STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2666

Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons

D. I. Ohlssen*, †, L. D. Sharples and D. J. Spiegelhalter

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

SUMMARY

Random effects models are used in many applications in medical statistics, including meta-analysis, cluster randomized trials and comparisons of health care providers. This paper provides a tutorial on the practical implementation of a flexible random effects model based on methodology developed in Bayesian non-parametrics literature, and implemented in freely available software. The approach is applied to the problem of hospital comparisons using routine performance data, and among other benefits provides a diagnostic to detect clusters of providers with unusual results, thus avoiding problems caused by masking in traditional parametric approaches. By providing code for Winbugs we hope that the model can be used by applied statisticians working in a wide variety of applications. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: Bayesian non-parametrics; Dirichlet process; institutional comparisons; WinBUGS

1. INTRODUCTION

Random effects models are used in a wide variety of biostatistical contexts, and can be analysed both from a classical and Bayesian viewpoint. We can distinguish two types of applications: those such as cluster randomized trials [1] in which the random effects are nuisance parameters and are not of direct interest, and those such as disease mapping [2] in which the individual effects are of paramount interest.

A prominent example of the latter type of application is the domain of institutional comparisons, in which 'units' such as schools or hospitals are each assigned an 'effect' to be estimated and compared. Sometimes such analyses arise in specific one-off investigations: for example, recent high-profile cases in the U.K. include the commissioned analysis of relevant data sources for the Bristol Royal infirmary inquiry [3, 4] and an investigation into heart and lung transplantation at a

*Correspondence to: D. I. Ohlssen, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, U.K.
†E-mail: david.ohlssen@mrc-bsu.cam.ac.uk

Received 4 July 2005
Accepted 19 June 2006

specific centre [5]. There is also increasing use of routinely collected data to compare providers [6, 7]. Such analyses may have a number of objectives, but there is a common basic need to measure and communicate variability in performance in a fair and flexible manner.

Routine data on health care providers is usually collected in a two-level hierarchical structure, with patients' results nested within providers. For example, we may have binary outcome data y_{ij} for a patient j in centre i, and assume

$$y_{ij} \sim \text{Bernoulli}(\rho_{ij}) \quad i = 1, \ldots, M, \;\; j = 1, \ldots, n_i$$

where ρ_{ij} represents the probability of the event in question, assumed to obey the logistic regression model

$$\text{logit}(\rho_{ij}) = \gamma_i + \beta' x_{ij}$$

for centred covariates x_{ij}.

Attention then focusses on the 'centre effects' γ_i, for which three standard assumptions may be made:

1. A *common* centre effect $\gamma_1 = \cdots = \gamma_M$.
2. A *fixed-effects* assumption where a separate intercept is included in the model for each of the centres.
3. The *random-effects* assumption that γ_i are drawn from a common distribution $f(.)$.

The first assumption will in many cases be unrealistic as there is likely to be many unmeasured covariates creating variability between the centres. The second, fixed-effects assumption represents the alternative extreme of complete independence of units. There are three main justifications for rejecting this hypothesis: first, an *intuitive* feeling that the effect in a new centre is not entirely unrelated to that in other centres; second, the *empirical* observation that plots of fixed effects will typically reveal a smooth distribution; and third, a more Bayesian *judgement* of 'exchangeability' of the centres, meaning that one's beliefs about the magnitude of each centre's effect does not depend on the particular labels given to the centres—in other words, any covariates that are expected to lead to predictable differences between centres have been included in the model. The assumption of exchangeability is essentially equivalent to assuming they are drawn at random from a common population distribution [8].

We are thus led naturally to a simple two-level random-effects model, which fits into both the hierarchical [9] and generalized linear mixed model frameworks [10]. For example Goldstein and Spiegelhalter [11] used hierarchical models in the context of 'league tables' for education and health providers, Christiansen and Morris [12] and Normand *et al*. [13] focussed on using hierarchical models to provide shrinkage of institutional effects towards the overall mean while Bronskill *et al*. [14] extended this modelling to longitudinal data: in fact the methodology can also be motivated by regarding it as an adjustment for regression-to-the-mean.

It is important to note that each of these examples use a normal random-effects distribution, $\gamma_i \sim \text{N}(\mu, \sigma^2)$, but without providing any particularly strong justification for this. Uncritical use of this assumption when, for example, there are outlying centres, can lead both to undue influence of larger outliers, and undue impact on smaller outlying centres as they are shrunk towards the overall mean. The response to these phenomena depends on whether the objective of the analysis is to (a) *identify* outliers using diagnostics and then remove them to leave the bulk of centres assumed to obey the normal assumption, or (b) to *accommodate* outliers and hence fit a more flexible distribution than the normal.

DOI: 10.1002/sim

Here we assume the latter objective. Our methods will be rooted in the Bayesian non-parametrics literature [15], as is recent work by Burr *et al*. [16, 17] for fitting flexible random-effects distribution. However, in contrast to most applications of Bayesian non-parametrics, our development emphasizes simplicity and interpretability, and in particular will be applicable in the freely available software WinBUGS [18].

Section 2 introduces our running example, chosen to illustrate many of the problems that occur with large routinely collected data sets. Section 3 briefly reviews basic Bayesian non-parametrics and describes research on the truncated Dirichlet process which forms the basis of our flexible random-effects distributions. Section 4 focuses on prior specification for the truncated Dirichlet process and practical implementation, and Section 5 describes application of the model to our example. Section 6 extends the initial model to a continuous random effects distribution by using mixtures of normal distributions. The alternative model is applied to the example and compared with fixed effects, normal random-effects and the flexible model used in Section 5. Section 7 examines how the in-built cluster algorithm can be used as a diagnostic for subgroups of centres with unusual outcomes while Section 8 tests the methodology with two simulated data sets. Section 9 discusses the main ideas in the paper and describes alternative applications for the models. Throughout we attempt a tutorial approach to using such models, and provide full WinBUGS code in Appendix A.

2. AN ILLUSTRATIVE EXAMPLE

The consultation document *A First Class Service*: *Quality in the New NHS* [19] set out a range of measures to help the U.K. National Health Service (NHS) deliver a high quality of care. The document described a three part approach to improving quality—new national standards; dependable local delivery systems; and improved monitoring and performance assessment. In this example we take a data set used as part of the third part of this plan, taken from the clinical outcome indicators [20]. This report included six clinical indicators based on routinely available data from NHS Trusts and Health Authorities, such as Hospital Episode Statistics (HES). It is acknowledged in the report that these data sources are less than perfect, a fact also highlighted in the Bristol inquiry [4].

We focus on Indicator 3 concerning mortality rates: specifically the rates of death in hospital within 30 days of an emergency admission for myocardial infarction, for patients over 50, for all acute NHS trusts. The data set is deliberately selected as being likely to feature all the problems associated with routine data, and consisted of a numerator 'Number of continuous inpatient spells following emergency admission for patients aged over 50 with primary diagnosis of AMI (ICD 10 codes 121 122), ending in death in hospital, within 30 days of admission', denoted y_i for centre i, and a denominator 'Number of continuous inpatient spells following emergency admission for patients aged over 50 with primary diagnosis of AMI (ICD 10 codes 121, 122)', denoted n_i. The data were reported annually from 1995 to 1998, and we concentrate on the most recent data covering the year 1997–1998. A number of the trusts had missing data in the numerator and denominator or simply had no patients for the denominator. We exclude all trusts with missing data or with less than 20 patients in total, leaving data for 200 hospitals. The number of patients treated by the hospitals ranged from 25 to 967 with a median value of 295 and an interquartile range of (201, 397).

The overall mortality rate was 18 per cent, based on 61 663 cases and 11 086 deaths. We assume a Binomial distribution $y_i \sim \text{Bin}[n_i, \rho_i]$, and fit fixed and normal random-effects models using the

