Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–21

DOI: 10.1177/0962280215620828
smm.sagepub.com
SAGE

Performance of informative priors skeptical of large treatment effects in clinical trials: A simulation study

Claudia Pedroza, Weilu Han, Van Thi Thanh Truong, Charles Green and Jon E Tyson

Abstract

One of the main advantages of Bayesian analyses of clinical trials is their ability to formally incorporate skepticism about large treatment effects through the use of informative priors. We conducted a simulation study to assess the performance of informative normal, Student-*t*, and beta distributions in estimating relative risk (RR) or odds ratio (OR) for binary outcomes. Simulation scenarios varied the prior standard deviation (SD; level of skepticism of large treatment effects), outcome rate in the control group, true treatment effect, and sample size. We compared the priors with regards to bias, mean squared error (MSE), and coverage of 95% credible intervals. Simulation results show that the prior SD influenced the posterior to a greater degree than the particular distributional form of the prior. For RR, priors with a 95% interval of 0.50–2.0 performed well in terms of bias, MSE, and coverage under most scenarios. For OR, priors with a wider 95% interval of 0.23–4.35 had good performance. We recommend the use of informative priors that exclude implausibly large treatment effects in analyses of clinical trials, particularly for major outcomes such as mortality.

Keywords

Bayesian analysis, informative priors, large treatment effects, binary data, clinical trial, robust priors

1 Introduction

Most clinical interventions have no more than small to moderate treatment benefits, particularly for major outcomes such as mortality or impairment.[1–3] Recent studies[4–6] show that clinical trials with smaller sample sizes report a much larger proportion of large effect sizes, i.e. relative risk (RR) less than 0.5 or greater than 2.0, than trials with larger sample sizes. When subsequent larger trials are carried out for the same intervention, disease, and outcome, observed treatment effect sizes will

Center for Clinical Research and Evidence-Based Medicine, University of Texas Health Science Center at Houston, Houston, TX, USA

Corresponding author:
Claudia Pedroza, Center for Clinical Research and Evidence-Based Medicine, University of Texas Health Science Center at Houston, 6431 Fannin St., MSB 2.106, Houston, TX 77030, USA.
Email: claudia.pedroza@uth.tmc.edu

typically diminish.[4] Investigators need to consider this empirical evidence in the reporting and interpretation of results from randomized controlled trials (RCT). However, the classical frequentist approach to analyzing and reporting individual clinical trials does not formally incorporate any external evidence.[7]

A Bayesian approach formally incorporates prior evidence through the use of a prior distribution. However, it is important that investigators move away from default "flat" or noninformative prior distributions that include extreme effects that there would be no credible reason to expect and instead use the evidence from previous trials to inform these distributions. Choices for informative prior distributions have been suggested for binary outcomes.[8,9] However, their operating characteristics have not been investigated. Of particular interest is the robustness of the priors when the prior information and the trial data are in conflict. In this paper, we present a simulation study of a two arm RCT design with a binary outcome. We investigate normal prior distributions and compare them to beta and Student-*t* distributions, which have been suggested as robust alternatives.[8–10] We motivate the choice of the informative priors from evidence from Cochrane reviews and guidelines for rating the quality of clinical evidence.

1.1 Bayesian methods

While standard frequentist methods have been the conventional paradigm to conduct analysis in clinical trials, Bayesian methods are becoming more prominent in both design and analysis of RCTs. Briefly, under a Bayesian framework, all unknown quantities are treated as random and are assigned probabilities in the form of a prior distribution.[11] This prior distribution is then combined with the observed data, in the form of a likelihood. The result, referred to as the posterior distribution, is updated evidence of the likelihood of benefit or harm from the treatment being studied. From this posterior distribution, we can provide point estimates of treatment effect, such as posterior median of the relative risk (RR), as well as 95% credible intervals (CI) that indicate the most probable value(s) of the RR. Probabilities of specific effect size can also be calculated, such as the probability that the RR exceeds a clinically important effect, e.g. $\Pr(RR < 0.8)$. These probabilities of treatment effect cannot be calculated with frequentist analyses since parameters of interest, e.g. RR, are treated as fixed. Another advantage of a Bayesian approach is that uncertainty from all parameter estimates is accounted for in summaries reported, which is particularly important when data is sparse.[12,13]

1.2 Vague reference priors

One of the main criticisms of a Bayesian approach has been the specification of the prior distribution and in particular its subjective nature. One way Bayesians have approached this criticism is by using vague priors that result in posterior estimates which are very close to those obtained from a frequentist analysis. Such priors are often used when the investigators want the data to dominate when no prior data for a particular intervention exists.[12] This approach is quite appealing since it avoids the criticism of the prior but retains the desirable properties of a Bayesian approach, mainly interpretability and coherence while maintaining good frequentist properties, i.e. unbiased parameter estimates and CIs with coverage close to the nominal level.

However, as Greenland and others have observed these flat or vague priors are "contextually absurd" in a clinical or epidemiological setting.[14–17] These vague priors ignore knowledge about the magnitude of plausible effects and will inevitably put too much weight in implausibly large values.[17] We agree with these criticisms and argue that even if no prior information exists for a particular intervention and/or population, informative priors can be specified that incorporate evidence from

other medical interventions about the magnitude of treatment effects typically reported. In particular, the informative 95% prior interval should exclude implausibly large effects almost never observed with clinical interventions.

1.3 Evidence of the magnitude of treatment effects from Cochrane reviews and GRADE criteria

One of the gold standards in evidence-based medicine is a Cochrane review, which seeks to evaluate all best available evidence for an intervention.[18] Sinclair et al.[19] report on a survey of 113 systematic reviews of neonatal therapies conducted from 1997 to 2001, which included 559 eligible RCTs. The authors report the median relative risk for the 90 reviews with a binary primary outcome to be 0.84 with interquartile range of 0.59–1.02. Only the most promising, potent, or widely used interventions are likely to be studied in sufficiently large trials to assess effects on mortality. For the 42 reviews reporting mortality outcome, the median RR is 0.86 with interquartile range of 0.77–1.05. These findings are consistent with the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) group's definition of large treatment effects as RR > 2 (<0.50), and very large effects as those RR > 5 (< 0.20).[20] They suggest a higher threshold for ORs since they can be larger in magnitude than RR for high outcome rates. Pereira et al.[4] define large effects as OR > 5 (< 0.20). Using this empirical evidence of the size of treatment effects, we can specify priors that put very little weight on unrealistic effects even for interventions with little or no prior data.

We have previously utilized informative priors in analyses of an RCT in the National Institute of Child Health and Human Development (NICHD) Neonatal Research Network (NRN) comparing the outcomes of extremely low birth weight infants (<1000 g) randomized to aggressive or conservative phototherapy.[21,22] We specified a neutral prior distribution (not favoring either intervention a priori of the observed trial data) for the RR centered at 1.0 with a 95% interval of 0.5–2.0 to provide conservative (since the estimates would be shrunk towards RR of 1.0) probabilities of reduced or increased mortality with aggressive phototherapy.

2 Methods

We assume a two-arm RCT with binary outcome y where the metric of interest is either the relative risk or odds ratio. We investigate the three most widely used models.

2.1 Log binomial regression model

Letting y_i be the observed outcome for subject i, the model is specified as

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i) \\ \log(p_i) &= \beta_0 + \beta_1 x_i \end{aligned} \tag{1}$$

where $x_i = 1$ for the treatment group and 0 for the control group. This model gives a direct estimate of the RR, $\exp(\beta_1)$, and it easily allows the inclusion of additional covariates such as any stratifying variables. Prior distributions need to be specified for the β's and a uniform distribution over the range of plausible log risk (β_0) and log RR (β_1) values (e.g. $\beta_1 \sim$ Uniform [−1,1]) could be used to serve as a flat prior.[7] However, normal prior distributions are usually specified for the β's with a very large variance, i.e. Normal $(0, 10^6)$, to make the prior vague.[23,24] To specify informative priors that give little weight to large effects, we need to specify a small prior variance or standard

deviation (SD). For example a normal prior for the log RR (β_1) with mean of 0 and SD of 0.35 implies a prior 95% interval of 0.5–2.0 for the RR. If we instead use a prior SD of 0.75, the 95% interval for the RR is 0.23–4.32. Hence, we can think of the prior SD as the level of skepticism of large effects, with smaller SDs indicative of more skepticism.

2.2 Logistic regression model

We can use the model

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i) &= \gamma_0 + \gamma_1 x_i \end{aligned} \tag{2}$$

to estimate OR with $\exp(\gamma_1)$, where x is again the indicator for treatment group. To complete the model, we specify prior distributions for the γ's. As for the log binomial model, normal priors with a large variance (e.g. 10^6) serve as vague priors in logistic regression.[25]

2.3 Beta-binomial model

We can alternatively specify a model for summary data

$$\begin{aligned} y_{\text{treat}} &\sim \text{Binomial}(n_{\text{treat}}, p_{\text{treat}}) \\ y_{\text{control}} &\sim \text{Binomial}(n_{\text{control}}, p_{\text{control}}) \end{aligned} \tag{3}$$

where y_{treat}, n_{treat}, and p_{treat} are respectively the observed number of outcomes, sample size, and the outcome probability in the treatment group, and y_{control}, n_{control}, and p_{control} are the corresponding quantities in the control group. Here, prior distributions would be specified for the probability parameters, p_{treat} and p_{control}. Beta(a, b) distributions are typically used due to their conjugacy with the binomial distribution. A Uniform(0,1) (beta with $a = 1$ and $b = 1$) has been used as the default flat prior for p_{treat} and p_{control}.[7] The posterior distribution for the relative risk (or odds ratio) can be easily computed from the draws of the posterior distributions of p_{treat} and p_{control} (see section 3).

