Bayesian Analysis (2007) **2**, Number 2, pp. 409–444

Identifying outliers in Bayesian hierarchical models: a simulation-based approach

E. C. Marshall[*] and D. J. Spiegelhalter[†]

Abstract. A variety of simulation-based techniques have been proposed for detection of divergent behaviour at each level of a hierarchical model. We investigate a diagnostic test based on measuring the conflict between two independent sources of evidence regarding a parameter: that arising from its predictive prior given the remainder of the data, and that arising from its likelihood. This test gives rise to a p-value that exactly matches or closely approximates a cross-validatory predictive comparison, and yet is more widely applicable. Its properties are explored for normal hierarchical models and in an application in which divergent surgical mortality was suspected. Since full cross-validation is so computationally demanding, we examine full-data approximations which are shown to have only moderate conservatism in normal models. A second example concerns criticism of a complex growth curve model at both observation and parameter levels, and illustrates the issue of dealing with multiple p-values within a Bayesian framework. We conclude with the proposal of an overall strategy to detecting divergent behaviour in hierarchical models.

Keywords: Hierarchical models, diagnostics, outliers, distributional assumptions.

1 Introduction

Introduction of Markov chain Monte Carlo (MCMC) methods has enabled researchers to fit a wide range of complex hierarchical models, in which observations are grouped within ‘units’ whose parameters (‘random’ effects) are assumed drawn from some population model. Applications are not, however, generally accompanied by the kind of model diagnostics that are standard practice when, say, carrying out regression analysis within traditional statistical packages. In this paper we consider one such class of diagnostic: simulation-based methods for detecting observations or units that do not appear to be drawn from the assumed underlying distributions.

An example of a context in which such diagnostics may be important is that of making comparisons between schools, hospitals or other institutions, in which hierarchical models are being increasingly used from both a non-Bayesian (Goldstein 1995) and a Bayesian (Normand et al. 1997) perspective. The default assumption for the random effects distribution is generally Gaussian, which is not only technically convenient but might also be justified by reasoning that there are inevitably many unmeasured institutional covariates whose total influence, by an informal central limit theorem argument, might approximate a normal distribution. It is of particular interest to identify institu-

[*] Civil Service, UK

[†] MRC Biostatistics Unit, Cambridge, UK, mailto:david.spiegelhalter@mrc-bsu.cam.ac.uk

tions whose effects appear to lie beyond the reasonable tails of the assumed distribution, and we shall term such institutions *divergent*, in distinction to merely *extreme* cases that are simply in the tails of the random effects distribution. While each of these situations may be said to represent an 'outlier', the latter behaviour does not lead us to question the model (although of course it still may attract attention to the institution). The interest in this paper will focus only on detection of divergent units, with some comments on detecting individual divergent observations.

We assume a general conditionally independent hierarchical model is being considered as our current null hypothesis H_0, where H_0 encompasses sampling likelihoods, forms of random effect distribution, and prior distributions. We assume a vector of observations y_{ij}, $j = 1, \ldots, n_i$ within each unit $i, i = 1, ..., K$, with

$$\begin{aligned} y_{ij} &\sim p(y_{ij}|\gamma, x_{ij}, \phi_i) \\ \phi_i &\sim p(\phi_i|\beta, z_i) \\ \beta, \gamma &\sim p(\beta, \gamma). \end{aligned}$$

This model is represented in graphical form in Figure 1 (Spiegelhalter 1998), showing the possibility of a vector of covariates $\{x_{ij}\}$ at observation and $\{z_i\}$ at the unit level, and γ represents, for example, error variances and regression coefficients. This can be extended to allow 'undirected links' between the random effects ϕ_i representing, for example, temporal or spatial association between the units (Marshall and Spiegelhalter 2003).

A number of features complicate the process of outlier detection in hierarchical models. First, as emphasised by Langford and Lewis (1998), it may be difficult to assign divergent behaviour to a particular level of the hierarchy – a unit may apparently be divergent but this finding may be driven by just a few observations, while conversely, sparse data within units may be labelled as individually divergent when in reality the entire unit is at fault. Second, 'shrinkage' of parameter estimates towards population means will tend to mask divergent behaviour, and so it may be misleading to rely on the size of observed residuals.

Furthermore, two responses are possible to the diagnosis of divergent observations or units. They might either be *identified* and considered separately, possibly as fixed effects, or *accommodated* by adopting, say, a heavier-tailed distribution. Which of these approaches is appropriate depends on the purpose of the investigation: for example, in the epidemiological context of disease mapping, on might be interested in 'hot-spots' (identification) or producing maps with reliable estimates of underlying disease risk (accommodation). Our interest will focus on identification.

The general approach to model criticism taken in the current paper is in line with well established themes in the Bayesian literature – most notably (i) those of predictive model criticism, considered by Gelfand et al. (1992), Gelman et al. (1996) and Bayarri and Berger (2000) amongst others; and (ii) the idea of measuring conflict between the likelihood and prior to highlight lack of fit (Box 1980; O'Hagan 2003). Section 2 reviews Bayesian predictive approaches to model checking, and provides a formal synthesis of the multitude of approaches to Bayesian model criticism. The cross-validatory

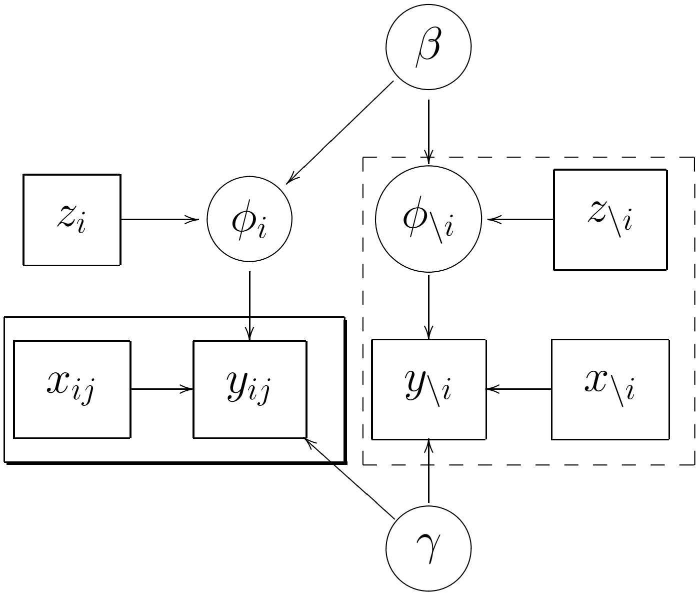

Figure 1: Null model shown as directed graphical model: circles represent parameters, squares are observations, arrows represent stochastic dependence, solid box indicates repeated observations in unit i, dashed box represents all units except i, and $\phi_{\backslash i}$ represents $\{\phi_k, k \neq i\}$.

method is described in Section 3 using strategies based on replicating both random effects and data, and comparing these predictions with the observations using 'mixed' p-values. Section 4 introduces an alternative, 'conflict' p-value that measures discordance between prior and data: we show that in many circumstances this will be exactly or approximately equal to the 'mixed' p-value, but can be applied much more generally. Section 5 contrasts these two approaches using data from an inquiry into excess surgical deaths. Since full cross-validation is not routinely feasible when using MCMC methods, approximate cross-validatory procedures are proposed in Section 6 and compared both analytically and using the previous example. In Section 7 a more complex growth-curve example illustrates the full range of methods to detection of both divergent data and units. Finally, in Section 8 we formulate a practical strategy for hierarchical model criticism. Appendices contain technical results on the distribution of p-values under null hypotheses, and exact analysis for normal hierarchical models. WinBUGS code is available from the authors on request.

2 A review of predictive model checks

2.1 General approaches

There is a large literature on checking hierarchical models, both from the classical and Bayesian viewpoints (Hodges 1998). One approach involves the embedding of the current model H_0 in a wider family, indexed by α, such that the simpler model arises when $\alpha = \alpha_0$. Criticism of H_0 is based on inferences on α within H_1 or model comparison between H_0 and H_1: Bayesian examples include the use of t instead of normal distributions (Wakefield and Bennett 1996), computing Bayes factors between H_0 and H_1 (Sharples 1990; Albert and Chib 1997) or comparing prior and posterior distributions for α within H_1 (Carota et al. 1996).

The approach taken in the paper is much more in the spirit of classical hypothesis testing, in which H_0 can be criticised without explicit consideration of an alternative. The standard approach is to specify a discrepancy measure T, and then contrast the observed discrepancy T^{obs} with some reference distribution $p_0(T)$ under the null hypothesis that H_0 is an appropriate model; 'extreme' values of T^{obs} relative to this distribution then cast doubt on H_0. The choice of the discrepancy function T and the associated reference distribution p_0 depend very much on the aspect of the model which is to come under scrutiny.

First, classical approaches treat the data as random and the parameters as fixed, and seek discrepancy measures $T(y)$ which are functions of the data y alone so that $T^{\text{obs}} = T(y^{\text{obs}})$, and whose sampling distribution $p_0(T)$ under the null does not depend, at least as an approximation, on unknown quantities. When checking the distributional form for the random effects ϕ_i, suggestions include setting $T(y)$ as the *fixed effects* estimates $\tilde{\phi}_i = \tilde{\phi}_i(y)$, with $p_0(T)$ as their estimated predictive distribution under H_0 (Dempster and Ryan 1985; Hardy and Thompson 1998), and setting $T(y)$ as the *random effects* estimates $\hat{\phi}_i = \hat{\phi}_i(y)$, again with a reference distribution $p_0(T)$ (Lange and Ryan

1989; Hilden-Minton 1995; Bryk and Raudenbush 1992; Goldstein 1995).

Second, Bayesian approaches to residuals treat the data as fixed and the parameters as random, and select discrepancy measures $T(y, \theta)$ which are functions of the data and all parameters $\theta = \{\beta, \phi, \gamma\}$, or of the parameters alone. The reference distribution is the posterior distribution $p_0(T|y)$ under the null: since $T(y^{\text{obs}}, \theta)$ is not wholly specified, an 'extreme' value for T must be judged relative to its prior distribution. This approach underlies the Bayesian residual analysis of Chaloner and Brant (1988), which was extended by Chaloner (1994) to residuals at higher levels of the model and adapted by Hodges (1998) to a reformulated normal hierarchical model.

Finally, what we shall term the 'Bayesian-predictive' approaches treat both the data and parameters as random. Considering first the situation in which $T(y)$ is chosen to be a function of the data alone, Gelman et al. (1996) review three approaches to the choice of reference distribution in the general context of checking hierarchical models:

1. Box (1980) proposes a *prior predictive* check in which $T(y^{\text{obs}})$ is compared with its predictive distribution

$$p_0(T(Y)) = \int p(T(Y)|\theta)p(\theta)d\theta.$$

   Reasonable application of this approach requires informative prior distributions for all model parameters and so may not be generally appropriate.

2. Gelman et al. (1996) suggest a *mixed predictive* check as relevant to the assessment of hierarchical models such as shown in Figure 1, in which the reference distribution is taken as

$$p_0^M(T(Y)|y^{\text{obs}}) \quad = \quad \int p(T(Y)|\beta)p(\beta|y^{\text{obs}})d\beta \tag{1}$$

   where $p(T(Y)|\beta) = \int p(T(Y)|\phi)p(\phi|\beta)d\phi$ is the predictive distribution for $T(Y)$ in a new set of replicate units. We assume γ is known at this stage. We may also express $p_0^M(T(Y)|y^{\text{obs}})$ as

$$p_0^M(T(Y)|y^{\text{obs}}) \quad = \quad \int p(T(Y)|\phi)p^M(\phi|y^{\text{obs}})d\phi \tag{2}$$

   where we shall call $p^M(\phi|y^{\text{obs}}) = \int p(\phi|\beta)p(\beta|y^{\text{obs}})d\beta$ the 'predictive prior' distribution for ϕ: we note the care necessary in not confusing $p^M(\phi|y^{\text{obs}})$ with the standard posterior distribution for ϕ given by $p(\phi|y^{\text{obs}}) = \int p(\phi|y^{\text{obs}}, \beta)p(\beta|y^{\text{obs}})d\beta$.