DOI: 10.1002/sim

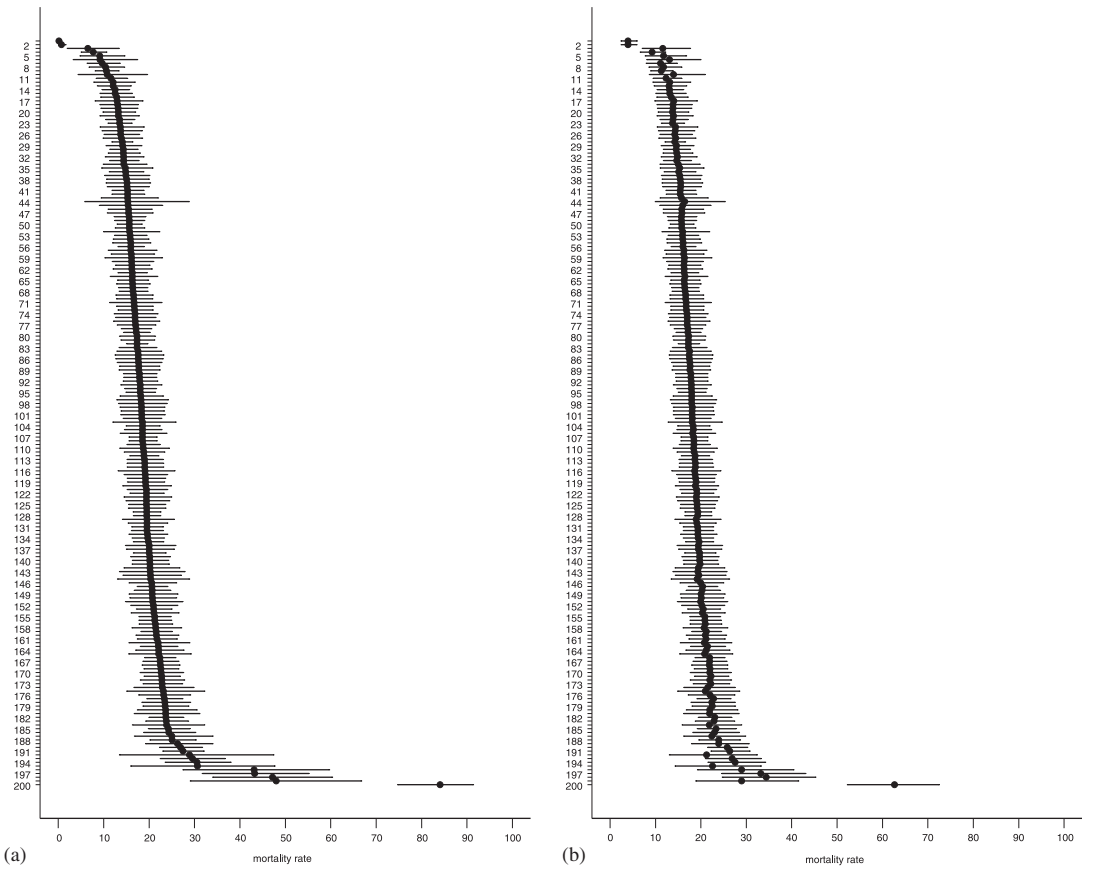

Figure 1. Caterpillar plots of 95 per cent credible intervals for fixed and random effects estimates of mortality rates for the NHS example. The hospitals were placed in the order of their observed mortality rate.

WinBUGS software: for the fixed effect model we assume a uniform prior for each ρ_i, while for the random-effects we assume $\text{logit}(\rho_i) \sim \text{N}[\mu, \sigma^2]$, with uniform priors on μ and σ. The prior on σ can have an important impact [21], so we considered three different ranges for the uniform prior (0, 10), (0, 100) and (0, 1000). The alternative choices made no difference (2SF) to the estimates and 95 per cent credible intervals for σ. Figures 1(a) and (b) show caterpillar plots for the estimated mortality rates: for the random-effects analysis the mean on the mortality rate scale was 0.179 with 95 per cent interval of (0.171, 0.188), σ had a posterior median of 0.36 with 95 per cent interval of (0.31, 0.42).

To provide a comparison of overall fit we calculate the standardized deviance

$$D(\hat{\rho}_i|y) = -2\sum_{i=1}^{M}[y_i \log(\hat{\rho}_i) + (n_i - y_i)\log(1 - \hat{\rho}_i)]$$

DOI: 10.1002/sim

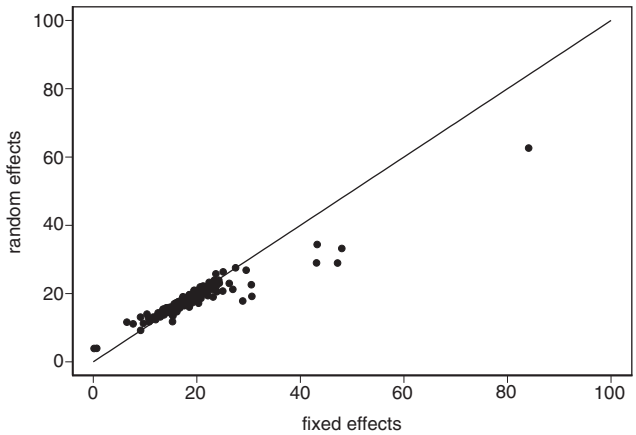

Figure 2. A plot of the fixed effect estimates of mortality rate against the normal random effect estimates. The line $y = x$ is superimposed to give an indication of shrinkage.

at each iteration, where $\hat{\rho}_i = y_i / n_i$. In the normal random effects model the posterior mean deviance was 260, compared with 200, the expected value, in the saturated fixed effects model. As a measure for model comparison the DIC criterion [22] penalizes $\overline{D(\rho|y)}$ for model complexity by adding p_d 'the effective number of parameter'.

$$p_d = \overline{D(\rho|y)} - D(\widetilde{\rho}|y)$$

where $D(\widetilde{\rho}|y)$ is the deviance evaluated at a suitable estimate, the posterior mean estimates for $\rho_1, \ldots, \rho_M$ work well when the marginal distributions for ρ_i are reasonably symmetrical and unimodal.

p_d was calculated to be 162 in the normal random effects model and 200 ($M = 200$) in the fixed effects model, thus estimating the actual number of free parameters correctly in the latter case. This gives overall DICs of 422 for the random effects and 400 for the fixed effects model, suggesting the superiority of the fixed effects model for these data when a normal random effects model is assumed.

There are a number of centres with extreme results, for example centre 1 has 0 deaths from 320 cases and centre 2 has 2 deaths from 370 cases. At the other extreme centre 200 has 60 deaths from 71 cases (85 per cent), and this is almost a 40 per cent higher death rate than any other centre: this mortality rate appears implausibly large and is probably related to a data quality problem, but nevertheless we desire a method that is able to automatically detect such apparently spurious findings.

The normal random effects model causes little shrinkage in the majority of hospitals due to the relatively high number of patients treated, but there is considerable shrinkage in a small number of hospitals with extreme mortality rates. For example, West Dorset general had an observed mortality rate of 47 per cent (25/53) but under a normal random-effects model this estimate is shrunk to 34 per cent. At the other extreme, Brompton had an observed mortality rate of 5.9 per cent (4/68) but under the normal random-effects model this was shrunk to 12 per cent. Figure 2 plots the fixed effects estimates of mortality rates against the random effect estimates of mortality rates and gives an overall indication of the shrinkage effect of the normal random effects model. In general

DOI: 10.1002/sim

shrinkage is a desirable feature because it guards against extreme estimates in small hospitals, however in this example the large difference in DIC between the fixed effects model and the normal random effects model suggests that the pattern of shrinkage imposed by a normal random effects model is questionable. This motivates investigation into modelling with less restrictive random effects assumptions.

3. USING A TRUNCATED DIRICHLET PROCESS

3.1. The Dirichlet process

Parametric inference assumes a distribution function $F(.)$ with a finite set of unknown parameters, whereas in Bayesian non-parametrics F is assumed unknown and a prior is placed over the infinite-dimension space of distribution functions. Here we require only the most readily applicable theory derived from an extensive literature [15, 23, 24]: we initially look at density estimation and then transfer the methods to modelling a random-effects distribution.

The Dirichlet process (DP) is a model for a random distribution function F, given a positive real α and a continuous distribution F_0 which is the baseline distribution around which F is centred. In practice, suppose we break the real line into k disjoint classes $(-\infty, x_1), [x_1, x_2), \ldots, [x_{k-2}, x_{k-1}), [x_{k-1}, \infty)$ where $-\infty = x_0 < x_2 < \cdots < x_{k-1} < x_k = \infty$, and that $p_1 = F(x_1)$, $p_2 = F(x_2) - F(x_1)$, $p_3 = F(x_3) - F(x_2), \ldots, p_{k-1} = F(x_{k-1}) - F(x_{k-2})$ and $p_k = 1 - F(x_{k-1})$ are the probabilities of lying in the intervals, and $p_{0,k-1} = F_0(x_{k-1}) - F_0(x_{k-2})$ are the corresponding probabilities for the baseline distribution. Then the p's have a Dirichlet distribution

$$(p_1, p_2, \ldots, p_k)' \sim \mathrm{Dir}(\alpha p_{0,1}, \alpha p_{0,2}, \ldots, \alpha p_{0,k})$$

where α is a parameter that measures the variability of F around F_0, high values of α corresponding to F being 'close' to F_0. Section A.3 shows that the DP naturally extends to any sample space Ω: this includes cases such as $\mathbb{R}^2$ and restricted portions of the real line.

The *constructive definition* of the DP [25] shows how we can simulate random distribution functions from a DP. We first generate a random sequence of draws $\theta_1, \theta_2, \ldots$ from F_0 and a random sequence of draws $\xi_1, \xi_2, \ldots$ from a Beta$(1, \alpha)$, so that $p(\xi_i) = \alpha \xi_i^{\alpha-1}$ and $E(\xi_i) = (1+\alpha)^{-1}$. The random distribution function $F(.)$ assigns probability $p_1 = \xi_1$ to the point θ_1, $p_2 = (1 - \xi_1)\xi_2$ to θ_2, $p_3 = (1 - \xi_1)(1 - \xi_2)\xi_3$ to θ_3 and so on. The generation of the masses p_k can be viewed as a *stick-breaking* prior [26], in that one can think of ξ_1 being 'broken off' a stick of length 1 leaving a remainder $q_1 = (1 - \xi_1)$, and then a proportion ξ_2 being broken off leaving $q_2 = (1 - \xi_1)(1 - \xi_2)$ and so on, hence

$$p_k = \xi_k \prod_{j<k} (1 - \xi_j) = \xi_k q_{k-1} \tag{1}$$

The fraction $1 - \xi_i$ left after each 'break' has expectation $\alpha/(1+\alpha)$ and hence after $N-1$ breaks there is expected to be a proportion

$$E\left[1 - \sum_{i=1}^{N-1} p_i\right] = E[q_{N-1}] = E\left[\prod_{i=1}^{N-1} (1 - \xi_i)\right] = \left(\frac{\alpha}{\alpha+1}\right)^{N-1} \tag{2}$$

left to assign.