2.4 Informative normal prior distributions

For the two regression models (1) and (2), we assess Normal$(0, \sigma^2)$ prior distributions for the treatment effect β_1 or γ_1 centered at 0, corresponding to a RR or OR of 1.0, and we use two values for the prior variance σ^2 to express the uncertainty about the treatment effect. The prior we have previously used for the analyses of the NRN trials is centered at RR of 1.0 with 95% interval of 0.5–2.0 and gives a small probability (5%) to large treatment effects. This corresponds to a Normal$(0, \sigma^2 = 0.35^2)$ prior distribution for β_1 (log RR) with a 95% interval of –0.69 to 0.69. We also investigate a normal prior with $\sigma = 0.75$ (RR 95% interval of 0.23–4.35) to compare to the prior distribution that Gelman et al.[8] found to perform the best (Cauchy with scale of 0.75 and 95% interval of 0.01–85 for RR). For the intercept terms, β_0 or γ_0, we use Normal$(0, 10^2)$ prior distributions. For the log-Binomial model, we also used normal vague priors ($\sigma^2 = 10^6$) for both the intercept and covariate coefficient to compare to the informative priors.

As noted by a reviewer, specification of these normal informative priors can be thought of in the context of sceptical priors introduced by Spiegelhalter et al.,[26] which are specified to be sceptical of an alternative hypothesis. Assuming the alternative hypothesis treatment effect is $\beta_{1,A}$ then the

Table 1. Prior distributions investigated in the simulation study. All priors are centered at RR or OR of 1.0 (0 in the log scale). Parameters for the beta prior depend on assumed p_{control}.

Prior distribution	Corresponding 95% interval for RR or OR
Log binomial model	
1) $\beta_1 \sim \text{Normal}(0, 10^6)$	$(0, \infty)$
2) $\beta_1 \sim \text{Cauchy}(0, 2.5^2)$	$(1.6 \times 10^{-14}, 6.25 \times 10^{13})$
3) $\beta_1 \sim \text{Cauchy}(0, 0.75^2)$	$(7.3 \times 10^{-5}, 1.38 \times 10^4)$
4) $\beta_1 \sim t_7(0, 0.75^2)$	(0.17, 5.89)
5) $\beta_1 \sim \text{Normal}(0, 0.75^2)$	(0.23, 4.35)
6) $\beta_1 \sim \text{Cauchy}(0, 0.35^2)$	(0.01, 85)
7) $\beta_1 \sim \text{Normal}(0, 0.35^2)$	(0.50, 2.0)
Beta-binomial model	
8) $p_{\text{control}}, p_{\text{treat}} \sim \text{Beta}(1, 1)$	(0.05, 20)
9) $p_{\text{control}}, p_{\text{treat}} \sim \text{Beta}(a, b)$	(0.23, 4.35)
Logistic model	
10) $\gamma_1 \sim \text{Cauchy}(0, 0.75^2)$	$(7.3 \times 10^{-5}, 1.38 \times 10^4)$
11) $\gamma_1 \sim t_7(0, 0.75^2)$	(0.17, 5.89)
12) $\gamma_1 \sim \text{Normal}(0, 0.75^2)$	(0.23, 4.35)
13) $\gamma_1 \sim \text{Cauchy}(0, 0.35^2)$	(0.01, 85)
14) $\gamma_1 \sim \text{Normal}(0, 0.35^2)$	(0.50, 2.0)

sceptical prior distribution $\beta_1 \sim \text{Normal}(0, \tau^2/n_0)$ is specified such that $p(\beta_1 > \beta_{1,A})$ is a small value δ. Our normal prior with SD of 0.35 ($\tau^2/n_0 = 0.35^2$) would match the sceptical prior where the alternative hypothesis is $|\beta_1| > \log(2)$ and $p(|\beta_1| > \log(2))$ is $\delta = 0.05$. The normal prior with SD of 0.75 can be thought of as being sceptical of an alternative hypothesis of $|\beta_1| > \log(4.35)$.

2.5 Informative robust prior distributions

We investigated Student-t distributions as robust priors for the β's and γ's as suggested by Gelman et al.[8] and Fúquene et al.[10] Gelman et al. recommend a weakly informative default prior of a Cauchy distribution (or t_1 distribution) which has heavier tails than a normal distribution and thus assigns greater probability to larger values of the log RR compared to a normal prior. Although the authors report that the Cauchy prior centered at 0 with scale of 0.75 performs the best in their study, they suggest a Cauchy with center 0 and scale 2.5 as a default prior for logistic regression models and other general linear models (i.e. log binomial or Poisson). This default weakly informative prior gives an implausibly large 95% interval (Table 1) for the RR that is unrealistic when compared to reported RRs in published RCTs. In fact, Gelman[27] has recently stated that a scale of 2.5 is too weak.

Here, we assess a Cauchy with scale of 0.75, a Cauchy with scale of 0.35 (to compare to the normal prior we have used in previous RCT analyses), and a t_7 with scale of 0.75 to serve as an intermediate choice between the normal and Cauchy priors. We compare these priors to the default Cauchy prior with scale of 2.5. For the intercept terms, β_0 or γ_0, we use the same distribution as for β_1 or γ_1 (Cauchy or t_7) with a scale of 10 as suggested by Gelman et al.[8]

2.6 Informative beta prior distributions

For the beta-binomial model, we can specify an informative prior distribution in different ways. If prior information exists on the most likely values for p_{control} and p_{treat}, then beta distributions could

be used where the parameters a and b are selected based on the method of moments to represent the prior evidence. If no information is available for the treatment arm or if we want to use neutral but informative priors (i.e. centered at RR of 1.0 but with very small probability of large values for the RR), a beta distribution similar to the one for p_{control} could be used for the treatment arm (i.e. centered around the same rate but perhaps with a wider interval). For example, suppose we know that p_{control} is expected to be 0.40 with a plausible range ($\pm$ 1 SD) of 0.25 to 0.55, then a beta distribution with $a = 3.87$ and $b = 5.8$ could be used for both p_{control} and p_{treat} to form a neutral prior. The implied prior for the RR is centered at 1.0 with 95% interval of 0.3–3.4.

Alternatively, we can specify a neutral prior for the relative risk centered at 1.0 with a range of plausible treatment effects, 95% interval of 0.5–2.0 and a best guess for p_{control}. We then numerically search for parameters of the beta distributions for p_{control} and p_{treat} that most closely match these constrains. For example, if the best guess for p_{control} is 40%, then a Beta(10.075, 15.112) distribution for both p_{control} and p_{treat} implies the wanted prior for RR. Here we evaluate this approach where we assume a reasonable estimate of p_{control} exists, and we set the 95% interval to 0.23–4.35 to correspond with the informative normal prior with $\sigma = 0.75$.

3 Simulation study

We assessed the performance of 14 prior distributions for binary outcome data (Table 1; Figure 1). All these priors are centered at RR or OR of 1.0 indicating no a priori difference between the two interventions. These priors can be divided into 3 sets: (1) Priors 1, 2 and 8 are default vague or weakly informative priors; (2) priors 3–5 and 9–12 use a scale or standard deviation of 0.75 to mirror the best performing prior (prior 3) reported by Gelman et al.; (3) priors 6–7 and 13–14 use a standard deviation of 0.35 which corresponds to a 95% interval of 0.5–2.0 for the RR or OR under a normal distribution in the log scale. For priors 2–7 and 10–14, the prior for the intercept had the same distributional form as for the log RR or log OR (since in a real-word data analysis it is likely that the same distribution would be used for all parameters in the model) centered at 0 and with scale of 10. For prior 1 the intercept prior was Normal(0,10^6).

We used a two group RCT design and simulated data from

$$\begin{aligned} y_i &\sim \text{Binomial}(n, p_i) \\ p_i &= p_{\text{control}}(1 - x_i) + p_{\text{treat}}\, x_i \end{aligned} \tag{4}$$

where $p_{\text{treat}} = p_{\text{control}}\, \text{RR}_{\text{true}}$, and the indicator variable x_i was sampled from a Bernoulli distribution with probability of 0.50. We varied the true control rate, $p_{\text{control}} = 0.10, 0.25, 0.50$.

To assess how the priors perform when there is true and plausible intervention effect (of either benefit or harm), we used values of $\text{RR}_{\text{true}} = 0.70, 1.5, 1.0$. A RR of 0.70 (or 30% reduction in the outcome) corresponds to the 25th percentile of treatment effects reported by Sinclair et al.[19] for reviews of neonatal interventions with mortality outcomes. A RR of 1.5 corresponds to the maximum effect from the same reviews. The implied true OR are 0.68, 0.64, and 0.54 for $\text{RR}_{\text{true}} = 0.70$, and 1.59, 1.8, and 3.0 for $\text{RR}_{\text{true}} = 1.5$. These effect sizes represent a range of plausible treatment effects from medical interventions.

The total number of subjects were $n = 50$, 250, 1000 reflecting sample sizes typically used in small pilot studies, small trials, and relatively large trials in pediatric patients. We ran a total of 24 scenarios from the combination of each p_{control}, RR_{true}, and sample size with the exception of $p_{\text{control}} = 0.10$ and $n = 50$ since we deemed this scenario too unrealistic. We generated 500 data sets for each scenario.