3. Rubin (1984) suggested the *posterior predictive* approach, in which the observed discrepancy measure $T^{\text{obs}} = T(y^{\text{obs}})$ is compared with its predictive distribution $p_0(T(Y)|y^{\text{obs}})$ given the observed data, which for the model in Figure 1 means

$$p_0(T(Y)|y^{\text{obs}}) = \int p(T(Y)|\phi)p(\phi|y^{\text{obs}})d\theta \tag{3}$$

> where $p(\phi|y^{\text{obs}})$ is the current posterior distribution. Contrasting (2) with (3), it is clear that the mixed predictive approach uses a predictive prior distribution for a new set of random effects, while the posterior predictive checks simply involve the prediction of new data conditional on the current posterior $p(\phi|y^{\text{obs}})$. While in the former mixed approach the data y^{obs} only influence the predictive distribution through the information they provide about β, in the posterior predictive approach y^{obs} influence the ϕ's directly. Hence we may expect strong conservatism in the latter case, in that the checking function being predicted is being directly influenced by the very data which is being checked. This conservatism is explored in detail by, for example, Bayarri and Berger (2000), Bayarri and Morales (2003) and Bayarri and Castellanos (2004). The posterior predictive approach may, however, be more appropriate for addressing a very specific question: conditional on the truth of the unit-level model, are there any individual data points which appear outlying? That is, is the likelihood model appropriate? We return to these ideas later in the paper.

The mixed and posterior predictive approaches have the advantage of being easy to compute using MCMC by generating replicate datasets Y^{rep}, but when as above they are not adopted in a full cross-validatory framework, come under criticism because the data are used twice, first to update the prior $p(\theta)$ into a posterior and then for computing the discrepancy measure. This is recognised by Gelman et al. (1995), who emphasise that "test quantities are commonly chosen to measure a feature of the data not directly addressed by the probability model". Bayarri and Berger (2000) suggest methods of ensuring this is the case, by either using a posterior distribution that is proportional to $p(y^{\text{obs}}|T^{\text{obs}},\theta)p(\theta)$ (the *partial* approach) or by conditioning the reference distribution on a statistic U that contains as much information about θ as possible (the *conditional* approach). The partial approach is clearly the 'ideal' in that it obeys the injunction of Gelman et al. (1995) in explicitly using a probability model that considers T^{obs} as fixed, but in practice the partial approach is complex to implement in non-trivial situations such as generalised linear mixed models.

Other suggestions have also been made: for example, Dey et al. (1998) consider the highly computationally-intensive approach of replicating data Y^{rep} from the prior distribution, as in the prior-predictive proposal (Box 1980), analysing each replicated dataset to obtain posterior distributions $p(\theta|Y^{\text{rep}})$, and finally seeing whether $p(\theta|y^{\text{obs}})$ is 'extreme' relative to the set of $p(\theta|Y^{\text{rep}})$.

In our specific context of identifying divergent units or individual data points, it seems appropriate to consider multiple discrepancy measures based on summary statistics for units or raw data respectively. We shall see in Section 3 that a cross-validatory approach leads to a convergence of all the proposals listed above. However, we first need to consider how a discrepancy measure might be compared to a reference distribution.

2.2 Comparing a discrepancy measure with a reference distribution

Suppose our discrepancy measure T is a function of data Y alone, and interest lies in comparing the observed value $T^{\text{obs}} = T(y^{\text{obs}})$ with a reference distribution $p_0(T)$ derived using one of the proposals outlined above. Since we are focussing on simulation-based methods in which replicate values of T will be generated from $p_0(T)$, we shall denote such values T^{rep}.

Our choice in this paper is based on the 'lower' tail-area $P_0^L = \Pr(T^{\text{rep}} \leq T^{\text{obs}})$, which has the double advantage over, say, the standardised Pearson residual or conditional predictive ordinate (Pettit and Smith 1985), of having a familiar calibration that does not rely on asymptotics, and being calculated directly as the proportion of times a simulated continuous discrepancy measure is less than or equal to an observed value (although for discrete measures care is required in defining the p-value). An 'upper' tail area is trivially defined as $P_0^U = 1 - P_0^L$. Small values of either P_0^L, P_0^U or both may be of interest depending on the context: in the last case the 2-sided p-value, $2 \times \min(P_0^L, P_0^U)$, can be reported.

If the discrepancy measure $T(y, \theta)$ is a function of both data and parameters, then Gelman et al. (1996) recommend calculating the posterior expectation of the p-value,

$$E_{\theta|y^{\text{obs}}}\left[\Pr\left(T(Y^{\text{rep}}, \theta) \leq T(y^{\text{obs}}, \theta)\right)\right];$$

this may be efficiently obtained by drawing θ^{rep} from $p(\theta|y^{\text{obs}})$, simulating Y^{rep} from $p(Y|\theta^{\text{rep}})$, and observing the proportion of times that $T(y^{\text{obs}}, \theta^{\text{rep}})$ exceeds $T(Y^{\text{rep}}, \theta^{\text{rep}})$.

Each unit or observation can give rise to its own p-value for which low values are of interest, leading to the classical problem that the interpretation of an apparently extreme p-value depends on the number that have been calculated, since the probability of a false positive result increases with increasing number of tests. It may be a useful guideline to multiply small p-values by the number calculated – the classic Bonferroni adjustment. Alternatively one can make use of the false discovery rate (FDR) procedure of Benjamini and Hochberg (1995), possibly adapted for dependent p-values (Benjamini and Yekutieli 2001): see Section 7 for an example. It may seem paradoxical that a Bayesian model-checking procedure can lead to classical adjustments for multiple comparisons, and this is discussed further in Section 8.

3 'Cross-validatory' methods for identifying divergent parameters

Cross-validation or 'leave-one-out' methods are well-established: the data being checked is left out, a prediction is made based on the remaining data, and divergence between observed and prediction casts doubt on the assumptions. See Gelfand et al. (1992) and Bernardo and Smith (1994) for extensive discussion of this approach from a Bayesian perspective. In the context of checking unit i, it is natural to leave y_i out of analysis, where y_i may be a vector, and use a discrepancy measure $T_i = T(y_i)$. Then, using the structure and notation of Figure 1, the reference distribution $p_0(T_i^{\mathrm{rep}})$ conditional on remaining data $y_{\setminus i}$, is obtained from

$$p(T_i^{\mathrm{rep}}|y_{\setminus i}) = \int p(T_i^{\mathrm{rep}}|\phi_i, \gamma)p(\phi_i|\beta), p(\beta, \gamma|y_{\setminus i})d\phi_i d\beta d\gamma. \tag{4}$$

now allowing γ to be unknown.

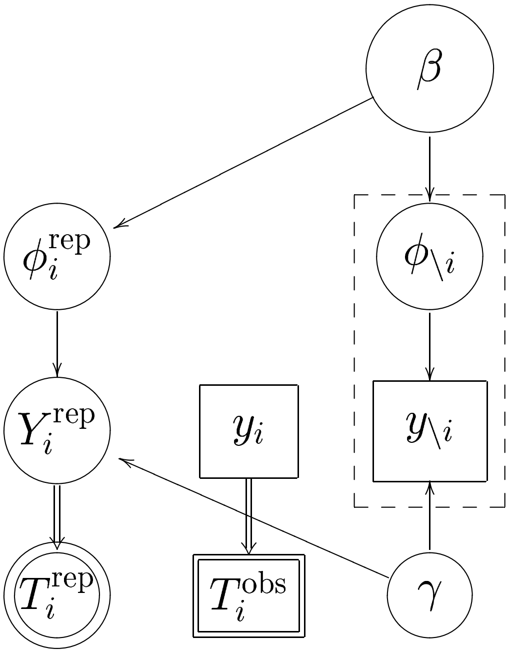

Figure 2: Cross-validatory 'mixed' replication of parameters and data: the double directed arrow indicates a logical function. The double ringed nodes represent quantities to be compared in order to assess divergence.

(4) may also be written as

$$p(T_i^{\mathrm{rep}}|y_{\setminus i}) = \int p(T_i^{\mathrm{rep}}|\phi_i, \gamma)p^M(\phi_i, \gamma|y_{\setminus i})d\phi_i d\gamma,$$

using the 'predictive prior'

$$p^M(\phi_i, \gamma|y_{\setminus i}) = \int p(\phi_i|\beta)p(\beta, \gamma|y_{\setminus i})d\beta.$$

Following this formulation, as reflected in Figure 2, we would obtain $\beta^{\mathrm{rep}}, \gamma^{\mathrm{rep}}|y_{\setminus i}$ as part of an MCMC simulation, then a replicate $\phi_i^{\mathrm{rep}}|\beta^{\mathrm{rep}}$, followed by a simulated

$Y_i^{\text{rep}}|\phi_i^{\text{rep}}, \gamma^{\text{rep}}$, and hence obtain $T_i^{\text{rep}} = T(Y_i^{\text{rep}})$ which can then be compared with T_i^{obs} using techniques described in Section 2. Since the reference distribution is obtained entirely independently of y_i, in most situations the resulting mixed p-values

$$P_{i,\text{mix}} = \Pr(T_i^{\text{rep}} \leq T_i^{\text{obs}}|y_{\setminus i}) \tag{6}$$

will be a 'proper' p-value, in that it will be drawn from a uniform distribution on (0, 1) under H_0 (although the p-values for each unit will not in general be independent): see Section A.1 for detailed discussion. This suggests a strategy of quantile-quantile (QQ) plots of p-values against order statistics of the uniform (0, 1) distribution, as well as taking into account multiple comparisons (Section 2.2).

We note that all the Bayesian-predictive approaches outlined in Section 2 (except the posterior predictive approach) coincide in this cross-validatory context: as pointed out by Carlin (1999), the reference distribution (4) corresponds to both a *partial posterior* and a *conditional* prediction where the posterior is based on data disjoint from T_i (Draper 1996; Bayarri and Berger 2000), a *prior* prediction where the prior conditions on other data, and a *mixed* prediction according to its definition. This approach can also be seen as an application of the cross-validatory suggestions of Gelfand et al. (1992) to hierarchical models.

The issue remains of selecting a discrepancy measure T_i for a vector y_i. If T_i is chosen to be the full data y_i, then the individual observations y_{ij} each contribute to the overall measure of divergence. However, the reference distribution is then a convolution of the *likelihood* $p(y_{ij}|\phi_i)$ with the *prior* $p(\phi_i|\beta)$, and so loses power if our interest is solely in checking for divergent ϕ_i and we are willing to assume the likelihood is correct. This is clearly illustrated if sufficient statistics s_i exist, since by definition the likelihood factorises $p(y_i|\phi_i)$ into $p(y_i|s_i)p(s_i|\phi_i)$. The first term contains no information about ϕ_i and hence its inclusion in a reference distribution for a discrepancy measure can only add noise to the procedure. Thus it will be more efficient to use s_i as a discrepancy measure or, more generally, if closed-form estimators $\hat{\phi}_i$ exist, to set $T_i = \hat{\phi}_i(y_i)$ and then compare $\hat{\phi}_i$ with $p(\hat{\phi}_i^{\text{rep}}|y_{\setminus i})$.