DOI: 10.1002/sim

The constructive definition shows that realizations of the DP can be represented as infinite mixtures of point masses [23], so that the resulting density function is of the form

$$f(.) = \sum_{k=1}^{\infty} p_k I_{\theta_k}, \quad \theta_k \sim F_0$$

where I_{θ_k} represents an indicator function at θ_k and $f(.)$ is the density function corresponding to the distribution function F. The form of the density function highlights the fact that random realizations from the DP are discrete probability distributions. This means that direct use of the DP for density estimation is unsuitable if there is prior belief that the underlying density is continuous. However, a simple extension is to extend the DP to form a mixture of continuous distributions

$$f(.) = \sum_{k=1}^{\infty} p_k h(.|\theta_k), \quad \theta_k \sim F_0$$

where $h(.|\theta_k)$ is a density function for a continuous random variable. This is often referred to as a mixture of DP model: and was originally applied with normal distributions to form a Bayesian approach to kernel density estimation [27].

In this paper interest lies in using the DP or a mixture of DP's to provide a flexible random-effects distribution. Previous work in this area has focused on extending existing computational methods [27] to hierarchical models: Bush and MacEachern [28] consider using the standard DP to create a prior for the random-effects distribution in a normal–normal hierarchical model, Kleinman and Ibrahim [29] extend this idea to generalized linear mixed models, while a mixture of DPs has also been applied in random effects ANOVA [30, 31]. Burr and colleagues [16, 17] use a combination of two DP to estimate the median treatment effect in a random effects meta-analysis.

All these approaches suffer from fairly severe computational complexity and/or restrictions to normal likelihoods. We therefore turn to a computationally straightforward approximation.

3.2. The truncated Dirichlet process

Recent research has focussed on using the constructive definition of the DP to produce practical MCMC algorithms [26, 32–35]. The principle is to approximate the full process by truncating the mixture at a maximum number of components N, so that

$$\sum_{k=1}^{\infty} p_k I_{\theta_k} \approx \sum_{k=1}^{N} p_k I_{\theta_k} \tag{3}$$

and such a truncated DP is denoted $F \sim \text{TDP}(\alpha, F_0, N)$. A restriction is placed on the final weight, $p_N = 1 - \sum_{k=1}^{N} p_k$, so that a proper distribution is formed. This idea could also be used to model a mixture of DP

$$\sum_{k=1}^{\infty} p_k h(.|\theta_k) \approx \sum_{k=1}^{N} p_k h(.|\theta_k)$$

The mixture model has the advantage of relaxing the assumption of a discrete distribution function, we investigate this approach further in Section 6. Both models reduce to a finite mixture model that can be fitted using standard MCMC methods and implemented in standard Bayesian software. In addition to providing a flexible distribution for the random-effects, there is the added advantage of an in-built cluster algorithm which could be used to detect groups of units with unusual results.

DOI: 10.1002/sim

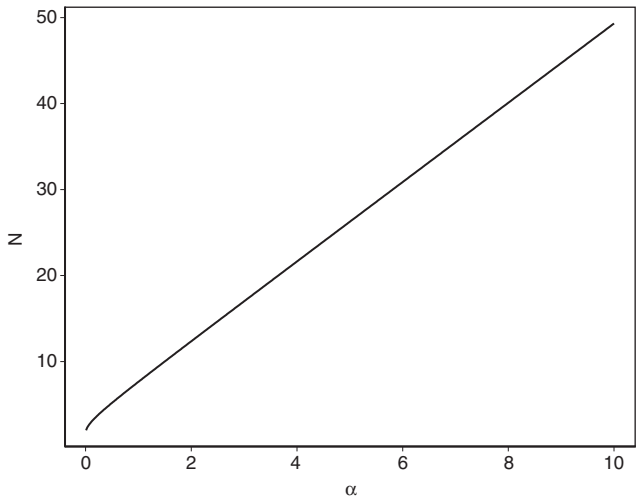

Figure 3. The value of N so that $E[\sum_{i=1}^{N-1} p_i] > 0.99$ against α.

3.3. Specifying N

The parameter N specifies the number of mass points used in the approximation of the DP, and the value of N must be closely related to the value of α which controls the amount of clustering between the centre effects. Models with smaller numbers of mass points are easier to compute, but this reduces the quality of the approximation to the full DP. A pragmatic approach to this problem is to set N so that the amount of probability assigned to the final mass point $p_N = 1 - \sum_{k=1}^{N-1} p_k$ is expected to be small, so that

$$E[p_N] \approx \varepsilon$$

From (2) we have

$$N \approx 1 + \frac{\log \varepsilon}{\log[\alpha/(1+\alpha)]}$$

Suppose we set $\varepsilon = 0.01$, then Figure 3 plots α against N and shows an almost linear relationship, which is to be expected since $\log[\alpha/(1+\alpha)] \approx -1/\alpha$ for moderate α, and hence

$$N \approx 1 - \alpha \log \varepsilon$$

In practice this means that, for $\varepsilon = 0.01$, a quick conservative approximation to the number of points required is $5\alpha + 2$. The choice of N is therefore essentially determined by the choice of α, discussed in Section 4.2. Ishwaran and Zarepour [33] used a similar approach for choosing an appropriate truncation point. They examined both the $E[p_N]$ and $V[\sum_{k=1}^{N-1} p_k]$ and concluded that for values of α up to 10 there was negligible tail probability (p_N) once N is reasonably large, such as 50.

DOI: 10.1002/sim

Whilst our approach for choosing N is essentially pragmatic it should produce a flexible model with similar properties to the DP. A more formal approach to selecting the N has been proposed in a series of papers [26, 32, 36]. This approach is based on comparing the total variation metric between the truncated DP and the full DP, a full explanation is given in Section A.4. Using this approach it is possible to set N so that the marginal results produced by the truncated DP are virtually identical to the full DP, although this usually requires a larger truncation value leading to an increase in computational time. In the examples we test our proposed approximation against a fuller approximation that should give virtually identical results to the full DP.

3.4. Computation of the model

WinBUGS code for both the DP and mixture of DP models is provided in Appendix A. Similar but more restricted models have been fitted using the freely available WinBUGS software [18] by Congdon [34], but not in the full generalization of allowing for estimation of the α parameter and formulating a relationship between α and N. Recently, Erkanli *et al*. [37] examined similar models for semi-parametric ROC analysis using Winbugs. They placed an informative prior for α and used a small number of mass points ($N = 10$), but did not investigate the impact of increasing the number of mass points on their estimates. Further details of this approach are discussed in Section 4.2.

An alternative way to approximate the DP involves using a Dirichlet distribution for the weights $p_1, \ldots, p_N$ [38, 39]:

$$(p_1, \ldots, p_N)' \sim \text{Dirichlet}\left(\frac{\alpha}{N}, \ldots, \frac{\alpha}{N}\right)$$

which forms a DP as $N \to \infty$. Ghosh and Norris [40] used this method to formulate a flexible bivariate distribution for animal capture and recapture probabilities. Their application was based on a small number of animal counts from 5 capture periods and only required a distribution with a small number of mass points. They used WinBUGS to fit an approximate DP with up to 5 mass points and a fixed value of α. We found this method to be numerically less stable than the stick-break approximation when applied with larger N.

The next section examines the prior specification of the DP for modelling the random effects in the NHS example.

