Journal of the Royal Statistical Society
SERIES C Applied Statistics
Appl. Statist. (2014)
63, *Part* 1, *pp.* 25–46

Semiparametric Bayesian models for clustering and classification in the presence of unbalanced in-hospital survival

Alessandra Guglielmi, Francesca Ieva and Anna M. Paganoni,

Politecnico di Milano, Italy

Fabrizio Ruggeri

Consiglio Nazionale di Ricerca, Milano, Italy

and Jacopo Soriano

Duke University, Durham, USA

[Received May 2012. Final revision February 2013]

Summary. Bayesian semiparametric logit models are fitted to grouped data related to in-hospital survival outcome of patients hospitalized with an *ST*-segment elevation myocardial infarction diagnosis. Dependent Dirichlet process priors are considered for modelling the random-effects distribution of the grouping factor (hospital of admission), to provide a cluster analysis of the hospitals. The clustering structure is highlighted through the optimal random partition that minimizes the posterior expected value of a suitable loss function. There are two main goals of the work: to provide model-based clustering and ranking of the providers according to the similarity of their effect on patients' outcomes, and to make reliable predictions on the survival outcome at the patient's level, even when the survival rate itself is strongly unbalanced. The study is within a project, named the 'Strategic program of Regione Lombardia', and is aimed at supporting decisions in healthcare policies.

Keywords: Bayesian clustering; Bayesian non-parametrics; Random-effects models; Random partitions; Unbalanced binary outcomes

1. Introduction

Bayesian non-parametrics provide extremely flexible models for fitting a variety of data sets. One of their most popular uses is in modelling distributions for random effects in hierarchical models for grouped data, as in the seminal paper of Kleinman and Ibrahim (1998). With such grouped data, the aim is usually to find clusters among groups which can capture the latent structure in the data that are assigned to each group. In this context, a natural way to achieve model-based clustering via Bayesian non-parametrics is to assume that the random-effects distribution is almost surely discrete, so that there will be ties in the posterior values of the random-effect parameters. In this way, two groups are in the same cluster if their corresponding sampled random-effects parameter values coincide. Dirichlet processes (DPs), which were introduced by Ferguson (1973), are the most popular discrete random probability mea-

Address for correspondence: Francesca Ieva, Dipartimento di Matematica, Politecnico di Milano, piazza Leonardo da Vinci 32, I-20133, Milano, Italy.
E-mail: francesca.ieva@mail.polimi.it

© 2013 Royal Statistical Society 0035–9254/14/63025

sures, used to represent population distributions. In particular, the discrete feature of DP-based models has been frequently exploited as a mechanism to generate clusters of subjects or groups (see De la Cruz-Mesía *et al.* (2007) and Green and Richardson (2001) among others).

In many applications, data include covariates besides the recorded responses. Recent efforts have produced interesting classes of random probability measures, dependent on such covariates, yielding dependent Dirichlet processes (DDPs) as described in MacEachern (1999, 2000) and Barrientos *et al.* (2012). Applications or extensions of such priors include covariate DDPs resembling traditional analysis-of-variance models as in De Iorio *et al.* (2004), DDPs with an additional probability model for group classification for longitudinal data as in De la Cruz-Mesía *et al.* (2007) and probit stick breaking random probability measures as in Rodriguez and Dunson (2011). See also the references therein.

In this paper we present two Bayesian semiparametric mixed models for the analysis of binary survival data coming from a clinical registry on ST-segment elevation myocardial infarction (STEMI), where statistical units (i.e. patients) are grouped by hospital of admission. In particular, in such hierarchical frameworks we adopt non-parametric DDP priors for modelling random effects superimposed on the grouping factor, to provide a proper methodological approach to the problem of profiling hospitals according to their effects on patients' outcomes. This topic is crucial within the context of healthcare planning, and proper methods for addressing such a problem are of great interest to people in charge of healthcare government (see Ash *et al.* (2012) and Spiegelhalter *et al.* (2012) for details on recent discussions and developments). Since the outcome of interest (in-hospital survival, i.e. whether a patient is discharged alive from hospital) is strongly unbalanced within the context of the disease that we focus on, any model will perform poorly in predicting the related adverse event. Therefore we propose a new method for classifying patients by using the whole predictive distributions of their outcome, labelling them as 'alive' or 'death' according to a criterion based on the posterior predictive credibility intervals.

