STATISTICS IN MEDICINE
Statist. Med. 2001; **20**:435–452

Bayesian methods for cluster randomized trials with continuous responses

David J. Spiegelhalter*

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.

SUMMARY

Bayesian methods for cluster randomized trials extend the random-effects formulation by allowing both the use of external evidence on parameters and straightforward relaxation of the standard normality and constant variance assumptions. Care is required in specifying prior distributions on variance components, and a number of different options are explored with implied prior distributions for other parameters given in closed form. Markov chain Monte Carlo (MCMC) methods permit the fitting of very general models and the introduction of parameter uncertainty into power calculations. We illustrate these ideas using a published example in which general practices were randomized to intervention or control, and show that different choices of supposedly 'non-informative' prior distributions can have substantial influence on conclusions. We also illustrate the use of forward simulation methods in power calculations with uncertainty on multiple inputs. Bayesian methods have the potential to be very useful but guidance is required as to appropriate strategies for robust analysis. Our current experience leads us to recommend a standard 'non-informative' prior distribution for the within-cluster sampling variance, and an independent prior on the intraclass correlation coefficient (ICC). The latter may exploit background evidence or, as a reference analysis, be a uniform ICC or a 'uniform shrinkage' prior. Copyright © 2001 John Wiley & Sons, Ltd.

1. INTRODUCTION

Methodology for the design and analysis of cluster randomized trials focuses on ways of dealing with the dependence introduced between subjects which have been randomized to an intervention as a group or cluster [1]. As in other areas in which dependencies must be taken into account, there are two broad approaches to analysis; see, for example, Liang and Zeger [2] in the context of repeated measures studies. The first is the population-average or marginal approach which essentially adjusts the variance of the estimate of the treatment effect for the clustering, while the second, cluster-specific or conditional, approach explicitly models the variability between the clusters. For outcomes following a generalized linear model with linear link function, so that the estimated treatment effect is directly proportional to the difference in expectation of the response between individuals given different interventions, there is no difference between the two approaches in the interpretation of the estimated coefficients.

* Correspondence to: David J. Spiegelhalter, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.

It is natural to embed the cluster-specific approach within the large literature on random-effects modelling [1], whether traditional mixed-model analysis of variance [3] or the more general area known as multilevel modelling [4–6]. In this paper we extend these hierarchical models by taking a Bayesian approach to the unknown parameters, thus allowing substantive knowledge external to a study to be used in both its design and analysis; such an extension is reasonably straightforward, and it is remarkable that there appear to be no published references on full Bayesian analyses of cluster randomized trials. This paper presents a basic analysis for continuous responses and should ideally be read in conjunction with Turner *et al.* [7] who focus on binary outcome data: they relate the between-cluster variance to the intraclass correlation coefficient, and discuss the choice of prior distributions and more general aspects of Bayesian modelling in this context.

In Section 2 we provide a very brief introduction to the use of Bayesian methods in randomized trials and their potential for use in clustered designs with continuous responses. Section 3 describes a published motivating example, while Section 4 examines the delicate issue of specifying priors for variance components in this context. The results for the example are given in Section 5, while Section 6 describes how uncertainty about the magnitude of plausible intraclass correlations can be incorporated into power calculations. Finally, Section 7 summarizes advantages and disadvantages of the methods covered in the paper.

Appendix A provides details of the form of the derived prior distributions. Bayesian approaches often do not permit closed-form or even iterative methods for estimating effects and their errors, and Appendix B describes the simulation approaches used for computation.

2. BAYESIAN APPROACHES TO RANDOMIZED TRIALS

Suppose we have data y and unknown quantities θ which might be model parameters, missing data, or events we did not observe directly such as censored survical times. Standard statistical methods base inferences on an assumed likelihood $p(y \mid \theta)$. Bayesian methods add to this a prior distribution $p(\theta)$ which expresses our uncertainty about θ before taking into account the data.

This development extends the use of random-effects or mixed models in clustered randomized trials, in which data y_i within each cluster i are assumed to depend on parameters θ_i, which in turn are assumed drawn from some population distribution with parameters ψ. Multilevel modelling or empirical Bayes techniques then seek to estimate ψ by likelihood based methods, whereas the full Bayesian extension is to assume a 'hyperprior' distribution for ψ and so form a 'full probability model' for both observations and parameters. The prior distribution should be based on evidence or judgement external to the study in question, and in cluster randomized trials a natural source is previous experience of intraclass correlation coefficients in similar contexts.

A posterior distribution $p(\theta \mid y)$ expresses the appropriate uncertainty for the unknown quantities after taking into account the data, and is obtained via Bayes theorem

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{\int p(\theta)p(y \mid \theta)\,\mathrm{d}\theta}$$

$$\propto p(\theta)p(y \mid \theta)$$

Obtaining a posterior distribution can be computationally difficult, largely since nuisance parameters have to be integrated out rather than maximized. Markov chain Monte Carlo (MCMC) simulation methods have largely taken over from analytic approximations and are used in this paper.

There have been a number of recent books on Bayesian methods in general [8–10], as well as biostatistics in particular [11], while review papers on Bayesian methods in clinical trials have focused on standard randomized designs [12–14].

3. AN EXAMPLE

Kinmonth *et al.* [15] describe a trial in which 41 general practices were randomized to receive additional training in patient-centred diabetes care or not. Here we just consider body mass index (BMI) at one year as a continuous response, which is available for 240 patients in 38 practices (18 in intervention group, 20 in control). Patient-level and practice-level covariates were recorded but are not explicitly considered in the analysis below, although we shall indicate how they could be incorporated.

Preliminary plotting of the data is shown in Figure 1 at the patient and practice level. Figure 1(a) indicates some heterogeneity of within-practice variances (although some of the clusters with high interquartile range contain very few cases) while (b) suggests some evidence of positive skewness in the distribution of residuals, with a few outliers. Figures 1(c) and (d) support a normality assumption for the practice effects, with a suggestion of higher BMI responses in the intervention group.

A reasonable initial model, following as far as possible the notation of Turner *et al.* [7], is to assume that the BMI measurement y_{ij} for patient j in practice i has a distribution

$$y_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$$
$$\mu_{ij} = u_i \tag{1}$$

The practice-specific random effects are assumed to have a normal distribution

$$u_i \sim \mathrm{N}(\phi_i, \sigma_u^2) \tag{2}$$
$$\phi_i = \alpha + \beta x_i \tag{3}$$

where x_i represents the treatment given to the ith practice, coded as $x_i = -0.5$ for control and $x_i = 0.5$ for intervention. This simple two-level variance component model can be written in many ways: this particular form has been chosen to illustrate how patient-level and practice-level covariates may be simply introduced into equation (1) and equation (3), respectively.