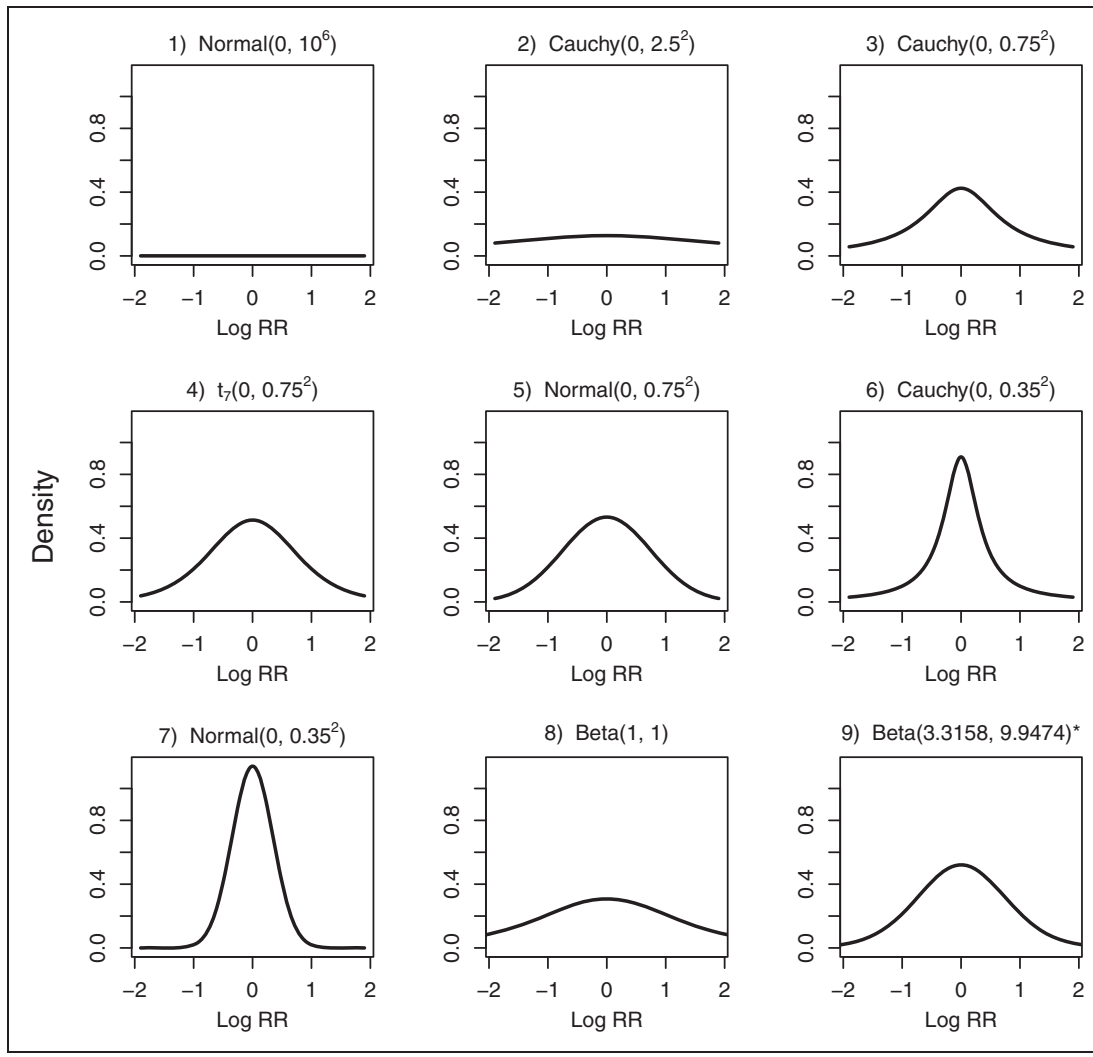

Figure 1. Density of the priors for the log RR investigated in the simulation study. Priors 3–7 correspond to priors 10–14 used for the log OR. *The beta prior shown corresponds to a $p_{\text{control}} = 0.25$.

We analyzed each data set using the 14 priors shown in Table 1. To obtain a 95% prior interval of 0.23–4.35 for the RR with prior 9, the implied (a, b) parameters of the beta distribution for $p_{\text{control}} = 0.10, 0.25, 0.50$ are $(3.779, 34.013)$, $(3.316, 9.947)$, and $(2.58, 2.58)$, respectively.

We fitted all regression models using Markov chain Monte Carlo methods (MCMC). All simulations and analyses were conducted in R[28] and OpenBUGS.[29] For each dataset we ran one MCMC chain with starting values set to the estimated parameters from a frequentist logistic model. A burn-in of 1000 iterations was used, with sampling from a further 10,000 iterations. To monitor for convergence, we visually inspected the trace plots of the parameters for the first 50 data sets in each scenario for each of the prior distributions. In practice, when using MCMC methods for a single model, multiple chains and more iterations should be used along with convergence

diagnostics.[30,31] However, for the large number of models fitted in this simulation study this was not practical.

For all models, we calculated the posterior median and 95% CI of the relative risk (e^{β_1}) or odds ratio (e^{γ_1}). For the beta-binomial models, the posterior distribution of the RR was calculated as

$$\mathrm{RR}^{(i)} = \frac{p^{(i)}_{\mathrm{treat}}}{p^{(i)}_{\mathrm{control}}} \tag{5}$$

where $p^{(i)}_{\mathrm{treat}}$, and $p^{(i)}_{\mathrm{control}}$ indicate the ith posterior draws of the parameters.

Performance of the 14 different priors was evaluated using the bias (posterior median RR [OR] $-\mathrm{RR}_{\mathrm{true}}$ [$\mathrm{OR}_{\mathrm{true}}$]) and root mean squared error (RMSE) of the posterior median of the RR and OR, and the width and coverage of the 95% credible intervals. These 95% intervals were calculated using the 2.5th and 97.5th percentiles of the posterior distribution.

We conducted a sensitivity analysis to assess the effect of the weakly informative intercept prior (scale of 10) on the posterior distribution of the treatment effect. We used a Normal(0,10^6) prior for the intercept in conjunction with priors 2–7 and 10–14 for the log RR or log OR. Since the prior will have the most influence in datasets with small sample size, we limited the scenarios to those with $\mathrm{RR}_{\mathrm{true}} = 1.5$, $n = 50$, $p_{\mathrm{control}} = 0.25, 0.50$ and $\mathrm{RR}_{\mathrm{true}} = 1.5$, $n = 250$, $p_{\mathrm{control}} = 0.10, 0.25, 0.50$.

4 Simulation results

We excluded results with posterior median of the RR or OR that were unrealistically large (> 150) or small ($< 1/150$). This resulted in exclusion of $< 1.5\%$ of data sets for the Normal($0, 10^6$) prior for scenarios with $n = 50$ and $p_{\mathrm{control}} = 0.25, 0.50$ (for RR of 1.5), and $< 1\%$ for Cauchy priors 2, 10, 13 and beta prior 8 for $n = 50$, $p_{\mathrm{control}} = 0.25$, and $\mathrm{RR}_{\mathrm{true}} = 0.70$. No data sets were excluded for scenarios with $n = 250$ or $n = 1000$.

The scenario with RR $= 1.5$, $p_{\mathrm{control}} = 0.50$, and $n = 50$ was the only one with convergence problems. For priors 1, 2, 3, 4, and 6 the MCMC chain appeared to not have converged for a small number of the datasets (10%) in this scenario. We show examples of trace plots exhibiting nonconvergence for these priors in Figure 2. No other scenarios or priors exhibited poor convergence for the inspected datasets.

4.1 Results for RR

Figures 3 to 5 (Supplementary Material; Tables S.1 and S.2) show the bias and RMSE of the posterior median and coverage of the 95% CI for RR for priors 1–9 for $\mathrm{RR}_{\mathrm{true}}$ of 0.70 and 1.50. In general, as the sample size n and p_{control} increase, both the bias and RMSE decreases for all priors. As the magnitude of the true effect size increases the RMSE also increases. For a given scenario, priors with similar scales (or spread) have similar bias and RMSE. The three vague or weakly informative priors (1, 2, and 8) give larger RMSEs compared to the other 6 priors. Compared to the normal or t_7 priors with the same scale, the Cauchy priors tend to have larger RMSEs for $n = 50$, 250 and even for $n = 1000$ for small p_{control} of 0.10 (Table S.1). However, under the Cauchy priors the point estimates for the relative risk are less biased than under a normal prior with the same scale. For example, for true RR $= 0.70$, $p_{\mathrm{control}} = 0.10$, and $n = 250$, the Cauchy prior with scale of 0.35 has bias of 0.15 and RMSE of 0.25 whereas the normal prior with the same scale of 0.35 has bias of 0.18 and RMSE of 0.23.

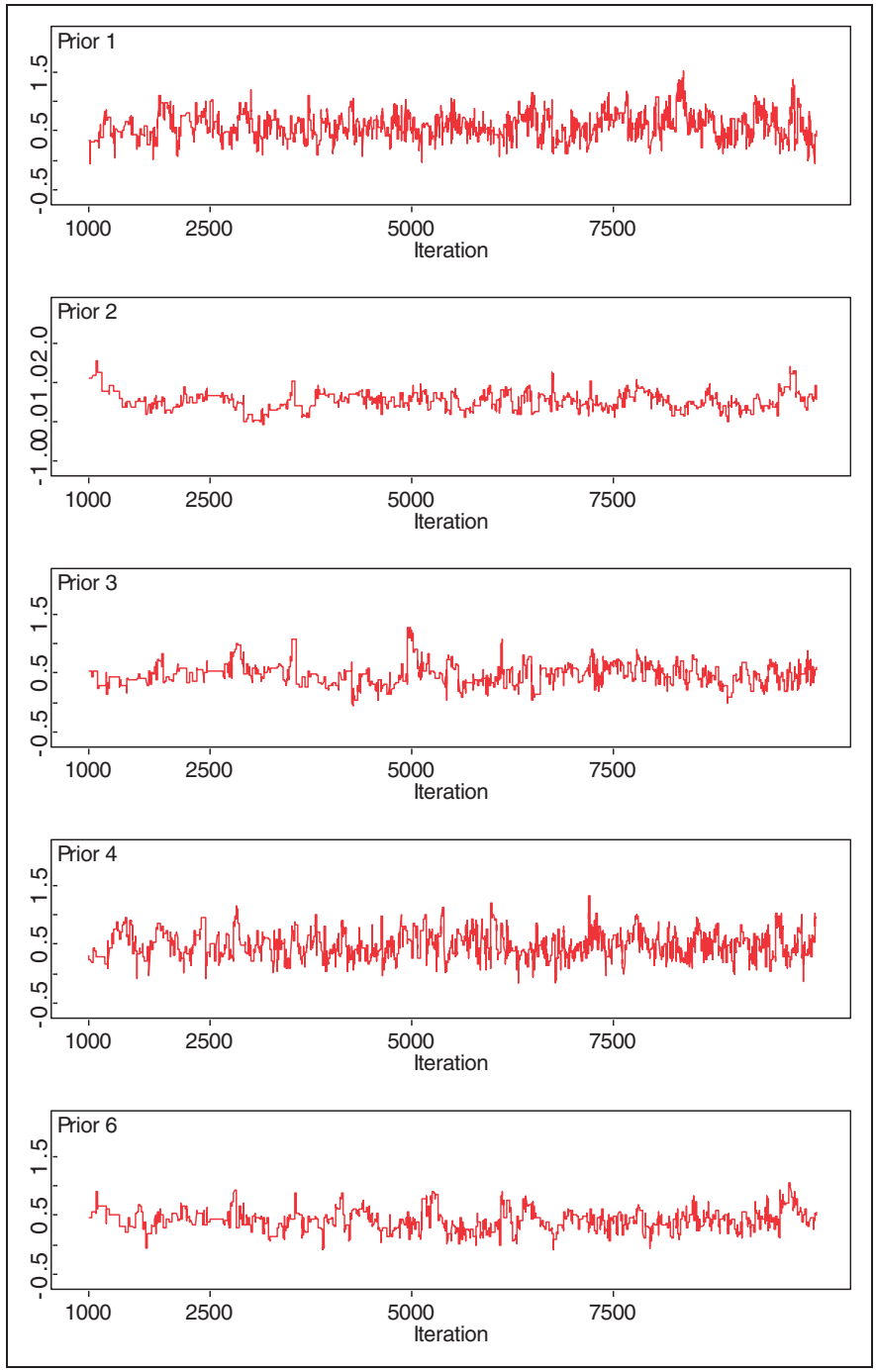

Figure 2. Example of trace plots of log RR for one simulated dataset where the MCMC chain exhibited nonconvergence for priors: Normal(0,10^6) (1), Cauchy(0,2.5^2) (2), Cauchy(0,0.75^2) (3), t_7(0,0.75^2) (4), and Cauchy(0,0.35^2) (6).

