with this approach, the boundaries depend on the clinically relevant effect size, which must be stated in advance and yields a maximum information value, $\mathcal{I}_{max}$.

Increased information corresponds to increased precision of the pooled estimate $\hat{\theta}$, as the information at analysis k is the inverse variance of $\hat{\theta}$. Hence, cumulative information intrinsically accounts for heterogeneity in random effects meta-analysis. For example, it is possible for the precision of a meta-analysis to decrease when the addition of a new study causes the estimate of τ^2 to increase substantially, such as when the results of the new study are substantially different from those seen in previous studies. This corresponds to a reduction of information and a backward path on the cumulative plot, which complicates the calculation of monitoring boundaries. Higgins *et al.* applied the non-corrected Whitehead boundaries ($d_k = H$) at backward steps [5].

All of the above methods that assume the random effects model require τ^2 to be estimated sequentially using the study data available at the time of each interim analysis. However, the most widely-used heterogeneity estimator, the DerSimonian-Laird estimator τ^2_{DL}, which is recommended in the Cochrane Handbook [24], is unbiased only when the study-level variances are known, and otherwise has been shown empirically to demonstrate substantial bias when the number of contributing studies is small [25]. This drawback hinders the effective application of the frequentist methods described above.

2.3. Semi-Bayes Sequential Method

To address the issue of estimating heterogeneity effectively in sequential random effects meta-analysis, Higgins *et al.* employed restricted Whitehead monitoring boundaries with a 'semi-Bayes' procedure for updating the heterogeneity

while still estimating θ using a frequentist method [5].

Evidence on the heterogeneity (τ^2) is updated sequentially, using an informative inverse gamma prior distribution, $\mathcal{IG}(\eta, \lambda)$, and assuming an independent prior for θ and normal likelihood as in (1). Hence, the joint posterior density for θ and τ^2 at the k-th interim analysis is:

$$p(\theta, \tau^2 | y_{1:t_k}, \sigma^2_{1:t_k}) \propto p(\theta)(\tau^2)^{-\eta-1} \exp\left(-\frac{\lambda}{\tau^2}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{t_k} \frac{(y_i-\theta)^2}{\sigma_i^2+\tau^2}\right) \times \prod_{j=1}^{t_k} \sqrt{\frac{1}{2\pi(\sigma_j^2+\tau^2)}}.$$

The posterior mean of τ^2 is chosen as the summary statistic used in the frequentist score and information calculations described in Section 2.2. For the required numerical integration, θ is replaced by its estimate $\hat{\theta}_{k-1}$ from the $(k-1)$th analysis:

$$\tau^2_{B,k} = \frac{\int (\tau^2)^{-\eta} \exp\left(-\frac{\lambda}{\tau^2}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{t_k} \frac{(y_i-\hat{\theta}_{k-1})^2}{\sigma_i^2+\tau^2}\right) \times \prod_{j=1}^{t_k} \sqrt{\frac{1}{2\pi(\sigma_j^2+\tau^2)}}\, d\tau^2}{\int (\tau^2)^{-\eta-1} \exp\left(-\frac{\lambda}{\tau^2}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^{t_k} \frac{(y_i-\hat{\theta}_{k-1})^2}{\sigma_i^2+\tau^2}\right) \times \prod_{j=1}^{t_k} \sqrt{\frac{1}{2\pi(\sigma_j^2+\tau^2)}}\, d\tau^2},$$

$$\mathcal{I}_k = \sum_{i=1}^{k} \frac{1}{\sigma_i^2+\tau^2_{B,k}}, \qquad S_k = \sum_{i=1}^{k} \frac{y_i}{\sigma_i^2+\tau^2_{B,k}}.$$

This method outperformed the analogous frequentist sequential methods in simulation studies, most importantly resulting in lower Type I error rates. In addition, an approximation to the semi-Bayes procedure – avoiding the numerical integration – is available.

2.4. Fully Bayesian Sequential Method

Evidently, Higgins' semi-Bayes method can be extended into a fully Bayesian sequential meta-analysis approach – with priors placed on both the effect size (θ) and heterogeneity (τ^2) and the corresponding posterior distributions accessed directly for stopping decisions and parameter estimation. Bayesian inference in meta-analysis is well established [17, 26], but has had limited application using sequential methods. Most notably, in 2014 Chen *et al.* reported a Bayesian sequential meta-analytic design for survival regression models, with emphasis on sample size determination [27].

The Bayesian paradigm is intrinsically suited to sequential testing regimes, as the posterior distribution from each interim analysis can be used as the prior for the next, either directly or by using a power prior approach to control the influence of the previous data [28]. Conclusions may be drawn from the resulting posterior distributions which could potentially recommend stopping of the experiment. While Bayesian decision theory can be applied [29, 30, 31], a common approach in clinical trials is to pre-define stopping rules based purely on the posterior density of the effect size θ [32, 33, 34]. For example, stopping for effect can be recommended if a credible interval of given coverage excludes zero.

At each interim analysis, the posterior distribution summarises the current information and uncertainty in θ, independent of the subsequent decision to stop or continue. Furthermore, as the inference relates to posterior probabilities rather than significance levels, conclusions from Bayesian inference are not explicitly affected by multiple hypothesis testing. Difficulties with this approach arise when Bayesian sequential methods are judged using frequentist criteria; for example, repeated application of a stopping rule can lead to extremely inflated Type I errors. As frequentist methods remain extremely widely used in applied medical statistics, it is desirable to show that analogous Bayesian approaches can be adapted to preserve overall error rates. Hence, we will monitor these frequentist properties in this study to allow for direct comparison between the methods.

We propose the following Bayesian sequential meta-analysis scheme. At the first interim analysis, Bayesian inference is undertaken using priors for θ and τ^2 and all study results available at that point, given the notation $\mathcal{D}_1$. This gives rise to posterior distributions for θ and for τ^2, which are sampled using MCMC methods.

The resulting posterior distribution for θ is then subjected to stopping rules, illustrated in Figure 1. These rules required three quantities to be specified: two threshold values, ϵ_e and ϵ_f, representing 'effect' and 'futility' respectively, and a

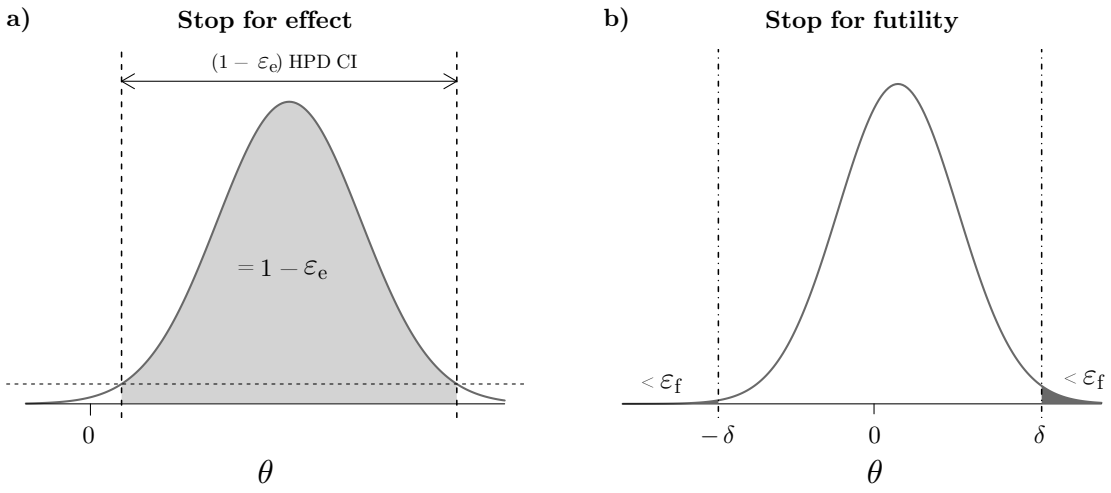

Figure 1. Stopping rules applied to the posterior distribution for the overall effect size during the Bayesian sequential meta-analysis. Specifically, a) stop for effect when the $(1 - \epsilon_e)$ highest posterior density credible interval (HPD CI) excludes zero, and b) stop for futility when $\mathbb{P}(\theta > \delta) < \epsilon_f$ and $\mathbb{P}(\theta < -\delta) < \epsilon_f$.

minimum clinically relevant effect size δ, such that:

- If the $(1 - \epsilon_e)$ highest posterior density (HPD) credible interval for θ excludes 0, the meta-analysis is stopped for effect, indicating either benefit or harm of the intervention under consideration, depending on the sign of the effect;
- If $\mathbb{P}(\theta > \delta \mid \mathcal{D}_1) < \epsilon_f$ and $\mathbb{P}(\theta < -\delta \mid \mathcal{D}_1) < \epsilon_f$, the meta-analysis is stopped for futility, indicating the probability that the intervention can achieve the minimum clinically relevant effect size is small.