There are many potential elaborations to this basic model. First, as illustrated by Turner *et al.* [7], we might relax the normality assumption at either the patient or practice level; Figure 1 provides limited support for this and it will not be pursued in this analysis. Second, again as illustrated by Turner *et al.* [7], we could relax the assumption of constant within-practice variances and common between-practice variances in the intervention and control groups. Figure 1(a) provides evidence that the former may be reasonable and we shall explore

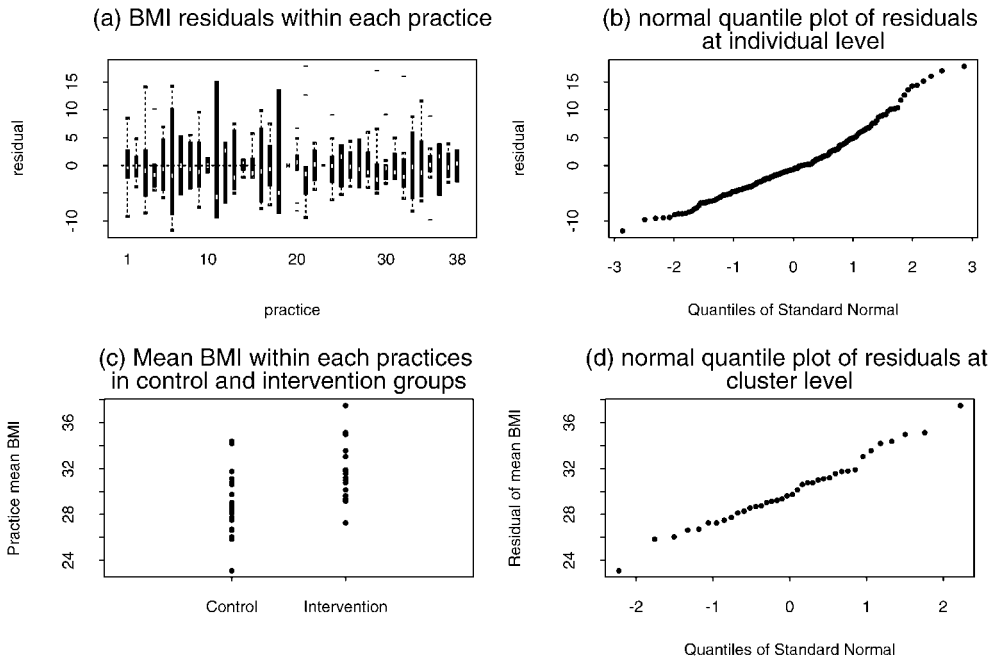

Figure 1. Preliminary plots of BMI data. (*a*) The unstandardized patient-level residuals formed by subtracting the practice means from the BMI — the shaded area is the interquartile range. (*b*) These residuals plotted against the quantiles of a standard normal distribution. (*c*) The BMI practice means within the control and intervention groups. (*d*) The practice-level residuals formed by subtracting the overall mean BMI from each of the practice means.

its effect later. However, the primary focus of this paper is on the appropriate prior distributions for the variance components σ^2 and σ_u^2.

4. PRIOR DISTRIBUTIONS FOR PARAMETERS IN CLUSTER RANDOMIZED TRIALS

The choice of prior distribution in a Bayesian analysis can be controversial. Ideally there should be an opportunity to include substantive external knowledge in an explicit and intuitive way, but also it is useful to have a suitable formulation for a 'non-informative' or 'reference' prior distribution for the parameters. For the regression coefficients α and β such a prior can be taken as a locally uniform distribution over a range that covers all plausible values, such as assuming

$$\alpha \sim \text{Unif}(-10000, 10000)$$

$$\beta \sim \text{Unif}(-10000, 10000)$$

More controversial is an appropriate prior for the variance components σ^2 and σ_u^2. The standard 'non-informative' prior distribution for a measurement variance σ^2 is to assume that $\log \sigma^2$ is uniformly distributed, which is equivalent to assuming that $p(\sigma^k) \propto \sigma^{-k}$ for any power k. This prior is 'improper', in the sense that it does not integrate to one. In single-level regression models, such a prior, combined with uniform priors for regression coefficients, leads to a Bayesian analysis giving equivalent results to the traditional approach [8], and in particular provides inferences that are scale-invariant. However, it is not always appreciated that assuming such a prior for the between-cluster variance σ_u^2 would lead to an improper posterior distribution for σ_u^2 (Reference [8], p. 140), and hence some care is required.

There is a substantial literature on prior distributions for variance components [16] and intraclass correlations coefficients [17], focusing particularly on choice of appropriate reference priors; in particular, Burch and Harris [17] review a number of suggestions and make recommendations.

We consider two groups of priors. From the relationship

$$\text{ICC} = \rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

it can be seen that placing independent prior distributions on σ^2 and σ_u^2 will imply a prior distribution for ρ, and it is important that this induced prior appears reasonable. This approach is taken in the first four of the options examined below.

The relationship may also be written as

$$\log \sigma_u^2 = \log \sigma^2 + \text{logit}(\rho) \tag{4}$$

which emphasizes how independent priors placed on ρ and σ^2 would imply a prior distribution on σ_u^2. For example, Burch and Harris use a pivotal likelihood argument to recommend a standard reference prior $p(\sigma^2) \propto \sigma^{-2}$ combined with an independent prior for ρ. The last four of our options illustrate this approach.

Eight prior options will therefore be investigated:

(a) and (b) *Inverse gamma priors for* σ^2 *and* σ_u^2. We may adopt 'just' proper priors for σ^2 and σ_u^2, by using the conjugate forms in which σ^{-2} and σ_u^{-2} are assumed to have gamma$(\varepsilon, \varepsilon)$ distributions, which approaches the 'non-informative' prior as ε tends to 0. Thus

$$p(\sigma^2) \propto \sigma^{2(-\varepsilon-1)} e^{-\varepsilon/\sigma^2}, \quad 0 < \sigma^2 < \infty$$

$$p(\sigma_u^2) \propto \sigma_u^{2(-\varepsilon-1)} e^{-\varepsilon/\sigma_u^2}, \quad 0 < \sigma_u^2 < \infty$$

which implies (Appendix A, (a)) that the ICC has a beta$(\varepsilon, \varepsilon)$ distribution

$$p(\rho) \propto \rho^{\varepsilon-1}(1-\rho)^{\varepsilon-1}, \quad 0 < \rho < 1$$

Figure 2 (*a*) shows these priors for $\varepsilon = 0.001$. These priors are fairly uniform for most of the range but then strongly support values of σ_u^2, σ^2 that are very near 0 and values of ρ very near 0 or 1. Kass and Wasserman [18] have warned that such priors, while admittedly avoiding an improper posterior distibution, may still result in unreasonable prior sensitivity. To investigate this we also consider $\varepsilon = 0.1$, as shown in Figure 2(*b*).