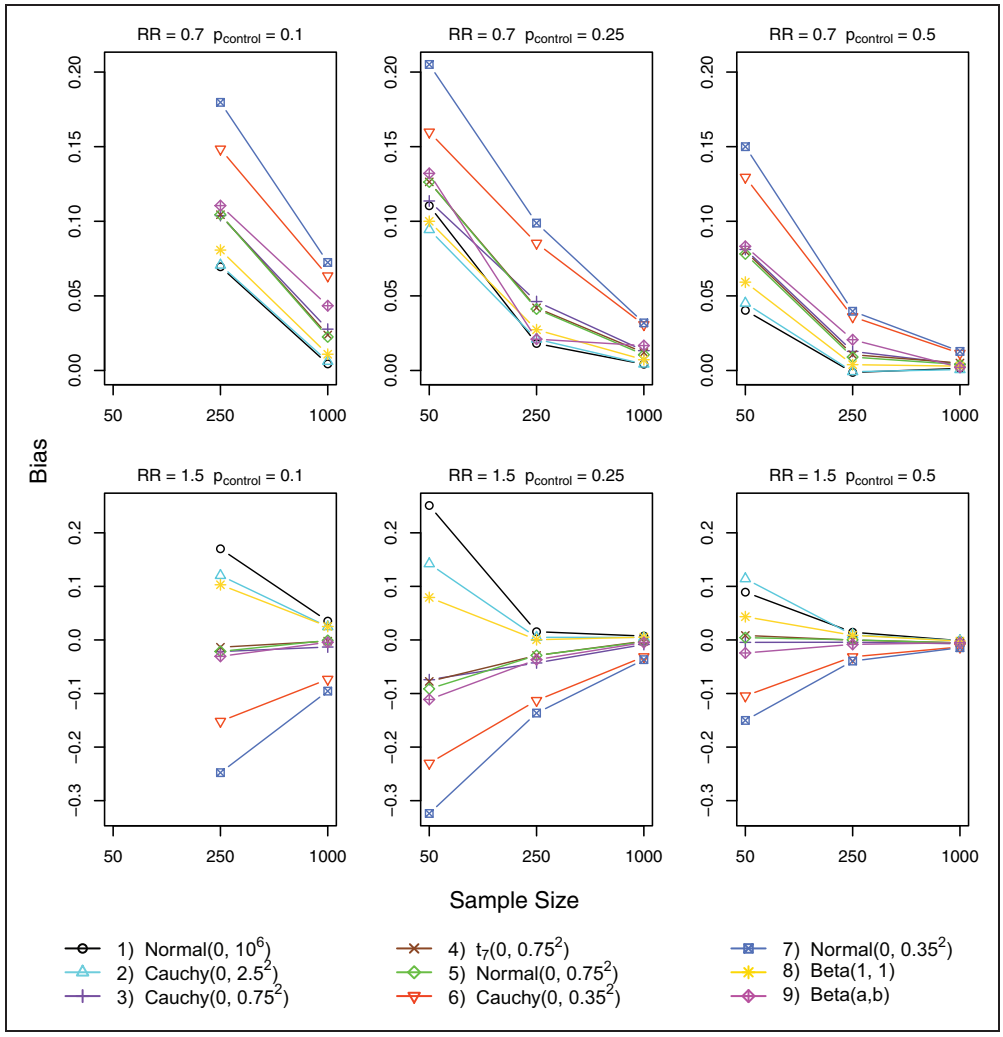

Figure 3. Bias of the posterior median of the RR using priors 1–9 (based on 500 simulated datasets).

The weakly informative Cauchy prior of Gelman et al. with scale of 2.5 tends to give very similar results to the default vague Normal(0,10^6), but under some scenarios the RMSE is two times larger than the RMSE from the default normal prior (Figure 4 and Table S.2, RR $= 1.5$ and $p_{\text{control}} = 0.50$). For the largest sample size, $n = 1000$ and p_{control} of 0.25 and 0.50, all 9 priors give very similar results.

Priors with the largest scale (1, 2, and 8) and with the smallest scale of 0.35 (priors 6 and 7) have coverage slightly below the nominal 95% level for $n = 50$ and for some scenarios with $n = 250$ (Figure 5 and Tables S.1 and S.2). For $n = 1000$, the coverage was very close to 95% for all priors in all corresponding scenarios. As expected, the average widths of the 95% CI were larger for priors with larger scales (data not shown). For the smallest sample size the Cauchy priors gave extreme average widths. For example, the average width was 2.4×10^8 for the scenario with

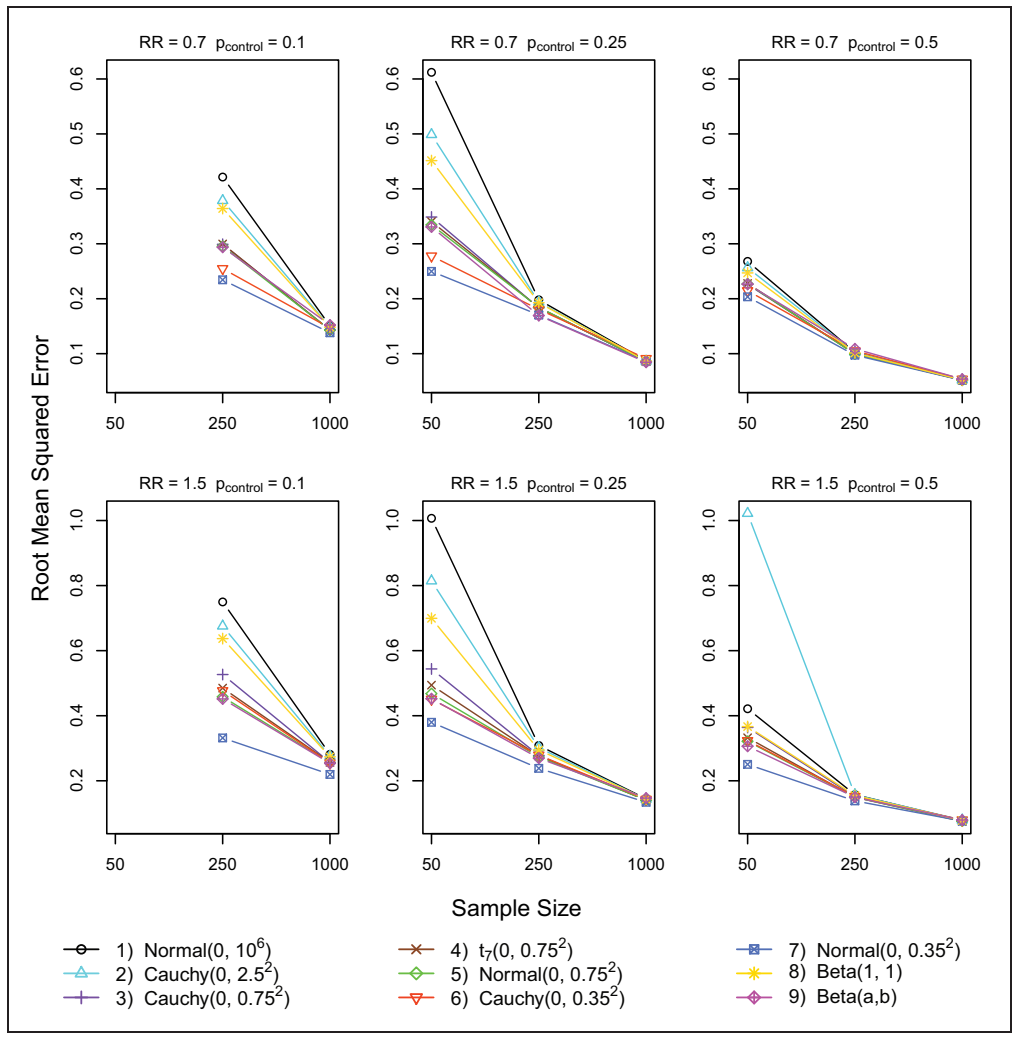

Figure 4. RMSE of the posterior median of the RR using priors 1–9 (based on 500 simulated datasets).

RR $= 1.5$ and $p_{\text{control}} = 0.25$ (for comparison, the Normal$(0, 10^6)$ prior had an average width of 1.79). Hence, we also computed the median widths which were comparable across priors with the same scale (1.54 and 1.50 for the Cauchy and normal priors in the previous example). For $n = 1000$, the average and median width were virtually identical for a given prior, and there was little difference across the 9 priors.

Results for scenarios with true RR of 1.0 follow the same patterns described above (Supplementary Material; Table S.3). As expected the estimates of the RR have little bias for all 9 priors, even for $n = 50$.

For scenarios with $n \geq 250$, priors with scale of 0.35 or 0.75 have good coverage with small bias and RMSE and can be seen as good default priors. For the smallest sample size of 50, priors with scale of 0.75 have good operating characteristics and may be good starting points.