4 Measuring conflict between prior and likelihood

In many circumstances closed form estimators $\hat{\phi}_i$ will not exist to use as discrepancy functions: for example in non-normal generalised linear models with covariates, and non-linear models such as used in pharmacokinetics. An alternative approach is motivated by the observation by Box (1980) that the 'mixed' approach described above can also be interpreted as measuring conflict between likelihood and prior: see also O'Hagan (1994) [p 179] and O'Hagan (2003).

A general implementation is illustrated in Figure 3. At each iteration a *predictive prior* replicate $\phi_i^{\text{rep}}|y_{\setminus i}$ is generated just as in the standard cross-validation approach described above. A second replicate $\phi_i^{\text{fix}}|y_i^{\text{obs}}$ is then generated from the posterior distribution for the 'fixed effect' estimate using only the data from the unit being tested and a 'non-informative' or 'reference' prior for ϕ_i: this can also be termed a *likelihood* replicate since it is essentially drawn from a normalised likelihood in a suitable parameterisation. These *prior* and *likelihood* replications can be considered as two independent sources of evidence about ϕ_i, and conflict between them suggests faults in the model.

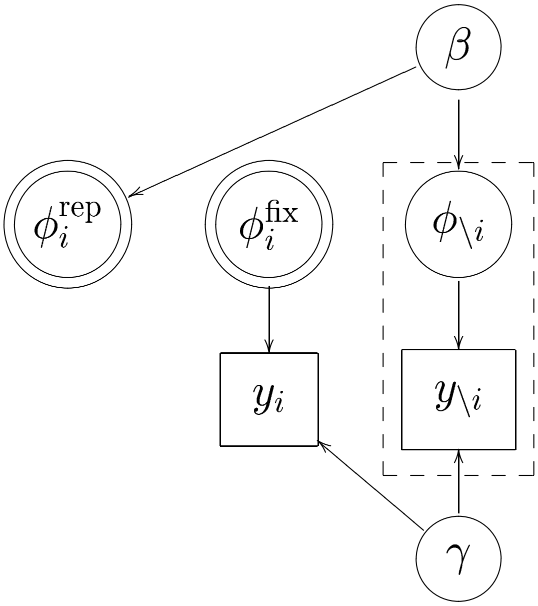

Figure 3: Checking for divergent ϕ_i based on comparing two parameter replicates: ϕ_i^{rep} simulated from the predictive prior and ϕ_i^{fix} simulated from, essentially, the normalised likelihood. The double ringed nodes represent quantities to be compared in order to assess divergence.

How may prior and likelihood replicates best be compared? O'Hagan (2003) suggests normalising the prior and the likelihood to have unit maximum height, and using the ordinate at which they cross as a measure of conflict. Our preference is to define a p-value. First define the difference in replicates as $\phi_i^{\text{diff}} = \phi_i^{\text{rep}} - \phi_i^{\text{fix}}$: then for scalar ϕ_i, we can calculate a 'conflict' p-value

$$P_{i,\text{con}} = \Pr(\phi_i^{\text{diff}} \leq 0|y) \tag{7}$$

from simulated values of ϕ_i^{diff}.

A primary reason for choosing this measure of conflict is to match, or closely approximate, the cross-validatory mixed p-value (6), when it exists. For example, consider the simple normal hierarchical model with likelihood

$$Y_i \sim \mathrm{N}(\phi_i, \sigma^2) \tag{8}$$

(so that in this instance $\gamma = \sigma$) and a predictive prior for ϕ_i (*i.e.* the posterior conditional on the remainder $y_{\setminus i}$ of the data) given by

$$\phi_i | y_{\setminus i} \sim \mathrm{N}(\tilde{\beta}, \tilde{\omega}^2), \tag{9}$$

where $\tilde{\beta}, \tilde{\omega}^2$ are functions of $y_{\setminus i}$ and known parameters - see A.2 in the Appendix for details. Then assuming a uniform prior for the fixed effect ϕ_i^{fix}, we have

$$\begin{array}{rcl} \phi_i^{\mathrm{fix}} | y_i^{\mathrm{obs}} & \sim & \mathrm{N}(y_i^{\mathrm{obs}}, \sigma^2) \\ \phi_i^{\mathrm{rep}} | y_{\setminus i} & \sim & \mathrm{N}(\tilde{\beta}, \tilde{\omega}^2) \end{array}$$

and so $\phi_i^{\mathrm{diff}} | y \sim \mathrm{N}(y_i^{\mathrm{obs}} - \tilde{\beta}, \sigma^2 + \tilde{\omega}^2)$. The p-value (7) measuring the conflict between prior and likelihood is therefore

$$P_{i,\mathrm{con}} = \Pr(\phi_i^{\mathrm{diff}} \leq 0 | y) = \Phi\left(\frac{y_i^{\mathrm{obs}} - \tilde{\beta}}{\sqrt{\sigma^2 + \tilde{\omega}^2}}\right). \tag{10}$$

However, from (8) and (9) we see that the predictive distribution is $Y_i^{\mathrm{rep}} | y_{\setminus i} \sim \mathrm{N}(\tilde{\beta}, \sigma^2 + \tilde{\omega}^2)$, and hence we obtain the cross-validatory mixed p-value (6)

$$P_{i,\mathrm{mix}} = \Pr(Y_i^{\mathrm{rep}} \leq y_i^{\mathrm{obs}} | y_{\setminus i}) = \Phi\left(\frac{y_i^{\mathrm{obs}} - \tilde{\beta}}{\sqrt{\sigma^2 + \tilde{\omega}^2}}\right), \tag{11}$$

so in these circumstances $P_{i,\mathrm{mix}} = P_{i,\mathrm{con}}$ exactly.

If ϕ_i is a vector of dimension n, let $\mathrm{E}[\phi_i^{\mathrm{diff}} | y] = m_{\phi_i}$, $\mathrm{Cov}[\phi_i^{\mathrm{diff}} | y] = \Sigma_{\phi_i}$. Then $m'_{\phi_i} \Sigma_{\phi_i}^{-1} m_{\phi_i}$ is a standardised discrepancy measure. If we were willing to assume a multivariate normal distribution for ϕ_i^{diff}, this discrepancy measure could be compared with a χ^2_n distribution. In Section A.2 we derive exact forms for the cross-validatory mixed discrepancy measures in general normal hierarchical models with known variances, show the resulting p-values have uniform distributions under H_0, and that they exactly match the p-values based on conflict between prior and likelihood. However, in non-normal models with multiple parameters per unit the resulting p-values should be interpreted with caution, since in general we might not expect a good χ^2 approximation which, in turn, would not give rise to a uniform distribution of p-values under the null hypothesis.

Two issues arise in generating the 'likelihood-based' replicates that represent fixed effect estimates: choice of a 'non-informative' prior and handling nuisance parameters.

First, the 'non-informative' prior for the fixed effect model should be chosen so that the conflict p-values match, as far as possible, the mixed p-values when they exist. In Section A.3 we show that if ϕ_i is a location parameter and an estimator $\hat{\phi}_i$ has a symmetric likelihood, then $P_{i,\mathrm{mix}} = P_{i,\mathrm{con}}$ when adopting a uniform prior for ϕ_i^{fix}. In other situations, Box and Tiao [p. 32] show that using Jeffreys' non-informative prior is equivalent to a locally uniform prior in a parameterisation that leads to approximate 'data-translated' likelihoods. We would therefore recommend the use of a Jeffreys' prior for ϕ_i when generating the fixed-effect posterior distribution: examples include $p(\phi) \propto \phi^{-\frac{1}{2}}$ for a Poisson mean ϕ, and $p(\phi) \propto \phi^{-\frac{1}{2}}(1-\phi)^{-\frac{1}{2}}$ for a Bernoulli mean ϕ: since we are concerned with parameters appearing in likelihoods for observed data the known difficulties with using Jeffreys' priors for hyper-parameters should not arise.

Second, problems arise in the presence of the nuisance parameters denoted γ in Figure 3, which may for example comprise error variances and regression coefficients at the observation level. In general these will not be estimable from individual units i. The simulated values from $p(\gamma|y_{\setminus i})$ may be used, as illustrated in Figure 3, which means that $y_{\setminus i}$ will slightly influence ϕ_i^{fix} and hence the two replications are not entirely independent. We would not, however, expect such an influence to be large. Care should also be taken that there is no 'feedback' from y_i when sampling γ, although again in practice we would not expect this to be very influential.

5 Example: Bristol Royal Infirmary Inquiry data

A public inquiry was set up following suggestions of excess mortality in complex paediatric cardiac surgery carried out at the Bristol Royal Infirmary prior to 1995. Part of the data presented to the Inquiry is shown in Table 1, based on national returns derived from patient administration systems: this source of data is treated with some suspicion by clinicians but data submitted to their professional register show much the same pattern. For detailed statistical discussion see Spiegelhalter et al. (2002). Publication of the final report of the Inquiry (BRI Inquiry Panel 2001) has led to substantial changes in health service monitoring in the UK.

The baseline analysis (Spiegelhalter et al. 2002) is a random-effects model

$$\begin{aligned} Y_i &\sim \text{Binomial}(\pi_i, n_i) \\ \text{logit}(\pi_i) &= \phi_i \\ \phi_i &\sim \text{N}(\beta, \omega^2). \end{aligned}$$

Independent uniform priors are initially assumed for β and ω. Each hospital is removed in turn from the analysis, the parameters re-estimated, and the observed deaths y_i^{obs} compared to the predictive distribution of $Y_i^{\text{rep}}|y_{\setminus i}$ to give the mixed p-value, defined as

$$P_{i,\text{mix}} = \Pr(Y_i^{\text{rep}} > y_i^{\text{obs}}) + \frac{1}{2}\Pr(Y_i^{\text{rep}} = y_i^{\text{obs}}) :$$

a 'mid' p-value is used given the discrete response. We note that, in contrast to (6), the upper tail-area is being used as a one-sided p-value since in this context these are the only departures of interest. All computations have been carried out in WinBUGS (Spiegelhalter et al. 2003) and are based on 100000 iterations following convergence. The left-hand side of Figure 4 illustrates these tail areas for Bristol (hospital 1) and Leicester (hospital 2), based on the 100000 simulations.

		Full data		'Tenth data'	
	Hospital	Operations n_i	Deaths y_i	Operations	Deaths
1	Bristol	143	41	14	4
2	Leicester	187	25	19	3
3	Leeds	323	24	32	2
4	Oxford	122	23	12	2
5	Guys	164	25	16	3
6	Liverpool	405	42	41	4
7	Southampton	239	24	24	2
8	Great Ormond St	482	53	48	5
9	Newcastle	195	26	20	3
10	Harefield	177	25	18	3
11	Birmingham	581	58	58	6
12	Brompton	301	31	30	3

Table 1: Numbers of open-heart operations and deaths for children under one year of age carried out in 12 hospitals in England between 1991 and 1995, as recorded by Hospital Episode Statistics. The 'tenth' data represents similar mortality rates but based on approximately one tenth of the sample size.