While there are several possible ways to define such stopping rules, we chose to maintain logic close to the analogous frequentist hypothesis tests, for example two-sided tests for the rejection of the null hypothesis of no treatment effect. If neither effect nor futility is concluded, the next sequential analysis is conducted when further study results are available. Additional analyses might take place after each future study is published, or the analyst might wait until several new results are available, for example at the time of the next update of a systematic review. At that time, the likelihood is re-calculated using $\mathcal{D}_2$ (the updated available data) and the model is fitted as previously, using the same priors. This is equivalent to employing the previous posterior distribution as the next prior, but was easier to implement. The procedure is repeated until it recommends stopping or all data have been used. It is theoretically possible, if somewhat unlikely, that both effect and futility conditions could be met simultaneously, for example by an extremely localised posterior away from $\theta = 0$ but within $-\delta < \theta < \delta$ (Supplementary Figure 1). In this case, futility seems the most appropriate conclusion.

To preserve the frequentist error rates, implementations of Bayesian sequential methods in clinical trials have often employed sceptical ('handicap') prior distributions of varying weights for the treatment effect, whilst choosing stopping rule thresholds to match the overall target error levels in non-Bayesian methods (for example $\alpha = 0.05$) [33, 35, 36]. In Bayesian meta-analysis, however, it is more common to employ non-informative (or weakly informative) prior distributions for θ [6, 17, 37, 38]. Hence, we will look to control the relevant frequentist properties empirically, by adjusting the subjective stopping rule thresholds. Specifically, we investigate the effect of systematically adjusting the values of ϵ_e and ϵ_f on the observed error rates and time to conclusion in a Bayesian sequential meta-analysis by simulation, as described in Section 3.2. The values of thresholds explored in this study are chosen with the aim of maintaining overall Type I and Type II errors within the typical target levels of $\alpha = 0.05$ and $\beta = 0.1$.

In contrast to θ, an informative prior distribution will be employed for τ^2. This is a key feature of the Bayesian approach as prior information on the likely between-study heterogeneity can be incorporated into the analysis, which is especially important when a meta-analysis contains only a small number of studies.

3. Simulation Methods

3.1. Simulated meta-analysis data

We simulated data for meta-analysis by broadly following the procedure described by Higgins *et al* [5]. Three parameters controlled the generation of the data – the true effect size θ (set as either 0 or 0.5, in different simulations), the heterogeneity τ^2 (0, 0.0625 or 0.25), and the expected number of studies in a frequentist, fixed effects sequential meta-analysis concluding for no true effect, t (5, 10 or 20). Hence, there were 18 different simulation scenarios.

In each single meta-analysis of n studies, the true effect sizes θ_i from different studies (for $i = 1, ..., n$) were sampled independently from $\mathrm{N}(\theta, \tau^2)$, as in (1). The mean within-study variance, σ^2, was determined by t:

$$\sigma^2 = \frac{t}{44.32}.$$

The denominator value is the maximum information size ($\mathcal{I}_{max}$) from the Whitehead boundaries in the semi-Bayes approach (see Section 3.3). As t increases, each simulated trial tends to have lower precision, and so higher values of t correspond to a scenario with a larger number of studies, each with smaller implied sample sizes. For each simulation scenario, a large enough value of n was chosen such that a conclusion of either effect or futility was always reached in both the semi-Bayes and the fully Bayesian methods described below. Hence, in our simulations the final result of a simulated meta-analysis was never inconclusive.

To model variation in the size (and precision) of individual studies within a meta-analysis, we drew values for σ_i^2 according to:

$$\sigma_i^2 \sim \mathrm{U}_{[0.25\sigma^2, 1.75\sigma^2]}.$$

As the study variances are sampled from a uniform distribution, the distribution of study information (inverse variance, a measure of study size) is right skewed. Therefore more smaller studies are simulated than larger studies, in agreement with a previous investigation on the distribution of study sample sizes within the CDSR [39]. Finally, we sample observed study effect sizes y_i from $\mathrm{N}(\theta_i, \sigma_i^2)$.

For each combination of θ, τ^2 and t parameters, we simulated data from 5,000 meta-analyses. The simulations with $\theta = 0$ would allow for estimation of the Type I error, whilst those with $\theta = 0.5$ would estimate the power corresponding to the minimum effect size of interest, set as $\delta = 0.5$. The values of τ^2 and t explore contrasting situations, namely homogeneous data compared with those exhibiting moderate or substantial heterogeneity relative to the effect size, and meta-analyses containing a few large trials, intermediate sized trials, or many smaller trials.

3.2. Bayesian method

We applied the fully Bayesian sequential meta-analysis approach described in Section 2.4 to the simulated data. The Appendix contains the relevant R code. An inverse gamma prior, $\mathcal{IG}(\eta, \lambda)$, was placed on the heterogeneity, τ^2, with parameters $\eta = 1.5$ and $\lambda = 0.08$. As in [5], this prior was chosen to reflect a plausible level of heterogeneity for a typical meta-analysis, and has a mean of 0.16, a mode of 0.03, and 2.5th and 97.5th percentiles of 0.02 and 0.45 respectively. For θ, a relatively non-informative $\mathrm{N}(0, 5)$ prior was employed. This prior is extremely wide compared with the true θ values of 0 and 0.5 used in the simulations.

We assumed that sequential analyses could be undertaken after each trial, including the first, and sampled the posterior distributions of θ and τ^2 using the Gibbs Sampler MCMC scheme *JAGS* [40]. Initial values were drawn from $\theta_0 \sim \mathcal{U}_{(-2,2)}$ and $\tau_0^2 \sim \mathcal{U}_{(0,0.5)}$, and three MCMC chains of 10,000 iterations were then generated, each with a burn-in period of 100 steps and subsequent thinning by 10.

At each analysis, the stopping rules described above were applied to the posterior distribution of θ for different pairs of threshold values $\{\epsilon_e,\ \epsilon_f\} \in \{0.010, 0.008, 0.006, 0.004, 0.002\}$. These threshold values are much lower than the target

α and β values of 0.05 and 0.1 used in the Whitehead boundaries in the semi-Bayes method (see Section 3.3), as these thresholds relate to posterior probabilities at each individual interim analysis rather than overall Type I and Type II errors. Hence, to control the overall error rates within the target levels empirically, a range of strong stopping thresholds were chosen.

Within each simulated meta-analysis, we conducted this process sequentially until stopping for effect or futility had occurred. For every threshold combination, this was repeated for 5,000 meta analyses and the following measures recorded:

- The mean number of studies at stopping;
- The observed probability of stopping for effect, corresponding to the observed Type I error when $\theta = 0$.
- The observed probability of stopping for effect with $\theta > 0$ (termed 'benefit'), corresponding to the observed power when $\theta = 0.5$.

In each simulated meta-analysis, we recorded point estimates of the effect size ($\hat{\theta}$) and heterogeneity ($\hat{\tau}^2$) as the medians of the relevant posterior distributions at the stopping point. We then calculated the mean $\hat{\theta}$ and $\hat{\tau}^2$ at stopping for each set of 5,000 simulated meta-analyses. Additionally, we investigated posterior coverage by considering whether 95% HPD credible intervals at the final analysis contained the true values of θ and τ^2.

3.3. Semi-Bayes method

The semi-Bayes sequential meta-analysis approach was employed as in Higgins *et al.* [5]. We did not consider the approximate method presented in that paper, as we judged exploiting the intended Bayesian framework to be more important than reducing the computation time required for numerical integration, especially as in practice interim analyses would be undertaken one at a time. The following parameters determine $\mathcal{I}_{max}$ and H for the restricted Whitehead boundaries: the target overall Type I and Type II errors α (0.05) and β (0.1), and the minimum intervention effect for detection δ (0.5), yielding $\mathcal{I}_{max} = 44.32$ and $H = 14.92$. The last analysis undertaken in this method corresponds to the first time that $\mathcal{I}_k > \mathcal{I}_{max}$. Higgins *et al.* decided to still apply the monitoring boundaries at this analysis (even though this is a region in which the sequential theory does not formally apply), and the calculation of Whitehead boundaries in this situation is straightforward, as the formulation described can be applied to $\mathcal{I}_k > \mathcal{I}_{max}$.