4. PRIOR SPECIFICATION FOR THE EXAMPLE

4.1. The form of F_0

We assume a normal distribution for the baseline F_0 with unknown parameters as a natural extension of the standard normal random-effects model, so that

$$\theta_k \sim \text{N}(\mu_{F_0}, \sigma^2_{F_0}), \quad k = 1, \ldots, N, \tag{4}$$

The priors for μ_{F_0} and σ_{F_0} may be chosen to be fairly *weak*, in the sense that they are flat well beyond the range of values that are supported by the data, such as

$$\mu_{F_0} \sim \text{N}(0, 10^2)$$

$$\sigma_{F_0} \sim \text{Uniform}(0, 10)$$

DOI: 10.1002/sim

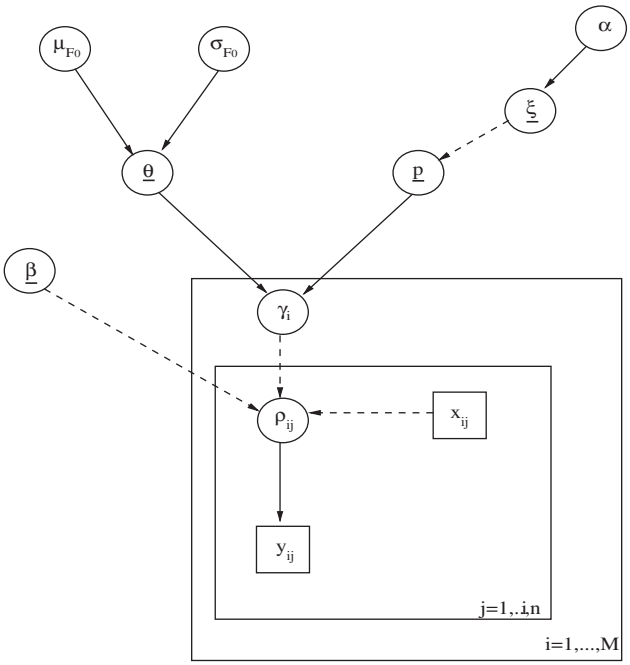

Figure 4. DAG of the flexible random-effects model. Unbroken arrows represent stochastic relationships, broken arrows represent deterministic relationships, circles represent parameters, squares represent data and large square plates represent loops.

On the logit scale these priors cover a very wide range, but this should not represent an automatic default choice of prior, since different data sets might require a different specification depending on the scale of the data and range of observed values. The formulation of the model used the latent variable approach to mixture modelling [41], and Figure 4 shows the model in the form of a direct acyclic graph (DAG).

4.2. Specifying α

The assumption of a normal baseline distribution has a useful interpretation in the context of hospital comparisons. As $\alpha \to 0$ all of the centre effects γ_i are forced into a single common cluster, which can be thought of as support for the common unit effect assumption and as $\alpha \to \infty$ the DP forces each of the unit effects into a separate cluster which is equivalent to a normal random-effects assumption. Thus support for small values of α shows the common mean model might be reasonable for the data, support for large values support the normal random-effects model, while intermediate support suggests a flexible alternative is required. The issue is then identifying what constitutes 'small' or 'large' values of α.

Ishwaran [32] suggests that a value of $\alpha = 3$ can be considered a 'large' value, but if the *true* distribution was normal there would be support for much larger values of α. The relationship between α and N derived in Section 3.3 shows that if $\alpha = 10$ then the random effects distribution might be modelled with around 52 mass points ($N = 52$): in contrast, a likelihood-based analysis

DOI: 10.1002/sim

in which quadrature is used to approximate a continuous distribution [42] will usually use 20–30 quadrature points, thus support for values of α greater than 10 suggests that the data might be better modelled with a continuous distribution rather than a discrete distribution, and hence, for example, a mixture of DPs may be more appropriate: see Section 5.

If we wish to allow α to adapt to the data, the problem then remains of specifying its prior distribution. Based on the observations above we have adopted a uniform prior on α with a lower bound of 0.3 and an upper bound of 10:

$$\alpha \sim \text{Uniform}(0.3, 10)$$

N was set as 52 to deal with the largest possible value for α. The lower bound of 0.3 was specified to avoid computational difficulties caused by small p_k's. The formal metric shown in Section A.4 suggests that in the example setting $N = 200$ should produce a model that is virtually indistinguishable from the full DP. We use this alternative choice of N to assess the quality of our proposed approximation.

The parameter α has a direct relationship with the number of clusters K (occupied components) [27]. It is therefore possible to use genuine prior information about K to form a prior for α. Liu [43] showed that given a fixed α and the number of centres M, $E[K|\alpha, M] \approx \alpha \log((\alpha + M)/\alpha)$ and $\text{SD}[K|\alpha, M] \approx \sqrt{\alpha(\log((\alpha + M)/\alpha) - 1)}$: these results were used by Erkanli *et al.* [37] to derive informative priors for α. The conditional expectation results can be used to give some idea of the impact of our priors on K: for example, if $\alpha = 10$ and $M = 200$, $E[K|\alpha, M] \approx 30$ and $\text{SD}[K|\alpha, M] \approx 4.5$, if $\alpha = 0.3$, $E[K|\alpha, M] \approx 2$ and $\text{SD}[K|\alpha, M] \approx 0.5$. These results suggest that using $N = 52$ points should be conservative and we would expect many unoccupied clusters.

5. EXAMPLE

We applied the flexible random-effects distribution developed in Sections 3 and 4 to the NHS data, in the main analysis we adopted a model with $N = 52$ points. In this example we were able to place weak priors on the parameters in the baseline distribution specified in the previous section, in the informal sense that the priors are relatively flat over the range that appears to be supported by the data. This is possible because the example is based on a large data set, but with smaller data sets it may be necessary to use informative priors possibly based on results from previous periods and subject to sensitivity analysis. Three parallel chains of equal length with different initial values were run for each model. The results were based on 1 in 10 thinning of the total 150 000 iterations which followed a burn-in period of 50 000 iterations for each chain. The estimates were consistent across the chains, as judged by the Gelman–Rubin diagnostic.

Histogram plots of the posterior means of the hospital effects are shown in Figure 5. The corresponding histograms for the fixed effects model and normal random-effects model are included in the plot for comparison. The plots for the random effects models have a kernel density estimate of the predictive distribution for a new hospital effect superimposed.

The plots show there is far less shrinkage in the flexible model than the normal random-effects model in hospitals that have the most extreme observed results. This is in contrast to the main group of hospital effects which are shrunk closer to the overall mean by the flexible random-effects model. The predictive distributions show similar results to the histograms of the hospital effects. The flexible random-effects model has a more peaked distribution than the normal random-effects model but also has much longer tails.

DOI: 10.1002/sim

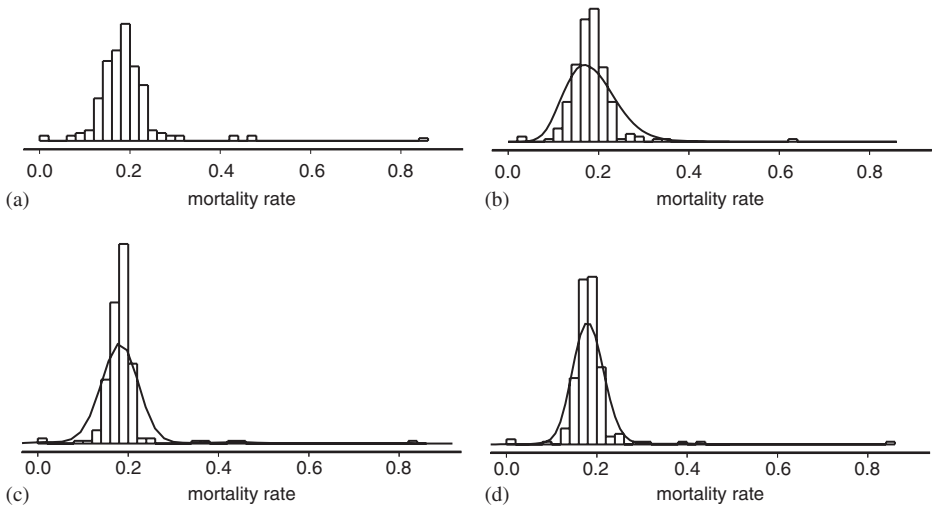

Figure 5. Histograms of the posterior means of hospital effects ρ_i on the mortality rate scale for the fixed effects model, normal random-effects model and both flexible random-effects models. Density estimates of the predictive distribution for a new hospital are superimposed: (a) fixed effects; (b) normal random effects; (c) flexible random effects; and (d) flexible mixture of normals.

Table I. Posterior mean deviance for each model and deviance information criterion (DIC) for selected models.

Model	$\overline{D(\rho\|y)}$	$D(\hat{\rho}\|y)$	p_d	DIC
Fixed effects	200	0	200	400
Normal random effects	260	98	162	422
Mixture of points	230	108	122	352
Mixture of normals	219	90	129	348

Figure 6 shows the marginal posterior distributions for a number of extreme centres. Figures 5 and 6 show that the flexible model causes virtually no shrinkage in the most extreme centres (1, 2 and 200). The estimates for centres 196 and 199 appear to be bimodal: this suggests the model is uncertain whether these centres form a local cluster with a few extreme centres or cluster with the main group.

Table I shows a table of deviances for the various alternative models. The posterior mean deviance for the flexible mixture of points was 230, formally indicating that the flexible model fits the data better than the normal random effects model. Some caution is required in estimating p_d for the flexible model, as mixture models can give rise to bi-modal marginal distributions for the ρ_i's, as demonstrated above. This can make the posterior mean a poor 'plug-in' estimator and so understate p_d, and for this reason WinBUGS will not automatically calculate DIC for such models. However, only a few bimodal distributions were observed and so we calculate the DIC outside WinBUGS giving a value of 122 for p_d and 352 for DIC, representing an improvement on both the fixed and random effects model in spite of its complexity.