Spiegelhalter et al. (2002) calculated such p-values, and also presented summaries of the prior and likelihood for Bristol, although no formal measure of their conflict was made. Here we shall use $P_{i,\mathrm{con}}$ as such a formal measure. We assume a Jeffreys' prior for π_i^{fix}, and the resulting posterior distribution for π_i^{fix} and the predictive distribution for π_i^{rep} are shown in Figure 4 for Bristol (hospital 1) and Leicester (hospital 2), with P_{con} given by (7). P_{mix} and P_{con} are remarkably close even though the model does not strictly obey the criteria for an exact match.

Figure 5 compares mixed and conflict cross-validatory p-values for all hospitals, and their close agreement is clear. Any measure of agreement between p-values should reflect the importance of close agreement at the extremes, and therefore we first carry out an inverse-normal transformation to the real line to obtain z-scores $Z_{i,\mathrm{con}} = \Phi^{-1}(P_{i,\mathrm{con}})$, $Z_{i,\mathrm{mix}} = \Phi^{-1}(P_{i,\mathrm{mix}})$. The relative agreement between mixed and conflict cross-validatory p-values is then derived from the percentage relative agreement between $Z_{i,\mathrm{con}}$ and $Z_{i,\mathrm{mix}}$, defined as $100\,|Z_{i,\mathrm{con}} - Z_{i,\mathrm{mix}}|/Z_{i,\mathrm{mix}}$, and then taking the average of this quantity over the 12 hospitals.

The 'baseline analysis' column of Table 2 summarises the 12 conflict p-values, for example, by transforming to

$$X^2_{\mathrm{con}} = \sum_i Z^2_{i,\mathrm{con}},$$

which would have a χ^2_{12} distribution were the p-values to have independent Uniform(0,1) distributions. X^2_{con} and X^2_{mix} closely agree. The lack of independence between the p-values is illustrated by the fact that the X^2's are not 'significant', even with the extreme contribution from Bristol, and so the X^2's should only be used for comparing methods rather than assessing 'significance'.

The good agreement between $P_{\rm mix}$ and $P_{\rm con}$ may, however, reflect the approximate normality due to the substantial sample size. We therefore consider the 'tenth data' shown in Table 1 obtained by dividing all numerators and denominators by ten and rounding to the nearest integer. Table 2 shows the p-values still agree well, illustrating the good data-translation property of the Jeffreys' prior, even with small sample sizes.

Finally, the wide interval around the posterior estimate of ω illustrates the difficulty of accurately estimating a random-effects standard deviation ω with data on only 11 units, especially when assuming a uniform prior for ω. In order to assess the effect of assuming an informative prior for ω, we can use the fact that 95% of hospitals will have odds on death within a range of $\exp(\pm\ 1.96\ \omega)$. Reasonable homogeneity between hospitals would therefore correspond to values of ω less than around 0.2, since then the odds ratio between a 'high' mortality (97.5 centile) and 'low' mortality hospital (2.5 centile) would be around $\exp(2 \times 1.96 \times 0.2) = 2.2$. An informative prior was therefore adopted by assuming a 'half-normal' distribution, so that $\omega = |W|$, where

$$W \sim \mathrm{N}(0, \psi^2),$$

with $\psi = 0.1$. The resulting prior for ω has a mode at 0; a median of $0.67\ \times\psi = 0.067$ and 95% point $1.96\ \times\psi = 0.20$.

The results in Table 2 show that this informative prior leads to a low estimate 0.09 for ω when Bristol is excluded, and hence makes its p-value somewhat more extreme. The agreement between the mixed and conflict p-values remains good.

6 Alternatives to cross-validation

MCMC techniques are well-suited to generating replicate parameters and data at each iteration, but are ill-suited to a full cross-validatory approach in which each observation or unit has to be removed in turn and the analysis re-run. A number of approximations can be proposed: see Marshall and Spiegelhalter (2003) for further discussion and an example in spatial modelling.

1. *Full-data mixed replications*, which involve repeating the cross-validation procedure of Section 3 but without leaving y_i out of the analysis, thus generating $\phi_i^{\rm rep}|y$ followed by $Y_i^{\rm rep}$. This is illustrated in Figure 6(a). The generation of a new $\phi_i^{\rm rep}$ can be viewed as 'ghosting': for each unit in turn, a ghost unit is created in a parallel universe, and observations generated. Some conservatism will be introduced, but this may be only moderate as $y_i^{\rm obs}$ only influences $\phi_i^{\rm rep}$ through β. We denote the resulting full-data p-values $P_{\rm mix.f}$.

2. *Full-data prior and 'likelihood' replications* in which $y_i^{\rm obs}$ contributes to the simulation of $\phi_i^{\rm rep}$ through influencing, to some extent, β. This is illustrated in Figure 6(b). We might expect this to have similar properties to full-data mixed replication. A replicate set of data is required to simulate $\phi_i^{\rm fix}$, which is then

Quantity	Baseline analysis	Tenth data	Informative prior on ω
P_{mix} for Bristol (hosp. 1)	0.00201	0.0556	0.00001
P_{con} for Bristol (hosp. 1)	0.00191	0.0483	0.00001
X^2_{mix}	12.9	5.7	29.4
X^2_{con}	13.0	6.2	28.2
Relative error			
(mixed *vs* conflict)	0.8%	4.1%	0.8%
Between-hospital sd ω:			
excluding Bristol (hosp. 1)			
median	0.19	0.17	0.09
95% interval	(0.01 to 0.46)	(0.01 to 0.68)	(0.004 to 0.22)
excluding Leicester (hosp. 2)			
median	0.44	0.19	0.25
95% interval	(0.02 to 0.46)	(0.01 to 0.75)	(0.14 to 0.36)

Table 2: Comparison of mixed and conflict p-values when carried out using full cross-validation, in which each hospital is removed in turn. X^2 statistics are a composite measure of the extremity of the p-values. 'Tenth data' corresponds to original counts divided by ten and rounded to the nearest integer, while the informative prior on ω reflects prior belief in reasonable homogeneity of centres.

compared with ϕ_i^{rep} to create a measure of conflict. Care is required when programming in WinBUGS (Spiegelhalter et al. 2003) or elsewhere, as there must not be feedback to γ from the replicate set of data otherwise the precision of γ will be over-estimated. We denote the resulting p-values $P_{\text{con.f}}$.

3. *Full-data posterior predictive p-values* in which data are replicated from $p(Y_i^{\text{rep}}|\phi_i, \gamma)$ (Figure 6(c)). This is likely to be very conservative as y_i^{obs} strongly influences ϕ_i and hence Y_i^{rep} will tend to agree too well with y_i^{obs}. Although Gelman et al. (1995) and Gelman et al. (1996) do not recommend this procedure for outlier detection, their recommended global measures of goodness-of-fit are based on it (Section 8). We denote the resulting p-values $P_{\text{post-pred.f}}$.

In Section A.4 we extend the analysis of Marshall and Spiegelhalter (2003) and show that the full-data mixed (and equivalently the prior and 'likelihood' replication) technique only induces mild conservatism in the normal hierarchical model, whereas, as noted by Bayarri and Berger (2000) and Robins et al. (2000), the full-data posterior predictive p-values can be very conservative.

Figure 7 shows the results from applying these three approximate methods to the Bristol data, comparing with the 'gold-standard' cross-validatory mixed p-values. Both the full-data mixed (a) and conflict (b) p-values are little influenced by the lack of cross-validation, as might be expected from the analysis in Section A.4. In contrast, Figure 7(c) reveals the predicted conservatism of the posterior predictive p-values.

Table 3 reproduces the previous results but using the the full data, and shows there is still good agreement between mixed and conflict p-values. All the results show the deep conservatism of the posterior predictive values, emphasising the danger of producing replicates that are heavily dependent on the data being checked.

7 Example: Longitudinal model following hepatitis vaccination

Spiegelhalter et al. (1996) considered the following example involving vaccination against hepatitis B (HB) in which 2 or 3 longitudinal log(anti-HB titre) observations are available from each of I =106 children, comprising 288 observations in all. The null model H_0 assumes normal errors and random coefficient linear growth curves with time t measured on a centred logarithmic scale, and a covariate μ_0 comprising the baseline log(titre). The covariate is not directly observed, but a baseline observation y_0 is assumed made with the same error as the subsequent observations.

Quantity	Baseline analysis	Tenth data	Informative prior on ω
for Bristol (hosp. 1):			
$P_{\text{mix.f}}$	0.024	0.068	0.0012
$P_{\text{con.f}}$	0.024	0.059	0.0013
$P_{\text{post-pred.f}}$	0.23	0.096	0.058
$X^2_{\text{mix.f}}$	7.9	5.0	17.4
$X^2_{\text{con.f}}$	7.9	5.5	17.3
$X^2_{\text{post-pred.f}}$	1.1	3.9	4.8
% Relative error *vs* cross-validatory mixed			
Full-data mixed	7.2%	8.0%	10.4%
Full-data conflict	7.0%	6.5%	10.3%
Full-data post-pred	95.5%	22.6%	69.6%
Between-hospital sd ω			
median	0.40	0.18	0.24
95% interval	(0.23 to 0.73)	(0.01 to 0.70)	(0.13 to 0.35)

Table 3: Comparison of mixed, conflict and posterior-predictive p-values when using the full data for prediction rather than leaving each centre out in turn. The cross-validatory mixed approach is considered the 'gold-standard'.

For the jth measurement on the ith individual we therefore assume

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \phi_{i1} + \phi_{i2} t_{ij} + \gamma \mu_{i0} \\ \phi_i &\sim \mathrm{N}_2(\beta, \Omega) \\ y_{i0} &\sim \mathrm{N}(\mu_{i0}, \sigma^2) \\ \mu_{i0} &\sim \mathrm{N}(\eta, \tau^2) \end{aligned}$$

where $\beta, \gamma, \log(\sigma), \eta$ and $\log(\tau)$ are given locally uniform prior distributions and a diffuse Wishart prior is specified for Ω^{-1}.

This example is important because the functional form has a strong scientific interpretation, and systematic deviations in individuals could cast doubt on the underlying model. The presence of measurement error on the covariate means that standard estimation and diagnostic procedures are not readily available.

Divergent children have been investigated using conflict p-values of both intercept ϕ_{i1} and gradient ϕ_{i2}, in which predictive prior replicates generated conditional on all observed data are contrasted with replicates generated conditional only on the data for that child, based on full data without cross-validatory removal of each case. Figure 8 shows that the intercepts show reasonable agreement with a uniform distribution. In contrast, the gradients exhibit a cluster of surprisingly small p-values: with false discovery rate (FDR) (Benjamini and Hochberg 1995) of 0.05, four gradients are identified as being divergent: no adjustment for non-independent p-values has been made since we view this full-data approach as exploratory. These cases are highlighted in Figure 9 and Table 4.

Although the focus in this paper has been on the detection of divergent units, the divergence of individual *observations* may also be of interest. This could be checked by an approximate cross-validation using mixed replicates which are sensitive to both child and observation divergence, or posterior-predictive replicates which target divergence *within* a child. However, neither approach identifies individually divergent observations using a FDR of 0.05, but this lack of sensitivity might be expected in this example given very limited data per child.

The posterior predictive approach essentially seeks to identify individual observations that do not fit the fitted straight line for each child: examination of Figure 9 reveals that, for example, the fitted line for Child 20 will be strongly shrunk towards the common gradient, and hence neither of the two observations will fit the fitted line for that child, the first observation being low and the second high. For Child 34 the second observation is identified as high, for Child 76 the first is high and for Child 95 the first is low. In contrast, the mixed approach simply identifies individual observations that lie outside the general 'cloud', without taking into account the other observations on that child.