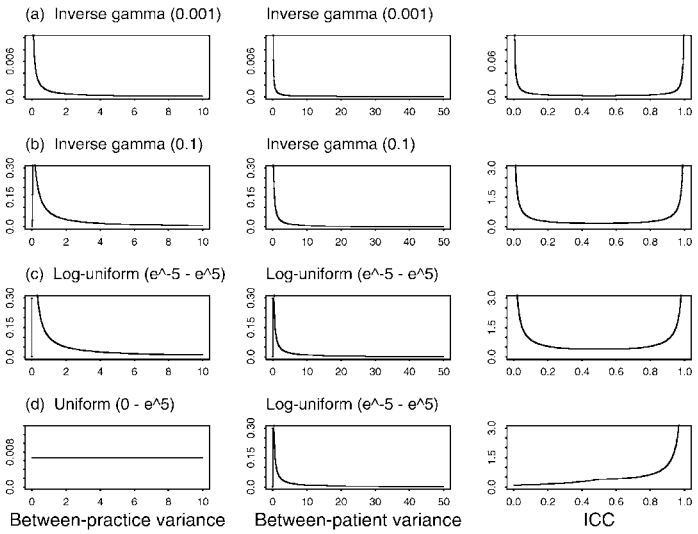

Figure 2. Four options (*a*) to (*d*) for prior distributions in which the two variance components are assumed *a priori* independent. The columns show the prior distributions for the between-practice variance σ_u^2, the between-patient (within-cluster) variance σ^2, and the intraclass correlation coefficient ICC $= \sigma_u^2/(\sigma_u^2 + \sigma^2)$. For each option, labels are specified for those given independent priors, and the prior for the remaining parameter is derived using the analysis shown in Appendix A. Note that the vertical scale is not kept fixed.

(c) *Log-uniform for* σ^2 *and* σ_u^2. An alternative means of making the prior distributions (and hence posterior distributions) proper is to let $\log \sigma^2$ and $\log \sigma_u^2$ be uniform on a bounded range $(-\log t, \log t)$, which we shall define as a log-uniform distribution on $(1/t, t)$. Thus

$$p(\sigma^2) \propto \sigma^{-2}, \quad 1/t < \sigma^2 < t$$

$$p(\sigma_u^2) \propto \sigma_u^{-2}, \quad 1/t < \sigma_u^2 < t$$

which implies (Appendix A, (c)) a prior distribution

$$p(\rho) \propto \frac{1 - |\text{logit}\,\rho|/\log t^2}{\rho(1-\rho)}, \quad \frac{1}{1+t^2} < \rho < \frac{1}{1+t^{-2}}$$

This will tend to $p(\rho) \propto \rho^{-1}(1-\rho)^{-1}$ for large t. Figure 2(*c*) shows these priors for a somewhat arbitrarily chosen value $\log t = 5$, $t = 148.5$. These priors support, to a very similar extent to Figure 2(*b*), values of σ_u^2, σ^2 that are near 0 and values of ρ near 0 or 1.

(d) *Uniform on* σ_u^2. Gelman *et al.* (reference [8], p. 139) suggest adopting a uniform prior for σ_u^2 while retaining the standard 'non-informative' prior for σ^2. A proper version of

these priors with bounded ranges leads to

$$p(\sigma^2) \propto \sigma^{-2}, \quad 1/t < \sigma^2 < t$$

$$p(\sigma_u^2) \propto 1, \qquad 0 < \sigma_u^2 < t$$

which implies (Appendix A, (d)) a prior distribution

$$p(\rho) \propto \begin{cases} \frac{1}{(1-\rho)^2}(1 - \frac{1}{t^2}), & 0 < \rho < \frac{1}{2} \\ \frac{1}{\rho(1-\rho)}(1 - \frac{\rho}{t^2(1-\rho)}), & \frac{1}{2} < \rho < 1 \end{cases}$$

Figure 2(d) shows these priors for $t = e^5 = 148.5$. The implied prior for ρ tends to support values near 1.

(e) *Uniform on ICC.* It may not be reasonable to assume that our prior beliefs about σ^2 and σ_u^2 are entirely independent, since the scale of measurement should give us some information in common. It therefore appears appropriate to place independent prior distributions directly on the ICC and measurement variance, and hence derive a prior distribution for σ_u^2. If we adopt a log-uniform $(1/t, t)$ prior for σ^2 and a uniform prior for ρ on (0,1), we obtain

$$p(\sigma^2) \propto \sigma^{-2}, \quad 1/t < \sigma^2 < t$$

$$p(\rho) \propto 1, \qquad 0 < \rho < 1.$$

which implies (Appendix A, (e)) a prior distribution

$$p(\sigma_u^2) \propto \frac{1}{(t^{-1} + \sigma_u^2)(t + \sigma_u^2)}, \quad 0 < \sigma_u^2 < \infty$$

which tends to $p(\sigma_u^2) \propto \sigma_u^{-2}$ as t becomes large. This prior was used by Chaloner [19] and recommended by Burch and Harris [17], who point out that the REML estimate of the ICC is the mode of the resulting posterior distribution. Figure 3(e) shows these priors for $t = e^5 = 148.5$. The implied prior for σ_u^2 tends, to a moderate extent, to support values near 0.

(f) *Approximate Jeffreys.* In the case of a balanced design with constant m cases per cluster, Box and Tiao (reference [20], p. 251) suggest the appropriate Jeffreys prior is

$$p(\sigma_u^2, \sigma^2) \propto \sigma^{-2}(\sigma^2 + m\sigma_u^2)^{-2} \tag{5}$$

Consider a new parameter

$$s = \frac{\sigma^2}{\sigma^2 + m\sigma_u^2} = \frac{1 - \rho}{1 + (m-1)\rho} \tag{6}$$

This is known as the 'shrinkage' parameter [21], as it represents the extent to which the cluster-specific means are shrunk towards the overall mean in a random effects estimation model. In Appendix A,(f), we show that the Box and Tiao version of the Jeffreys prior is equivalent to assuming independent log-uniform distributions for σ^2 and s, so that $p(s) \propto 1/s$, $0 < s < 1$. Further developments in reference priors, without

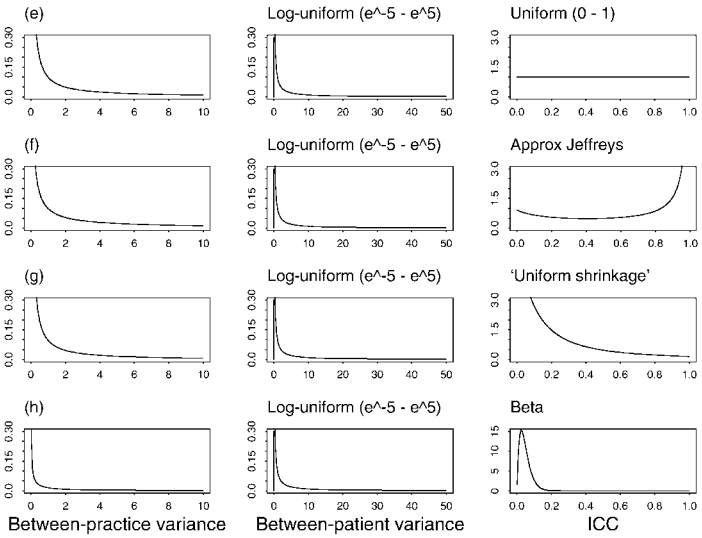

Figure 3. Four options (*e*) to (*h*) for prior distributions in which the ICC and the between-patient variance are assumed *a priori* independent.