4. Simulation Results

4.1. MCMC diagnostics

Figure 2 and Supplementary Figures 2 and 3 show diagnostic plots for representative runs of the MCMC routine. While both trace plots (θ and τ^2, Supplementary Figure 2) indicate good convergence, the traces for τ^2 possess more frequent extreme values (for example $\tau^2 > 1$), indicating the typical positive skew in the pesterior distribution of this parameter. The corresponding density plots (Figure 2) show good agreement between parallel MCMC chains and well-defined posterior distributions. Moreover, autocorrelation plots (Supplementary Figure 3) display no significant correlations, and consequently the effective sample sizes for θ and τ^2 were close to 3,000 (three chains of 1,000 samples).

As it was not viable to consider these diagnostic plots for each run of the full simulation procedure of 5,000 sequential meta-analyses, we instead recorded the Brooks-Gelman-Rubin multivariate potential scale reduction factor for each MCMC run. For this parameter, values less than 1.2 are considered satisfactory indicators of convergence [41], and this condition was met in all of the simulations.

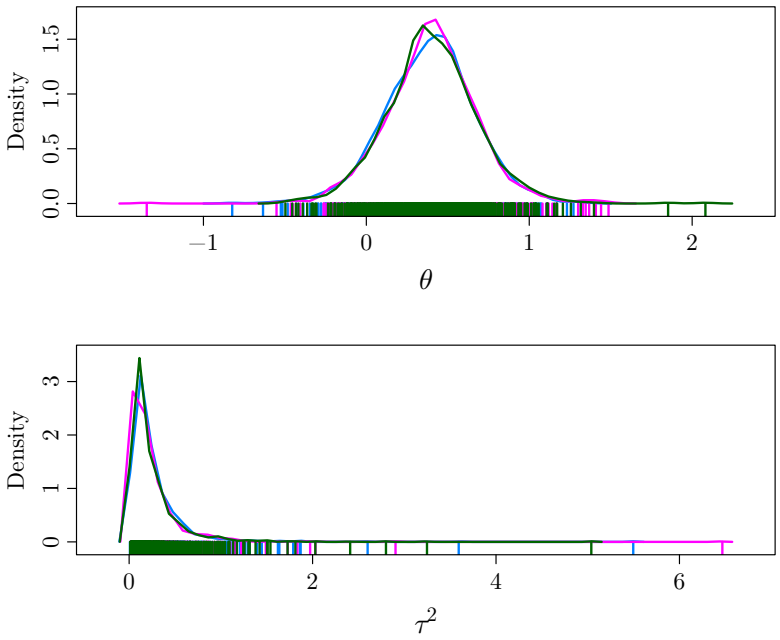

Figure 2. Representative posterior densities for θ and τ^2 (from an analysis with simulation parameters $t = 5, \tau^2 = 0.25, \theta = 0.5$ and $n = 5$, and prior distributions $\theta \sim \mathcal{IG}(1.5, 0.08)$ and $\tau^2 \sim \text{N}(0, 5)$). Different colours represent different MCMC chains.

4.2. Effect of posterior credible thresholds

Figures 3–5 show the results of the Bayesian sequential meta-analysis simulations. The graphs represent the results from applying each pair of threshold values to the 18 simulation scenarios, with the axes corresponding to the mean number of studies required when $\theta = 0$ (x-axis) and $\theta = 0.5$ (y-axis), and the size of the crosses depicts the observed Type I error (horizontal) or Type II error (vertical). Supplementary Table 1 shows the interquartile range and median number of studies required under different parameter combinations for a representative pair of threshold values.

The figures show that strengthening the stopping criteria by reducing the threshold values increases the number of studies required before a conclusion is reached. Specifically, reducing ϵ_e increases the average number of studies required before stopping, substantially so when there is a true effect and to a lesser extent when no effect exists. Similarly, reducing ϵ_f increases the average number of studies required before stopping when there is no effect, with only a slight increase when a true effect exists.

In general, strengthening both stopping criteria reduces the observed Type I and Type II errors (for more detailed graphs displaying the observed errors, see Supplementary Figures 4–6). Reducing ϵ_e while holding ϵ_f constant disfavours early stopping for effect, resulting in a considerable reduction in Type I error. Minor increases in Type II error also occur, as lower values of ϵ_e mean that there are more opportunities for the sequential scheme to terminate incorrectly for futility. Conversely, any futility conclusions are delayed by reducing ϵ_f, and so the Type II error is lowered, with minor increases in the Type I error.

4.3. Comparison with semi-Bayes method and target error levels

Figures 3–5 also depict the results from simulations employing the semi-Bayes method (blue crosses). Scenarios in which the fully Bayesian approach yielded greater errors than the semi-Bayes method are marked on the graphs (red lines).

When $\theta = 0$, the mean number of studies required in the semi-Bayes simulations lies within the observed range from the different fully Bayesian thresholds – i.e. some threshold pairs resulted in the fully Bayesian method taking longer than

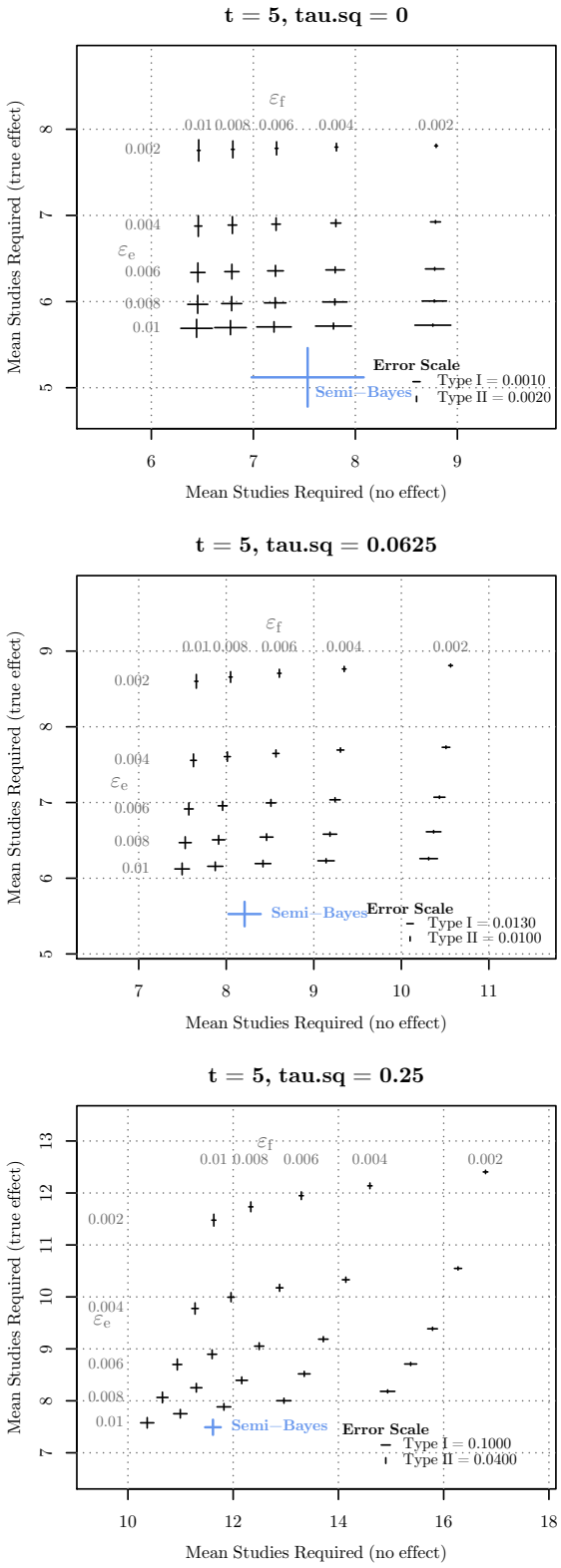

Figure 3. Summary graphs for the Bayesian sequential meta-analysis simulations with $t = 5$. The plotted values correspond to the mean studies required when $\theta = 0$ (x-axis) and $\theta = 0.5$ (y-axis) for different pairs of posterior credible threshold values (ϵ_e and ϵ_f). The length of the crosses represent the observed Type I error (horizontal) and Type II error (vertical). The blue cross is the corresponding result using the semi-Bayes method, and any red lines highlight the situations in which the Bayesian simulations yielded errors larger than those from the semi-Bayes.