DOI: 10.1002/sim

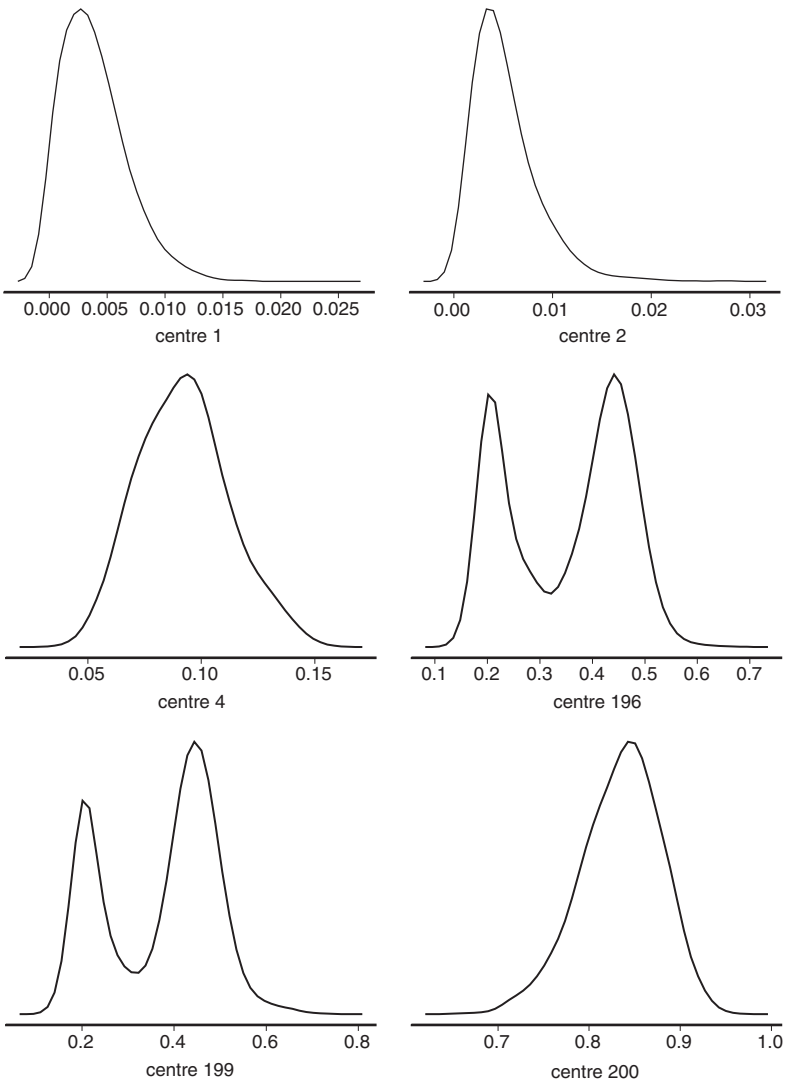

Figure 6. Kernel density estimate of the hospital effects for selected hospitals on the probability of death scale.

A summary of the estimates of the important parameters for models using 52 points and 200 points is shown in Table II. The results are very similar suggesting that in the example our proposed level of truncation closely approximates the full DP. Below we present some analysis of the results from the 52 point model.

The 95 per cent credible interval for α is $(0.85, 5.38)$, which suggests that both the common mean model and the normal random-effects assumption model are unsuitable. A kernel density plot of α is shown in Figure 7(a), which shows the distribution is clearly peaked away from 0

DOI: 10.1002/sim

Table II. Posterior median estimates and 95% credible intervals for selected parameters from both of the flexible models.

Parameter	Mixture of points $N = 52$	Mixture of points $N = 200$	Mixture of normals $N = 52$
μ_{F_0}	-1.61 $(-3.50, 0.031)$	-1.60 $(-3.67, 0.14)$	-1.81 $(-5.94, 1.80)$
σ_{F_0}	2.16 (1.21, 5.02)	2.19 (1.24, 4.99)	3.37 (1.63, 8.80)
K	10 (7, 17)	10 (7, 17)	6(4, 10)
α	2.38 (0.89, 5.56)	2.24 (0.85, 5.52)	1.27 (0.42, 3.28)
b	—	—	0.09 (0.001, 0.29)
ρ^{new}	0.18 (0.11, 0.27)	0.18 (0.11, 0.27)	0.18 (0.11, 0.27)
$E[F_0]$	-1.53 $(-1.63, -1.44)$	-1.53 $(-1.63, -1.45)$	-1.53 $(-1.66, -1.44)$
$\text{SD}[F_0]$	0.57 (0.36, 1.00)	0.57 (0.36, 0.95)	0.62 (0.36, 1.27)

K denotes the number of clusters and ρ^{new} is the predicted mortality rate for a new centre.

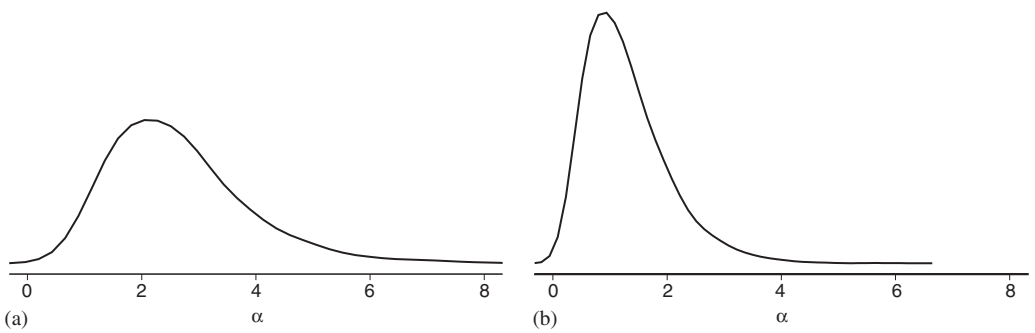

Figure 7. Kernel density estimates of α for the mixture of points model and the mixture of normals model: (a) mixture of points; and (b) mixture of normals.

and 10. K denotes the number of occupied points in the random-effects distribution, which can also be interpreted as the number of clusters. Summaries for the baseline distribution parameters show that the data provided information about all of the hyper parameters. The expectation of the random effects distribution $E[F_0] = \sum_i^N p_i \theta_i$ was calculated at each iteration of the Markov chain: on the mortality rate scale the posterior median was 0.178 which is almost identical to the normal random effects model. The standard deviation of the random effects distribution, $\text{SD}[F_0]$, was also monitored. This had a slightly larger posterior median (0.57) than that of the normal random effects standard deviation σ (0.36), reflecting the flexible model accounting for a greater spread of data.

6. EXTENSION TO MIXTURE OF NORMALS

6.1. Model specification

Section 3 noted that a possible problem with using the DP was the assumption of a discrete random effects distribution. It is possible to extend the model to allow a flexible continuous random-effects

DOI: 10.1002/sim

distribution based on a mixture of normals with a large number of components. In this model a mean and a variance parameter must be defined for each component of the mixture; we shall denote these parameters by θ_k and τ_k^2 $(k=1,\ldots,N)$, respectively. The model for the institution effects γ_i, $i=1,\ldots,M$ becomes

$$\gamma_i \sim \mathrm{N}(\theta_k, \tau_k^2) \text{ with probability } p_k \tag{5}$$

where p_k is defined by the *stick-breaking* algorithm given in equation (1). When computing the model it is convenient to parameterize equation (5) with an allocation variable z_i where

$$z_i = k \text{ with probability } p_k$$

$$\gamma_i \sim \mathrm{N}(\theta_{z_i}, \tau_{z_i}^2)$$

We adopted the same priors for the location parameters θ_k that were used for the discrete points (equation (4)).

To avoid problems associated with empty components [44], it is usual to assume either a single common variance parameter $\tau_k^2 = \tau_{\mathrm{mix}}^2$ or to impose a hierarchical structure on the variance parameters. Richardson and Green [45] proposed a hierarchical structure to express the belief that the τ_k^2 are similar, without being informative about their absolute size. For example,

$$\tau_k^{-2} | b \sim \mathrm{Gamma}(a, b)$$

$$b \sim \mathrm{Gamma}(c, d)$$

They suggest taking $a>1>c$ and taking d to be a small multiple of $1/R^2$, where R represents the range of the observed log odds. This prior has been employed when assuming a mixture of two normals for a random-effects distribution in the context of hospital performance [46] and when using a mixture of normals with a DP prior to identify clusters in univariate data [38].

We adopt this prior for the NHS example and assume $a=3$, $c=0.03$ and $d=0.03$ reflecting the large spread of centre effects on the odds scale. A kernel density estimate of the prior for each τ_k, based on 10 000 simulated values, is shown in Figure 8, and shows that the prior places a large amount of weight on values less than 1 but the long tail ensures the prior is relatively uninformative about the absolute size of each τ_k. The sensitivity of the final results to this prior were checked by considering setting both c and d to 0.1 and 0.5 and by setting $c=0.5$ $d=0.03$. We found there was little difference between results.