making the independence argument, lead to somewhat complex forms depending on an order of importance for parameters [17, 22]. For unbalanced models the prior becomes even more unmanageable and non-intuitive. We therefore follow others [16, 21] in making an approximation to the Jeffreys prior, in this case by using the balanced form of the prior but with an average cluster size $\bar{m}$ substituted for the constant cluster size. Proper versions of the priors for σ^2 and s may be obtained by restricting their ranges to $(1/t, t)$ and $(1/t, 1)$, respectively, leading to marginal priors (Appendix A,(f))

$$p(\sigma^2) \propto \sigma^{-2}, \qquad 1/t < \sigma^2 < t$$

$$p(\rho) \propto \frac{1}{1-\rho}\,\frac{1}{1+(\bar{m}-1)\rho}, \qquad 0 < \rho < 1 - \bar{m}/(t+\bar{m}-1)$$

$$p(\sigma_u^2) \propto \frac{1}{\sigma_u^2}\,\log\frac{1+t\bar{m}\sigma_u^2}{1+t^{-1}\bar{m}\sigma_u^2}, \qquad 0 < \sigma_u^2 < (t-1)t/\bar{m}$$

Figure 3(*f*) shows these priors for $t = e^5 = 148.5$. The prior for ρ is somewhat unusual, but the implied prior for σ_u^2 is very similar to that of option (e).

(g) *Uniform shrinkage.* Daniels [16] and Natarajan and Kass [21] have suggested putting a uniform prior on the shrinkage parameter s in equation (6) with the aim of producing robust estimators with good sampling properties. In Appendix A,(g), we show that this implies a prior on ρ

$$p(\rho) \propto \frac{1}{(1+(\bar{m}-1)\rho)^2}, \quad 0 < \rho < 1$$

With an independent log-uniform$(1/t, t)$ prior for σ^2, the marginal distribution for σ_u^2 is

$$p(\sigma_u^2) \propto \frac{1}{(t^{-1} + \bar{m}\sigma_u^2)(t + \bar{m}\sigma_u^2)}, \quad 0 < \sigma_u^2 < \infty$$

We note the close connection between this prior and that obtained under option (e), and Figure 3(f) shows that for $t = e^5 = 148.5$ there is little difference in the implied priors for σ_u^2.

(h) *Informative beta for ICC.* Finally, we consider the situation in which external information exists about the ICC arising, say, from clinical trials carried out in a similar context. A beta(g, h) prior distribution, with an independent log-uniform$(1/t, t)$ prior for σ^2, leads to

$$p(\sigma^2) \propto \sigma^{-2}, \qquad 1/t < \sigma^2 < t$$

$$p(\rho) \propto \rho^{g-1}(1-\rho)^{h-1}, \quad 0 < \rho < 1$$

which implies (Appendix A) a prior distribution

$$p(\sigma_u^2) \propto \sigma_u^{-2} \left(F_{g,h}\left[\frac{\sigma_u^2}{t^{-1} + \sigma_u^2}\right] - F_{g,h}\left[\frac{\sigma_u^2}{t + \sigma_u^2}\right] \right), \quad 0 < \sigma_u^2 < \infty$$

where $F_{g,h}$ denotes the cumulative probability function of a beta(g, h) distribution. For example, suppose previous evidence suggested that an ICC of around 0.05 would be reasonable, and that an ICC above 0.15 would be quite surprising. It turns out that a beta$(2, 38)$ distribution has these characteristics, as shown in Figure 3(h) (Assessing g and h is helped by knowing that the mean of the beta distribution is $m = g/(g+h) = 0.05$, and the standard deviation is $s = \sqrt{m(1-m)/(g+h+1)} = 0.034$.) The implied prior for σ_u^2 tends to support values near 0.

5. RESULTS FOR THE EXAMPLE

Figure 4 shows the results for the eight prior options outlined in Section 4, calculated using the methods described in Appendix B. The choice has little influence on the overall intercept α and the within-practice variance σ^2; in the latter case it is clear that the strong evidence regarding σ^2 easily overcomes the prior weight towards values near 0, and further analyses reveal these conclusions are insensitive to choice of t. However, as may have been predicted from Figure 2 and Figure 3, priors (a) (b) (c), and to a lesser extent (g), lead to substantially smaller estimates of the ICC, corresponding to smaller estimates of the between-practice variance σ_u^2, and tighter confidence intervals for the treatment effect β. Futhermore, additional analyses (not shown) indicate option (c) is sensitive to the choice of t in the prior for σ_u^2. Prior options (d) and (e), which do not favour low values of the ICC, lead to wider intervals for β.

Table I shows the results for β and the intraclass correlation coefficient ρ. In spite of the considerable variation in the estimates of the ICC and β, the 'significance' of β, that is, $p(\beta > 0 \mid y)$, does not vary materially. The classical analysis carried out by Kinmonth *et al.* [15] led to a one-sided p-value for β of 0.015 and an estimated ICC of 0.045

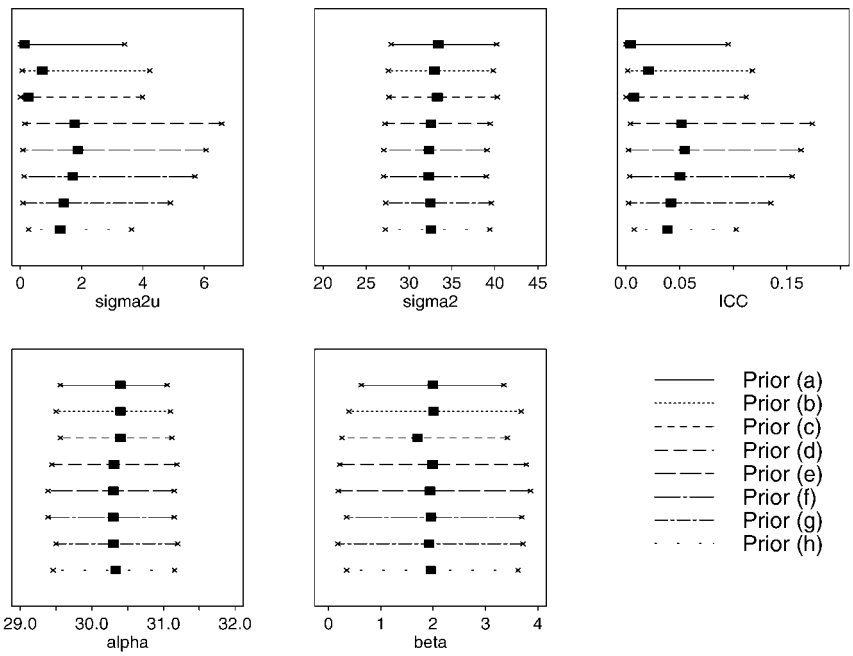

Figure 4. Posterior medians and 95 per cent intervals resulting from five choices of prior distributions: `sigma2u` $= \sigma_u^2 =$ between-practice variance; `sigma2` $= \sigma^2 =$ between-patient within-practice variance, `ICC` $= \rho =$ intraclass correlation coefficient, `alpha` $= \alpha =$ mean overall response, `beta` $= \beta =$ treatment effect.