We adopt a Bayesian semiparametric approach since it has a twofold advantage. First, Bayesian semiparametric models allow great flexibility in modelling data, avoiding critical dependence on parametric assumptions; see Müller and Quintana (2004). Secondly, Bayesian non-parametric priors on the random-effects parameters selecting discrete probability measures yield a random partition of the group indices set; consequently, cluster estimates will be based on the posterior distribution of the random partition itself. A common way of estimating the unknown true partition is to observe the maximum *a posteriori* estimate. However, since the number of partitions is large even for moderate sizes of the indices set and the posterior is usually spread out, the maximum *a posteriori* estimate may not be a good choice, and therefore different summary statistics of the posterior distribution of the random partition are needed. Formal decision-theoretic-based procedures for choosing one single estimate based on posterior expectations of appropriate loss functions were discussed in Lau and Green (2007) and Quintana and Iglesias (2003). Since one of the main focuses of the paper is to exploit model-based clustering of groups, we shall pursue this issue by providing a Bayesian estimate, as proposed in Lau and Green (2007), looking for *a posteriori* clustering structure, optimal with respect to a specified loss function. However, our interest here is also focused on classification and prediction of binary responses in situations where the chance of success is strongly unbalanced. We then propose a new rule for the classification of patients, which is based on the posterior credibility intervals of patients' survival probability, instead of point estimates, discussing how the classification that is obtained depends on the choice of a reference threshold, according to what was suggested in Cramer (1999). A discussion on performances of threshold criteria for binary classification based on pointwise outcome estimates is presented in Freeman and Moisen (2008).

We apply these methods to a data set arising from a clinical registry (the STEMI archive;

see Direzione Generale Sanità—Decreto Regione Lombardia (2009) and Ieva (2013)) on patients affected by STEMI and admitted to any hospital in Lombardia, which is an Italian region whose capital is Milan. Specifically, the binary outcome of interest is measured at patients' level, and patients are grouped according to the hospital of admission. Hence, there is a hierarchical structure in the data set: hospitals at a higher level and patients at a lower. It is known from the literature (see Cannon *et al.* (2000) among others) that STEMI is characterized by a strongly unbalanced share of success in terms of in-hospital survival; in our data set, in fact, 97% of patients were discharged alive from the hospital. It is also known (see Bradley *et al.* (2006) and Ieva and Paganoni (2011) for instance) that, for such disease, reducing treatment times and optimizing pre- and intra-hospital patterns of care strongly improve patients' prognosis. In particular, we are interested in profiling healthcare providers, investigating whether any clustering of the hospitals has a meaning. Since clustering is obtained through estimates of the posterior distribution of the random partition of the hospital index set, we shall be able to assess the effect of groups of healthcare providers with 'similar' behaviour on patients' outcomes, as well as to evaluate the quality of their performances in treating STEMI patients, adjusting for casemix and all other known sources of variability that induce overdispersion in the outcomes distribution.

We shall consider two logit models for the in-hospital survival probability. We adopted this link function because it enables a straightforward clinical interpretation of parameters and results, and, since our study is motivated by a clinical problem, it is also important to ease the communication of results. In both models that we consider, the random-effect parameters are given a non-parametric prior, similarly to Kleinman and Ibrahim (1998), whereas lower level covariates are treated parametrically. Specifically, the random-effect parameters are assumed as a sample from a Dirichlet process. In our case, since a random effect is superimposed on the grouping factor that is represented by the hospital of admission of patients, we model the dependence across random distributions through the hospitals' covariates, so that the priors can be interpreted as DDP distributions. The two Bayesian models differ in the choice of covariates that are included in the likelihood and in the non-parametric components of the random-effect parameters (see Section 2).

The novelty of this work lies in exploiting a model-based clustering, provided by the optimal partition of the random effects estimated through a Bayesian semiparametric hierarchical model, for profiling providers in a real clinical problem. In fact, the method that we propose in this paper yields a model-based ranking of hospitals, based on the evolution of the optimal partition of the random effects. Moreover, using posterior credibility intervals for classifying patients as dead or alive instead of pointwise estimates, we identify a classification rule that proves to be less sensitive to the choice of the threshold discriminating groups of alive and dead patients.

The paper is organized as follows. In Section 2 we present the models and the methodology that is developed for hospital clustering and patients' classification. Goodness-of-fit indices for comparing the models are also considered, and details on random-effects clustering that is carried out through the optimal random partition are provided. Section 3 presents the results of the inference for the STEMI archive data. Finally, some conclusions are drawn and discussed in Section 4. All the analyses have been carried out with R (see R Development Core Team (2009)) and JAGS (see Plummer (2003)).