To allow comparison between this model and a mixture of points model the prior for α was uniform (0.3, 10) and the truncation level was set at $N=52$. The extra parameters in the mixture of normals model make the interpretation of α more complex. When $\alpha \to 0$ the model will force each of the centre effects into a single normal component with mean μ_{F_0} and variance $\sigma_{F_0}^2$, which produces a normal random effects model. When $\alpha \to \infty$ the situation is more difficult to assess as it depends on σ_{F_0} and τ_k, $k=1,\ldots,N$: for example, as $\tau_k \to 0$ the model will act in the same way as a mixture of points (a normal random effects model), but when τ_k is larger the model will behave more like a fixed effects model as each of the centre effects will be forced into separate clusters with very little dependence. These cases essentially highlight a difficulty identifying α. Further identifiability problems are caused by the relationship between α and σ_{F_0}: in a mixture of normals model it is well known that parameters relating to the variation between

DOI: 10.1002/sim

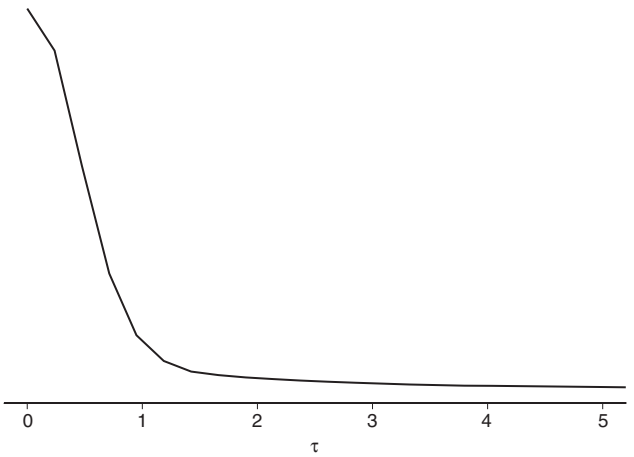

Figure 8. A kernel density plot of 10 000 simulated values from the prior for τ, the within component standard deviation, when $a = 3$, $c = 0.03$ and $d = 0.03$.

clusters (σ_{F_0}) are closely correlated with parameters relating to the number of clusters (α) [45]. This discussion suggests that α is more difficult to identify in the mixture of normals model than in the mixture of points model. The WinBUGS code to fit the DP mixture of normals is shown in Appendix A.

6.2. Applying the mixture of normals to the NHS data

Figure 5(d) displays the posterior mean estimates of the hospital effects and the predictive distribution of a new hospital effect in the mixture model. The flexible mixture of normals has similar estimates for the hospital effects and a similar predictive distribution to the flexible mixture of points. However, there is slightly greater heterogeneity among the main group of hospitals, which can be explained by the extra variation allowed around the mean of each cluster. Both models show very different results to the normal random-effects distribution which appears to apply too much shrinkage. We also examined the marginal posterior distributions for the selected centres with unusual results: the distributions are very similar to the model based on a mixture of points so we have omitted the plots.

Table II shows the posterior mean and 95 per cent credible interval for a number of important parameters. The posterior distribution for α (see Figure 7(b)) shows little support for small values of α, which suggests the normal random-effects model is unsuitable. There is also little support for large values of α, which suggests the fixed effects model is inappropriate. The parameter α provides a natural way to compare the flexible random-effects model to the standard normal random-effects models. In both flexible formulations the posterior distribution for α provides evidence that the normal random effects model is a poor model for the data. We also examined the results using a 200 point approximation but the results were virtually identical so have not been reported.

The $E[F_0]$ is the expectation of the random effects distribution calculated over the 50 component mixture of normals. Converting the estimate to the mortality rate scale shows it is equivalent to

DOI: 10.1002/sim

0.175 (0.16, 0.19), which is very similar to the posterior median estimate for μ in a normal random effects model. The posterior median of SD$[F_0]$ is similar to the mixture of points model but there is slightly more uncertainty around this estimate.

Table I shows the deviance for all of the models. The posterior mean deviance for the flexible mixture of normals model was 220, which is 10 less than the mixture of points model; however the extra variation in the clusters leads to a value of 129 for p_d and hence 348 for the DIC. The difference in DIC between the two flexible models is small and can be interpreted as only a slight preference for the mixture of normals model.

To get an idea of the variation within the components we monitored the average within component standard deviation $\overline{\tau} = (1/N)\sum_k \tau_k$ and standard deviation $\mathrm{SD}(\tau) = \sqrt{(1/N)\sum(\tau_k - \overline{\tau})^2}$. The posterior median for $\overline{\tau}$ was 0.19 (0.07, 0.35) and the posterior median for SD(τ) was 0.07 (0.03, 0.14). These results show that the posterior distribution for the τ_k's is in a region well supported by the prior shown in Figure 8, but it is much more concentrated suggesting it is well identified.

7. IDENTIFICATION OF SUBGROUPS AND CLUSTERS

7.1. Summary statistics as a diagnostic for subgroups

When using a full DP the random-effects distribution is intractable because it has an infinite number of parameters, and the algorithms used essentially integrate out the random-effects distribution. As a result labelling of components in the mixture model is impossible, and this has led to the use of other methods to identify subgroups in the data. For example Escobar and West [27] applied the DP to a mixture of normal distributions with the aim of density estimation, and studied the *predictive density*, to look for evidence of multi-modality and clusters. They also looked at the observed number of components occupied by the data (K). We have already seen in the examples from Sections 5 and 6 that these statistics are easy to monitor in our implementation of the flexible model.

A further useful diagnostic is to monitor pairwise clustering between centres by creating a $M \times M$ matrix of indicator variables $[I_{ij}]_{M\times M}$, where

$$I_{ij} = \begin{cases} 1 & \text{if } z_i = z_j \\ 0 & \text{if } z_i \neq z_j \end{cases} \tag{6}$$

The indicator variable is a marker of whether two units occupy the same component of the mixture at a particular iteration. A high probability might indicate a common unidentified covariate shared by two centres. A similar statistic which involves less storage is to monitor the total number of other centres in the same component as centre i. This can be calculated by monitoring $\sum_{i\neq j} I_{ij}$ and is particularly useful for identifying an outlying centre that is an isolated cluster or centres that are generally in small clusters.

7.2. Application of summary statistics

Table III shows the average number of hospitals in a cluster for the previously selected centres, for the mixture of points and mixture of normals models. The table also lists other centres with a high posterior probability (>0.4) of being in the same cluster. Centre 200 appears to have

DOI: 10.1002/sim

Table III. Summaries from monitoring $\Sigma_{i \neq j} I_{i,j}$ when applying both the mixture of points and mixture of normals.

	Mixture of points		Mixture of normals	
Centre	Median 95% interval	$j : p(z_i = z_j) > 0.4$	Median 95% interval	$j : p(z_i = z_j) > 0.4$
1	2 (1, 2)	2	2 (1, 2)	2
2	2 (1, 2)	1	2 (1, 2)	1
4	3 (1, 8)	9	5 (1, 196)	3, 5, 6, 7, 9, 11
196	4 (1, 115)	197, 198	8 (1, 71)	187, 190, 194, 197, 198, 199
197	3 (1, 8)	196, 197, 199	4 (1, 195)	196, 198, 199
198	3 (1, 8)	196, 197, 199	4 (1, 195)	196, 197, 199
199	4 (1, 115)	197, 198	4 (1, 114)	196, 197, 198
200	1 (1, 1)	none	1 (1, 2)	none

no other centres within the same cluster. When examining the vector $I_{200, j=1,\ldots,200}$ we found that the posterior probability of a hospital being in the same cluster as centre 200 is less than 0.022 in the mixture of normals model and 0 in the mixture of points model. Centres 1 and 2 appear to form a cluster, $p(z_1 = z_2) = 0.71$ in the points model and 0.59 in the mixture of normals. The results also show that centres 199-196 have a high probability of being in the same cluster.

The results from the diagnostic checks could lead to further investigation into the reasons for the apparent clustering. This may involve examining covariates among hospitals that appear to cluster together or further, more detailed investigation into each hospital. The diagnostics for potential clusters provide a further advantage of using the flexible models as an alternative to parametric approaches.

8. TESTING THE METHODOLOGY WITH SIMULATED DATA

To test the methodology we fitted the models to two simulated data sets (using 52 points), both of which assumed the same denominator structure as the NHS example. The first data set assumed the true model was a normal random effects model with a mean $\mu = \text{logit}(0.2) = -1.39$ and a variance of 1. The second data set assumed the first 180 centres followed a binomial$(n_i, 0.2)$ distribution and the remaining 20 centres followed a binomial$(n_i, 0.5)$ distribution.

The results from fitting the two flexible models to the data simulated from the normal random effects model are shown in Table IV. The mixture of points model shows evidence for the normal random effects model as the estimate of α is close to the upper limit of 10. Figure 9(a) displays a plot of the posterior distribution of α and shows a mode of 10. The estimates for the baseline parameters μ_{F_0} and σ_{F_0} are close to the underlying normal distribution and the true estimates are well within the 95 per cent credible intervals. In the mixture of normals model we might have expected strong support for a single normal component and small values of α, but the marginal posterior distribution for α (Figure 9(b)) is almost identical to the prior distribution. We found when the posterior distribution supported a small number of clusters the component variances τ_i were relatively large, conversely when a larger number of components were occupied the variance components were relatively small. This suggests that the model was able to fit the data

DOI: 10.1002/sim

Table IV. Estimates for the first simulated data set from the mixture of points and mixture of normals models.