Table I. Results for alternative prior specifications. Options (a) to (d) place independent priors on the variance components σ^2 and σ_u^2, while (e) to (h) place independent priors on σ^2 and the ICC.

Summary of prior	Inference on β			Inference on ICC	
	Estimate	95 per cent interval	$p(\beta > 0 \mid y)$	Estimate	95 per cent interval
(a) Gamma $(0.001, 0.001)$	1.99	0.63 to 3.35	0.005	0.004	0.000 to 0.095
(b) Gamma $(0.1, 0.1)$	2.01	0.39 to 3.68	0.008	0.021	0.002 to 0.118
(c) Log-uniform for σ^2 and σ_u^2	1.70	0.26 to 3.42	0.010	0.007	0.000 to 0.112
(d) Uniform on σ_u^2	1.99	0.22 to 3.79	0.014	0.004	0.002 to 0.051
(e) Uniform on ICC	1.94	0.19 to 3.87	0.014	0.054	0.002 to 0.163
(f) Approximate Jeffreys	1.96	0.35 to 3.70	0.011	0.050	0.003 to 0.155
(g) Uniform on shrinkage	1.93	0.18 to 3.72	0.014	0.042	0.002 to 0.135
(h) Informative beta for ICC	1.96	0.35 to 3.62	0.009	0.038	0.008 to 0.103

(95 per cent interval 0 to 0.165), which are similar to the results of both the uniform prior for ICC (option (e)) or the uniform shrinkage prior (option (g)).

Among the 'non-informative' priors (a) to (g), our strong preference is for either the uniform prior for ICC (option (e)) or the uniform shrinkage prior (option (g)). These produce inferences that are not sensitive to an arbitrary range constraint t, are based on sensible prior

independence assumptions, and have plausible distributional forms for ICC. In this particular example their conclusions are very similar to the classical analysis, but they have the additional flexibility of a full Bayesian approach.

The apparent heterogeneity in the within-cluster variances shown in Figure 1(a) suggests allowing different variances in each practice, but assuming they are drawn from a distribution such as a log-normal. We have investigated this by assuming

$$y_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma_i^2)$$

$$\log \sigma_i^2 \sim \mathrm{N}(\mu_v, \sigma_v^2)$$

and placing 'non-informative' priors on μ_v, σ_v^2, and a uniform prior on σ_u^2 as in option (d). This model estimates α to be 30.0 $(29.2, 30.9)$ and β 2.08 $(0.35, 3.85)$ – no single definition of ICC exists. Thus, in this particular example, this elaboration has negligible effect on the results.

6. INCORPORATING PRIOR EVIDENCE INTO POWER CALCULATIONS

Power calculations for cluster randomized trials can be carried out by assuming there is no cluster effect but inflating the variance of the response by the 'variance inflation factor' $1 + (m - 1)\mathrm{ICC}$, where m is the number of observations per cluster [1]. Thus if we wish to have $100(1 - \beta)$ power to detect a difference of δ for a response with variance σ^2, using a two-sided test with size 100α, the number of patients in each group will need to be

$$n = 2\frac{\sigma^2}{\delta^2}(Z_{1-\beta} - Z_{\alpha/2})^2(1 + (m - 1)\mathrm{ICC}) \tag{7}$$

where $Z_{0.025} = -1.96$, for example. In the example described in Section 3, Kinmonth *et al.* assumed $\sigma = 2.22$, $\delta = 1.00$, $\mathrm{ICC} = 0.047$, $m = 5$, 80 per cent power and 5 per cent size (that is, $(Z_{1-\beta} - Z_{\alpha/2})^2 = 7.84$) to give 92 patients (around 18 practices) per group [15].

Such power calculations require specifications of parameters that are necessarily unknown, and hence some sensitivity analysis is appropriate. However, Bayesian methods allow us to go beyond this by explicitly taking into account uncertainty about the ICC and other parameters when planning new studies.

For example, suppose from background information we thought the ICC to be around 0.05, but with there being a some chance of it being up to 0.15. In Section 4 we modelled these assumptions using a beta distribution, whose ordinate returns to zero for an ICC of zero. It might be considered more reasonable to use a distribution that does not decline to zero, such as a normal distribution with mean and standard deviation 0.05 (precision = 1/variance = 400), but truncated at 0 so as not to take on negative values. Values for the ICC can then be simulated from this prior distribution, and the resulting number of patients per group calculated from equation (7) for each value, producing a distribution over the number needed. In addition, we might acknowledge uncertainty about σ by giving it a normal distribution with mean 2.22 and standard deviation 0.5 (precision 4). Finally, we can assume a fixed number of patients per

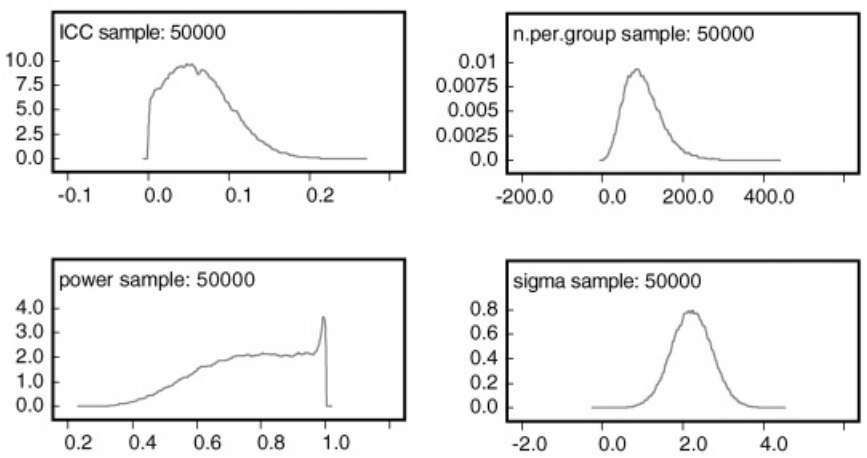

Figure 5. Prior distributions ICC and for σ used in power calculations, power when using an overall sample size of 92 patients in each group, and resulting predictive distribution of sample size per group necessary to achieve 80 per cent power. Distributions obtained as kernel density estimate from WINBUGS based on 50 000 samples.

Table II. Assumptions about ICC and σ, with consequences on the predicted power when using 92 patients per group, and the sample size n required to obtain 80 per cent power.

