STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:1417–1430
Published online 11 August 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2639

Trying to be precise about vagueness

Stephen Senn*, †

Department of Statistics, University of Glasgow, U.K.

SUMMARY

A previous investigation by Lambert *et al*., which used computer simulation to examine the influence of choice of prior distribution on inferences from Bayesian random effects meta-analysis, is critically examined from a number of viewpoints. The practical example used is shown to be problematic. The various prior distributions are shown to be unreasonable in terms of what they imply about the joint distribution of the overall treatment effect and the random effects variance. An alternative form of prior distribution is tentatively proposed. Finally, some practical recommendations are made that stress the value both of fixed effect analyses and of frequentist approaches as well as various diagnostic investigations.

KEY WORDS: meta-analysis; fixed effects; random effects; Bayesian methods; profile likelihood; graphical representation; HGLM

1. INTRODUCTION

Markov chain Monte Carlo (MCMC) approaches to the analysis of hierarchical data sets are making Bayesian analyses so popular that an investigation into the robustness of their conclusions with respect to choice of priors, such as recently carried out by Lambert *et al*. [1] in the pages of this journal, must surely be welcome. I also applaud one of their principal conclusions, namely, 'we feel that it is an important message that the use of vague prior distributions should be treated with a degree of caution' (p. 2424). Nevertheless, it seems to me, upon reflecting upon this energetically prosecuted study, that perhaps simulation does not advance our understanding of these matters as far as one might hope and that some basic and more 'philosophical' considerations may be useful.

A brief reminder of the contents of the paper by Lambert *et al.* may be helpful. First, they introduced an example involving a meta-analysis of five trials in otitis media. They analysed this example using a Bayesian random effects approach with a diffuse Normal prior distribution

*Correspondence to: Stephen Senn, Department of Statistics, University of Glasgow, 15 University Gardens, Glasgow, G12 8QQ, U.K.
†E-mail: stephen@stats.gla.ac.uk

Received 18 August 2005
Accepted 24 May 2006

Table I. Different prior distributions.

Number	Target parameter	Distribution	Form of probability density function	Prior parameters
1	$1/\tau^2$	Gamma	$\frac{\mu^r x^{r-1} e^{-\mu x}}{\Gamma(r)},\ x>0$	$\mu=0.001, r=0.001$
2	$1/\tau^2$	Gamma		$\mu=0.1, r=0.1$
3	$\log(\tau^2)$	Uniform		$a=-10, b=10$
4	$\log(\tau^2)$	Uniform		$a=-10, b=1.386$
5	τ^2	Uniform		$a=1/1000, b=1000$
6	τ^2	Uniform	$\frac{1}{b-a},\ a<x<b$	$a=1/1000, b=4$
9	τ	Uniform		$a=0, b=100$
10	τ	Uniform		$\alpha=0, b=2$
7	$1/\tau^2$	Pareto	$\alpha c^\alpha x^{-(a+1)},\ x>c$	$\alpha=1, c=0.001$
8	$1/\tau^2$	Pareto		$\alpha=1, c=0.25$
11	τ	Half-Normal	$\sqrt{\frac{2\lambda}{\pi}}\, e^{-(\lambda/2)x^2}, 0<x<\infty$	$\lambda=100$
12	τ	Half-Normal		$\lambda=1$
13	$1/\tau^2$	Logistic	$\frac{\beta e^{x\beta}}{(1+e^{x\beta})^2},\ -\infty<x<\infty$	$\beta=\sqrt{v_h}$

for the pooled log-odds ratio Θ with mean 0 and variance 10 000 and illustrated the use of 13 different prior distributions for the random effects variance, τ^2. These are listed in Table I, where the numbering is the same as given by Lambert *et al*. but the order has been changed to permit grouping by distribution. Note that the parameterisation follows that used in the winBUGS manual [2], which is eccentric in many respects compared to standard statistical convention. For that reason the distributions are given explicitly here. Second, Lambert *et al*. then simulated various possible collections of trials with a rather similar common odds ratio to that estimated for their example but with three possible values (0.001, 0.3 and 0.8) of the standard deviation of the log-odds ratio between studies and considering in addition to collections of 5 studies, the cases of 10 and 30 studies also. These nine different combinations were replicated 1000 times and each replication was analysed using each of the 13 prior distributions. Thus, in all results from $13 \times 9 \times 1000 = 117\,000$ simulated meta-analyses were run.

In this note I shall consider some issues of a 'philosophical' and practical nature that the choice of prior distributions and the assessment of that choice through simulation raise.

2. THE OTITIS MEDIA EXAMPLE

Before raising these more philosophical issues, however, it is worth spending a little time looking at the example given by Lambert *et al*. [1], which concerned a comparison of short *versus* long course treatment for acute otitis media. There are two aspects of this example that are unsatisfactory. The first, serious from the point of view of drawing conclusions about treatment for patients with otitis media but less serious from the point of view of using the study as a statistical guinea-pig, is that the original Cochrane Collaboration analysis is incorrect. The second point, to be dealt with below, is that the example is not well suited for investigating the robustness of choice of prior.

DOI: 10.1002/sim

Table II. Data for point estimates.