In order to assess the impact of the divergent children on the conclusions of interest, we removed their observations and re-analysed the data. Applying approximate

	Divergent children?		Divergent observations?		
	Intercept	Gradient			
Child	$P_{\text{con}}(\phi_{i1})$	$P_{\text{con}}(\phi_{i2})$	observation	$p_{\text{post-pred}}$	p_{mix}
20	0.0014	0.0000	1	0.999	0.966
			2	0.002	0.018
34	0.584	0.0003	1	0.065	0.664
			2	0.025	0.012
76	0.456	0.0014	1	0.018	0.169
			2	0.893	0.907
			3	0.889	0.921
95	0.568	0.0004	1	0.968	0.935
			2	0.758	0.692
			3	0.254	0.059

Table 4: Children from hepatitis data with any p-value less than 0.0025 or greater than 0.9975. Diagnostics for divergent children are based on replication of intercepts and gradients separately. Divergent observations are examined by posterior-predictive and mixed p-values.

cross-validatory predictive checks did not highlight any further divergent children or observations (data not shown). In this particular example, the conclusions are fairly insensitive to the outliers – Table 5 shows the data are compatible with $\beta_2 = -1, \gamma = 1$ which are important and interpretable values corresponding to titre/baseline being inversely proportional to time.

Parameter	Complete data		Outliers removed	
β_1	6.04	(5.70,6.38)	6.06	(5.72,6.39)
β_2	-1.07	(-1.34,-0.79)	-1.12	(-1.35,-0.90)
γ	0.98	(0.65,1.17)	0.82	(0.59,1.06)

Table 5: Posterior means and 95% credible intervals for the parameters of interest, with and without data from the four divergent children.

8 Some conclusions and a recommended strategy

In this paper we have attempted to develop a practical approach to model checking in Bayesian hierarchical models, that does not rely on analytical results or approximations and could be widely implemented in MCMC software. This has required a careful approach to handling predictive distributions through generating replicate parameters and observations, in particular making a clear distinction between the mixed approach, in which full 'ghost' sets of random effects are generated, and the posterior-predictive approach, in which the generated random effects depend directly on the observed data. The mixed approach can be approximated by a 'conflict' p-value, and leads us naturally to consider measures of conflict between prior and likelihood as the basis for model diagnostics.

Ideally one would carry out a full cross validation for individual units or sets of units, or alternatively use the partial approach of Bayarri and Berger (2000) in order to avoid any conservatism and ensure proper p-values with Uniform (0,1) distribution under the null hypothesis that the model assumptions are appropriate. However the partial approach appears only to be easily implemented in rather simple models, and full cross-validation approach is generally impractical within an MCMC analysis, except as a response to a preliminary screen. Our analytic and practical results suggest a full-data mixed approach provides a reasonable basis for preliminary screening.

As mentioned previously, it may at first appear strange that a Bayesian modelling procedure should lead to considerations of multiple comparisons and adjustment of p-values and so on. A similar debate has been going on in other areas in which many hypotheses are being tested, such as functional magnetic resonance imaging (Genovese et al. 2002) and microarray data (Efron and Tibshirani 2002), in which it has been suggested that a more appropriate Bayesian procedure would be to model an alternative hypothesis and hence produce posterior probabilities of the null and alternative hypothesis, rather than p-values (Efron et al. 2001). However, while this may be possible within a restricted domain such as microarrays, it does not seem feasible within generic hierarchical modelling and so a multiple p-value procedure, in which no precise alternative hypothesis is specified, seems reasonable.

The theoretical and empirical investigations described in this paper lead us to make the following recommendations in each of three potential scenarios in the criticism of hierarchical models.

1. If concern lies solely with the random-effects prior distribution $p(\phi_i|\beta)$ (*i.e.* we are prepared to believe the form of likelihood) then there are two potential strategies depending on whether or not estimates of the random effects ϕ are available in closed form. If closed form estimates are available, *mixed* replication can be used in which the unit-specific parameter estimates themselves are the discrepancy functions (i.e. $T_i(y) = \hat{\phi}_i\,;\, i = 1, .., I$). The realised discrepancy statistics are then compared to their predictive distribution, where the latter may be the true cross-validatory predictive distribution $p(\hat{\phi}_i^{\text{rep}}|y_{\setminus i})$ (Section 3), or the mixed predictive distribution $p(\hat{\phi}_i^{\text{rep}}|y)$ (Section 6) as a preliminary screen, followed by the full cross-validation for 'suspicious' units.

   If closed form estimates do not exist then *conflict* p-values can be used in which the (predictive) prior, ϕ_i^{rep}, and 'likelihood', ϕ_i^{fix} replicates are contrasted. Note that the former may be derived ignoring that from unit i (Section 4) or, less accurately, conditional on *all* the data (Section 6).

2. If concern lies solely with the likelihood $p(y_i|\phi_i)$ for a vector y_i, then posterior predictive replication of individual observations may be appropriate – see Section 6 (point 3). However, even in this context there may be considerable conservatism and so ideally a more precise approach, such as the partial procedure of (Bayarri and Berger 2000), should be used.

3. If we are concerned with both prior *and* likelihood, then the strategy depends on whether y_i is a scalar or vector: (i) if a scalar, then *mixed* replication of individual observations can be used, although it will not be possible to then distinguish failures in the specification of the prior or the likelihood; (ii) if a vector, we recommend a two stage 'bottom-up' process in which strategy 2 is first used to check the likelihood assuming independent parameters (fixed effect estimates) (Hilden-Minton 1995), and then strategy 1 is used to check the prior.

A Appendices

A.1 The distribution of cross-validatory p-values

For p-values to be useful they should have a sound calibration, and it is well-known that a classic 'proper' p-value will have a uniform (0,1) distribution under the null hypothesis. However a number of interpretations can be given to this statement depending on the null hypothesis underlying the reference distribution: below we consider three null hypotheses comprising a conditional distribution, a full joint distribution of prior and parameters, and a 'true' sampling model.

First, we consider a cross-validatory p-value for a scalar y_i

$$\mathcal{P}(y_i) = P_0(Y_i^{\text{rep}} \leq y_i|y_{\setminus i})$$

as a function of y_i alone. Then if Y_i^{rep} is truly generated from $p_0(y_i|y_{\setminus i})$, then $\mathcal{P}$ has a uniform distribution since

$$\text{E}_{Y_i|Y_{\setminus i}}[\mathcal{P}(Y_i) \leq \alpha] = \alpha.$$

Second, we take a pre-posterior Bayesian approach, in which

$$\mathcal{P}(Y) = P_0(Y_i^{\text{rep}} \leq Y_i|Y_{\setminus i})$$

is considered as a function of the entire future data Y. Let θ denote the parameters of the model that must be provided with prior distributions, and assume that $p(\theta)$ is proper and hence $p_0(y) = \int p(y|\theta)p(\theta)d\theta$ is proper for any y. Then, as a special case of Theorem 1 of Bayarri and Berger (2000), it follows that

$$P(\mathcal{P}(Y) \leq \alpha) = \text{E}_{Y_{\setminus i}}\text{E}_{Y_i|Y_{\setminus i}}[\mathcal{P}(Y) \leq \alpha] = \text{E}_{Y_{\setminus i}}[\alpha] = \alpha, \qquad (12)$$

and hence the p-value also has a uniform distribution with respect to the predictive distribution before sampling any data, but conditional on the truth of the prior $p(\theta)$ and the likelihood $p(y|\theta)$, and hence the joint distribution $p_0(y)$.

Third, we consider the crucial question addressed by Bayarri and Berger (2000) and Robins et al. (2000) concerning whether the p-values derived from a Bayesian argument in fact are frequentist p-values, in the sense that they have uniform (0,1) distributions under the 'true' model $p(y|\theta^T)$. Suppose the distribution of $\mathcal{P}(Y)$ does not depend on θ, which will certainly be the case if we are concerned with checking aspects of the model that do not depend on θ. Then Bayarri and Berger (2000) show that $\mathcal{P}$ will be a frequentist p-value when $p(\theta)$ is proper, and also if $p(\theta)$ is improper provided certain conditions are fulfilled (as is the case for reference priors in location-scale problems). This follows from the observation that, reversing the argument of (12),

$$\alpha = P(\mathcal{P}(Y) \leq \alpha) = \mathrm{E}_\theta\left[\mathrm{E}_{Y|\theta}[\mathcal{P}(Y) \leq \alpha]\right],$$

and so, since $\mathrm{E}_{Y|\theta}[\mathcal{P}(Y)]$ does not depend on θ, then $\mathrm{E}_{Y|\theta^T}[\mathcal{P}(Y) \leq \alpha] = \alpha$.

Thus the cross-validatory p-values should in general have uniform (0,1) distributions under the 'true' sampling model, as well as under the assumed Bayesian joint distribution of parameters and data. Robins et al. (2000) show that, under certain conditions, posterior predictive p-values do not have uniform (0,1) distributions even asymptotically, but their condition concerning normality of the test statistic is not fulfilled in our context.

A.2 Cross-validatory mixed replication in the Normal hierarchical model

Although in practice the relevant p-values and standardised discrepancy measures may be obtained using MCMC methods, it is illuminating to consider circumstances in which analytic forms are available. Consider the normal hierarchical model

$$\begin{aligned} \phi_i &\sim \mathrm{N}_p(\beta, \Omega) \\ Y_i &\sim \mathrm{N}_{n_i}(X_i\phi_i, \sigma^2 I_{n_i}) \end{aligned}$$

where X_i is a known design matrix, I_{n_i} is the $n_i \times n_i$ identity matrix, and σ^2 and Ω are assumed known.

Suppose we take $T_i = y_i$, a vector of length n_i. Then conditional on β,

$$Y_i^{\mathrm{rep}}|\beta \sim \mathrm{N}_{n_i}(X_i\beta, X_i\Omega X_i' + \sigma^2 I_{n_i}).$$

Assuming a normal prior for β results in a normal posterior distribution, conditional on the remaining data $y_{\setminus i}$, denoted $\beta|y_{\setminus i} \sim \mathrm{N}_p(\tilde{\beta}_{\setminus i}, \tilde{\Gamma}_{\setminus i})$. Algebraic forms for $\tilde{\beta}_{\setminus i}, \tilde{\Gamma}_{\setminus i}$ can be obtained, for example, from Lindley and Smith (1972). Thus the predictive distribution for Y_i^{rep} is

$$Y_i^{\mathrm{rep}}|y_{\setminus i} \sim \mathrm{N}_{n_i}(X_i\tilde{\beta}_{\setminus i}, X_i(\Omega + \tilde{\Gamma}_{\setminus i})X_i' + \sigma^2 I_{n_i})$$

and this forms the reference distribution with which to compare $T_i^{\mathrm{obs}} = y_i^{\mathrm{obs}}$.