2. Bayesian semiparametric models for random-effects clustering

In this section, we present the two models that we shall use to analyse the data in Section 3. In what follows, the model formulation is already intended for the application of interest, where the outcome is the in-hospital survival after an STEMI event, and patients are grouped by hospital

of admission. Among all the variables that are available at patients' level, we considered age, total ischaemic time (OB, symptom onset to balloon time), presence of chronic kidney disease (CKD = 1 if the patient had a loss of renal function; CKD = 0 otherwise) and Killip class (which is an ordinal variable indicating the severity of infarction, from 1, lowest, to 4, highest). Moreover, we included in the models hospital exposure, i.e. the number of patients who were treated with primary angioplasty per year, and a binary variable (Milano) indicating whether the hospital is in (Milano = 1) or outside (Milano = 0) Milan. Then the mathematical framework of loss functions for the evaluation of the optimal partition is briefly described, to cluster the hospitals. Finally, a classification rule based on posterior credibility intervals will be introduced.

As we mentioned in Section 1, we assume DP priors for the random-effects distributions in the logit likelihood. An equivalent representation yields that, in the models that we are considering, the random-effect parameters $\mathbf{b}_j$, corresponding to the jth hospital effect, are distributed according to a DDP prior P_{v_j}, which depends on a covariate v_j in its definition. Hence, marginally $\mathbf{b}_j$ has still a DP prior, with the property that 'P_{v_j} varies smoothly with v_j' (see MacEachern (2000)). This implies that P_{v_j} and $P_{v'_j}$ are correlated for $v_j \neq v'_j$ and, at least where continuous covariates are present, that $P_{v'_j}$ reaches P_{v_j} as long as v'_j approaches v_j. Of course, DDPs can adopt many different and quite elaborate forms, but here we analyse only two such specifications, which retain interpretability of model parameters.

2.1. Dependent Dirichlet process priors on random effects

For statistical unit $i = 1, \ldots, n_j$ in group $j = 1, \ldots, J$, let Y_{ij} be a Bernoulli random variable with mean p_{ij}. In our application, p_{ij} represents the probability that patient i treated in hospital j is discharged alive after an STEMI event. The p_{ij}s are modelled through a multivariate logistic regression with fixed effects α and β, and a random effect $\mathbf{b}$ superimposed on the covariates referred to the grouping factor, i.e.

$$\begin{aligned} Y_{ij}|p_{ij} &\overset{\text{ind}}{\sim} \text{Be}(p_{ij}), \\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \sum_{l=1}^{4} \alpha_l u_{ijl} + \sum_{k=1}^{5} \beta_k x_{ijk} + b_{0j} + b_{1j} z_j. \end{aligned} \tag{1}$$

Within the context of the application motivating this study, $u_{ij} = (u_{ij1}, \ldots, u_{ij4}) = (\text{Killip1}, \ldots, \text{Killip4})_{ij}$ is a vector of dummy variables, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij5}) = (\text{age}, \log(\text{OB}), \text{CKD}, \text{exposure}, \text{Milano})_{ij}$ and z_j is the exposure of the jth hospital. All continuous covariates have been centred and standardized (so that their range is between -1 and 1) to obtain a better mixing of the Markov chains arising from simulations. A null covariate vector represents a patient with 'average' age and total ischaemic time, not at risk in terms of CKD and treated in a structure dealing with an 'average' number of STEMI patients per year, also. In what follows, we shall refer to such a patient as a 'standard reference' and compare hospital effects once adjustments for all fixed effects have been carried out in the standard reference setting. The prior distribution that is assumed for the parameters of the model is

$$\begin{aligned} \alpha &= (\alpha_1, \ldots, \alpha_4) \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2 \mathbb{I}_4), \\ \beta &= (\beta_1, \ldots, \beta_5) \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2 \mathbb{I}_5), \end{aligned} \tag{2}$$

$$(b_{0j}, b_{1j})|P \overset{\text{IID}}{\sim} P \qquad j = 1, \ldots, J, \quad P|a, P_0 \sim \text{DP}(a, P_0). \tag{3}$$

Independence between α, β and P is assumed. By $P \sim \text{DP}(a, P_0)$ we mean that P, the (con-

ditional) distribution of the bivariate random-effects parameter $\mathbf{b}_j$, has a DP prior with total mass parameter $a>0$ and base probability measure parameter P_0; see Ferguson (1973) for details on the definition and standard notation of DPs. The base probability measure on $\mathbb{R}^2$ for this model, P_0, will be chosen as the product measure $\mathcal{N}(0,\sigma_0^2)\times\mathcal{N}(0,\sigma_1^2)$, being σ_0 and σ_1 independent and uniformly distributed. Moreover, a is assumed to be random with prior $\pi(a)$; in Section 3 a truncated exponential distribution is chosen as prior distribution for a.