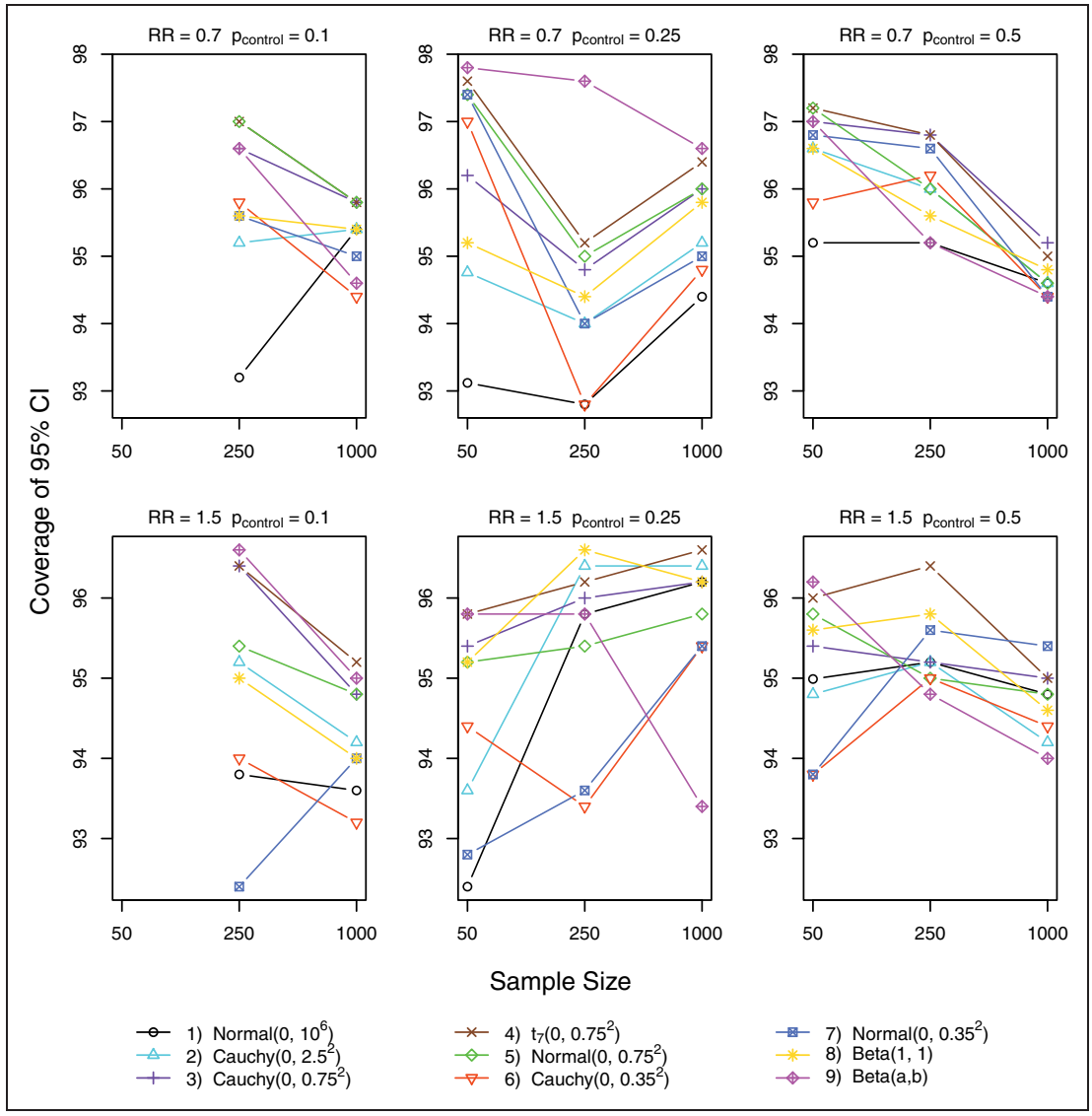

Figure 5. Coverage of the 95% CI of the RR using priors 1–9 (based on 500 simulated datasets).

4.2 Results for OR

For the logistic models estimating ORs, we see the same pattern as for RRs where the prior scale has a greater impact on the posterior than the distributional form of the prior (Figures 6 to 8 and Tables S.4 to S.6). Priors with the same scale or SD have very similar bias, RMSE and coverage, except for the larger true OR of 1.8 and 3.0. For these scenarios with $n = 50$, we note some differences for priors with scale of 0.75 where the Cauchy prior has the largest RMSE followed by the t_7 and normal prior (Figure 7). Differences among these 3 priors diminish as the sample size increases and the size of the treatment effect decreases.

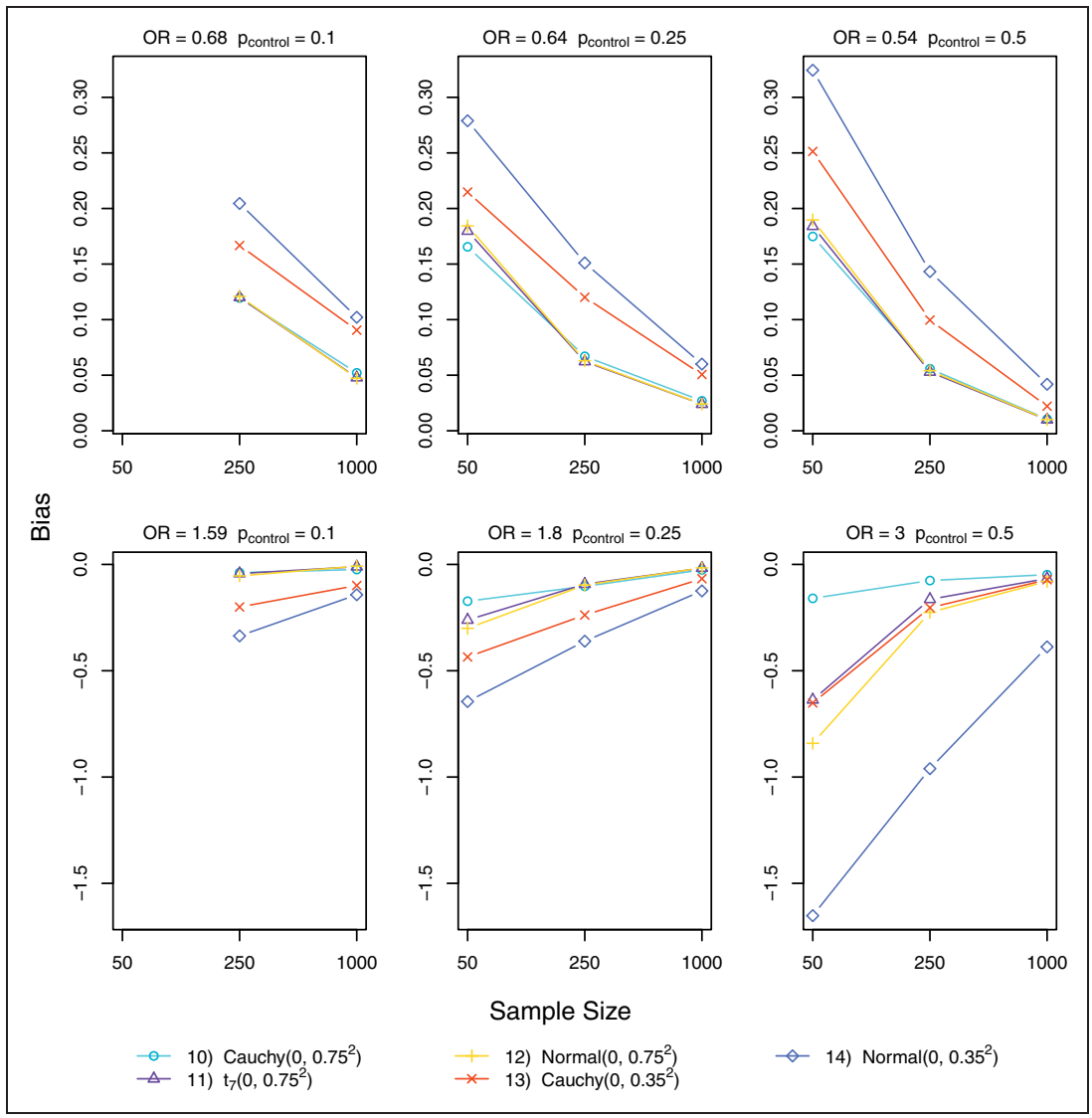

Figure 6. Bias of the posterior median of the OR using priors 10–14 (based on 500 simulated datasets).

Priors with a scale of 0.35 (priors 13–14) give point estimates with large bias and poor coverage (Figures 6 and 8). The worst performance is for the true OR = 3.0 (Table S.5). For this value of OR and $n = 50$, the Normal$(0, 0.35^2)$ prior has bias of –1.65 for the OR and coverage of 6.0%. The Cauchy with scale of 0.35 has smaller bias (−0.65) but its coverage (85.8%) is still well below the nominal level. The performance of these 2 priors improves as the sample size increases but even for $n = 1000$ the coverage for the normal prior is below the 95% level. Hence for estimating ORs a scale of 0.35 appears to be too small.

The t_7 prior distribution with scale of 0.75 seems like a good compromise. It has smaller RMSE than the Cauchy$(0,0.75^2)$ and has better coverage than the normal with the same scale.