ICC		σ		Power (per cent) with 92 per group		n for 80 per cent power	
Median	(95 per cent int)	Median	(95 per cent int)	Median	(95 per cent int)	Median	(95 per cent int)
0.047	—	2.22	—	80	—	92	—
0.060	(0.004–0.15)	2.22	—	78	(67–86)	96	(78–124)
0.060	(0.004–0.15)	2.22	(1.24–3.20)	78	(46–99.8)	96	(29–209)

group (here taken to be $n = 92$) and calculate the power of such a trial at each iteration as

$$\text{Power} = \Phi\left(\sqrt{\left\{\frac{n\delta^2}{2\sigma^2(1+(m-1)\text{ICC})+Z_{\alpha/2}}\right\}}\right)$$

Such simulations can be carried out using a variety of software tools including spreadsheets; the necessary code for WINBUGS is shown in Appendix B.

Figure 5 shows the prior distribution specified for the ICC and σ, and the predictive distribution for the power assuming 92 patients per group, and the number of patients needed per group to obtain 80 per cent power – these plots were obtained from WINBUGS after 50 000 iterations taking around 15 seconds on a laptop computer. Table II shows the summary of these distributions; when incorporating uncertainty about ICC alone, there is some chance that the power could be considerably below the projected 80 per cent, and that up to 120 patients per group should be entered. Allowing additional uncertainty about σ would warn us that

plausible values of ICC and σ would lead to very low power and a need for substantially increased sample sizes.

The design implications of such analyses might be to increase sample size until we could be reasonably confident that the trial would have sufficient power.

7. DISCUSSION

The Bayesian approach outlined in this paper has both advantages and disadvantages. Favourable aspects include placing cluster randomized trials within the established literature on variance components models, while being able to use substantive prior opinion in design and analysis. Complex structures for variance components can be explored, including non-normal distributions for both responses and random effects, while incorporating uncertainty in power calculations extends sensitivity analysis by explicitly allowing for the relative plausibility of the assumptions. Newly available MCMC tools allow one to concentrate on realistic modelling rather than analytic techniques, while graphical representation may help communication of complex models.

However, problems can include sensitivity to prior assumptions (although when results of a Bayesian analysis are sensitive to prior beliefs, this may be the correct conclusion), and in particular we note that apparently 'non-informative' priors can be strongly influential. MCMC analysis can be slow when fitting many models, and caution may be required in checking convergence. Finally, there is little experience of the acceptability of Bayesian analyses to regulatory authorities, although they certainly are not ruled out [23].

In conclusion, the development of powerful MCMC methods means that computational issues are no longer a major obstacle to Bayesian inference. However, there is still considerable work needed in establishing robust strategies for Bayesian modelling that will provide convincing and generally acceptable results. In the meantime, we would recommend using background knowledge to produce an informative prior on the ICC where appropriate, and reference analyses based on either uniform priors on the ICC (option (e)) or the shrinkage parameter s (option (g)).

APPENDIX A

The implied prior distributions for the eight different prior options are obtained as follows:

(a) and (b) If $X \sim \text{gamma}(\psi, \gamma)$ and $Y \sim \text{gamma}(\pi, \gamma)$, then it is a standard result that $X/(X+Y) \sim \text{beta}(\psi, \pi)$. Thus if σ^{-2} and σ_u^{-2} are assumed to have $\text{gamma}(\varepsilon, \varepsilon)$ distributions

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2} = \frac{\sigma^{-2}}{\sigma^{-2} + \sigma_u^{-2}}$$

has a $\text{beta}(\varepsilon, \varepsilon)$ distribution.

(c) If X is log-uniform on (a, b) (that is, $\log X$ is uniform on $(\log a, \log b)$), then

$$p(x) = \frac{1}{\log \frac{b}{a}} \frac{1}{x}, \quad a < x < b$$

Suppose σ^2 and σ_u^2 have independent log-uniform(a,b) distributions, and we change the variables to $\rho=\sigma_u^2/(\sigma_u^2+\sigma^2)$ and $v=\sigma_u^2+\sigma^2$ with Jacobean v. The joint distribution for ρ and v is then

$$p(\rho,v)=\frac{1}{\log^2\frac{b}{a}}\ \frac{1}{\rho(1-\rho)}\ \frac{1}{v},\quad \begin{cases} \frac{a}{a+b}<\rho<\frac{b}{a+b} \\ a\max(\rho^{-1},(1-\rho)^{-1})<v<b\min(\rho^{-1},(1-\rho)^{-1}) \end{cases}$$

and 0 elsewhere. Integrating out v reveals a marginal distribution

$$p(\rho)=\frac{1}{\log\frac{b}{a}}\ \frac{1}{\rho(1-\rho)}\left(1-|\text{logit}\,\rho|\Big/\log\frac{b}{a}\right),\quad \frac{a}{a+b}<\rho<\frac{b}{a+b}$$

(d) Suppose σ^2 and σ_u^2 are independent with log-uniform(a,b) and uniform$(0,c)$ distributions, respectively, and we change the variables to $\rho=\sigma_u^2/(\sigma_u^2+\sigma^2)$ and $v=\sigma_u^2+\sigma^2$ with Jacobean v. The joint distribution for ρ and v is then

$$p(\rho,v)=\frac{1}{c\log\frac{b}{a}}\ \frac{1}{(1-\rho)},\quad \begin{cases} 0<\rho<\frac{c}{(a+c)} \\ a(1-\rho)^{-1}<v<\min(c\rho^{-1},b(1-\rho)^{-1}) \end{cases}$$

and 0 elsewhere. Integrating out v reveals a marginal distribution

$$p(\rho)=\begin{cases} \frac{1}{\log\frac{b}{a}}\ \frac{1}{(1-\rho)^2}\left(\frac{b}{c}-\frac{a}{c}\right), & 0<\rho<\frac{c}{(b+c)} \\ \frac{1}{\log\frac{b}{a}}\ \frac{1}{\rho(1-\rho)}\left(1-\frac{a\rho}{c(1-\rho)}\right), & \frac{c}{(b+c)}<\rho<\frac{c}{(a+c)} \end{cases}$$

(e) Suppose σ^2 and ρ are independent, σ^2 has a log-uniform(a,b) distribution and ρ has density $p(\rho)$ and cumulative distribution function F_ρ. Change the variables to $\sigma_u^2=\sigma^2\rho/(1-\rho)$ and ρ, with Jacobean $\rho/(1-\rho)$. The joint distribution for σ_u^2 and ρ is then

$$p(\sigma_u^2,\rho)=\frac{1}{\log\frac{b}{a}}\ \frac{1}{\sigma_u^2}\,p(\rho),\quad \begin{cases} 0<\sigma_u^2<\infty \\ \frac{\sigma_u^2}{b+\sigma_u^2}<\rho<\frac{\sigma_u^2}{a+\sigma_u^2} \end{cases}$$

and 0 elsewhere. Integrating out ρ reveals a marginal distribution

$$p(\sigma_u^2)=\frac{1}{\log\frac{b}{a}}\frac{1}{\sigma_u^2}\left[F_\rho\left(\frac{\sigma_u^2}{a+\sigma_u^2}\right)-F_\rho\left(\frac{\sigma_u^2}{b+\sigma_u^2}\right)\right],\quad 0<\sigma_u^2<\infty \tag{A1}$$