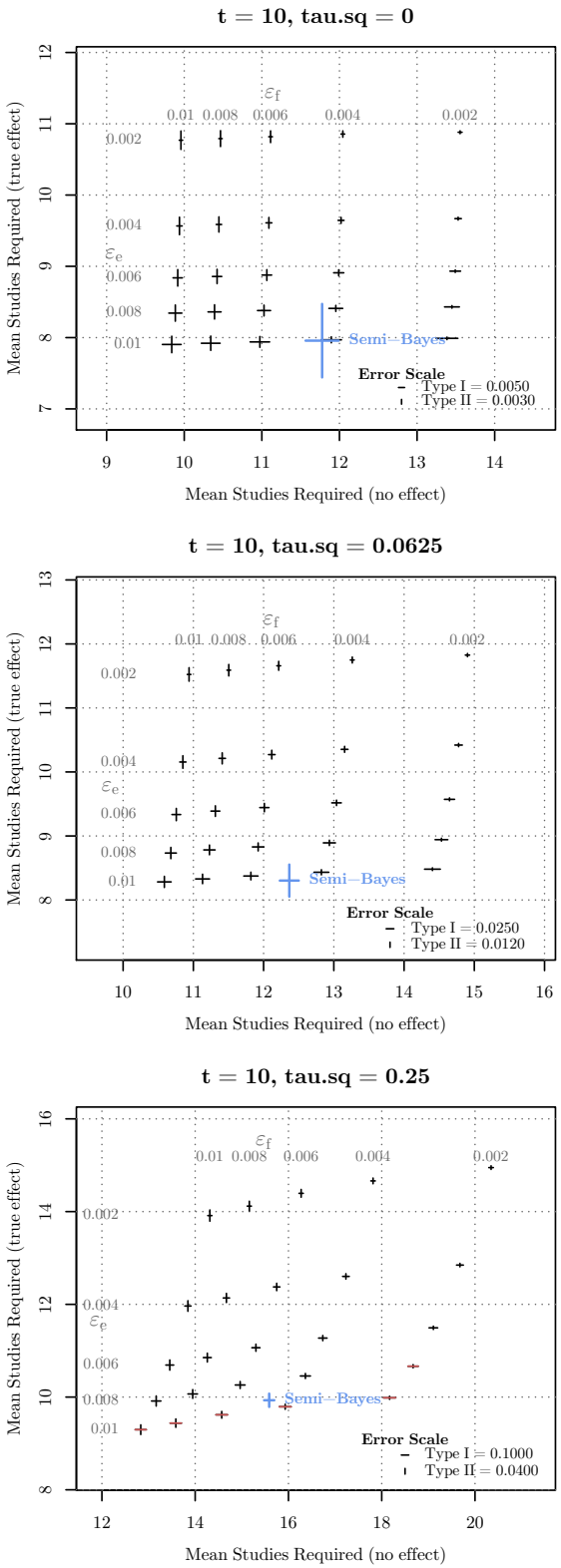

Figure 4. Summary graphs for the Bayesian sequential meta-analysis simulations with $t = 10$. The plotted values correspond to the mean studies required when $\theta = 0$ (x-axis) and $\theta = 0.5$ (y-axis) for different pairs of posterior credible threshold values (ϵ_e and ϵ_f). The length of the crosses represent the observed Type I error (horizontal) and Type II error (vertical). The blue cross is the corresponding result using the semi-Bayes method, and any red lines highlight the situations in which the Bayesian simulations yielded errors larger than those from the semi-Bayes. The dashed line on the middle graph represents an average number of studies across $\theta = 0$ and 0.5 equal to that from the semi-Bayes simulations.

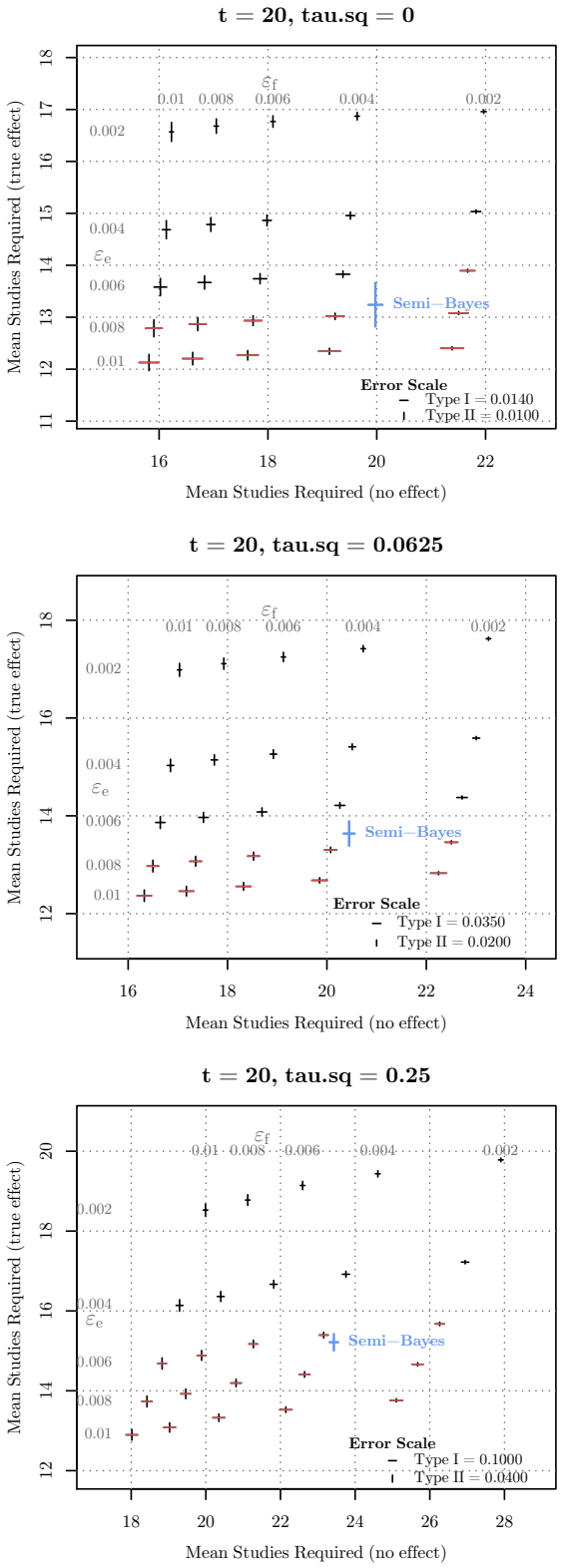

Figure 5. Summary graphs for the Bayesian sequential meta-analysis simulations with $t = 20$. The plotted values correspond to the mean studies required when $\theta = 0$ (x-axis) and $\theta = 0.5$ (y-axis) for different pairs of posterior credible threshold values (ϵ_e and ϵ_f). The length of the crosses represent the observed Type I error (horizontal) and Type II error (vertical). The blue cross is the corresponding result using the semi-Bayes method, and any red lines highlight the situations in which the Bayesian simulations yielded errors larger than those from the semi-Bayes.

the semi-Bayes approach, others resulted in faster conclusions. For $\theta = 0.5$, the semi-Bayes method took on average fewer studies than all of the Bayesian thresholds when $t = 5$, approximately equal numbers of studies compared to the weakest effect threshold (ϵ_e = 0.010) when $t = 10$, and similar numbers to the intermediate effect thresholds when $t = 20$.

Importantly, the fully Bayesian method exhibits smaller Type II errors compared with the semi-Bayes method across all threshold pairs and considerably smaller Type I errors in most cases. The observed Type I error rates are lower for all simulation scenarios when $t = 5$, and in the two scenarios with no or moderate heterogeneity when $t = 10$ ($\tau^2 = 0$ and 0.0625). When $t = 10$ and heterogeneity are high ($\tau^2 = 0.25$), six threshold pairs led to higher Type I errors than in the semi-Bayes simulations. When $t = 20$, the fully Bayesian simulations resulted in higher errors than the semi-Bayes for 10–15 threshold pairs at each of the different heterogeneity levels.

The observed error rates are compared with the target Type I error of 0.05 and Type II error of 0.1 in Supplementary Figures 4–6. In each meta-analysis scenario, all of the threshold pairs yielded Type II errors within the target level. For the Type I error, the observed error rates were all within the target level for the simulations with no heterogeneity (for all t values), and with moderate heterogeneity ($\tau^2 = 0.0625$) when $t = 5$. Whilst the same still holds for the majority of thresholds when $\tau^2 = 0.0625$ and t =10 or 20, Type I errors of greater than 0.05 were observed for almost all of threshold pairs when extreme heterogeneity ($\tau^2 = 0.25$) was simulated. It is unsurprising that extreme heterogeneity and a large number of interim analyses increases the probability of false negative conclusions.