Study	Short course	Long course
Boulesteix, 1995	11/124	11/118
Cohen, 1997	26/186	31/184
Hendrickse, 1988	14/74	6/77
Hoberman, 1997a	**57/197**	24/178
Hoberman, 1997b	**57/197**	40/189

The study does not in fact, appear to be, as claimed by Lambert *et al.* [1], the Cochrane review carried out by Glasziou *et al.* [3] but rather one carried out by Kozyrskyj *et al.* [4]. The data from which the point estimates and standard errors have been calculated must be those given by Figure 3 of that study [4] and are as reproduced in Table II. When this table is studied, what immediately strikes the eye is the fact that for short course therapy the last two entries are the same in terms of treatment failures and numbers treated and that the principle author, Hoberman, is the same. (These figures have been highlighted in bold in the table.) In fact, despite apparently different references, 1997a and 1997b, the same paper is being referred to and when this is checked it turns out to be a three-armed trial [5] comparing, according to the abstract, 'a new formulation of amoxicillin/clavulanate potassium (augmentin) oral suspension providing 45/6.4 mg/kg/day and administered twice daily (bid) for 5 and 10 days, respectively, with the safety and efficacy of the original formulation providing 40/10 mg/kg/day and administered three times daily (tid) for 10 days'. Thus, one of the arms, augmentin oral suspension bid for five days has been used twice in the meta-analysis, a clearly illegitimate procedure that every statistician will condemn. It also seems that the original meta-analysts, in their enthusiasm to summarize everything have overlooked the fact that the treatments being compared differ not only in terms of duration, but also in terms of formulation, dose and dosing schedule. This is a far more serious issue, say, than one of pooling all trials in which antibiotics, identical within a trial, but varying from trial to trial had been compared. In a fixed effects analysis, such a collection of trials could answer (in principle) the question, 'is there at least one antibiotic that can show different effects when given over a short course compared to a long course', although it is difficult to see what sensible question a random effects analysis could answer. Here, on the other hand, for one of the contrasts considered by Hoberman *et al.* [5], difference between length of treatments is confounded with treatment. A similar criticism applies to one of the other studies, that of Boulesteix [6], which compares cefpodoxime proxetil and cefixime.

In my opinion, this sort of inappropriate analysis is rife outside of the pharmaceutical industry and is encouraged by quality scores [7, 8] that place all the emphasis on identifying trials and none on correct analysis [9].

Of course, Lambert *et al.* [1] are not responsible for the original meta-analysis. However, their complaint that, 'The original meta-analysis used a fixed effects model even though there was strong evidence of heterogeneity of study effects using Cochran's test' [1] (p. 2403) is wide of the mark. There were far more serious issues regarding this example and in any case, the choice of fixed effects or random effects meta-analysis should not be made on the basis of perceived heterogeneity but on the basis of purpose [10, 11]. Furthermore, when such heterogeneity has been detected, there is a strong case for investigating its possible origin [12, 13], a *minimal* requirement for which is actually looking up the list of studies, and in any case, in view of the inherent heterogeneity of the studies in terms of *objectives*, it is doubtful that any useful question can be formulated that

DOI: 10.1002/sim

a random effects analysis would answer. (Readers who doubt this might like to try.) However, in what remains, I shall accept the otitis media example as Lambert *et al*. [1] present it and examine it using random effects approaches. Nevertheless, the example in my view is in any case not a good one for investigating the consequences of choice of prior, as I shall explain below.

3. WEIGHTING IN META-ANALYSIS

As is well known, point estimates, $\hat{\Theta}$, produced by conventional meta-analytic techniques can be expressed as a weighted linear combinations of the point estimates from individual studies as follows:

$$\hat{\Theta} = \sum_{i=1}^{k} w_i \hat{\theta}_i \tag{1}$$

where k is the number of estimates, $\hat{\theta}_i$ is the point estimate from study i and w_i is a weight (depending on method) assigned to study i. For example, for a conventional fixed effects estimator we have $w_i \propto 1/v_i$, where v_i is the within-study variance of study i and for a random effects estimator we have $w_i \propto 1/u_i$, where $u_i = v_i + \tau^2$ and τ^2 is the between-study variance. Since, for all methods, we have $\sum_{i=1}^{k} w_i = 1$ and that the mean of these weights is $1/k$, expression (1) can in turn be written in terms of deviations of estimates for individual studies from the arithmetic mean, $\bar{\hat{\theta}}_A$ as

$$\hat{\Theta} = \bar{\hat{\theta}}_A + \sum_{i=1}^{k} w_i(\hat{\theta}_i - \bar{\hat{\theta}}_A) = \bar{\hat{\theta}}_A + \sum_{i=1}^{k} (w_i - 1/k)(\hat{\theta}_i - \bar{\hat{\theta}}_A) \tag{2}$$

which is to say, as the arithmetic mean of the estimates plus the corrected sum of cross-product of weights and estimates. (Note that the simple arithmetic mean is the form originally proposed by Yates and Cochran in their pioneering paper [14].) It therefore follows that an equivalent expressions for (2) is

$$\bar{\hat{\theta}}_A + k\rho_{w\hat{\theta}}\sigma_w\sigma_{\hat{\theta}} \tag{3}$$

where $\sigma_{\hat{\theta}}$ is the standard deviation of the observed points estimates, σ_w is the standard deviation of the system of weights and $\rho_{w\hat{\theta}}$ is the correlation between the two.

Now consider two different estimators, the first using a system of weights, w and the second using a system w^*. Making an obvious extension of notation, it follows from (3) that the difference between them will be

$$k\sigma_{\hat{\theta}}(\rho_{w\hat{\theta}}\sigma_w - \rho_{w^*\hat{\theta}}\sigma_{w^*}) \tag{4}$$

This may seem to be an unnecessarily complicated way of expressing some well-known facts about meta-analysis but in fact it highlights the importance of the correlation coefficient between weights and estimates as regards sensitivity of resulting calculation of the overall treatment estimate and this, as I shall discuss below, will be relevant for discussing the simulation carried out by Lambert *et al*. [1].

Returning for the moment, however, to the otitis media example as reported, if a system of weights suitable for a fixed effects analysis on the log-odds ratio scale is constructed, then the results given in Table III may be found. The correlation of the weights with the estimates is

DOI: 10.1002/sim

Table III. Results of weights on log-odds ratio.