Thus if ρ is uniform(0,1), $F_\rho(r)=r$, and so

$$p(\sigma_u^2)=\frac{1}{\log\frac{b}{a}}\frac{b-a}{(a+\sigma_u^2)(b+\sigma_u^2)},\quad 0<\sigma_u^2<\infty$$

(f) Transforming the balanced-design Jeffreys prior (equation (5)) for (σ^2, σ_u^2) to variables (σ^2, s) with Jacobean $\sigma^2/(ms^2)$ provides a joint distribution

$$p(\sigma^2, s) \propto \frac{1}{\sigma^2}\frac{1}{s}, \quad \begin{cases} 0 < \sigma^2 < \infty \\ 0 < s < 1 \end{cases}$$

showing the Jeffreys prior is equivalent to independent log-uniform distributions on σ^2 and s. For an unbalanced design we adopt an approximate Jeffreys prior obtained by substituting the mean cluster size $\bar{m}$ for m and limiting the range. Specifically we assume a log-uniform$(\varepsilon, 1)$ distribution for s

$$p(s) = \frac{1}{\log \frac{1}{\varepsilon}}\frac{1}{s}, \quad \varepsilon < x < 1$$

Then ρ has a marginal prior distribution

$$p(\rho) = \frac{\bar{m}}{\log \frac{1}{\varepsilon}}\frac{1}{1-\rho}\frac{1}{1+(\bar{m}-1)\rho}, \quad 0 < \rho < 1 - \bar{m}/(t + \bar{m} - 1)$$

Giving σ^2 an independent log-uniform(a, b) and transforming to a joint distribution on (σ^2, σ_u^2) leads to

$$p(\sigma_u^2, \sigma^2) = \frac{1}{\log \frac{b}{a}}\frac{1}{\log \frac{1}{\varepsilon}}\frac{1}{\sigma^2}\frac{1}{(\sigma^2 + \bar{m}\sigma_u^2)^2}, \quad \begin{cases} a < \sigma^2 < b \\ 0 < \sigma_u^2 < (\varepsilon^{-1} - 1)b/\bar{m} \end{cases}$$

an approximate proper version of Jeffreys prior equation (5). Integrating out σ^2 reveals a marginal prior distribution

$$p(\sigma_u^2) = \frac{1}{\log \frac{b}{a}}\frac{1}{\log \frac{1}{\varepsilon}}\frac{1}{\sigma_u^2}\log\frac{1 + \bar{m}\sigma_u^2/a}{1 + \bar{m}\sigma_u^2/b}, \quad 0 < \sigma_u^2 < (\varepsilon^{-1} - 1)b/\bar{m}$$

(g) This option places a uniform$(0, 1)$ distribution on the shrinkage parameter s, which in our approximation is defined to be $s = (1 - \rho)/(1 + (\bar{m} - 1)\rho)$. Transforming to a distribution on $\rho = s/(1 + (\bar{m} - 1)s)$ with Jacobean $\bar{m}/(1 + (\bar{m} - 1)\rho)^2$ gives

$$p(\rho) = \frac{\bar{m}}{(1 + (\bar{m} - 1)\rho)^2}, \quad 0 < \rho < 1$$

Assuming an independent log-uniform (a, b) for σ^2 and adapting the development for the Jeffreys prior shown above leads to a joint prior

$$p(\sigma_u^2, \sigma^2) = \frac{1}{\log \frac{b}{a}}\frac{\bar{m}}{(\sigma^2 + \bar{m}\sigma_u^2)^2}, \quad \begin{cases} a < \sigma^2 < b \\ 0 < \sigma_u^2 < \infty \end{cases}$$

Integrating out σ^2 reveals a marginal prior distribution

$$p(\sigma_u^2) = \frac{\bar{m}}{\log \frac{b}{a}} \frac{b-a}{(a+\bar{m}\sigma_u^2)(b+\bar{m}\sigma_u^2)}, \quad 0 < \sigma_u^2 < \infty$$

(h) Using the results of (A1), and assuming ρ has a beta(g,h) with cumulative distribution function $F_{g,h}$, leads to

$$p(\sigma_u^2) = \frac{1}{\log \frac{b}{a}} \frac{1}{\sigma_u^2} \left[F_{g,h}\left(\frac{\sigma_u^2}{a+\sigma_u^2}\right) - F_{g,h}\left(\frac{\sigma_u^2}{b+\sigma_u^2}\right) \right], \quad 0 < \sigma_u^2 < \infty$$

APPENDIX B

None of the analyses described above can be carried out in closed form, and even analytic approximations can be difficult. Recent applied Bayesian analysis has made extensive use of Markov chain Monte Carlo (MCMC) methods, in which plausible values for the parameters are simulated from their joint posterior distribution. Each new simulated value depends only on immediately preceding ones and hence the iterations form a Markov chain. There are many algorithms for simulating appropriate values, as well as techniques for checking convergence of the Markov chain to its stationary distribution [24, 25].

We have used the WINBUGS program for the computations [26]. This is a general program for MCMC analysis in which models may be described by a language or through a graphical user interface. Appropriate sampling methods for each unknown quantity are automatically derived and there are facilities for on-line monitoring and output analysis. The WINBUGS software is available from `http://www.mrc-bsu.cam.ac.uk/bugs`.

Figure A1 shows the WINBUGS graphical input for prior option (e), the uniform prior on the ICC; the corresponding description in the WINBUGS language is shown below. In each case the model is described in a slightly different way from that shown in Section 3; the y_{ij} notation used previously is suited to equal numbers of patients per practice, while unbalanced models are more efficiently handled by using j to index all patients 1 to $J = 240$, and including the practice of the jth patient as a factor `prac[j]` taking values $i = 1$ to $I = 38$.