Parameter	Mixture of points	Mixture of normals
α	8.85 (5.46, 9.95)	5.03 (0.57, 9.72)
μ_{F_0}	−1.28 (−1.81, −0.75)	−1.43 (−2.21, −0.66)
σ_{F_0}	1.15 (0.86, 1.61)	0.86 (0.20, 2.06)
K	32 (23, 39)	19 (3, 33)

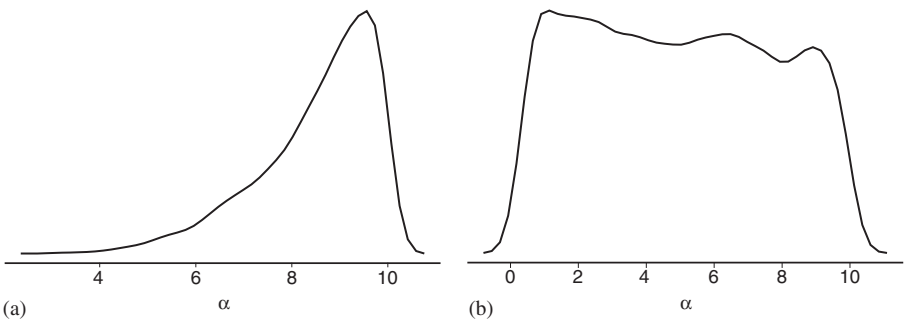

Figure 9. Kernel density plots of α in the normal random effects simulation for both flexible models: (a) mixture of points; and (b) mixture of normals.

equally well with a small or large number of clusters: thus the data provided little information about α.

In the second simulated data set both models showed strong evidence of two distinct clusters. In the mixture of points model the posterior median for the number of clusters K was 2 (2, 5), similarly in the mixture of normals the posterior median for K was 2 (2, 5). The posterior median for α was 0.56 (0.31, 1.67) in the mixture of points model and 0.52 (0.31, 1.48) in the mixture of normals. An examination of the centre effect estimates suggested that both models successfully identified the two groups of centres.

The first set of simulated data shows evidence that the mixture of points model can provide a useful diagnostic test to examine whether the underlying random effects distribution is normal. The mixture of normals did not appear to provide any evidence for an underlying normal random effects model. This may be a result of the number of components of a mixture being generally difficult to identify, particularly when the mixture is placed at the random effects level and there is only binary outcome data. Both models appeared to perform well when data were simulated from a mixture of two mortality rates.

9. DISCUSSION

This paper has described two flexible random effect models for the analysis of routine hospital outcome data, motivated by the need to provide a robust estimation approach when

DOI: 10.1002/sim

measuring the performance of health service providers. The methods are designed for simple routine data two-level structures that may or may not include covariates for case mix adjustment, and used methodology developed in Bayesian non-parametrics. Computation was carried out with the freely available software WinBUGS. Estimates provided by the flexible models can be compared with externally set standards, or standards based on the performance of other centres: a further advantage of the models is an in-built diagnostic for clusters of units with unusual outcomes. The models can also be compared with the three standard modelling approaches by monitoring the Dirichlet mass parameter α, which gives an indication of support for each model.

The approach was applied to an NHS indicator for death rates in hospital within 30 days of an emergency admission for myocardial infarction, for patients over 50, and highlighted a number of centres which appeared to be outliers or centres that formed small clusters. In practice the modelling should be seen as a starting point for further investigation into the reasons for unusual results or apparent clustering. This might include checking the data for coding errors, investigation of covariates of centres which appear to form clusters and more detailed inspection of hospitals with unusual outcomes. The example considered data with no case-mix adjustment which was typical of NHS routine data up until 2003: some of the more recent data have included an expected result based on indirect standardization for age and sex but it would be straightforward to incorporate these factors into the modelling.

In the example weak priors were used for each of the parameters in the truncated DP prior but with smaller data sets there might not be enough information in the data to provide information for such a complex hierarchical structure. A partial solution to this problem would be to limit the flexibility in the model by fixing either the parameters in the baseline distribution or more typically the parameter α. Alternatively informative priors could be used for some of the parameters. In the application of provider profiling such information may be based on historical data from previous audits. Section 3 highlighted some of the computational limitations of the approach in the mixture of points model however if there appears to be support for a model with $\alpha>20$ it would seem sensible to use the mixture of normal formulation, which is unlikely to require as many clusters to model the data.

An important technical point is the number of mass points N that are used in the flexible model. Our proposed approach is pragmatic and based on the sum of the first $N-1$ mass points being close to 1. We feel this approach will always produce a flexible model suitable for practical purposes but it may not give exactly the same inference as the full DP. If the aim of the analysis is to mimic a full DP then it is possible to increase the number of mass points and use the result described in Section A.4.

The model has a number of other applications in medical statistics including meta-analysis, cluster randomized trials, bioinformatics, measurement error problems [47], spatial processes and pharmacology [48]. In meta-analysis the model would provide a way of modelling the predictive distribution for the effect of a treatment in a new trial and Ohlssen [49] showed that a bimodal predictive distribution could be used to highlight conflicts of evidence between trials in a meta-analysis. In cluster randomized trials the random-effects are essentially nuisance parameters so the flexible model would provide a robust way to model the random effects distribution. In the analysis of microarrays the clustering algorithm in the flexible random-effects model could be used to look for clusters of genes. By providing the method in freely available software we hope that the model can be used by applied statisticians working in a wide variety of applications.

DOI: 10.1002/sim

APPENDIX A

A.1. Program for the mixture of points model

$\theta_k = \text{theta}[k]$, $p_k = p[k]$, $Z_k = Z[k]$ $k = 1, \ldots, N$, $\rho_i = \text{rho}[i]$ $i = 1, \ldots, N$, $\sigma_{F_0} = \text{sigmaF0}$, $\alpha = \text{alpha}$, $\mu_{F_0} = \text{basemu}$, $1/\sigma^2_{F_0} = \text{basetau}$

```
Model
   {
   # Random effects logistic regression part of model
for( i in 1: M ) {
     logit(rho[i]) <- theta[Z[i]]
     y[i] ~ dbin(rho[i],n[i])

   Z[i] ~ dcat(p[]) # integer variable
     }

# Constructive DPP

#stick-breaking prior
 p[1] <- r[1]
 for (j in 2:N-1) {p[j] <- r[j]*(1-r[j-1])*p[j-1]/r[j-1]}
 for (k in 1:N-1){ r[k] ~ dbeta(1,alpha)}

#assumption to ensure sum p[] is 1 Ishwaran truncation
 ps <- sum(p[1:N-1])
 for(k in N:N){p[k]<-1-ps}


 # Baseline distribution
 for(k in 1:N){theta[k] ~ dnorm(basemu,basetau) }
basemu~dnorm(0,0.01)
basetau<-pow(sigmaF0,-2)
sigmaF0~dunif(0,10)

# DPP parameter prior
 alpha~dunif(0.3,10)

# Programming for calculating summary statistics #
 for(i in 1:M) {for (j in 1:N) {
   SC[i,j] <- equals(j,Z[i])
} }

# total clusters K#
for (j in 1:N) {cl[j] <-step(sum(SC[,j])-1)}
K<-sum(cl[])
```


DOI: 10.1002/sim

```
# Random effects distribution mean#
 for(i in 1:N){mean2[i]<-p[i]*theta[i]}
 poptrue<-sum(mean2[])

# Random effects distribution variance#
 for(i in 1:N){mom2[i]<-p[i]*theta[i]*theta[i]}
mom2.true<-sum(mom2[])
var.true<-mom2.true-(poptrue*poptrue)


#Number of Centres in same cluster#
 for(i in 1:M){
    for(j in 1:M){
        equalsmatrix[i,j]<-equals(rho[i],rho[j])
    }
   equalsres[i]<-sum(equalsmatrix[i,])
}
}#END
```

A.2. Program for mixture of normals with priors for baseline distribution parameters

$\theta_k = \text{theta}[k]$, $\tau_k^2 = \text{tau2}[k]$, $\tau^{-2} = \text{tauinv}[k]$, $p_k = p[k]$, $Z_k = Z[k]$ $k = 1, \ldots, N$, $\rho_i = \text{rho}[i]$ $i = 1, \ldots, N$, $\sigma_{F_0} = \text{sigmaF0}$, $\alpha = \text{alpha}$, $\mu_{F_0} = \text{basemu}$, $1/\sigma_{F_0}^2 = \text{basetau}$

```
#change the following sections of code

# Random effects logistic regression part of model
for( i in 1: M ){
     logit(rho[i]) <- gamma[i]
     gamma[i]~dnorm(theta[Z[i]],tauinv[Z[i]])
     y[i] ~ dbin(rho[i],n[i])

  Z[i] ~ dcat(p[]) # allocation variable
     }

 # Baseline distribution
 for(k in 1:N){theta[k] ~ dnorm(basemu,basetau)
       tau2[k]<-pow(tauinv[k],-1)
            tauinv[k]~dgamma(3,b)}
basemu~dnorm(0,0.01)
basetau<-pow(sigmaF0,-2)
sigmaF0~dunif(0,10)
b~dgamma( 0.03,0.03)

# Random effects distribution variance#
```


DOI: 10.1002/sim

```
for(i in 1:N){
        var1[i]<-p[i]*tau2[i]
        sq[i]<-theta[i]-poptrue
        var2[i]<-p[i]*sq[i]*sq[i]
        }
# Estimate of the random effects variance
vartrue<-sum(var1[])+sum(var2[])
```

A.3. The general definition of a Dirichlet process

A DP can form a prior for a distribution function F defined over any space Ω, given positive real α and a baseline distribution F_0 (a finite measure over Ω) . The DP is defined when any partition $S_1, S_2, \ldots, S_k$ of Ω has the following property

$$(p_1, p_2, \ldots, p_k)' \sim \mathrm{Dir}(\alpha p_{0,1}, \alpha p_{0,2}, \ldots, \alpha p_{0,k})$$

where $p_1 = F(S_1), p_2 = F(S_2), \ldots, p_k = F(S_k)$, and $p_{01} = F_0(S_1), p_{02} = F_0(S_2), \ldots, p_{0k} = F_0(S_k)$.