We have argued, however, that it will generally be more powerful to take $T_i = \hat{\phi}_i = (X_i'X_i)^{-1}X_i'y_i$, the least squares estimate and also the posterior mean under a locally uniform prior for ϕ_i. Then conditional on β,

$$\hat{\phi}_i^{\mathrm{rep}}|\beta \sim \mathrm{N}_p(\beta, \Omega + \sigma^2(X_i'X_i)^{-1}).$$

Thus the predictive prior distribution for $\hat{\phi}_i^{\mathrm{rep}}$ is

$$\hat{\phi}_i^{\mathrm{rep}}|y_{\backslash i} \sim \mathrm{N}_p(\tilde{\beta}_{\backslash i}, \Omega + \tilde{\Gamma}_{\backslash i} + \sigma^2(X_i'X_i)^{-1}) \tag{13}$$

and this forms the reference distribution with which to compare $T_i^{\mathrm{obs}} = \hat{\phi}_i^{\mathrm{obs}}$. In this situation the standardised cross-validatory mixed discrepancy measure is

$$X^2_{\mathrm{cross:mixed}} = (\hat{\phi}_i^{\mathrm{obs}} - \tilde{\beta}_{\backslash i})'(\Omega + \tilde{\Gamma}_{\backslash i} + \sigma^2(X_i'X_i)^{-1})^{-1}(\hat{\phi}_i^{\mathrm{obs}} - \tilde{\beta}_{\backslash i}) \tag{14}$$

which is distributed as a χ^2_p random variable under the full Bayesian model H_0. The sampling distribution of $X^2_{\mathrm{cross:mixed}}$ is independent of the location parameter β which has been given a reference prior distribution, and hence by Theorem 1 of Bayarri and Berger (2000) (Section A.1) the resulting p-values should have uniform (0,1) distributions under the true sampling model in H_0.

To obtain the conflict measure, the predictive distribution for $\hat{\phi}_i^{\mathrm{rep}}$ is given by (13). Assuming a locally uniform prior for ϕ_i provides a fixed-effect distribution

$$\phi_i^{\mathrm{fix}} \sim \mathrm{N}_p(\hat{\phi}_i^{\mathrm{obs}}, \sigma^2(X_i'X_i)^{-1}).$$

The standardised cross-validatory conflict discrepancy between these distributions is denoted

$$X^2_{\mathrm{cross:conflict}} = (\hat{\phi}_i^{\mathrm{obs}} - \tilde{\beta}_{\backslash i})'(\Omega + \tilde{\Gamma}_{\backslash i} + \sigma^2(X_i'X_i)^{-1})^{-1}(\hat{\phi}_i^{\mathrm{obs}} - \tilde{\beta}_{\backslash i}), \tag{15}$$

precisely the measure (14) obtained when directly predicting $\hat{\phi}_i^{\mathrm{obs}}$: this was illustrated in the simple normal case in Section 4. Theil (1963) essentially obtains (15) and terms it a prior/data 'compatibility statistic'.

Thus, in normal models with known variance, comparison of the observed parameter estimates with their cross-validatory predictive distribution under the null model gives exactly the same conclusions as comparing the two independent sources of evidence, prior and likelihood, concerning the individual parameters. This was illustrated in Section 4 using the simplest univariate case, in which $n_i = 1, X_i = 1, p = 1, \phi_i \sim \mathrm{N}(\beta, \omega^2)$. Assuming a locally uniform prior for β, it follows that $\tilde{\beta}_{\backslash i} = \overline{y}_{\backslash i}$ and $\tilde{\Gamma}_{\backslash i} = (\omega^2 + \sigma^2)/(I-1)$, and hence $\phi_i^{\mathrm{rep}}|y_{\backslash i} \sim \mathrm{N}\left(\overline{y}_{\backslash i}, (I\omega^2 + \sigma^2)/(I-1)\right)$. Hence in the notation of (9), $\tilde{\beta} = \overline{y}_{\backslash i}, \tilde{\omega}^2 = (I\omega^2+\sigma^2)/(I-1)$ - the equivalence of the two approaches is shown in (10) and (11).

A.3 When will the cross-validatory conflict p-value be the same as the mixed p-value ?

Apart from the normal model described in Section 4, we can identify other circumstances in which the conflict and mixed approaches provide identical results. Suppose ϕ_i is scalar and an estimator $\hat{\phi}_i$ exists based on a sufficient statistic, then the two p-values may be written as

$$\begin{aligned} P_{\text{mix}} &= \int \Pr(\hat{\phi}_i^{\text{rep}} \leq \hat{\phi}_i^{\text{obs}} | \phi_i^{\text{rep}}) p(\phi_i^{\text{rep}} | y_{\setminus i}) d\phi_i^{\text{rep}} \\ P_{\text{con}} &= \int \Pr(\phi_i^{\text{fix}} > \phi_i^{\text{rep}} | \hat{\phi}_i^{\text{obs}}) p(\phi_i^{\text{rep}} | y_{\setminus i}) d\phi_i^{\text{rep}}. \end{aligned}$$

Thus a sufficient condition for equality of P_{mix} and P_{con} is that $\Pr(\hat{\phi}_i^{\text{rep}} \leq \hat{\phi}_i^{\text{obs}} | \phi_i^{\text{rep}}) = \Pr(\phi_i^{\text{fix}} > \phi_i^{\text{rep}} | \hat{\phi}_i^{\text{obs}})$ for all values of ϕ_i^{rep}.

We show that this will be the case under the conditions (i) ϕ_i is a location parameter and $\phi_i - \hat{\phi}_i$ is a pivotal quantity *i.e.* $p(\hat{\phi}_i | \phi_i) = f(\hat{\phi}_i - \phi_i)$, (ii) the likelihood $p(\hat{\phi}_i | \phi_i)$ is symmetric around ϕ_i so that $f(\hat{\phi}_i - \phi_i) = f(\phi_i - \hat{\phi}_i)$, and (iii) ϕ_i is given a locally uniform prior for the fixed effect estimate. These conditions imply that, as functions of $\hat{\phi}_i$ and ϕ_i, $p(\hat{\phi}_i | \phi_i) = f(\hat{\phi}_i - \phi_i) = f(\phi_i - \hat{\phi}_i) = p(\phi_i | \hat{\phi}_i)$. Hence

$$\begin{aligned} \Pr(\phi_i^{\text{fix}} > \phi_i^{\text{rep}} | \hat{\phi}_i^{\text{obs}}) &= \int_{\phi_i^{\text{rep}}}^{\infty} p(\phi_i^{\text{fix}} | \hat{\phi}_i^{\text{obs}}) d\phi_i^{\text{fix}} \\ &= \int_{\phi_i^{\text{rep}}}^{\infty} f(\phi_i^{\text{fix}} - \hat{\phi}_i^{\text{obs}}) d\phi_i^{\text{fix}} \\ &= \int_{\phi_i^{\text{rep}} - \hat{\phi}_i^{\text{obs}}}^{\infty} f(u) du \end{aligned}$$

where $u = \phi_i^{\text{fix}} - \hat{\phi}_i^{\text{obs}}$. Now u has the same distribution as $-v$, where $v = \hat{\phi}_i^{\text{rep}} - \phi_i^{\text{rep}}$. So we have

$$\begin{aligned} \Pr(\phi_i^{\text{fix}} > \phi_i^{\text{rep}} | \hat{\phi}_i^{\text{obs}}) &= \int_{-\infty}^{\hat{\phi}_i^{\text{obs}} - \phi_i^{\text{rep}}} f(v) dv \\ &= \int_{-\infty}^{\hat{\phi}_i^{\text{obs}}} f(\hat{\phi}_i^{\text{rep}} - \phi_i^{\text{rep}}) d\hat{\phi}_i^{\text{rep}} \\ &= P(\hat{\phi}_i^{\text{rep}} \leq \hat{\phi}_i^{\text{obs}} | \phi_i^{\text{rep}}). \end{aligned}$$

Hence for location parameters in symmetric distributions with known scale parameters the conflict and mixed approaches will give identical results. As discussed in Section 3, this should also be approximately true in 'data-translated' likelihoods and hence argues for the use of Jeffreys' prior for the fixed-effect replications.

A.4 Full data in the Normal hierarchical model

For the mixed replication, the development follows that of Section A.2 except that $\tilde{\beta}, \tilde{\Gamma}$ are based on the full data. Thus

$$X^2_{\text{full:mixed}} = (\hat{\phi}_i^{\text{obs}} - \tilde{\beta})'(\Omega + \tilde{\Gamma} + \sigma^2 (X_i' X_i)^{-1})^{-1} (\hat{\phi}_i^{\text{obs}} - \tilde{\beta}) \tag{16}$$

which will not have an exact χ^2_p distribution under the null hypothesis. To illustrate the conservatism resulting from using the full data instead of leaving out unit i, we consider the fully balanced case in which $X_i = X, i = 1, ..., K$. Then it is straightforward to show that

$$\begin{aligned} K\tilde{\Gamma} &= (K-1)\tilde{\Gamma}_{\setminus i} = \Omega + \sigma^2(X_i'X_i)^{-1} \\ K\tilde{\beta} &= (K-1)\tilde{\beta}_{\setminus i} + \hat{\phi}_i^{\text{obs}} = \sum_{j=1}^{K} \hat{\phi}_j^{\text{obs}}. \end{aligned}$$

Substituting into (16), rearranging, and comparing with (14) reveals that,

$$X^2_{\text{full:mixed}} = X^2_{\text{cross:mixed}} \left(\frac{K-1}{K+1} \right).$$

Using the 'incorrect' full data X^2 statistic therefore reduces the 'correct' X^2 by a factor $(K-1)/(K+1)$ in this balanced case, introducing a fairly moderate (and potentially correctable) conservatism. The results are exactly the same for the conflict procedure shown in Figure 6(b).

For the posterior replication shown in Figure 6(c), we first note that ϕ_i has a prior $\phi_i | y_{\setminus i} \sim \mathrm{N}_p(\tilde{\beta}_{\setminus i}, \Omega + \tilde{\Gamma}_{\setminus i})$ and likelihood $\hat{\phi}_i^{\text{obs}} | \phi_i \sim \mathrm{N}_p(\phi_i, \sigma^2(X_i'X_i)^{-1})$, which entails a posterior distribution of ϕ_i given all the data

$$\phi_i | y \sim \mathrm{N}_p(\tilde{\phi}_i, (\sigma^{-2}X_i'X_i + (\Omega + \tilde{\Gamma}_{\setminus i})^{-1})^{-1}),$$

where $\tilde{\phi}_i = \sigma^{-2}X_i'X_i\hat{\phi}_i^{\text{obs}} + (\Omega + \tilde{\Gamma}_{\setminus i})^{-1}\tilde{\beta}_{\setminus i}$. Thus the replicate distribution with which to compare $\hat{\phi}_i^{\text{obs}}$ is

$$\hat{\phi}_i^{\text{rep}} | y \sim \mathrm{N}_p(\tilde{\phi}_i, \sigma^2(X_i'X_i)^{-1} + (\sigma^{-2}X_i'X_i + (\Omega + \tilde{\Gamma}_{\setminus i})^{-1})^{-1}). \tag{17}$$

Considering the fully balanced case $X_i = X, i = 1, ..., K.$, it can be shown that

$$X^2_{\text{full:post}} = (\hat{\phi}_i^{\text{obs}} - \tilde{\beta}_{\setminus i})' D (\Omega + \tilde{\Gamma}_{\setminus i} + \sigma^2(X_i'X_i)^{-1})^{-1} (\hat{\phi}_i^{\text{obs}} - \tilde{\beta}_{\setminus i}) \tag{18}$$

where $D = (K-1)(2K\Omega\sigma^{-2}X_i'X_i + (K+1)I_p)^{-1}$. Comparison with (14) reveals D as a measure of conservatism which will increase as the between-unit variability Ω increases. In the simplest situation in which $n_i = 1, X_i = 1, \phi_i \sim \mathrm{N}(\beta, \omega^2)$, then $D = (K-1)/(2K\omega^2/\sigma^2 + K + 1) \approx (2\omega^2/\sigma^2 + 1)^{-1}$ for large K, showing the strong conservatism induced when the between-unit to within-unit variability increases.