Observe that in model (1) the random-effect parameter of hospital j appears linearly as $b_{0j}+b_{1j}z_j$. Moreover, each $\mathbf{b}_j=(b_{0j},b_{1j})$, given P, has distribution

$$P=\sum_{h=1}^{\infty} w_h\delta_{\theta_h} \tag{4}$$

where θ_h are independent and identically distributed (IID) according to P_0 and $\{w_h\}$ are the weights in the stick breaking representation (see Sethuraman (1994)). It is straightforward to see that $b_{0j}+b_{1j}z_j$, given $\tilde{P}$, is distributed as $\tilde{P}$, where

$$\tilde{P}=\sum_{h=1}^{\infty} w_h\delta_{\tilde{\theta}_h(z_j)}. \tag{5}$$

Here $\tilde{\theta}_h$s are IID according to $\tilde{P}_0$ which is the distribution of $b_{0j}+b_{1j}z_j$ if $\mathbf{b}_j$ is distributed according to P_0. Therefore, by equation (5), the random-effect contribution to the likelihood in model (1) is distributed according to a DDP. This is a rather simple case of a DDP, called a 'single-p linear DDP' (see MacEachern (2000)), since the weights in the stick breaking construction do not depend on covariates, whereas the location points do, in a linear way.

The other semiparametric model that we consider is

$$\begin{aligned} Y_{ij}|p_{ij} &\overset{\text{ind}}{\sim} \text{Be}(p_{ij}),\\ \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \sum_{l=1}^{4}\alpha_l u_{ijl}+\sum_{k=1}^{3}\beta_k x_{ijk}+b_{v_j j}, \end{aligned} \tag{6}$$

where α, β and $\mathbf{b}$ are the parameter vectors corresponding to the fixed and random effects, as in the previous case. Referring to the motivating application, $\mathbf{u}_{ij}$ is the Killip dummy vector as in the previous model, whereas $\mathbf{x}_{ij}=(x_{ij1},x_{ij2},x_{ij3})=(\text{age},\ \log(\text{OB}),\text{CKD})_{ij}$. Finally, $b_{v_j j}$ is the random intercept depending on values that are assumed by the location dummy Milano ($v_j=0$ or $v_j=1$). Note that here we distinguish the random-intercept parameter according to the geographical origin of the hospital: in fact, $b_{v_j j}$ is the parameter referring to the jth hospital, which will be b_{1j} if the jth hospital is in Milan, and b_{0j} otherwise. We assume the following prior:

$$\begin{aligned} \alpha&=(\alpha_1,\ldots,\alpha_4)\sim\mathcal{N}(\mu_\alpha,\sigma_\alpha^2\mathbb{I}_4),\\ \beta&=(\beta_1,\ldots,\beta_3)\sim\mathcal{N}(\mu_\beta,\sigma_\beta^2\mathbb{I}_3), \end{aligned} \tag{7}$$

$$(b_{0j},b_{1j})'|P\overset{\text{IID}}{\sim}P \qquad j=1,\ldots,J, \quad P|a,P_0\sim\text{DP}(a,P_0). \tag{8}$$

Independence between α, β and P is assumed. For our application, we shall assume that the base probability measure on $\mathbb{R}^2$, P_0, is chosen as the product measure $P_{00}\times P_{01}\equiv\mathcal{N}(0,\sigma_0^2)\times\mathcal{N}(\mu_1,\sigma_1^2)$. Moreover, σ_0 and σ_1 will be assumed to be uniformly distributed. Finally, a Gaussian distribution will be considered for μ_1 and a truncated exponential distribution for a. Observe that the number of random-effect parameters in model (6) is J (and not $2J$), since, if j is the index of a statistical unit with $v_j=1$, then the corresponding random-effect parameter is b_{1j}; in

contrast, if j is the index corresponding to a group with $v_j = 0$, the corresponding random-effect parameter is b_{0j}.

In this case the non-parametric prior component that is assumed for the random-effects parameters can be interpreted as an analysis-of-variance DDP prior with one factor and two levels (see De Iorio *et al.* (2004)), where the v_j-covariate ruling the prior assumes only values in $\{0, 1\}$, here representing the Milano effect. In fact, we could equivalently assume

$$b_{v_j j} | P, v_j \overset{\text{ind}}{\sim} P_{v_j},$$
$$P_{v_j} | P_{0v_j} \sim \text{DP}(a, P_{0v_j}),$$

where, for v equal to 0 or 1,

$$P_v = \sum_{h=1}^{\infty} w_h \delta_{\theta_{vh}},$$
$$(\theta_{0h}, \theta_{1h})' \overset{\text{IID}}{\sim} P_{00} \times P_{01}. \tag{9}$$

Observe that P_v is marginally $\text{DP}(a, P_{0v})$, and the dependence between P_0 and P_1 is induced by the presence of common weights in their stick breaking representation.