A.4. Approximations based on the total variation metric

Ishwaran [32] used the following standard metric to assess the difference between the TDP and the full DP in the context of generalized linear mixed models:

$$D(P_1, P_2) = \sup_B |P_1(B) - P_2(B)|$$

$D(P_1, P_2)$ is the total variation distance between two probability measures P_1 and P_2 defined over the same measure space $(\Omega, \Im)$, where $\Im$ is the Borel algebra for Ω (the set of all possible subsets of Ω) and $B \in \Im$. Informally, this distance can be thought of as the largest possible difference between the probabilities that two probability distributions can assign to the same event.

Theorem 1 of Reference [32] states that the distance between $P_1 = \mathrm{DP}(\alpha, F_0)$ and $P_2 = \mathrm{TDP}(\alpha, F_0, N)$ is

$$D(P_1, P_2) \leqslant 2M \exp(-(N-1)/\alpha)$$

Thus if $N = 200$ and $\alpha = 10$ then $D(P_1, P_2) \leqslant 9.1 \times 10^{-7}$.

REFERENCES

1. Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomized trials: an application of meta-analysis techniques. *Statistics in Medicine* 1997; **16**:2063–2079.
2. Clayton DG, Kaldor J. Empirical Bayes estimates of age-standardized relative risk for use in disease mapping. *Biometrics* 1987; **42**:671–681.
3. The Bristol Royal Infirmary Inquiry. The inquiry into the management of care of children receiving complex heart surgery, 2001. Available at http://www.bristol-inquiry.org.uk
4. Spiegelhalter DJ, Aylin P, Best NG, Murray GD. Commissioned analysis of surgical performance by using routine data: lessons from the Bristol inquiry (with discussion). *Journal of the Royal Statistical Society*, *Series A* 2002; **165**:191–231.
5. Commission for Health Improvement. Report on the investigation into heart and lung transplantation at St George's Healthcare NHS Trust, 2001. Available at http://www.chi.nhs.uk/eng/organisations/london/st_georges

DOI: 10.1002/sim

6. Healthcare Commission. NHS performance ratings 2003/2004, 2004. Available at http://www.healthcarecommission.org.uk/InformationForServiceProviders/PerformanceRatings
7. New York State Department of Health. *Adult Cardiac Surgery in New York State 1998–2000*. New York State Department of Health: Albany, 2004.
8. Bernardo JM, Smith AFM. *Bayesian Theory*. Wiley: Chichester, 1994.
9. Bryk AS, Raudenbush SW. *Hierarchical Linear Models*. Sage: Newbury Park, 1992.
10. Breslow NE, Clayton DG. Approximate inference for generalized linear mixed models. *Journal of the American Statistical Association* 1993; **66**:9–25.
11. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society*, *Series A* 1996; **159**:385–443.
12. Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine* 1997; **127**:764–768.
13. Normand SL, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association* 1997; **92**:803–814.
14. Bronskill SE, Normand S-L, Landrum MB, Rosenheck RA. Longitudinal profiles of health care providers. *Statistics in Medicine* 2002; **21**:1067–1088.
15. Walker SG, Damien P, Laud PW, Smith AFM. Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1999; **61**:485–527.
16. Burr D, Doss H, Cooke GE, Goldscmidt-Clermont PJ. A meta-analysis of studies on the association of platelet P1A polymorphisms of glycoprotein IIIa and risk of coronary heart disease. *Statistics in Medicine* 2003; **22**:1741–1760.
17. Burr D, Doss H. A Bayesian semi-parametric model for random effects meta-analysis. *Journal of the American Statistical Association* 2005; **100**:242–251.
18. Spiegelhalter DJ, Thomas A, Best NG. *WinBUGS 1.4 User Manual*. MRC Biostatistics unit and Imperial College, 2003. Available from http://www.mrc-bsu.cam.ac.uk/bugs
19. NHS Executive. A first class service: quality in the new NHS, 1998. Available at http://www.doh.gov.uk/newnhs/quality.htm
20. NHS Executive. Quality and performance in the NHS clinical indicators, 1999. Available at http://www.doh.gov.uk/indicat/nhsci.htm
21. Lambert PC, Sutton AJ, Burton PR, Abrams KR, Jones DR. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using winbugs. *Statistics in Medicine* 2005; **24**:2401–2428.
22. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of complexity and fit (with discussion). *Journal of the Royal Statistics Society*, *Series B* 2002; **64**:1–34.
23. Muller P, Quintana FA. Nonparametric Bayesian data analysis. *Statistical Science* 2004; **19**:95–110.
24. O'Hagan A, Forster J. *Kendall's Advanced Theory of Statistics Volume 2B Bayesian Inference* (2nd edn, Chapter 13, Nonparametric models). Arnold: New York, 2004; 374–400.
25. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; **4**:639–650.
26. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; **96**:161–173.
27. Escobar MD, West M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 1995; **90**:577–588.
28. Bush CA, MacEachern SN. A semiparametric Bayesian model for randomised block designs. *Biometrika* 1996; **83**:275–285.
29. Kleinman KP, Ibrahim JG. A semi-parametric Bayesian approach to generalised linear mixed models. *Statistics in Medicine* 1998; **17**:2579–2596.
30. Muller P, Rosner GL. A Bayesian population model with hierarchical mixture priors applied to blood count data. *Journal of the American Statistical Association* 1997; **92**:1279–1292.
31. De Iorio M, Muller P, Rosner GL, MacEachern SN. An ANOVA model for dependent random measures. *Journal of the American Statistical Association* 2004; **99**:205–215.
32. Ishwaran H. Inference for the random effects in Bayesian generalized linear mixed models. *ASA Proceedings of the Bayesian Statistical Science Section*, 2000; 1–10. Available at http://www.bio.ri.ccf.org/Resume/Pages/Ishwaran/publications.html
33. Ishwaran H, Zarepour M. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* 2000; **87**:371–390.
34. Congdon P. *Bayesian Statistical Modelling*. Wiley: Chichester, 2001.
35. Gelfand AE, Kottas A. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 2002; **11**:289–305.

DOI: 10.1002/sim

36. Ishwaran H, James LF. Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* 2002; **11**:508–532.
37. Erkanli A, Sung M, Jane Costello E, Angold A. Bayesian semi-parametric roc analysis. *Statistics in Medicine* 2006.
38. Green PJ, Richardson S. Modelling with and without the Dirichlet process. *Scandinavian Journal of Statistics* 2001; **28**:355–375.
39. Ishwaran H, Zarepour M. Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* 2002; **12**:941–963.
40. Ghosh SK, Norris JL. Bayesian capture-recapture analysis and model selection allowing for heterogeneity and behavioral effects. *Journal of Agricultural Biological and Environmental Statistics* 2005; **10**:35–49.
41. Diebolt J, Robert CP. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B* 1994; **56**(2):363–375.
42. Aitkin M. A general maximum likelihood estimation of variance components in generalized linear models. *Biometrics* 1999; **55**:117–128.
43. Liu JS. Nonparametric hierarchical Bayes via sequential imputations. *Annals of Statistics* 1996; **24**:911–930.
44. Robert CP. Mixtures of distributions: inference and estimation. In *Markov Chain Monte Carlo in Practice*, Gilks WR, Richardson S, Spiegelhalter DJ (eds). Chapman & Hall: London, 1996; 441–464.
45. Richardson S, Green P. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1997; **59**(4):731–792.
46. Marshall EC, Spiegelhalter DJ. Comparing institutional performance using Markov chain Monte Carlo methods. In *Statistical Analysis of Medical Data. New Developments*, Everitt B, Dunn G (eds). Arnold: Paris, 1998.
47. Richardson S, Leblond L, Jaussent I, Green PJ. Mixture models in measurement error problems, with reference to epidemiological studies. 2002; **165**:549–566.
48. Rosner GL, Muller P. Bayesian population pharmacokinetic and pharmacodynamic analyses using mixture models. *Journal of Pharmacokinetics and Biopharmaceutics* 1997; **25**:209–233.
49. Ohlssen DI. Methodological issues in the use of random effects models for comparison of health care providers. *Ph.D. Thesis*, Department of Public Health and Primary Care, University of Cambridge, 2004.

DOI: 10.1002/sim