We also note that WINBUGS parameterizes the normal distribution in terms of its precision $\tau = 1/\sigma^2$, so that, in the obvious notation

$$\tau_u = \tau \frac{(1-\rho)}{\rho}$$

```
model;
{
   for( j in 1 : 240 ) {
      y[j] ~ dnorm(mu[j],tau)
      mu[j] <- u[prac[j]]
   }
   log(tau) <- logtau

   for( i in 1 : 38 ) {
      u[i] ~ dnorm(phi[i],tau.u)
```


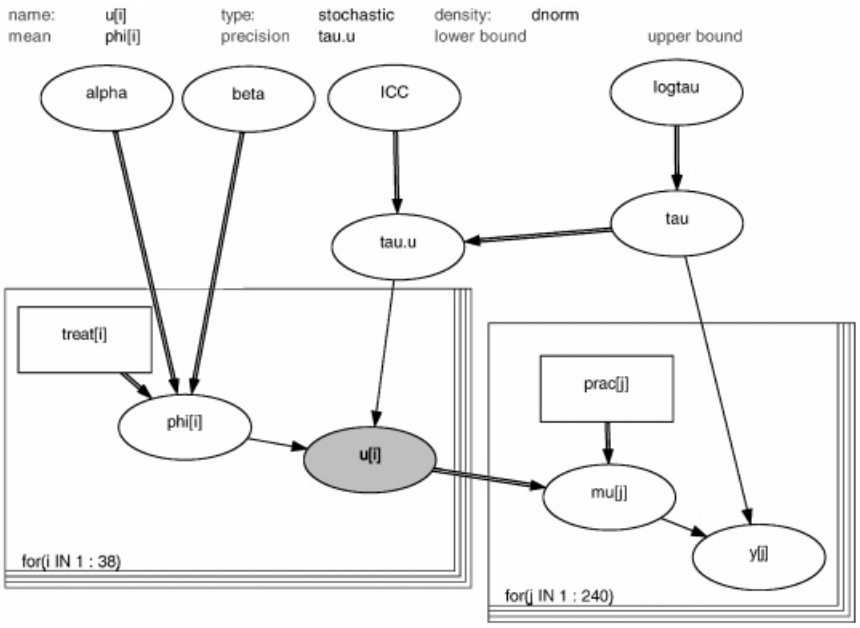

Figure A1. Conditional independence graph for prior option (e), in which each node represents a quantity and the links represent stochastic (single line) or deterministic (double line) dependence. Independent uniform priors are placed on `alpha`, `beta`, `ICC` and the logarithm of the within-practice precision denoted `logtau`. The cluster-specific random effects `u[i]` are highlighted and their description shown above the graph. The 'plates' in the graph represent repeated 'submodels', for patients on the right and for practices on the left. The program can be run directly from this graph or from the automatically-constructed model description shown in the text. The graph can be elaborated in many ways, by adding covariates in the patient or practice submodels, changing distributional or prior assumptions, placing structure on the variance components and so on.

```
      phi[i] <- alpha + beta * (treat[i] -  0.5)
   }

   tau.u <- (tau * (1 - ICC)) / ICC
   beta ~ dunif(-10000,10000)
   alpha ~ dunif(-10000,10000)
   logtau ~ dunif(-5,5)
   ICC ~ dunif(0,1)
}
```

No difficulties with obtaining convergence was encountered in these models, and all analyses were based on a sample of 5000 iterations following a burn-in of 5000.

The WINBUGS code for the power calculation with uncertain ICC and σ is given by

```
model;
{
   n.per.group <- (2 * sigma * sigma * 7.84 * (1 + (m - 1) * ICC))/
                  (delta * delta)
   ICC ~ dnorm(0.05,400)I(0,)
```



```
   sigma ~ dnorm(2.22, 4 )

   power <- phi( sqrt( (92 * delta * delta)/
                  (2 * sigma * sigma * (1 + (m - 1) * ICC)) ) - 1.96 )
}
```

ACKNOWLEDGEMENTS

I would like to thank Ann Louise Kinmonth and Michael Campbell for making the data available to me, and Simon Thompson and Rebecca Turner for many useful discussions.

REFERENCES

1. Donner A. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics* 1998; **47**:95–113.
2. Liang KY, Zeger SL. Longitudinal data analysis using generalized estimating equations. *Biometrika* 1986; **73**:13–22.
3. Murray DM. *Design and Analysis of Group-Randomized Trials.* Oxford University Press: New York, 1998.
4. Bryk AS, Raudenbush SW. *Hierarchical Linear Models*. Sage: Newbury Park, 1992.
5. Goldstein H. *Multilevel Models in Educational and Social Research*, 2nd edn. Edward Arnold: London, U.K., 1995.
6. Omar RZ, Thompson SG. Analysis of a cluster randomised trial with binary outcome data using a multilevel model. *Statistics in Medicine* 2000; **19**:2675–2688.
7. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomised trials with binary outcome data. *Statistics in Medicine* 2000; **19**(submitted).
8. Gelman A, Carlin J, Stern H, Rubin DB. *Bayesian Data Analysis*. Chapman and Hall: New York, 1995.
9. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall: London, U.K., 1996.
10. Bernardo JM, Smith AFM. *Bayesian Theory*. Wiley: Chichester, U.K., 1994.
11. Berry DA, Stangl DK (eds.), *Bayesian Biostatistics*. Marcel Dekker: New York, 1996.
12. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials (with discussion). *Journal of the Royal Statistical Society*, *Series A* 1994; **157**:357–416.
13. Kadane JB. Prime time for Bayes. *Controlled Clinical Trials* 1995; **16**:313–318.
14. Spiegelhalter DJ, Myles JM, Jones DR, Abrams KR. An introduction to Bayesian methods in health technology assessment. *British Medical Journal* 1999; **319**:508–512.
15. Kinmonth AL, Woodcock A, Griffin S, Spiegal N, Campbell MJ. Randomised controlled trial of patient centred care of diabetes in general practice: impact on current well being and future disease risk. *British Medical Journal* 1998; **317**:1202–1208.
16. Daniels MJ. A prior for the variance components in hierarchical models. *Canadian Journal of Statistics* 1999; **27**:569–580.
17. Burch BD, Harris IR. Bayesian estimators of the intraclass correlation coefficients in the one-way random effects model. *Communications in Statistics — Theory and Methods* 1999; **28**:1247–1272.
18. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 1996; **91**:1343–1370.
19. Chaloner K. A Bayesian approach to the estimation of variance components for the unbalanced one-way random effects model. *Technometrics* 1987; **29**:323–337.
20. Box GEP, Tiao GC. *Bayesian Inference in Statistical Analysis*. Addison-Wesley: Reading, Mass, 1973.
21. Natarajan R, Kass RE. Reference Bayesian methods for generalised linear mixed models. *Journal of the American Statistical Association* 2000; **95**:227–237.
22. Berger J, Bernardo JM. Reference priors in a variance components problem. In *Proceedings of the Indo-USA Workshop on Bayesian Analysis in Statistics and Econometrics*. Springer: New York, 1992; 323–340.
23. International Conference on Harmonisation E9 Expert Working Group. Statistical Principles for Clinical Trials: ICH Harmonised Tripartite Guideline. *Statistics in Medicine* 1999; **18**:1905–1942. Available on `http://www.ich.org/ich5e.html`.
24. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo Methods in Practice*. Chapman and Hall: New York, 1996.
25. Brooks SP. Markov chain Monte Carlo method and its application. *Statistician* 1998; **47**:69–100.
26. Spiegelhalter DJ, Thomas A, Best NG. *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit: Cambridge, 1999.