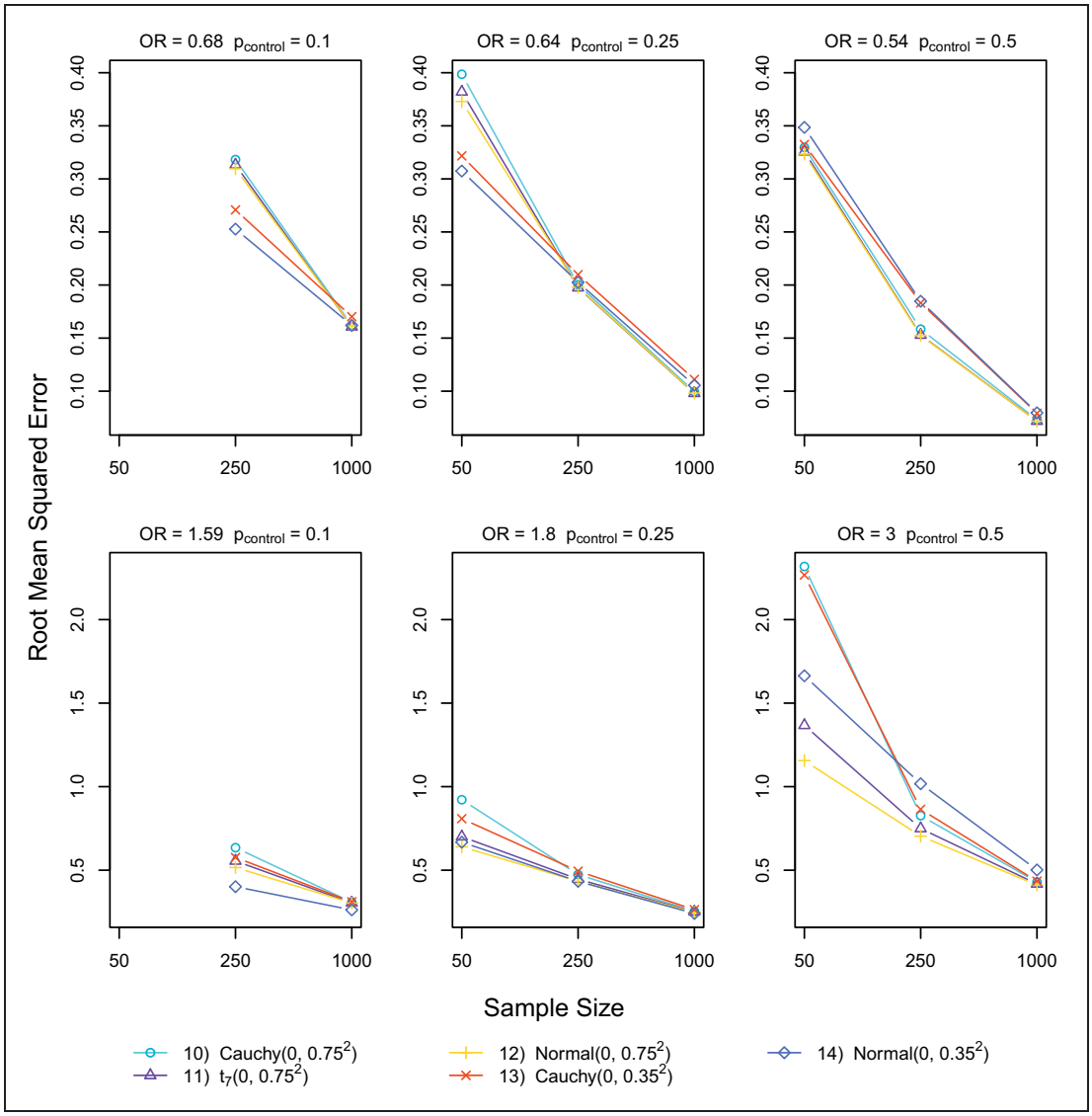

Figure 7. RMSE of the posterior median of the OR using priors 10–14 (based on 500 simulated datasets).

4.3 Sensitivity analysis results

For the analyses with priors 2–7 and 10–14 under a Normal(0,10^6) prior for the intercept, we observe the same patterns as with the informative intercept priors. Priors with the same scale had similar performance in terms of bias, RMSE, and coverage (Tables S.7 and S.8). The main difference observed was for $n = 50$ and $p_{\text{control}} = 0.25$ where the RMSE of the RR with Cauchy priors 3 and 6 (scale of 0.75 and 0.35) was more than 3-fold with the vague normal prior than with the priors with scale of 10 (values shown in bold in Table S.7). Similarly, the RMSE also increased by more than 3-fold for Cauchy priors 10 and 13 (scale of 0.75 and 0.35) when estimating true OR of 1.8 and 3.0

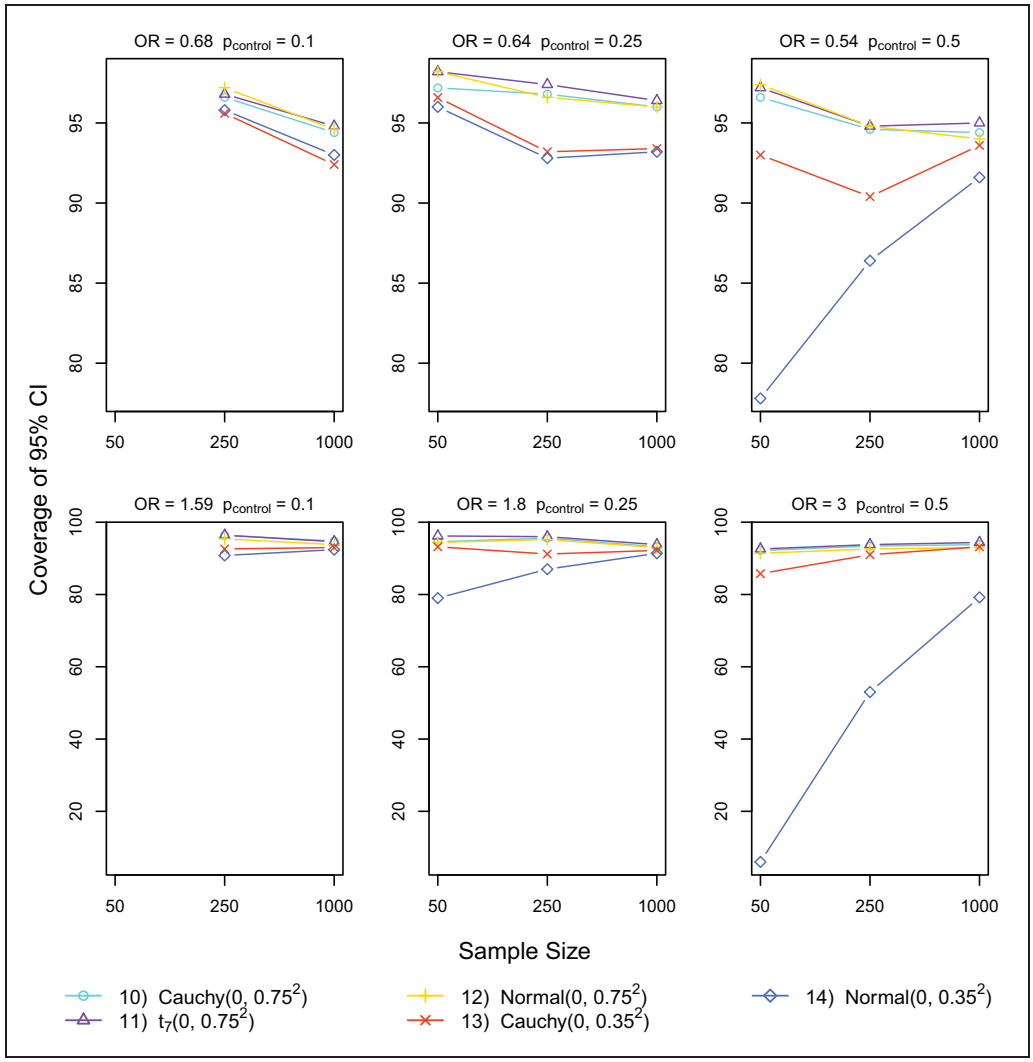

Figure 8. Coverage of the 95% CI of the OR using priors 10–14 (based on 500 simulated datasets).

with a sample size of 50 (Table S.8). Except for these two Cauchy priors, we see little difference between the results with informative intercept priors and the vague priors for both sample sizes of 50 and 250.

5 Examples

5.1 NICHD hypothermia trial

In this highly influential trial, Shankaran et al.[32] investigated the effect of hypothermia on death or disability among term infants with hypoxic ischemic encephalopathy (HIE). Within the NRN, they randomized 102 infants to whole-body cooling at 33.5°C for 72 h and 106 infants to the control group. The predefined primary outcome, death or moderate or severe disability occurred in 44% in

the hypothermia group and 62% in the control group (RR, 0.72; 95% CI, 0.54–0.95). An important issue was the effect on mortality as the most accurately measured and easily interpreted outcome. Death occurred in 24 (24%) infants in the hypothermia group and 38 (37%) in the control group. For the sake of illustration, we use the unadjusted RR which was 0.66, 95% CI 0.43–1.01. Although this trial was the largest feasible in the 16-center Network, this trial is not large, and a one-third reduction in mortality among infants with HIE can be viewed as implausibly high. Hence without confirmation in other trials, skepticism about these results would not have been unexpected when the trial was first reported. Nevertheless, even a 10% reduction in the RR would be considered clinically important.

Here, we conducted a Bayesian analysis of the mortality outcome using a log binomial model and 3 informative priors, Normal(0,0.35^2), Cauchy(0, 0.35^2), and Cauchy(0,0.75^2), to calculate posterior distributions of the effect of hypothermia. Under the normal prior, the posterior median of the RR is 0.74 (95% CI: 0.51–1.04), 0.74 (95% CI: 0.48–1.05) with the Cauchy(0, 0.35^2), and 0.68 (95% CI: 0.44–1.02) with Cauchy(0,0.75^2) prior. These estimates are in agreement with the RR estimate from the last updated Cochrane review (RR, 0.73; 95% CI: 0.61–0.89).[33] The posterior probability of a clinically important effect of a relative risk reduction of 10%, Pr(RR < 0.90), based on the NICHD trial and either the normal, Cauchy(0, 0.35^2), or Cauchy(0,0.75^2) is 87%, 85%, or 91% respectively (area to the left of the dashed line in each panel in Figure 9). Here we see that although the prior distributions differ in the range of plausible treatment effects, the resulting posterior distributions are very similar, and the conclusions drawn from the trial data would not differ.