4.4. Parameter estimation

In the simulations with the most realistic level of heterogeneity ($\tau^2 = 0.0625$) and intermediate sized trials ($t = 10$), six threshold pairs have a lower average number of required studies than the semi-Bayes approach, considering both $\theta = 0$ and 0.5 (see dashed line in Figure 4). Of those, $\epsilon_e = 0.006, \epsilon_f = 0.010$ have the lowest total error rate for Type I and II combined. These threshold values appear to be a reasonable choice for examining the performance of the fully Bayesian method in more detail because they yielded lower errors than the semi-Bayes method in all but one of the 18 simulation scenarios (the extreme case of $t = 20, \tau^2 = 0.25$). In real situations, the choice of a single pair of threshold values would more greatly depend on the desired balance between the length of time taken to draw a conclusion and the chance of an incorrect conclusion (either false positive or false negative) being reached.

Tables 1 and 2 show results from simulations with this pair of threshold values. The parameters were estimated within each simulated meta-analysis using the posterior medians, and coverage refers to the observed probability that the true parameter value lies within the 95% HPD credible interval at the final analysis. The probability of benefit, $\mathbb{P}(\theta > 0|\text{data})$, corresponds to approximately half the observed Type I error in Table 1 ($\theta = 0$) and is equal to the observed power in Table 2 ($\theta = 0$).

Across the simulations, increased heterogeneity does lead to increased numbers of studies required, increased observed errors and reduced θ coverage. Posterior coverage for θ was between 83.4 and 99.2%, with the lowest coverage occurring when the meta-analysis consists of a small number of studies possessing extremely high heterogeneity. Posterior coverage for the heterogeneity cannot occur when $\tau^2 = 0$ as this parameter value has zero support under the inverse gamma prior, but coverage is close to 100% when $\tau^2 = 0.0625$ and 73–77% when $\tau^2 = 0.25$.

For $\theta = 0.5$ (Table 2), the mean effect size estimate is greater than this true value, as early stopping for effect is more likely when extreme results are observed (i.e. $\theta > 0.5$ compared with $\theta < 0.5$). For the meta-analyses that require more studies, the resulting estimates most often lie closer to the true value (see Figure 7 in Supplementary Material). Overall, the combined results from these situations yield inflated effect estimates. The difference between θ and $\hat{\theta}$ increases with τ^2, as increased heterogeneity increases the chance of observing more extreme early results.

For the simulations with zero or mild heterogeneity ($\tau^2 = 0$ and 0.0625), the observed Type I and Type II error rates increase with t but all lie within the target error levels ($\alpha = 0.05$ and $\beta = 0.1$). For example, the probability of benefit increases from 0.010 ($t = 5$) to 0.014 ($t = 10$) to 0.020 ($t = 20$) when $\theta = 0$ and $\tau^2 = 0.0625$ (Table 1), compared with the target $\alpha/2 = 0.025$. The number of interim analyses undertaken increases with t, thus increasing the potential for false

Table 1. Bayesian simulation results for $\theta = 0$ with $\epsilon_e = 0.006$ and $\epsilon_f = 0.01$. The posterior medians are used as estimates for θ and τ^2, coverage relates to the 95% highest posterior density credible intervals, and the probability of benefit ($\mathbb{P}(\theta > 0|\text{data})$) corresponds to approximately half the observed Type I error.

$t = 5, \quad \theta = 0$			
	$\tau^2 = 0$	$\tau^2 = 0.0625$	$\tau^2 = 0.25$
Studies at stopping	6.5	7.6	10.8
Mean $\hat{\theta}$ at stopping	0.001	−0.002	−0.005
Mean $\hat{\tau}^2$ at stopping	0.051	0.065	0.148
θ coverage	0.992	0.965	0.884
τ^2 coverage	0.000	0.999	0.736
$\mathbb{P}(\theta > 0\|\text{data})$	0.001	0.010	0.048

$t = 10, \quad \theta = 0$			
	$\tau^2 = 0$	$\tau^2 = 0.0625$	$\tau^2 = 0.25$
Studies at stopping	9.8	10.7	13.4
Mean $\hat{\theta}$ at stopping	−0.000	0.001	−0.001
Mean $\hat{\tau}^2$ at stopping	0.055	0.066	0.135
θ coverage	0.986	0.962	0.894
τ^2 coverage	0.000	1.000	0.740
$\mathbb{P}(\theta > 0\|\text{data})$	0.003	0.014	0.047

$t = 20, \quad \theta = 0$			
	$\tau^2 = 0$	$\tau^2 = 0.0625$	$\tau^2 = 0.25$
Studies at stopping	16.0	16.7	19.1
Mean $\hat{\theta}$ at stopping	0.001	−0.003	−0.000
Mean $\hat{\tau}^2$ at stopping	0.058	0.067	0.120
θ coverage	0.974	0.948	0.897
τ^2 coverage	0.000	1.000	0.732
$\mathbb{P}(\theta > 0\|\text{data})$	0.011	0.020	0.049

conclusions and increased error rates. Although the observed Type I error is considerably above the target level in the extreme case of $\tau^2 = 0.25$ (when the heterogeneity is of the same order as the mean within-study variance, σ^2), the power still exceeds the desired value (0.9) and both errors are stable across the different values of t.

5. Application

We illustrate the Bayesian sequential meta-analysis method using two systematic reviews updated in the CDSR in 2012, with review numbers CD003407 [42] and CD007176 [43]. Within each review, we selected the meta-analysis of the binary outcome containing the most contributing studies. Table 3 shows details of the meta-analyses selected.

The meta-analysis from review CD003407 investigates the effect of erythropoiesis-stimulating agents (ESAs) in reducing the need for red blood cell transfusion in patients with cancer who require treatment to prevent or control anaemia, and has 88 contributing studies. The review concludes a clear benefit of using ESAs compared to a control

Table 2. Bayesian simulation results for $\theta = 0.5$ with $\epsilon_e = 0.006$ and $\epsilon_f = 0.01$. The posterior medians are used as estimates for θ and τ^2, coverage relates to the 95% highest posterior density credible intervals, and the probability of benefit ($\mathbb{P}(\theta > 0|\text{data})$) corresponds to the observed power.

$\boldsymbol{t = 5, \quad \theta = 0.5}$

	$\boldsymbol{\tau^2 = 0}$	$\boldsymbol{\tau^2 = 0.0625}$	$\boldsymbol{\tau^2 = 0.25}$
Studies at stopping	6.4	6.8	8.9
Mean $\hat{\theta}$ at stopping	0.539	0.567	0.630
Mean $\hat{\tau}^2$ at stopping	0.051	0.063	0.129
θ coverage	0.989	0.950	0.834
τ^2 coverage	0.000	1.000	0.760
$\mathbb{P}(\theta > 0|\text{data})$	0.995	0.976	0.927

$\boldsymbol{t = 10, \quad \theta = 0.5}$

	$\boldsymbol{\tau^2 = 0}$	$\boldsymbol{\tau^2 = 0.0625}$	$\boldsymbol{\tau^2 = 0.25}$
Studies at stopping	8.9	9.3	10.6
Mean $\hat{\theta}$ at stopping	0.566	0.590	0.649
Mean $\hat{\tau}^2$ at stopping	0.055	0.065	0.116
θ coverage	0.971	0.946	0.848
τ^2 coverage	0.000	1.000	0.760
$\mathbb{P}(\theta > 0|\text{data})$	0.985	0.974	0.927

$\boldsymbol{t = 20, \quad \theta = 0.5}$

	$\boldsymbol{\tau^2 = 0}$	$\boldsymbol{\tau^2 = 0.0625}$	$\boldsymbol{\tau^2 = 0.25}$
Studies at stopping	13.5	13.8	14.5
Mean $\hat{\theta}$ at stopping	0.601	0.610	0.675
Mean $\hat{\tau}^2$ at stopping	0.059	0.065	0.104
θ coverage	0.953	0.935	0.856
τ^2 coverage	0.000	1.000	0.773
$\mathbb{P}(\theta > 0|\text{data})$	0.975	0.965	0.932

treatment for this outcome variable, with a random effects meta-analysis of these studies leading to a pooled odds ratio of 0.44 (or risk ratio of 0.65) and 95% confidence interval excluding unity.