Study	Estimate	Estimated standard error	Weight
Boulesteix, 1995	−0.05	0.45	0.0957
Cohen, 1997	−0.22	0.29	0.2304
Hendrickse, 1988	1.02	0.52	0.0717
Hoberman, **1997a**	0.96	0.27	0.2658
Hoberman, **1997b**	0.42	0.24	0.3364

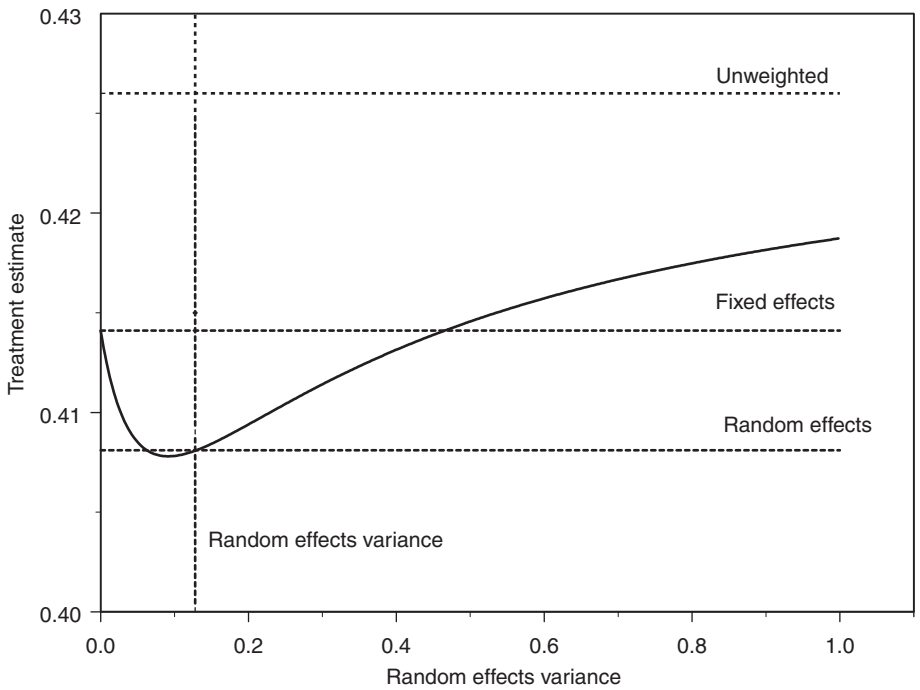

Figure 1. Example used by Lambert *et al*. Treatment estimate, $\hat{\Theta}$, as a function of the random effects variance. The position of the estimate variance τ^2, and three estimators, fixed, random and unweighted are indicated.

extremely modest at −0.05, and hence it is hardly a surprise, given (3) that the fixed effects estimate at 0.414 is very close to the arithmetic mean at 0.426. (The correlation being negative, it is, of course, lower.)

In general, random effects estimators tend to weight studies more equally, since a common between-study variance contributes to the weight [13]. Hence, when the fixed effects estimate is scarcely different from the arithmetic mean, then the random effects estimate is unlikely to differ very much either. An analysis using the method of Dersimonian and Laird [15] produces a random effects estimate of 0.409 and that of Hardy and Thompson [16] yields 0.408. Figure 1 plots the meta-analysis estimate that would result by treating the random effects variance, τ^2, as a constant

DOI: 10.1002/sim

known *a priori* and allowing it to vary from 0 to 1. It will be seen that the maximum departure from the arithmetic mean, happens for this example to be close to that for the random effects estimator, but that whatever value of this variance that is chosen, the overall estimate of the treatment effect is close to the fixed effects estimate. Note that assuming that the value of τ^2 is known is equivalent to having a totally informative prior distribution on its value.

It thus follows, that it is hardly surprising that the various priors chosen for this example should yield similar point estimates. It also follows that *no reassurance as regards this finding should be taken from this example*. It may or may not be a feature of Bayesian random effects meta-analysis that choice of prior distribution for the random effects variance has little impact on the point estimate provided that it is sufficiently vague but this example is too well behaved to advance understanding of this issue. It is a point that is likely to apply in general for examples in which the within-study precisions (the reciprocals of the squares of the standard errors) are weakly correlated with the estimates. Since it is the observed correlation that matters, this in turn is more likely to be the case when there are many studies, but this is precisely the case under which the random-effects variance is likely to be well estimated. If it were not the case from a large collection of studies then it would call into question the very notion of exchangeability of studies and imply that the conventional random effects estimate (Bayesian or otherwise) would be of little interest [10, 11].