References

Albert, J. and Chib, S. (1997). "Bayesian tests and model diagnostics in conditionally independent hierarchical models." *Journal of the American Statistical Association*, 92: 916–925. 412

Bayarri, M. J. and Berger, J. O. (2000). "P-values for composite null models." *Journal of the American Statistical Association*, 95: 1127–1142. 410, 414, 417, 424, 428, 429, 430, 431

Bayarri, M. J. and Castellanos, M. E. (2004). "Bayesian checking of hierarchical models." Technical report, Univ. Valencia. 414

Bayarri, M. J. and Morales, J. (2003). "Bayesian Measures of Surprise for outlier detection." *Journal of Statistical Planning and Inference*, 111: 3 22. 414

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society - Series B*, 57: 289–300. 415, 426

Benjamini, Y. and Yekutieli, D. (2001). "The Control of the False Discovery Rate Under Dependency." *The Annals of Statistics*, 29: 1165–1188. 415

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester, England: John Wiley and Sons. 415

Box, G. E. P. (1980). "Sampling and Bayes inference in scientific modelling and robustness (with discussion)." *Journal of the Royal Statistical Society - Series A*, 143: 383–430. 410, 413, 414, 417

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison–Wesley. 420

BRI Inquiry Panel (2001). *Learning from Bristol: The report of the Public Inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995*. London, UK: The Stationery Office. 420

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Newbury Park: Sage. 413

Carlin, B. P. (1999). "Discussion of 'Quantifying surprise in the data and model verification' by Bayarri and Berger." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 6*, 73–74. Oxford, UK: Oxford University Press. 417

Carota, C., Parmigiani, G., and Polson, N. G. (1996). "Diagnostic measures for model criticism." *Journal of the American Statistical Association*, 91: 753–762. 412

Chaloner, K. (1994). "Residual analysis and outliers in Bayesian hierarchical models." In Freeman, P. R. and Smith, A. F. M. (eds.), *Aspects of Uncertainty: a Tribute to D V Lindley*, 149–157. Chichester: Wiley. 413

Chaloner, K. and Brant, R. (1988). "A Bayesian approach to outlier detection and residual analysis." *Biometrika*, 75: 651–659. 413

Dempster, A. P. and Ryan, L. M. (1985). "Weighted normal plots." *Journal of the American Statistical Association*, 80: 84–850. 412

Dey, D. K., Gelfand, A. E., Swartz, T. B., and Vlachos, P. K. (1998). "A simulation-intensive approach for checking hierarchical models." *Test*, 7: 325–346. 414

Draper, D. (1996). "Comment: utlity, sensitivity analysis, and cross-validation in Bayesian model checking. Discussion of 'Posterior predictive assessment of model fitness via realized discrepancies' by Gelman, Meng and Stern." *Statistica Sinica*, 6: 760–767. 417

Efron, B. and Tibshirani, R. (2002). "Empirical Bayes methods and false discovery rates for microarrays." *Genetic Epidemiology*, 23: 70–86. 428

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). "Empirical Bayes analysis of a microarray experiment." *Journal of the American Statistical Association*, 96: 1151–1160. 428

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model determination using predictive distributions, with implementation via sampling-based methods." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*, 147–168. Oxford, UK: Oxford University Press. 410, 415, 417

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall. 414, 424

Gelman, A., Meng, X.-L., and Stern, H. S. (1996). "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, 6: 733–760. 410, 413, 415, 424

Genovese, C. R., Lazar, N. A., and Nichols, T. E. (2002). "Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate." *NeuroImage*, 15: 870–878. 428

Goldstein, H. (1995). *Multilevel Models in Educational and Social Research, Second Edition*. London, U.K.: Edward Arnold. 409, 413

Hardy, R. and Thompson, S. (1998). "Detecting and describing heterogeneity in meta-analysis." *Statistics in Medicine*, 17: 841–856. 412

Hilden-Minton, J. A. (1995). *Multilevel Diagnostics for Mixed and Hierarchical Linear Models*. Ph.D. Thesis, University of California, Los Angeles. 413, 429

Hodges, J. S. (1998). "Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion)." *Journal of the Royal Statistical Society, Series B*, 60: 497–521. 412, 413

Lange, N. and Ryan, L. (1989). "Assessing normality in random effects models." *The Annals of Statistics*, 17: 624–642. 412

Langford, I. H. and Lewis, T. (1998). "Outliers in Multilevel Data." *Journal of the Royal Statistical Society, Series A*, 161: 121–153. 410

Lindley, D. V. and Smith, A. F. M. (1972). "Bayes estimates for the linear model." *Journal of the Royal Statistical Society - Series B*, 34: 1–41. 430

Marshall, E. C. and Spiegelhalter, D. J. (2003). "Approximate cross-validatory predictive checks in disease-mapping models." *Statistics in Medicine*, 22: 1649–1660. 410, 422, 424

Normand, S.-L., Glickman, M. E., and Gatsonis, C. A. (1997). "Statistical methods for profiling providers of medical care: issues and applications." *Journal of the American Statistical Association*, 92: 803–814. 409

O'Hagan, A. (1994). *Bayesian Inference*. London: Edward Arnold. 417

— (2003). "HSSS Model Criticism." In Green, P. J., Hjort, N. L., and Richardson, S. T. (eds.), *Highly Structured Stochastic Systems*. Oxford: Oxford University Press. 410, 417, 418

Pettit, L. I. and Smith, A. F. M. (1985). "Outliers and influential observations in Linear models." In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.), *Bayesian Statistics 2*, 473–494. Amsterdam: North-Holland. 415

Robins, J. M., van der Vaart, A., and Ventura, V. (2000). "Asymptotic distribution of P-values in composite null models." *Journal of the American Statistical Association*, 95: 1143–1156. 424, 430

Rubin, D. B. (1984). "Bayesian justifiable and relevant frequency calculations for the applied statistician." *Annals of Statistics*, 12: 1151–1172. 413

Sharples, L. D. (1990). "Identification and accommodation of outliers in general hierarchical models." *Biometrika*, 77: 445–53. 412

Spiegelhalter, D. J. (1998). "Bayesian graphical modelling: a case-study in monitoring health outcomes." *Applied Statistics*, 47: 115–134. 410

Spiegelhalter, D. J., Aylin, P., Evans, S. J. W., Murray, G. D., and Best, N. G. (2002). "Commissioned analysis of surgical performance using routine data: lessons from the Bristol Inquiry (with discussion)." *Journal of the Royal Statistical Society, Series A*, 165: 191–232. 420

Spiegelhalter, D. J., Best, N. G., Gilks, W. R., and Inskip, H. (1996). "Hepatitis: a case study in MCMC methods." In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.), *Markov Chain Monte Carlo Methods in Practice*, 21–44. New York: Chapman and Hall. 424

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. MRC Biostatistics Unit, Cambridge, Available from `www.mrc-bsu.cam.ac.uk/bugs`. 420, 424

Theil, H. (1963). "On the use of incomplete information in regression analysis." *Journal of the American Statistical Association*, 58: 401–414. 431

Wakefield, J. and Bennett, J. (1996). "The Bayesian modeling of covariates for population pharmacokinetic models." *Journal of the American Statistical Association*, 91: 917–927. 412

Acknowledgements

The authors would like to express their thanks to Sylvia Richardson for her valuable comments on an earlier draft of this paper. They are also indebted to the participants in the HSSS 'Research Kitchen' on model criticism held in Oslo: Tony O'Hagan, Arnoldo Frigessi, Roger Cooke, Antti Penttinen and Sylvia Richardson.

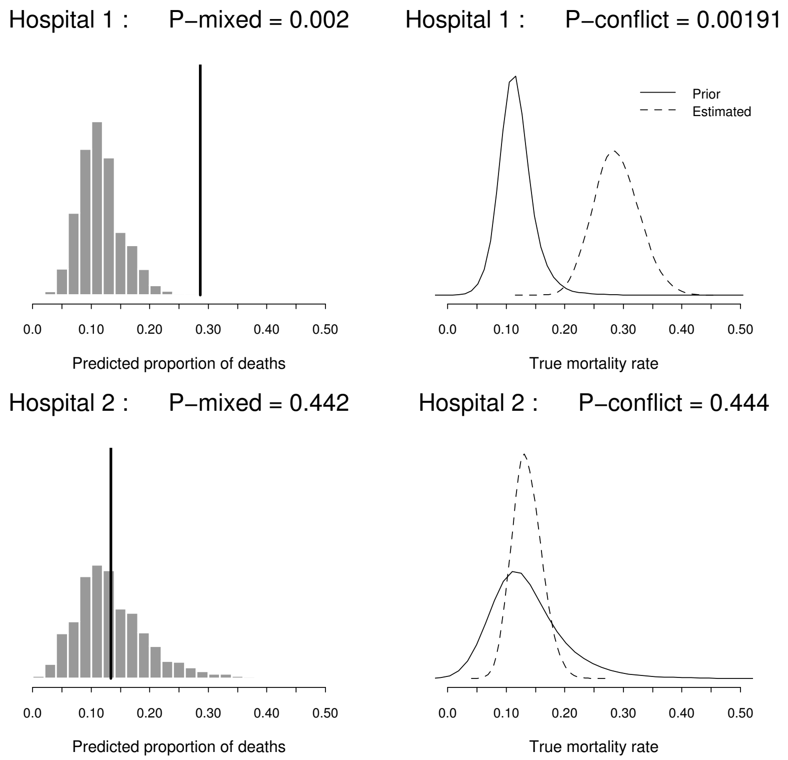

Figure 4: For hospital 1 (Bristol) and hospital 2 (Leicester), the left-hand plot shows the cross-validatory 'mixed' prediction of proportions of deaths, and P_{con} is the area to the right of the observed proportion of deaths as indicated by the vertical line. The right-hand plot shows the predictive prior (for the true mortality rate based on the remainder of the data) and the 'likelihood' (the posterior for the true mortality rate using Jeffreys prior and the observed data for that hospital alone), with P_{con} being the probability of a random draw from the 'likelihood' exceeding a draw from the prior. p-values near 0 correspond to higher than expected mortality, near 1 to lower.

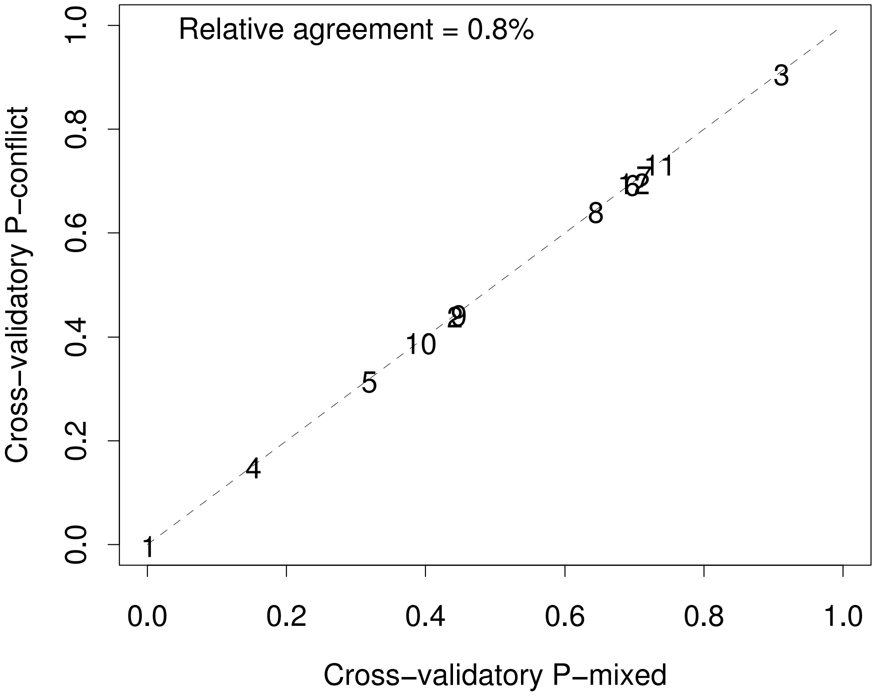

Figure 5: Comparison of cross-validatory P_{mix} and P_{con} for the 12 hospitals, with mean relative agreement measured on the inverse-normal scale.