Review CD007176 describes the effect of antioxidant supplementation on all-cause mortality, compared to placebo, during follow-up (mean 3.4 years in the contributing trials). We focus on the subgroup of 56 trials classified as having low risk of bias. In this subgroup, a random-effects meta-analysis produces a pooled odds ratio estimate of 1.04 (or risk ratio of 1.04). The lower limit of the 95% confidence interval is 1.01, indicating a statistically significant result with the intervention associated with increased mortality risk (the opposite direction to the one expected), although the effect size appears small clinically.

The large number of trials contributing to each of these meta-analyses is an advantage when investigating the effect of using a sequential meta-analysis scheme. For the purposes of the current paper, we assume that the use of the sequential meta-analysis methods after each available study had been pre-planned and that the data are analysed using a random effects model for the log-odds ratio. Given these assumptions, we show how the conclusions of these reviews may have evolved had our proposed sequential meta-analysis methods been adopted.

We compare three sequential methods: a frequentist method using the DerSimonian-Laird heterogeneity estimator

Table 3. Details of Cochrane reviews chosen to illustrate the sequential meta-analysis methods.

	CD003407
Review title	Erythropoietin or darbepoetin for patients with cancer
Review group	Haematological Malignancies
Comparison	Erythropoietin or darbepoetin vs. control
Outcome	Need for red blood cell transfusion

	CD007176
Review title	Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases
Review group	Hepato-Biliary
Comparison	Antioxidants vs. placebo
Outcome	Mortality
Subgroup	Trials with low risk of bias

in the construction of restricted Whitehead boundaries, the semi-Bayes method described in Section 2.3, and the fully Bayesian approach from Section 2.4. In the fully Bayesian approach, we used 10^6 iterations for posterior sampling in the MCMC scheme and the same R code given in the Appendix (a full code file is provided online). For an initial heterogeneity prior in the semi-Bayes and fully Bayesian methods, the same $\mathcal{IG}(1.5, 0.08)$ distribution was employed as in the simulations (Section 4). In the Bayesian method we also compared the results from using two further inverse gamma priors – $\mathcal{IG}(0.001, 0.001)$ and $\mathcal{IG}(1.5, 1)$. The former is a commonly used 'vague' prior, and latter could be more appropriate if extreme heterogeneity were suspected as it possesses a mode of 0.4. Furthermore, we selected appropriate prior distributions from Turner's empirical study [6]: for the meta-analysis from review CD003407 – a pharmacological vs. control comparison, with a "resource use / hospital process" outcome – this was a log-normal distribution with a log-mean of -2.34 and a log-variance of 1.74^2; and for CD007176 – pharmacological vs. control, and all-cause mortality – a $\mathcal{LN}(-3.95, 1.34^2)$. Both of these log-normal distributions have considerable prior density at lower heterogeneity values than $\mathcal{IG}(1.5, 0.08)$ (Supplementary Figure 8).

For both meta-analyses, we set the clinically relevant beneficial effect size (δ) to be a log-odds ratio of -0.5, equivalent to an odds ratio of 0.61, given that negative log-odds ratios favour the treatment arm in both meta-analyses considered. In practice, suitable δ values would be decided for each individual meta-analysis by clinical experts. We set target overall error levels for the semi-Bayes method at $\alpha = 0.05$ and $\beta = 0.1$, and used the stopping rule thresholds in the Bayesian method of $\epsilon_e = 0.006$ and $\epsilon_f = 0.01$.

Table 4 and Figure 6 show the results from these analyses. All of the sequential methods recommended stopping before the end of the meta-analysis and agree with each other in their final conclusions – in favour of the treatment in CD003407, and for futility in CD007176. In CD003407, the frequentist method concluded first (after the seventh study), while the stopping points for the semi-Bayes and fully Bayesian method occur later as all of the heterogeneity priors allow greater potential for between-study variation. Whilst later stopping could be seen as disadvantageous, the simulation studies here and in reference [5] indicate that these methods would be expected to have lower error rates when applied to many meta-analyses. The later stopping is particularly pronounced for the more disparate prior ($\mathcal{IG}(1.5, 1)$): fifteen studies were required in CD003407 for the fully Bayesian method with this prior. The other heterogeneity priors employed with the Bayesian method all resulted in stopping being recommended after nine studies, as did the semi-Bayes method.

For CD007176, both the frequentist method and the fully Bayesian approach using the log-normal prior recommended stopping after three studies. The use of the inverse gamma prior distributions with the Bayesian method delayed conclusion – by a single study when employing $\mathcal{IG}(1.5, 0.08)$, a further study with $\mathcal{IG}(0.001, 0.001)$, and again more substantially

Table 4. Results from applying sequential meta-analysis methods to the selected meta-analyses from CD003407 and CD007176. $\hat{\theta}$ values are maximum likelihood estimates for the frequentist and semi-Bayes methods and posterior medians for the fully Bayesian method. $\hat{\tau}^2$ values are DerSimonian-Laird estimates for the frequentist method, posterior means for the semi-Bayes method and posterior medians for the fully Bayesian method.

CD003407

	τ^2 **prior**	**Result**	**Studies**	$\hat{\theta}$	$\hat{\tau}^2$
Frequentist	-	Benefit	7	-0.575	0.033
Semi-Bayes	$\mathcal{IG}(1.5, 0.08)$	Benefit	9	-0.603	0.072
Fully Bayes	$\mathcal{IG}(1.5, 0.08)$	Benefit	9	-0.598	0.056
Fully Bayes	$\mathcal{IG}(0.001, 0.001)$	Benefit	9	-0.591	0.020
Fully Bayes	$\mathcal{IG}(1.5, 1)$	Benefit	15	-0.665	0.326
Fully Bayes	$\mathcal{LN}(-2.34, 1.74^2)$	Benefit	9	-0.595	0.048

CD007176

	τ^2 **prior**	**Result**	**Studies**	$\hat{\theta}$	$\hat{\tau}^2$
Frequentist	-	Futility	3	-0.011	0.000
Semi-Bayes	$\mathcal{IG}(1.5, 0.08)$	Futility	6	-0.014	0.046
Fully Bayes	$\mathcal{IG}(1.5, 0.08)$	Futility	4	-0.038	0.049
Fully Bayes	$\mathcal{IG}(0.001, 0.001)$	Futility	5	-0.029	0.012
Fully Bayes	$\mathcal{IG}(1.5, 1)$	Futility	11	-0.019	0.253
Fully Bayes	$\mathcal{LN}(-3.95, 1.34^2)$	Futility	3	0.008	0.014

with $\mathcal{IG}(1.5, 1)$ (concluding after a total of 11 studies). The semi-Bayes method stopped two studies later than the corresponding Bayesian method with the same prior (after six and four studies respectively). Interestingly, the authors of CD007176 applied Trial Sequential Analysis (TSA) to this meta-analysis within their review, concluding at a much later point (44 studies) for harm, rather than futility, albeit having used the equivalent of a smaller δ. The standard pooled odds ratio estimate after this number of studies is 1.06 (a log odds ratio of 0.058). Reducing the value of δ in the fully Bayesian approach did delay conclusion, however still yielding a conclusion of futility – for example, for $\delta = 0.1$ (an odds ratio of 1.11) stopping was recommended after 53 studies again for no effect.

Effect size estimates ($\hat{\theta}$) are shown in Table 4 – maximum likelihood point estimates at stopping for the frequentist and semi-Bayes methods, and the posterior median in the fully Bayesian method. In both meta-analyses, these estimates are similar in magnitude for all methods given the variation in stopping points, although with the estimates from CD007176 fluctuating above or below zero (see Figure 6, bottom). However, the heterogeneity estimates exhibit greater variation. For this parameter, the frequentist method uses the DerSimonian-Laird estimator, while the semi-Bayes method employs the posterior mean of τ^2 (calculated by integration) as a point estimate in the monitoring boundary calculations. In contrast, an approximation to the entire posterior distribution is sampled in the fully Bayesian method and the posterior median is presented in the table. Bayesian heterogeneity estimates are known to be sensitive to the choice of prior distribution, and as expected, the fully Bayesian $\hat{\tau}^2$ values are substantially increased when utilising the $\mathcal{IG}(1.5, 1)$ prior compared with the other inverse gamma or the relevant log-normal distributions.