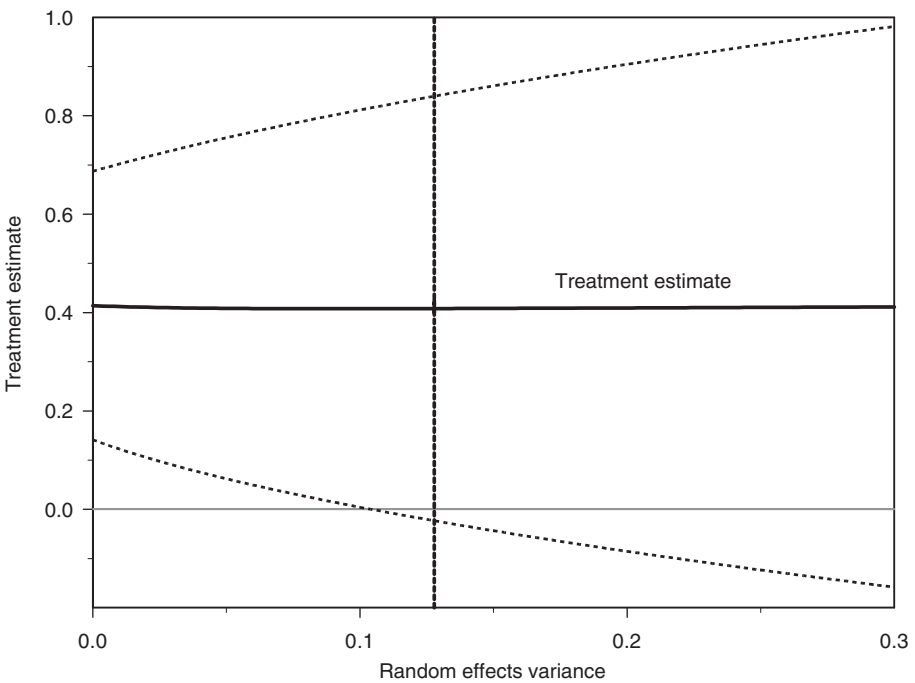

Figure 2. Treatment estimate $\hat{\Theta}$ (solid line) and 95 per cent confidence intervals (dashed lines) as a function of the random effects variance τ^2 when this is assumed known. Note that this scale for $\hat{\Theta}$ is far larger than that of Figure 1 and the dependence of $\hat{\Theta}$ on τ^2 is no longer evident, a consequence of the example chosen by Lambert *et al.*

DOI: 10.1002/sim

Of course the variance of the overall treatment estimate can differ considerably according to the random effect. The effect of this is plotted in Figure 2, which gives the point estimate and upper and lower confidence intervals using a simple crude limit of the form $\hat{\Theta} \pm 1.96 \times \mathrm{SE}(\hat{\Theta})$ assuming the value of τ^2 known, as a function of that value and using the formula for the variance of $\hat{\Theta}$

$$\mathrm{var}(\hat{\Theta}) = \frac{\overline{u}_h}{k} \tag{5}$$

where $\bar{u}_h$ is the *harmonic* mean [17] of the terms $u_i = v_i + \tau^2$. When $\tau^2 = 0$ then we have $u_i = v_i$, and hence the same estimator as in the fixed effects case and the variance becomes

$$\mathrm{var}(\hat{\Theta}) = \frac{\overline{v}_h}{k} \tag{6}$$

That being so one must expect, of course, that the crucial issue in determining the width of the confidence interval is the estimated between-study variance, $\hat{\tau}^2$ although, as Lambert *et al*. [1] discuss, the precision with which it is estimated will also have some influence.

4. PROBLEMS IN SIMULATING

The investigation that Lambert *et al*. [1] carried out was supplemented by an extensive simulation using the following model

$$\begin{aligned} &\delta_i : \mathrm{N}(0, \tau^2) \\ &\mathrm{logit}(p_{0i}) = \alpha \\ &\mathrm{logit}(p_{1i}) = \alpha + \Theta + \delta_i \\ &r_{0i} = \mathrm{Binomial}(n_{0i}, p_{0i}) \\ &r_{0i} = \mathrm{Binomial}(n_{1i}, p_{1i}) \end{aligned} \tag{7}$$

Here $\delta_i = \theta_i - \Theta$ in terms of the model previously given and is a 'random effect'. The number of subjects for the trials were 100, 200, 300, 400 and 500. Note that this model does not allow for a main effect of trial. The trials are homogenous as regards control group response. However, the model used for analysis by Lambert *et al*. does allow for differences from trial to trial in the control group effect. This point will be taken up below.

The simulation was run in such a way that study variances and point estimates would be largely independent. The variances will have been almost entirely driven by the study sizes. (In fact, apart from sample size, the only thing that differs as a parameter from trial to trial is δ_i.) Therefore, all that the simulation did was repeat the general situation that applied in the otitis media example. Admittedly, especially for the case of five studies, some runs will have produced strong correlations but this feature will disappear averaging over all runs.

Another more philosophical puzzle attends simulation, however. The Bayesian inference about the overall treatment estimate is one that is averaged over the prior distribution, which is known with absolute certainty, although any given value from it is, of course, not. Thus, take, for example, prior distribution 1. This is gamma (0.001, 0.001) for the precision, that is to say $1/\tau^2$. Conceptually, what happens in analysis, is that the particular set of studies for meta-analysis is regarded as having been drawn at random from an infinity of sets with this particular distribution. Given that this infinite

DOI: 10.1002/sim

collection of sets has been sampled to produce a single set, Bayesian analysis then summarizes what is believed about the particular unobserved $1/\tau^2$ given the observed data from the trials.

Now if this posterior belief is assessed via simulation against the situation that applies when τ^2 is known to be some particular value, say 0.001, 0.3 or 0.8 as the case may be then this is like comparing chalk and cheese. These are totally informative priors on the parameter values for the statistician *as all-knowing simulator* against which the statistician *as ignorant inferencer* assesses his or her performance. You might as well expect two Bayesians, with radically different priors, one very informative and the other not at all to come to similar conclusions. In a frequentist context, statisticians have come to grief in failing to distinguish between the two cases of simulation and inference. For example, suppose in a randomized trial the baselines are observed with error (as must always be the case). By simulating from a known *true difference* between groups at baseline you can show that analysis of covariance using the observed baselines is conditionally biased [18]. However, this fact is completely irrelevant. It simply means that a particular analysis which conditions perfectly reasonably on that which is observed, could have been outperformed if more information, which is not and never could be available, had been [19]. Human inference does not match up to divine omniscience.

In other words, the simulation is unfair and irrelevant to any Bayesian who truly believed what the prior distributions represented. Of course, by the same token one could say that what Lambert *et al*. [1] are doing is providing the Bayesian analysis with a rigorous test. However, the reply would be that there is no point in expecting a Bayesian analysis to pass this test. The whole point about the Bayesian approach is that it is subjective. Bayesians are bound to disagree since there is no such thing as necessary agreement [20].