5.2 Trials of magnesium sulfate in acute myocardial infarction

The example of magnesium sulfate after myocardial infarction is one of the well-known cases where early trials with small sample sizes showed large benefit (55% reduction in odds of death),[34] but were later contradicted by the mega-trial ISIS-4 which found no evidence of benefit.[35] A previous Bayesian analysis incorporated a skeptical prior distribution in the meta-analysis of these trials.[23] Here, we performed a Bayesian analysis of the trials reported by Sterne et al.[36] that had large treatment effects of RR < 0.50 (> 2.0). For each of these 10 trials, we used a log binomial model with a vague normal prior and 3 informative priors for the log RR: (1) Normal(0,10^6); (2) Normal(0,0.35^2); (3) Cauchy(0, 0.35^2); and (4) Cauchy(0, 0.75^2). The informative priors express the belief that large treatment effects as reported by the 10 magnesium trials are unlikely, with the informative normal prior expressing the strongest skepticism. Figure 10 (and Table S.9) shows the posterior median of the RR and 95% CI under these four priors. We note that there is complete separation[8] for the smallest trial (Bertschat 1989) where there were 0 deaths in the treatment group and 1 death in the control group. For this study the model with the vague normal prior fails to give a valid solution, which is a problem that has been noted before.[8] As expected, the Bayesian analysis under the Normal(0,0.35^2) prior tempers the treatment effects the most, and all 10 posterior medians are larger than 0.50. For 3 of the 10 trials, the 95% posterior interval from this prior includes 1.0 and may have resulted in different conclusions than with the other 3 priors.

6 Discussion

Most clinical interventions have no more than small to moderate treatment benefits, particularly for major outcomes such as mortality or impairment.[1–4] Results from clinical trials need to be interpreted in the context of typically small to moderate effects, and clinicians, investigators, and

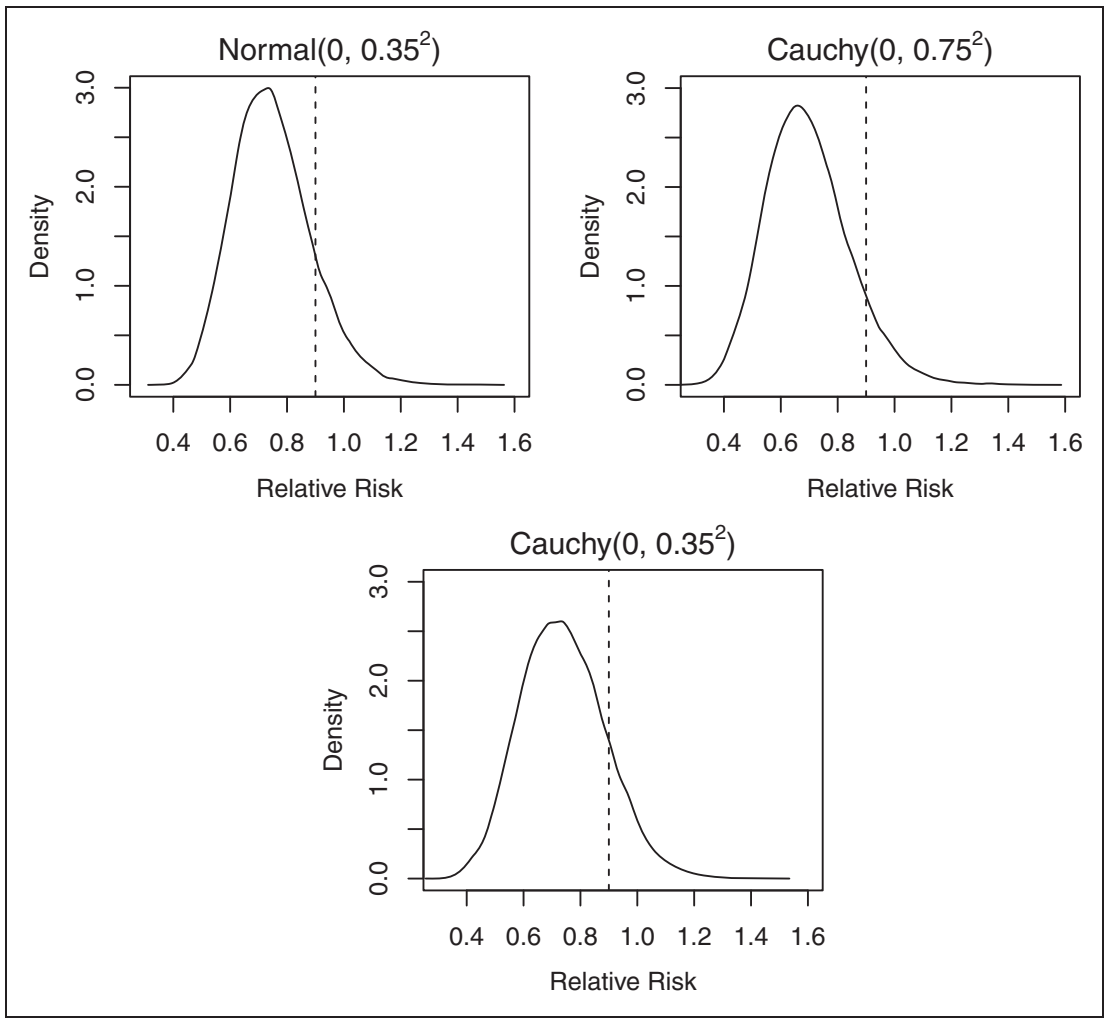

Figure 9. Posterior densities of the RR for the NICHD hypothermia trial using 3 informative priors for the RR. Area to the left of the dashed line corresponds to the posterior probability of RR < 0.90.

patients need to be skeptical of implausibly large treatment effects that are "too good to be true".[3] Bayesian analyses provide a natural way to formally incorporate empirical evidence on the magnitude of treatment effects and have been proposed as a way to temper large effects.[14,17,26,37,38]

Some informative priors have been suggested but their operating characteristics have not been studied. In particular, we are interested in whether priors with small scales reflective of empirical treatment effects perform well in terms of bias, RMSE, and coverage of the 95% CI. Our simulation study investigated these properties for 14 priors, 2 of them default vague priors and 12 informative ones. In general, we note that priors with same scale or SD perform similarly regardless of the underlying distributional form (i.e. Cauchy compared to normal). As expected, as the sample size increases, the prior distributions have less weight and all 14 priors give very similar results. For the smaller sample sizes, the more vague priors exhibit larger RMSEs and poorer coverage. Hence we

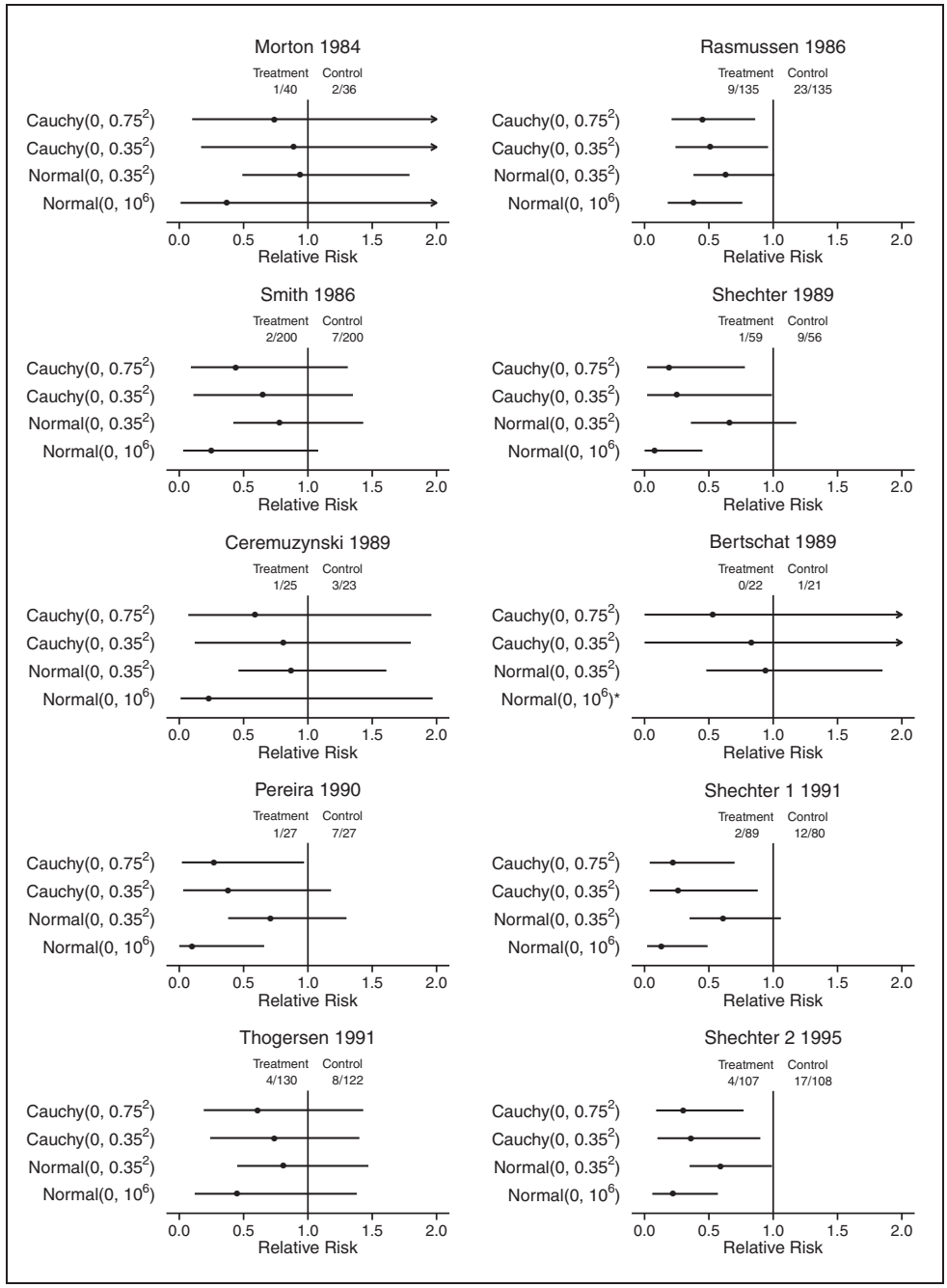

Figure 10. Bayesian estimates of RR for 10 trials from Sterne et al.[36] reporting large treatment effects (RR < 0.50) of intravenous magnesium sulfate on mortality after myocardial infarction. *There is complete separation in the Bertschat 1989 trial, and the model with the Normal(0,10^6) prior fails to give a valid solution.

would recommend that they not be used in these settings. The main concerns with using informative priors like the ones we investigated are the possibilities of obtaining biased estimates or missing a true effect. Here, we have shown that for plausible treatment effects in clinical trials the proposed informative priors provide estimates with little bias for $n \geq 250$. For small n of 50, both the noninformative and informative priors give biased estimates, but all of the informative priors have better coverage than the noninformative ones.