6. Discussion

This paper outlines a Bayesian approach to sequential random effects meta-analysis which incorporates prior information about the level of heterogeneity between studies and applies stopping rules directly to the posterior distribution of the effect size. This contrasts with previously reported sequential meta-analysis methods, which used frequentist and semi-Bayesian

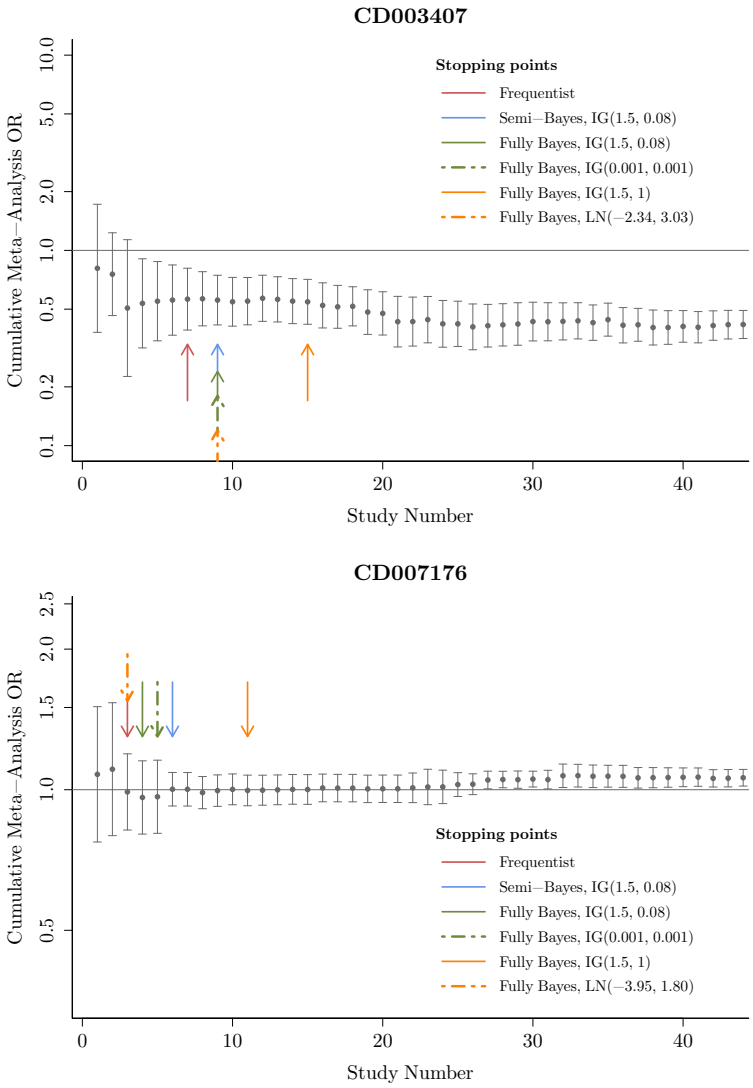

Figure 6. Cumulative meta-analysis plots for the selected meta-analyses from CD003407 and CD007176. Arrows depict stopping points using different sequential meta-analysis methods. Error bars correspond to 95% confidence intervals for pooled effect sizes, calculated using a random effects meta-analysis model. Note that the x-axes are truncated as both meta-analyses contain more than 50 studies.

approaches [3, 4, 5, 20]. While frequentist methods are limited by difficulties in estimating heterogeneity reliably at early stages within a meta-analysis, Higgins's semi-Bayes method applies Bayesian inference and places an informative prior on the heterogeneity parameter. In employing only a point estimate for τ^2 in the subsequent frequentist monitoring boundary calculations, however, the semi-Bayes method does not make allowance for the evident uncertainty in this estimate.

A fully Bayesian approach places prior distributions on both the heterogeneity and effect size parameters, and samples both posterior distributions accordingly. Stopping decisions, or recommendations, are based around posterior probabilities calculated directly from the effect size posterior distribution. At any step in which the additional data induce an increased estimate for the heterogeneity present, the effect size posterior can widen, discouraging stopping. This favourable behaviour causes no additional difficulties in the Bayesian sequential monitoring scheme, unlike other methods in which adaptations may be necessary to take into account 'backward' information steps.

In general, Bayesian inference is highly suited to sequential updating, with the previous posterior distributions playing the role of priors for the next analysis. Concerns with this approach often arise from the lack of direct control over the

overall Type I and Type II error rates, which is more traditionally offered by standard frequentist methods. Our simulation results demonstrate, however, that by strengthening the stopping rule thresholds, overall error rates comparable to the target levels are obtained. Additionally, MCMC methods are necessary to access the desired posterior distributions. Whilst this requires extra computation, the increase in time is not a significant disadvantage when undertaking an individual analysis, and to obtain the results in our two applications took about a minute of run time each for all of the interim analyses.

We made a number of standard assumptions. The effect size estimates were assumed to be normally distributed according to the random effects model, with independent and identical sampling. It was assumed that this model adequately describes variation within and between studies, and that the within-study variances can be reliably estimated from the observed standard errors. The normality assumption determines the likelihood components within the Bayesian inference, in which the effect sizes and variances were assumed to be uncorrelated, and the prior distributions for θ and τ^2 were assumed independent.

Our simulation studies investigated the effect of different stopping rule thresholds and their performance compared to the semi-Bayes method. When there was a true non-zero effect in the simulated data, the fully Bayesian method generally took longer to conclude than the semi-Bayes, but always with lower observed Type II errors. For meta-analyses with no true effect, appropriately chosen thresholds for the fully Bayesian method led to earlier stopping and almost always had lower Type I errors. The Type I errors from the fully Bayesian approach became closer to those from the semi-Bayes method as smaller studies were included (and correspondingly more analyses undertaken) in a sequential meta-analysis. Overall, our conclusion is that the fully Bayesian method exhibits both lower Type I and Type II errors in most of the simulation scenarios, with similar numbers of studies required before termination occurs. The Bayesian treatment of both θ and τ^2, therefore, appears promising in balancing the need to conclude for effect or futility, and adding further studies to the meta-analysis.

Further development of this approach is required, specifically relating to the choices of stopping rule thresholds and prior distributions. The choice of threshold values determine the stopping points for effect or futility, and would be influenced by an investigator's target overall error levels, anticipated heterogeneity, and their relative weighting of false positives, false negatives and speed of conclusion. In practice these considerations are likely to vary in importance, depending on the disease area and the nature of the interventions. From our simulation studies, it appears that the threshold values of $\epsilon_e = 0.006$ and $\epsilon_f = 0.01$ would perform favourably in comparison to overall error targets of $\alpha = 0.05$ and $\beta = 0.1$ when moderate or no heterogeneity is anticipated. If extreme heterogeneity is suspected, further strengthening of ϵ_e would be recommended. Similar factors determine the choice of a minimum clinically relevant effect size, although one could envisage exploring whether different conclusions are drawn for a range of δ values. We chose to employ a relatively high-variance prior for the effect size parameter, and more informative heterogeneity priors in an attempt to improve the estimates of τ^2 during the early stages of the sequential meta-analysis. We recommend the use of empirically-derived heterogeneity priors, such as those developed by Turner *et al.* [6], to further refine the proposed method, as these provide an objective way to inform the Bayesian inference based on previous, relevant meta-analysis.

In conclusion, we have investigated using a fully Bayesian method for sequential meta-analysis method using simulation studies and data from two applications from the Cochrane Database of Systematic Reviews. The results demonstrates the potential of the Bayesian approach. We propose that this approach can be implemented using the concise code provided in the Appendix, and would benefit from further work as a method for sequential meta-analysis for wider application.