As a minor point, it is also worth noting that the simulation also mismatches prior belief in the opposite direction as regards within-study standard errors. These will vary randomly from study to study. Yet the formulation adopted for them in the model implemented in BUGS treats them as being known parameters, which is to say with a perfectly informative prior distribution. In fact, it is rather curious that of the six unknown variances for this example, it is only the random effects variance that is treated as unknown. For small trials, treating standard errors as if they were known overestimates precision and can also lead to a loss of efficiency [10, 11, 14]. This is a problem with conventional frequentist analyses, whether fixed effect or random.

Yet a further minor point regarding the simulation is that whereas the estimation procedure, by reducing the data to estimates per trial, implicitly treated the main effect of trials as fixed (this is the part random model 5.3, given by Senn [11]) which is equivalent to using uninformative non-hierarchical prior distributions on their values, the simulation procedure (as noted above) treated them as homogenous, which is equivalent to an informative prior distribution that the effects are zero. However, this disparity will have a very minor impact on inferences and is unimportant.

A possible frequentist analysis would be to apply the model given by (7) to the data in Table II but with the addition of a fixed main effect per trial and implement it, for example, in proc nlmixed® of SAS®. Such a model (which is a form of model 5.1 in Senn [11]) yields estimates for Θ of 0.39 with a standard error of 0.20 and 95 per cent confidence limits of -0.18 to 0.96, which are slightly wider than those of -0.12 and 0.95 for a point estimate of 0.41 with a standard error of 0.22 given by the method of Hardy and Thompson [16] applied to the summary statistics. Alternatively, the hierarchical generalized linear model (HGLM) approach of Lee and Nelder [21] implemented in GenStat® and applied to the data in Table II yields a point estimate of 0.41 with a standard error of 0.25. For SAS, the estimate of τ^2 is 0.10, as it is for the HGLM approach, whereas for the Hardy and Thompson [16] method it is 0.13.

DOI: 10.1002/sim

In fact, it is not necessary to simulate to expose the problems with the 13 prior distributions. They *all* clearly have unsatisfactory features.

5. PRIOR DISTRIBUTIONS

The magnificent logical development that was the subjective Bayesian theory put together by Ramsey [22], DeFinetti [23, 24], Savage [25], Lindley [26] and others was surely one of the great intellectual achievements of the last century. (Good [27] also belongs to this list, and also has some very practical and eclectic tendencies that set him apart.) In fact, of the four systems of inference identified by Barnard [28] (the other three being Fisherian, Neyman-Pearson and that of Jeffreys) it seems to me that it is in many ways the most impressive [20]. The important work of Gelfand and Smith [29] and others since 1990, including the authors of BUGS [2], has turned Bayesian statistics from being an extremely insightful way of *looking* at problems to being a potentially powerful way of *analysing* them. However, the programme of this system is coherence and it is surely a matter of coherent logic that there is little point in coherently updating priors that are themselves incoherent, yet this is what all 13 prior distributions considered by Lambert *et al*. [1] are, at least by the test of finding anyone who is prepared to use them as the basis for bets. These prior distributions have already been listed in Table I.

I agree with Lambert *et al*. [1] that vague prior distributions are problematic. Such prior distributions are *insufficiently informative*. The main problem with a frequentist random effects meta-analysis of a small collection of trials is that it is liable to produce an estimated value of τ^2 that nobody believes. A Bayesian analysis cannot rescue this situation by using or attempting to find uninformative priors. Take for example, prior distribution 3. Any value of $\log(\tau^2)$ between -10 and 10 is deemed equally likely and the median is 0 and the quartiles are -5 and $+5$. On the scale of τ, these five figures correspond (to two significant figures) to 0.0067, 0.082, 1, 12, 150. Now imagine that you have to bet a large sum of money on the result of a meta-analysis of a huge collection of trials on the log-odds scale, choosing one of the four intervals between these numbers. Can there be any statistician on the planet who believes that there is nothing to choose between the interval 0.082–1 and 12–150?

A further problem arises for meta-analyses that do not involve binary outcomes. Then neither Θ nor τ will be a unit-free quantity. This implies that no automatic specification of priors without careful reflection on scale will be possible but careful reflection on this point leads to the conclusion that a similar phenomenon must apply to some extent for the analysis of binary outcomes.

In my view, although 'uninformative' priors can be justified for the parameter of main interest, in this case Θ, in the spirit of seeing what the data will show, provided that result is not taken too seriously, the case with nuisance parameters is different. They are just too far from the problem of direct interest for the subjunctive and artificial nature of what is being fed into the analysis to be taken on board when the results are considered. One is likely to be misled. The analysis of the AB/BA cross-over is a case point. If you put an uninformative prior distribution on the carry-over parameter then you end up using the first period data only for the estimate of the treatment effect and you should not have run a cross-over in the first place [30]. Other examples abound. Any Bayesian regression model implicitly has an infinite number of *informative* priors declaring all factors not in the model to have zero effect. No progress in inference could be made otherwise.

One of the 13 prior distributions is particularly inappropriate. It is number 13, originally proposed by DuMouchel and Normand [31]. It uses the square root of the harmonic mean of the within-study

DOI: 10.1002/sim

variances of the treatment effects as the parameter for the logistic distribution. This is illogical since it implies that ones prior belief about the *true* variation in effects between studies depends on how precisely the individual within-study effects have been measured. See also Skene and Wakefield [32] for another example of prior distributions inappropriately based on observed results.

All 13 prior distributions on τ^2 are independently specified from the prior distribution for the treatment effect, Θ, itself but no applied statistician believes that the two are independent.