If we wish to exclude large treatment effects, defined by GRADE as RRs of magnitude greater than 2.0 or less than 0.50, which are very unlikely given the empirical evidence, then a Bayesian analysis with a prior such as the Normal$(0,0.35^2)$ distribution should be used, particularly for major outcomes such as mortality where reported effects outside of this range have rarely been confirmed in well-conducted large RCTs. We have used this prior in the design of two ongoing NRN trials investigating hypothermia to treat HIE in term infants. A third NRN trial of hypothermia for HIE in preterm infants (ClinicalTrials.gov identifier: NCT01793129) will use a similar prior distribution but with a wider 95% interval (0.33–3.0) reflecting greater uncertainty of the treatment effect in this population. Incorporation of informative priors is straightforward using the `bayesglm` function in the arm R package[39] or OpenBUGS (sample code is given in the Supplementary Material).

We note that for the estimation of ORs a prior distribution with scale of 0.35 may be too restrictive. The priors with scale of 0.75 performed reasonably well in most scenarios which agrees with the results reported by Gelman et al.[8] for logistic models. The coverage for the priors for the logistic model remained somewhat low even for large n. This may be due to model misspecification since the data were simulated under the log binomial model.

We also investigated informative g-priors as proposed by Hanson et al.[9] for the logistic model. An advantage of these informative multivariate normal priors is that they preserve the correlation among the predictors in the model. However, for our simulation settings the performance of the g-priors was almost identical to that of the independent normal priors with the same SD and did not appear to offer any advantages over these priors (data not shown).

The sensitivity analysis showed that for most priors on the log RR or log OR, the effect of the prior on the intercept had little impact on the posterior distribution of RR and OR. In fact, the weakly informative intercept priors resulted in overall smaller RMSEs for the RR and OR than with the vague normal intercept prior.

We chose to focus on the estimation of relative risks and odds ratios since these are the measures most commonly reported in clinical trials and observational studies.[40] More recently, some investigators have argued that the absolute risk difference (ARD) is a more clinically meaningful treatment measure and should be reported along with RR.[40–45] Thus, it would also be important to identify plausible treatment effects in terms of the ARD to derive informative priors that exclude implausible ARDs.

All priors we investigated can be considered as neutral or equivalent meaning that a priori they favor neither intervention. However, if prior evidence or even strong beliefs from investigators exist for a particular intervention then enthusiastic or optimistic priors can also be formed where the range of possible treatment effects is still part of the prior. It is important to realize that investigators with different a priori beliefs of an intervention's potential benefits and hazards may use different prior distributions.[7] The resulting posterior distributions may also differ, particularly for small sample sizes, and the investigators may reach different conclusions for a particular study. While this subjectivity of the prior is often offered as a main drawback of a Bayesian approach, we see it as an advantage since it formalizes how experts with differing pre-existing opinions will view the results.

In conclusion, we strongly recommend the use of Bayesian analyses with informative priors that incorporate evidence on the magnitude of treatment effects of medical intervention in the analyses

and interpretation of RCTs. Informative priors such as a Normal(0,0.35^2) for RR or Normal(0,0.75^2) for OR should be used for analyses of major outcomes. The robust alternatives such as the Cauchy or t_7 priors with the same scale can also be used or considered in sensitivity analyses.

Acknowledgements

The authors gratefully acknowledge the helpful comments from two reviewers, which have improved the content and presentation of this paper.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. Peto R, Collins R and Gray R. Large-scale randomized evidence: Large, simple trials and overviews of trials. *J Clin Epidemiol* 1995; **48**: 23–40.
2. Yusuf S. Meta-analysis of randomized trials: Looking back and looking ahead. *Control Clin Trials* 1997; **18**: 594–601.
3. Yusuf S and Flather M. Magnesium in acute myocardial infarction. *BMJ* 1995; **310**: 751–752.
4. Pereira TV, Horwitz RI and Ioannidis JA. Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 2012; **308**: 1676–1684.
5. Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology* 2008; **19**: 640–648.
6. Pereira TV and Ioannidis JPA. Statistically significant meta-analyses of clinical trials have modest credibility and inflated effects. *J Clin Epidemiol* 2011; **64**: 1060–1069.
7. Spiegelhalter DJ, Abrams KR and Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, UK: John Wiley & Sons, 2004.
8. Gelman A, Jakulin A, Pittau MG, et al. A weakly informative default prior distribution for logistic and other regression models. *Ann Appl Stat* 2008; **2**: 1360–1383.
9. Hanson TE, Branscum AJ and Johnson WO. Informative *g*-priors for logistic regression. *Bayesian Anal* 2014; **9**: 597–612.
10. Fúquene JA, Cook JD and Pericchi LR. A case for robust Bayesian priors with applications to clinical trials. *Bayesian Anal* 2009; **4**: 817–846.
11. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov* 2006; **5**: 27–36.
12. Lambert PC, Sutton AJ, Burton PR, et al. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using winbugs. *Stat Med* 2005; **24**: 2401.
13. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. *Int J Epidemiol* 2007; **36**: 195–202.
14. Greenland S. Putting background information about relative risks into conjugate prior distributions. *Biometrics* 2001; **57**: 663–670.
15. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol* 2006; **35**: 765–775.
16. Greenland S and Poole C. Living with p values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology* 2013; **24**: 62–68.
17. Gelman A. P values and statistical practice. *Epidemiology* 2013; **24**: 69–72.
18. Higgins JP and Green S. *Cochrane handbook for systematic reviews of interventions*. Version 5.1.0, The Cochrane Collaboration, 2011. www.cochrane-handbook.org.
19. Sinclair JC, Haughton DE, Bracken MB, et al. Cochrane neonatal systematic reviews: a survey of the evidence for neonatal therapies. *Clin Perinatol* 2003; **30**: 285–304.
20. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol* 2011; **64**: 1311–1316.
21. Morris BH, Oh W, Tyson JE, et al. Aggressive vs. conservative phototherapy for infants with extremely low birth weight. *N Engl J Med* 2008; **359**: 1885–1896.
22. Tyson JE, Pedroza C, Langer J, et al. Does aggressive phototherapy increase mortality while decreasing profound impairment among the smallest and sickest newborns? *J Perinatol* 2012; **32**: 677–684.
23. Higgins JPT and Spiegelhalter DJ. Being sceptical about meta-analyses: a Bayesian perspective on magnesium trials in myocardial infarction. *Int J Epidemiol* 2002; **31**: 96–104.
24. Wijeysundera DN, Austin PC, Hux JE, et al. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. *J Clin Epidemiol* 2009; **62**: 13–21.
25. Turner RM, Omar RZ and Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Stat Med* 2001; **20**: 453–472.

26. Spiegelhalter DJ, Freedman LS and Parmar MKB. Bayesian approaches to randomized trials. *J R Stat Soc Ser A Stat Soc* 1994; **157**: 357–416.
27. Gelman A. Informative priors for logistic regression. http://andrewgelman.com/2014/11/05/firth-bias-correction-penalization-weakly-informative-priors-case-log-f-priors-logistic-related-regressions (accessed April 2015).
28. R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. http://www.R-project.org/.
29. Thomas A, O'Hara B, Ligges U, et al. Making BUGS Open. *R News* 2006; **6**: 12–17.
30. Brooks SP and Gelman A. General methods for monitoring convergence of iterative simulations. *J Comp Graph Stat* 1998; **7**: 434–455.
31. Gelman A, Carlin JB, Stern HS, et al. *Bayesian data analysis*, 3rd ed. Boca Raton, FL: CRC Press, 2014.
32. Shankaran S, Laptook AR, Ehrenkranz RA, et al. Whole-body hypothermia for neonates with hypoxic-ischemic encephalopathy. *N Engl J Med* 2005; **353**: 1574–1584.
33. Jacobs SE, Berg M, Hunt R, et al. Cooling for newborns with hypoxic ischaemic encephalopathy. *Cochrane Database Syst Rev* 2013; **1**: CD003311.
34. Teo KK, Yusuf S, Collins R, et al. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. *BMJ* 1991; **303**: 1499–1503.
35. (Fourth International Study of Infarct Survival) Collaborative Group ISIS-4. ISIS-4: A randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. *Lancet* 1995; **345**: 669–685.
36. Sterne JAC, Sutton AJ, Ioannidis JPA, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ* 2011; **343**: d4002.
37. Kass RE and Greenhouse JB. Investigating therapies of potentially great benefit: ECMO: Comment: A Bayesian perspective. *Stat Sci* 1989; **4**: 310–317.
38. Fayers PM, Ashby D and Parmar MK. Bayesian data monitoring in clinical trials. *Stat Med* 1997; **16**: 1413–1430.
39. Gelman A and Su YS. *arm: Data analysis using regression and multilevel/hierarchical models*, 2015. http://CRAN.R-project.org/package=arm. R package version 1.8-03.
40. Schechtman E. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat–which of these should we use? *Value Health* 2002; **5**: 431–436.
41. Laupacis A, Sackett DL and Roberts RS. An assessment of clinically useful measures of the consequences of treatment. *N Engl J Med* 1988; **318**: 1728–1733.
42. Ayton P. Number needed to treat. Risk measures expressed as frequencies may have a more rational response. *BMJ* 1995; **310**: 1269.
43. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987; **125**: 761–768.
44. Cook RJ and Sackett DL. The number needed to treat: A clinically useful measure of treatment effect. *BMJ* 1995; **310**: 452–454.
45. Sinclair JC and Bracken MB. Clinically useful measures of effect in binary analyses of randomized trials. *J Clin Epidemiol* 1994; **47**: 881–889.