Appendix: R-code

```
# R function to perform a fully Bayesian sequential meta-analysis
# Priors: N(0, var) for theta, IG(eta, lambda) for tau.sq
# Input:
#   y.i, sigma.sq.i = effect estimate and variance from each trial
#                     (e.g. on log odds ratio scale)
```

```
#   delta = minimally relevant clinical effect size (e.g. logs odds ratio)
#   epsilon.e, epsilon.f = posterior decision thresholds
#   var, eta, lambda = prior parameters
#   iterations, chains, burn, thin = MCMC settings
# Requires JAGS download: http://mcmc-jags.sourceforge.net
#   and 'rjags' R package: https://cran.r-project.org/web/packages/rjags/

### Define model and priors ### ================================================
write("
      model {
      for (i in 1:n) { y.i[i] ~ dnorm(theta.i[i], 1 / sigma.sq.i[i])
      theta.i[i] ~ dnorm(theta, inv.tsq)}  # JAGS uses precision not variance
      theta ~ dnorm(mu, prec)
      inv.tsq ~ dgamma(eta, lambda)  # Appropriate prior for precision
      tsq <- 1 / inv.tsq
      }
      ", "BayesMA.bug")

### Bayesian sequential meta-analysis function ### ============================
Bayesian.seqMA <- function(y.i, sigma.sq.i, delta, epsilon.e=0.006,
                        epsilon.f=0.010, var=5, eta=1.5, lambda=0.08,
                        iterations=1E5, chains=3, burn=100, thin=10){

  i <- 1; result <- "Uncertain"; diag <- c()

  while(result == "Uncertain" & i <= length(y.i)) {
    Y <- y.i[1:i]
    v <- sigma.sq.i[1:i]

    # Set-up and run simulation
    BayesMA.jags <- jags.model("BayesMA.bug", n.chains=chains, quiet=TRUE,
                            list(y.i=Y, sigma.sq.i=v, n=i, eta=eta,
                                 lambda=lambda, mu=0, prec=1 / var),
                            inits=list(theta=runif(1, -2, 2),
                                       inv.tsq=1 / runif(1, 0, 0.5)))
    update(BayesMA.jags, burn, progress.bar="none") # Burn-in
    BayesMA.sims <- coda.samples(BayesMA.jags, c("theta", "tsq"),
                                 n.iter=iterations, thin=thin,
                                 progress.bar="none")

    # Check for convergence
    diag[i] <- gelman.diag(BayesMA.sims)$mpsrf

    # Theta HPD credible interval for decision rule
    all.chains <- as.mcmc(do.call(rbind, BayesMA.sims))
    theta.HPD.eps <- HPDinterval(all.chains, 1 - epsilon.e)[1, ]

    # Status
    if (theta.HPD.eps[[1]] > 0) result <- "Harm"
    if (theta.HPD.eps[[2]] < 0) result <- "Benefit"
    if (quantile(all.chains[, 1], 1 - epsilon.f)[[1]] < abs(delta) &
        quantile(all.chains[, 1], epsilon.f)[[1]] > -abs(delta)) {
      result <- "Futility"
    }
    studies <- i
    i <- i + 1
  }
  # Warning if possible problem with convergence
  if (sum(diag > 1.2) > 0) {
    cat("Warning: problem with convergence")
  }

  # Summary measures
  theta.median <- median(all.chains[, 1])
  theta.HPD.95 <- HPDinterval(all.chains)[1, ]
  tsq.median <- median(all.chains[, 2])
  tsq.HPD <- HPDinterval(all.chains)[2, ]
  list(result=result, studies=studies,
```

```
       theta=theta.median, theta.hpd=theta.HPD.95,
       tau.sq=tsq.median, tau.sq.hpd=tsq.HPD)
}
```

Acknowledgement

This work was supported by an NIHR Research Methods Fellowship (NIHR-RMFI-2015-05-015).

References

1. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Wiley, 1997.
2. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, 1999.
3. van der Tweel I, Bollen C. Sequential meta-analysis: an efficient decision-making tool. *Clinical Trials* 2010; **7**(2):136–146.
4. Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. *User manual for Trial Sequential Analysis (TSA).* Copenhagen Trial Unit, Centre for Clinical Intervention Research, Copenhagen, Denmark., 2011. Available from www.ctu.dk/tsa [accessed on 1 February 2016].
5. Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. *Statistics in Medicine* 2011; **30**(9):903–921.
6. Turner RM, Jackson D, Wei Y, Thompson SG, Higgins JPT. Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine* 2015; **34**(6):984–998.
7. Cochrane library. Http://www.cochranelibrary.com [Accessed on 1 February 2016].
8. Barrowman NJ, Fang M, Sampson M, Moher D. Identifying null meta-analyses that are ripe for updating. *BMC Medical Research Methodology* 2003; **3**(1):13.
9. Lau J, Schmid CH, Chalmers TC. Cumulative meta-analysis of clinical trials builds evidence for exemplary medical care. *Journal of Clinical Epidemiology* 1995; **48**(1):45–57.
10. Sørensen H. Small sample distribution of the likelihood ratio test in the random effects model. *Journal of Statistical Planning and Inference* 2008; **138**(6):1605–1614.
11. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**(3):177–188.
12. Thompson SG, Pocock SJ. Can meta-analyses be trusted? *Lancet* 1991; **338**(8775):1127–1130.
13. Biggerstaff BJ, Tweedie RL. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine* 1997; **15**(16):753–768.
14. Pullenayegum EM. An informed reference prior for between-study heterogeneity in meta-analyses of binary outcomes. *Statistics in Medicine* 2011; **30**(26):3082–3094.
15. Rhodes KM, Turner RM, Higgins JPT. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Research Synthesis Methods* 2015; doi:10.1002/jrsm.1193.
16. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* 2002; **21**(11):1539–1558.
17. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine* 1995; **14**(24):2685–2699.
18. Warn DE, Thompson SG, Spiegelhalter DJ. 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.
19. Pogue JM, Yusuf S. Cumulating evidence from randomized trials: Utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials* 1997; **18**(6):580–593.
20. Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. *BMC Medical Research Methodology* 2009; **9**(1):1–12.
21. Kulinskaya E, Wood J. Trial sequential methods for meta-analysis. *Research Synthesis Methods* 2014; **5**(3):212–220.
22. Whitehead A. A prospectively planned cumulative meta-analysis applied to a series of concurrent clinical trials. *Statistics in Medicine* 1997; **16**(24):2901–2913.
23. Stallard N, Facey KM. Comparison of the spending function method and the Christmas tree correction for group sequential trials. *Journal of Biopharmaceutical Statistics* 1996; **6**(3):361–373.
24. Higgins JPT, Green S, (editors). *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0*. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org [accessed 1 February 2016].
25. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; **30**(3):261–293.

26. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**(4):325–337.
27. Chen MH, Ibrahim JG, Amy Xia H, Liu T, Hennessey V. Bayesian sequential meta-analysis design in evaluating cardiovascular risk in a new antidiabetic drug development program. *Statistics in Medicine* 2014; **33**(9):1600–1618.
28. Ibrahim JG, Chen MH, Gwon Y, Chen F. The power prior: theory and applications. *Statistics in Medicine* 2015; **34**(28):3724–3749.
29. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Wiley, 1997; p202–212.
30. Lewis RJ, Berry DA. Group sequential clinical trials: A classical evaluation of bayesian decision-theoretic designs. *Journal of the American Statistical Association* 1994; **89**(428):1528–1534.
31. Lewis RJ, Lipsky AM, Berry DA. Bayesian decision-theoretic group sequential clinical trial design based on a quadratic loss function: a frequentist evaluation. *Clinical Trials* 2007; **4**(1):5–14.
32. Freedman LS, Spiegelhalter DJ. Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials* 1989; **10**(4):357–367.
33. Freedman LS, Spiegelhalter DJ. Application of bayesian statistics to decision making during a clinical trial. *Statistics in Medicine* 1992; **11**(1):23–35.
34. Berry DA. Bayesian clinical trials. *Nature Reviews Drug Discovery* 2006; **5**(1):27–36.
35. Freedman LS, Spiegelhalter DJ, Parmar MKB. The what, why and how of bayesian clinical trials monitoring. *Statistics in Medicine* 1994; **13**(13-14):1371–1383.
36. Grossman J, Parmar MKB, Spiegelhalter DJ, Freedman LS. A unified method for monitoring and analysing controlled trials. *Statistics in Medicine* 1994; **13**(18):1815–1826.
37. Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine* 1996; **15**(24):2733–2749.
38. 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**(15):2401–2428.
39. Davey J, Turner M, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology* 2011; **11**(1):160.
40. Plummer M. *rjags: Bayesian graphical models using MCMC* 2014. R package version 3-14, http://CRAN.R-project.org/package=rjags [accessed 1 February 2016].
41. Gelman A. Chapter 8: Inference and monitoring convergence. *Markov Chain Monte Carlo in Practice*, Gilks W, Richardson S, Spiegelhalter D (eds.). Taylor & Francis, 1996.
42. Tonia T, Mettler A, Robert N, Schwarzer G, Seidenfeld J, Weingart O, Hyde C, Engert A, Bohlius J. Erythropoietin or darbepoetin for patients with cancer. *Cochrane Database of Systematic Reviews* 2012; **12**. Art. No.: CD003407.
43. Bjelakovic G, Nikolova D, Gluud L, Simonetti R, Gluud C. Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. *Cochrane Database of Systematic Reviews* 2012; **3**. Art. No.: CD007176.