Can anything be done? I suspect it will take a great deal of hard work to come up with anything reasonable. I offer without any great enthusiasm, as a tentative possibility, a *conditional* prior distribution of the form

$$f(\tau|\Theta;\beta)=\frac{1}{\beta|\Theta|}\exp\left\{-\frac{\tau}{\beta|\Theta|}\right\} \tag{8}$$

This is a form of exponential that allows dependence of τ on Θ. However, implementation of (8) may be problematic since the probability goes to zero as $|\Theta|$ goes to zero. A possible modification might be

$$f(\tau|\Theta;\beta,\alpha)=\frac{1}{\alpha+\beta|\Theta|}\exp\left\{-\frac{\tau}{\alpha+\beta|\Theta|}\right\} \tag{9}$$

with α positive but small, although this has an objection from a frequentist point of view that it allows for the possibility of an interaction when there is no main effect [33].

Figure 3 is a plot of the resulting posterior distributions for Θ applying the conditional prior distribution for τ given by (9) and assuming a locally uniform prior [34] for Θ. This is calculated as follows. First the joint probability of τ and the data, conditional on Θ is formed as the product of (9) and the random effects likelihood [16] given by (10) below

$$\exp\left(\sum_{i=1}^{k}\left\{-\frac{1}{2}\ln[2\pi(v_i+\tau^2)]-\frac{(\theta_i-\Theta)^2}{2(v_i+\tau^2)}\right\}\right) \tag{10}$$

Second, from the resulting joint density, the nuisance parameter τ^2 is integrated out. (Here, calculations have been done using numerical integration in Mathcad® 12 [35], which has been used for most of the calculations in this note.) Third, this integrated likelihood is rescaled to form a reference posterior distribution by dividing by its integral over plausible ranges of Θ. This is equivalent to assuming uniform prior support for the integrated (over τ^2) likelihood in the domain of Θ.

From Figure 3, it can be seen that as the value of α or of β increases the distribution becomes more diffuse. The posterior distribution for the case $\alpha=0.05$, $\beta=1$ given by the solid line will be seen to have a definite kink at the value $\Theta=0$ presumably because of the strong dependence of the prior value of τ on Θ. In practice, choosing values of α and β will be difficult. Values of $\beta<1$ would seem generally appropriate. Small values of α would also be appropriate. However, if values of $\alpha=0.05$, $\beta=0.5$ are used, then a 95 per cent posterior credible interval for Θ of (0.10, 1.03)) with a posterior maximum of 0.41 results, which is much narrower than any of those considered by Lambert *et al*. or, indeed, the frequentist random effects limits. One way of pushing anxiety about β to a further remove would be to have a prior distribution on its value at a higher level of the hierarchy. This general approach was proposed by Good some 40 years ago in connection with the analysis of multinomial distributions [36]. (See also, Reference [27, Chapter 9].)

DOI: 10.1002/sim

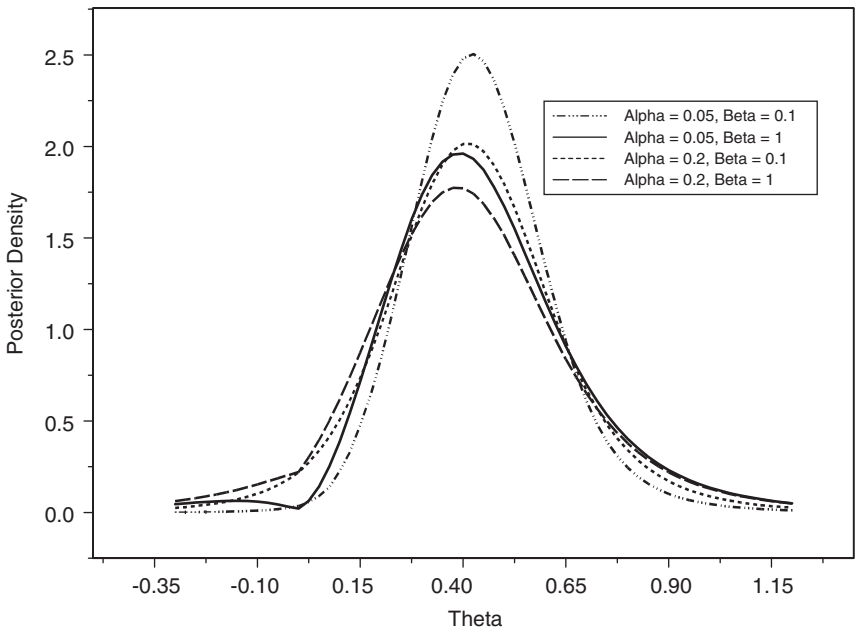

Figure 3. Marginal posterior distributions for Θ for four prior distributions for τ of the type given by (9).

However, I repeat that it is far from clear that a prior distribution of the form given by (9) would be a good one to use. It has the advantage from the point of view of estimating τ^2 of linking this to Θ but the reverse effect also applies and it may be that this is not such a good idea and in any case it may incorporate implausible assumptions about the relationship between τ and Θ.

6. CONCLUSIONS AND RECOMMENDATIONS

Based on the previous examination I offer the following recommendations.

First, it is always valuable to perform a fixed effects meta-analysis. This tests the null-hypothesis that treatments were identical in all trials. When and if this is rejected, then the alternative hypothesis that may be asserted is, 'there is at least one trial in which the treatments differed'. To go beyond this causal 'finding' requires strong assumptions [11]. In any case, any person who insists that the only valuable inference from a meta-analysis is a random effects one is incapable of drawing any useful inferences from a single trial, however large and well planned, unless informative prior distributions on the random effect variance exist. That being so, the position of simultaneously holding that fixed effects meta-analyses are inappropriate but that 'uninformative' prior distributions for random effect variances must be used in connection with random effects meta-analyses seems untenable.