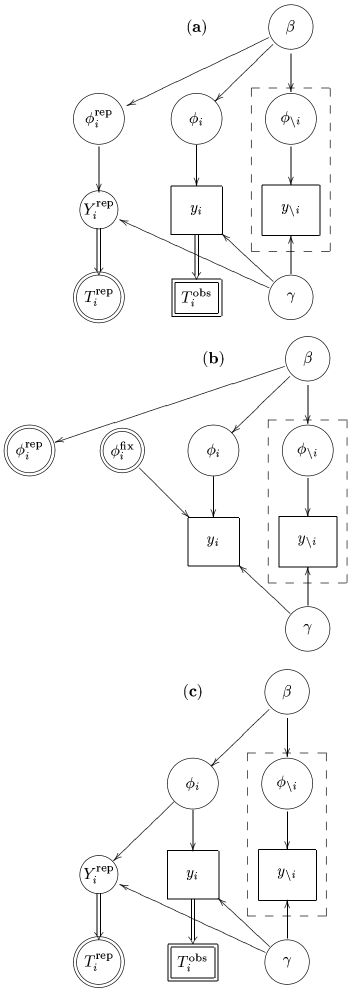

Figure 6: Alternative replication approaches when using full data (not cross-validation). (a) 'Mixed' replication of random effects and data, (b) Conflict between 'predictive prior' replication of random effects and 'likelihood' replication of fixed effects, (c) posterior-predictive replication of data alone. In each case, the double ringed nodes represent quantities to be compared in order to assess divergence.

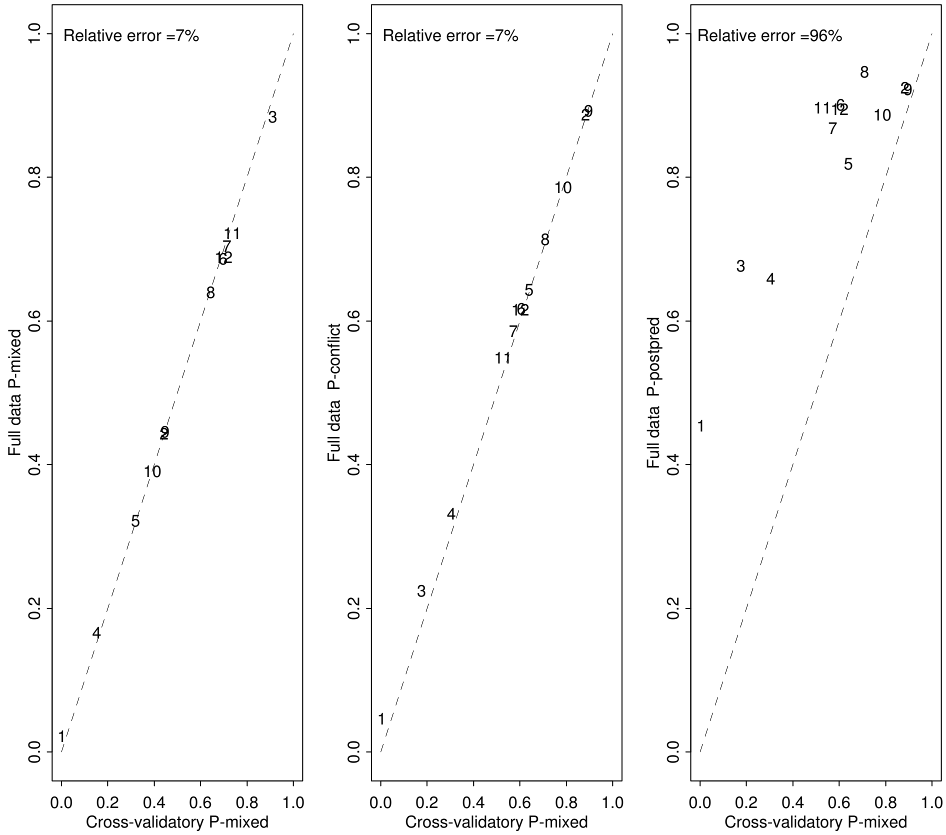

Figure 7: Comparison of cross-validation and full-data p-values for Bristol data calculated using three different approximations: (a) mixed, (b) conflict, (c) posterior predictive.

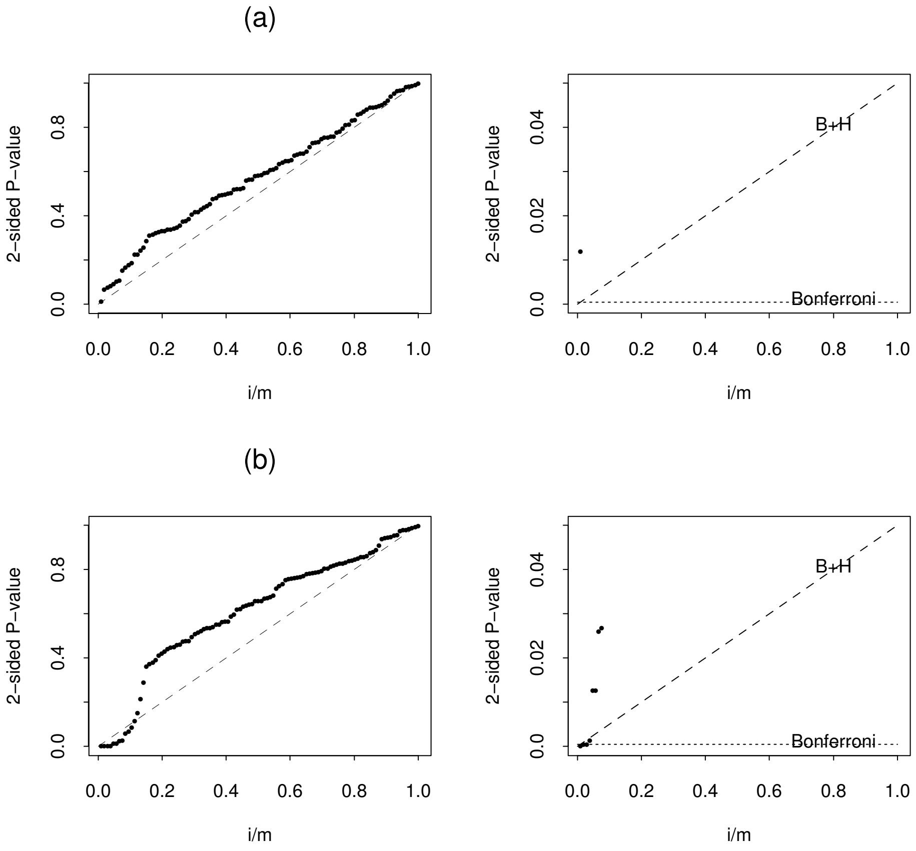

Figure 8: QQ plot of 2-sided conflict p-values for $I = 106$ children, where the plotted p-values correspond to (a) intercepts, (b) slopes. The Benjamini and Hochberg criterion is superimposed on p-values of less than 0.05, identifying four gradients.

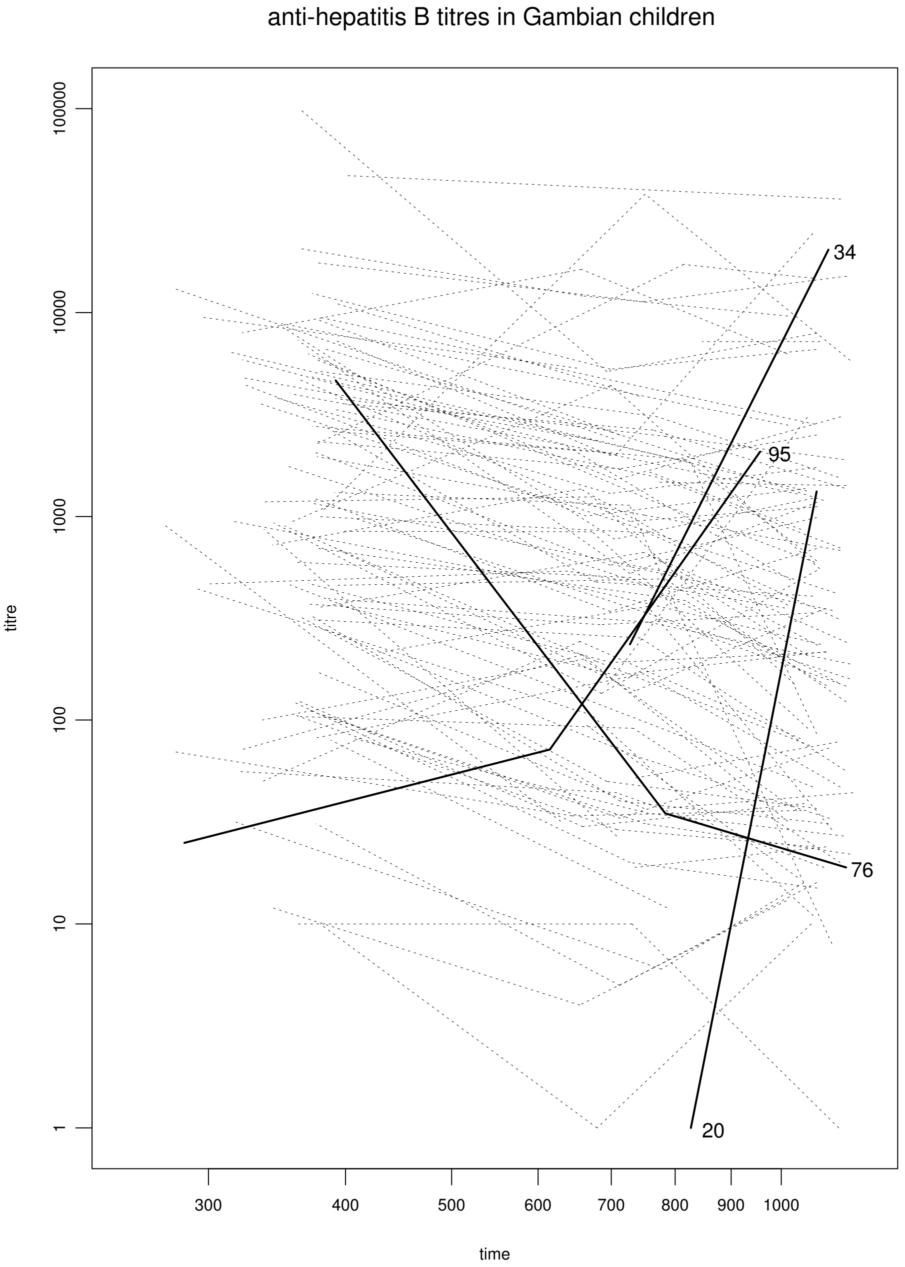

Figure 9: Trajectories for 106 children, highlighting children with gradients identified as divergent under a false discovery rate of 0.05.