Second, if it is decided that a wider purpose should be met, that of saying something about trials in general, it may become sensible to perform a random effects meta-analysis but not until something is understood about the differences between the various trials in the meta-analysis

DOI: 10.1002/sim

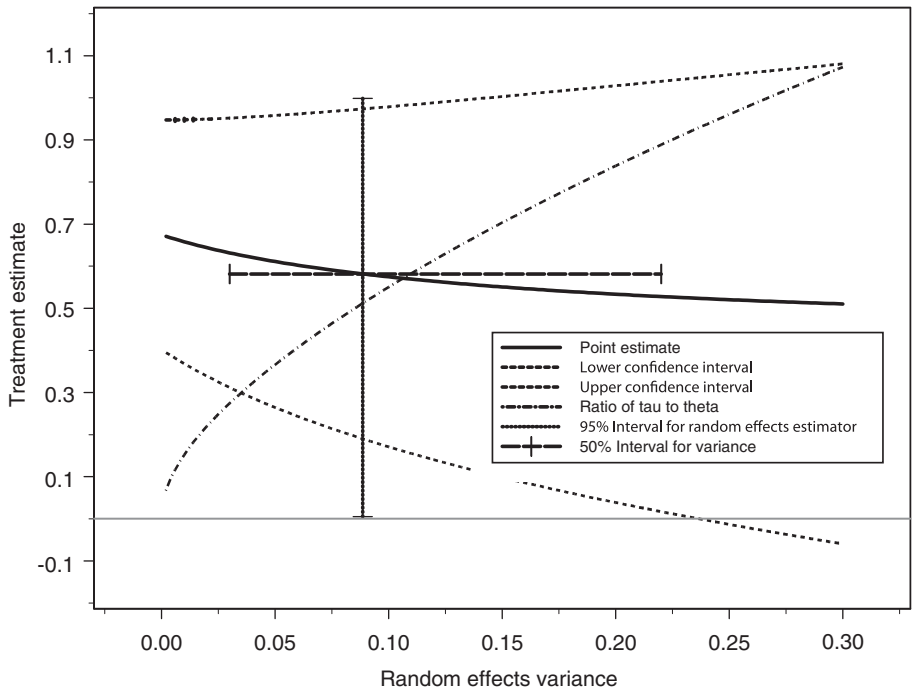

Figure 4. Artificial example created from that given by Lambert *et al*. (see text). Point estimates for Θ and upper and lower confidence intervals are given as a function of the random effects variance τ^2. Profile likelihood 95 per cent confidence limits for Θ are given, as are 50 per cent limits for τ^2 and the ratio τ/Θ.

in terms of their purpose. There seems to be generally a great deal of confusion about what meta-analysis can deliver here.

Third, if a random effects meta-analysis is desired, it will be valuable to attempt a frequentist analysis of this. There are a number of possible approaches. I personally like the Hardy and Thompson [16] approach and the H-Likelihood system of Lee and Nelder [21].

Fourth, it should be appreciated that if a random effects analysis is performed the usual approach delivers a prediction for the true effect for a future random trial of the sort already run. This is not at all the same as the true effect for a future randomly chosen patient, since trials already run may not be exchangeable with each other and may not be exchangeable with the target patient population [10, 11].

Fifth, some sort of diagnostic presentation of the results may be useful. For example, Figure 4 gives points estimates and confidence intervals for a modified example of the Kozyrskyj [4] meta-analysis to make it more interesting. The modification is to rearrange the standard errors so that there is a rank correlation of -1 with the point estimates. In the figure are plotted point estimates and 95 per cent confidence intervals for the treatment estimate as a function of the random effects estimate, assumed known. Also given is the 95 per cent profile likelihood confidence interval for the point estimate. In addition, two further aids to interpretation of the results are provided. The first is a 50 per cent confidence interval for the random effects variance. This narrower limit has

DOI: 10.1002/sim

been chosen to reflect the nuisance parameter status of the variable on the assumption that it is not of direct interest. The second is the ratio of random effects standard deviation to the corresponding point estimate. The value of this, being a dimensionless number, is not commensurate with the other quantities and it is somewhat fortuitous that it can be plotted on the same graph. More generally, a further right-hand axis would be needed.

Sixth it is unreasonable to expect a Bayesian analysis to produce some automatic answer that is acceptable to all as the relevant posterior distribution to use. Indeed a Bayesian analysis is effectively unusable by anyone not sharing the prior distribution that has been used to produce it. For example, although Lambert *et al*. were able to use the frequentist summary statistics for the five trials as the raw input to their Bayesian meta-analysis, *they would not have been able to use Bayesian posterior summaries from these trials as inputs* [30] and it is rather ironical that of all of the various types of summary of a trial that can be used as input to a Bayesian meta-analysis one of the least usable is a Bayesian analysis. The Bayesian theory is a theory of subjective inference. Many Bayesians would argue very strongly that this is the only theory that makes sense. The case they make is impressive.

Seventh, this is not to argue that the sort of technically Bayesian analyses considered by Lambert *et al*. [1] are not useful provided that they are not confused with the sort of Bayesian analysis a decision maker ought to use for personal purposes. As Lambert *et al*. put it 'In addition to the philosophical advantages of the Bayesian approach, the use of these methods has led to increasingly complex, but realistic, models being fitted' (p. 2402). The technical advantages of Markov chain Monte Carlo methods, especially when implemented in user-friendly software such as BUGS, is permitting users to tackle problems of greater complexity than is currently possible using frequentist mixed models are, indeed, considerable. However it seems to me that the philosophical advantages belong to the truly subjective Bayesian approach and have little to do with the sorts of analyses illustrated by Lambert *et al*. and that however complex these models are, they are only realistic in their treatment of likelihoods as opposed to prior distributions. It may, indeed, be useful to perform a range of possible 'Bayesian' analyses for sensitivity purposes, rather in the spirit of Spiegelhalter *et al*. [37]. I am firmly convinced, however, that where this is done it is the comparison of the Bayesian solutions with each other that is useful, not their evaluation compared to some simulated 'reality'.

ACKNOWLEDGEMENTS

I am extremely grateful to Andy Garrett, Agostino Nobile, Youngjo Lee and an anonymous referee for helpful comments on an earlier draft and to John Nelder and Roger Payne for advice on HGLMs. The responsibility for content of the paper and views expressed is mine.

REFERENCES

1. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine* 2005; **24**:2401–2428.
2. Spiegelhalter D, Thomas A, Best NG. *WinBUGS*, *Version 1.4*, *User Manual*, MRC Biostatistics Unit: Cambridge, 2001.
3. Glasziou PP, Del Mar CB, Sanders SL, Hayem M. Antibiotics for acute otitis media in children. *Cochrane Database Syst Rev* 2004: CD000219.
4. Kozyrskyj AL, Hildes-Ripstein GE, Longstaffe SE, Wincott JL, Sitar DS, Klassen TP, Moffatt ME. Short course antibiotics for acute otitis media. *Cochrane Database Syst Rev* 2000: CD001095.

DOI: 10.1002/sim

5. Hoberman A, Paradise JL, Burch DJ, Valinski WA, Hedrick JA, Aronovitz GH, Drehobl MA, Rogers JM. Equivalent efficacy and reduced occurrence of diarrhea from a new formulation of amoxicillin/clavulanate potassium (Augmentin) for treatment of acute otitis media in children. *Pediatric Infectious Disease Journal* 1997; **16**:463–470.
6. Boulesteix J, Dubreuil C, Moutot M, Rezvani Y, Rosembaum. Cefpodoxime proxetil 5 jours versus cexime 8 jours, dans le traitement des otites moyennes aigues de l'enfant. *Medicine Maladies Infectieuses* 1995; **25**:534–539.
7. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. *Journal of Clinical Epidemiology* 1991; **44**:1271–1278.
8. Jadad AR, Moher M, Browman GP, Booker L, Sigouin C, Fuentes M, Stevens R. Systematic reviews and meta-analyses on treatment of asthma: critical evaluation. *British Medical Journal* 2000; **320**:537–540.
9. Senn SJ. Review is biased. *British Medical Journal* 2000; **321**:297.
10. Senn SJ. *Statistical Issues in Drug Development*. Wiley: Chichester, 1997.
11. Senn SJ. The many modes of meta. *Drug Information Journal* 2000; **34**:535–549.
12. Thompson SG. Systematic review—why sources of heterogeneity in metaanalysis should be investigated. *British Medical Journal* 1994; **309**:1351–1355.
13. Cox DR, Solomon PJ. *Components of Variance*. Chapman & Hall: Boca Raton, 2003.
14. Yates F, Cochran WG. The analysis of groups of experiments. *Journal of Agricultural Science* 1938; **28**:556–580.
15. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; **7**:177–188.
16. Hardy RJ, Thompson SG. A likelihood approach to meta-analysis with random effects. *Statistics in Medicine* 1996; **15**:619–629.
17. Senn SJ, Stevens L, Chaturvedi N. Repeated measures in clinical trials: simple strategies for analysis using summary measures. *Statistics in Medicine* 2000; **19**:861–877.
18. Chambless LE, Roeback JR. Methods for assessing difference between groups in change when initial measurements is subject to intra-individual variation. *Statistics in Medicine* 1993; **12**:1213–1237.
19. Senn SJ. Methods for assessing difference between groups in change when initial measurement is subject to intra-individual variation [letter; comment] [see comments]. *Statistics in Medicine* 1994; **13**:2280–2285.
20. Senn SJ. *Dicing with Death*. Cambridge University Press: Cambridge, 2003.
21. Lee Y, Nelder JA. Hierarchical generalized linear models. *Journal of the Royal Statistical Society Series B—Methodological* 1996; **58**:619–656.
22. Ramsey F. Truth and probability. In *The Foundations of Mathematics and other Logical Essays*, Braithwaite RB. (ed.). Harcourt Brace and Company: New York, 1931; 156–198.
23. de Finetti BD. *Theory of Probability*, vol. 1. Wiley: Chichester, 1974.
24. de Finetti BD. *Theory of Probability*, vol. 2. Wiley: Chichester, 1975.
25. Savage J. *The Foundations of Statistics*. Wiley: New York, 1954.
26. Lindley DV. Theory and practice of Bayesian statistics. *Statistician* 1983; **32**:1–11.
27. Good IJ. *Good Thinking*: *The Foundations of Probability and its Applications*. University of Minnesota Press: Minneapolis, 1983.
28. Barnard GA. Fragments of a statistical autobiography. *Student* 1996; **1**:257–268.
29. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**:398–409.
30. Senn SJ. Consensus and controversy in pharmaceutical statistics (with discussion). *The Statistician* 2000; **49**:135–176.
31. DuMouchel W, Normand SL. Computer modelling and graphical strategies for meta-analysis. In *Meta-Analysis in Medicine and Health Policy*, Stangl D, Berry DA (eds). Marcel Dekker: New York, 2000, 127–178.
32. Skene AM, Wakefield JC. Hierarchical models for multicentre binary response studies. *Statistics in Medicine* 1990; **9**:919–929.
33. Senn SJ. Added values: controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine* 2004; **23**:3729–3753.
34. Lee PM. *Bayesian Statistics*: *An Introduction*. Edward Arnold: London, 1989.
35. Mathsoft. Mathcad 12, Mathsoft, Cambridge, MA, 2004.
36. Good IJ. A Bayesian significance test for multinomial distributions (with discussion). *Journal of the Royal Statistical Society Series B* 1967; **29**:399–431.
37. Spiegelhalter DJ, Freedman LS, Parmar MKB. Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine* 1993; **12**:1501–1511 (discussion 1513–1507).

DOI: 10.1002/sim