The main difference between the priors of the two models that were described in the previous section lies in the Milano covariate effect, which is included directly in the locations of the stick breaking representation in expression (9) in the second model.

In what follows, we shall refer to the model that is defined by equations (1)–(3) as 'model A', and to the model that is defined by equations (6)–(8) as 'model B'.

2.2. Model comparison

Since the data set that we deal with in the motivating application is complex and rich in covariates, there are many Bayesian models that could be fitted to the data. In particular, the covariates' dependence could be included in the DDP in many different ways. We focused on likelihoods containing the most significant covariates indicated in previous work (see Guglielmi *et al.* (2012) and Ieva and Paganoni (2011)) by some variable selection methods, and we tried various ways of combining hospital covariates within the non-parametric priors. Some covariates (at both patient and hospital level) are included to allow us to investigate specific topics related to clinical enquires and health analytics.

However, we fitted two more models: one is a simplified version of model A, where we removed the hospital exposure (fixed and random) from model (1) and assumed a univariate DP prior for the random intercept. The inference that we obtained from the two models was similar, but we preferred to consider the likelihood as in model (1), since it allowed us to draw conclusions on the relationship between goodness of performance and hospital exposure, as reported in Section 3. In contrast, as a second alternative, we fitted a model with a DDP prior for the bivariate vector of the random-effects parameter $\mathbf{b}_{v_j j}$ representing the effect of the intercept and the exposure for each hospital in and outside Milan. The analysis showed that there is no need to introduce this more complex model, since the posterior inference that we obtained was very similar to that given by model B, which is reported with details in Section 3.

To compare models A and B with respect to their estimates of the random effects, we must match them up to some extent, e.g. matching the marginal distribution of the random intercepts under the two models. Table 1 reports the random-intercept parameters, up to the Killip parameter α, of the two models for a hospital in or outside Milan. As we mentioned before, since we deal with standardized covariates, the random intercepts that are reported in Table 1

Table 1. Random-intercept parameters in model A and model B

Model	*Parameters for the following hospital locations:*	
	In Milan	*Outside Milan*
A	$\beta_5 + b_{0j}$	b_{0j}
B	b_{1j}	b_{0j}

represent the in-hospital survival probability on the logit scale for a 'standard reference' patient (without the Killip effect). As we shall see in Section 3, we fixed hyperparameters so that the prior marginal distributions of random intercepts of hospitals in Milan are equal, as well as those of random intercepts of hospitals outside Milan. Anyway, even if denoted with the same symbols, the intercepts have a different interpretation, according to the different likelihoods that they refer to. Moreover, the covariances between the random intercepts differ under the two models. It is easy to show that, for model A, for a hospital h outside Milan and a hospital l in Milan,

$$\operatorname{cov}(b_{0h}, b_{0l} + \beta_5) = \operatorname{cov}(b_{0h}, b_{0l}) = \frac{\sigma_0^2}{a+1},$$

whereas for the model B

$$\operatorname{cov}(b_{0h}, b_{1l}) = \frac{\operatorname{cov}(P_{00}, P_{01})}{a+1} = 0.$$

To evaluate model goodness of fit, we compute an index that was introduced by Gelman and Pardoe (2006), who proposed a Bayesian generalization of the R^2-index for linear models. In a frequentist framework, the coefficient of determination R^2 estimates the proportion of variance that is explained by the linear model. Here we apply it to the first level of the logistic regression, which can be rewritten in terms of a latent variables formulation (see Albert and Chib (1993)) as follows:

$$Y_{ij} = \mathbf{1}_{\{Z_{ij} \geqslant 0\}} \qquad Z_{ij} = \mu_{ij} + \varepsilon_{ij}. \tag{10}$$

Here the μ_{ij}s are the linear predictors, as in model (1) or in model (6), and the ε_{ij} are IID standard logistic random variables. We assume that, conditioning on the latent variables Z_{ij}, the Y_{ij} are independent.

Starting from the latent variable representation of the model provided in expression (10), a Bayesian generalization of the R^2-index for linear models can be defined as

$$R^2 = 1 - \frac{\mathbb{E}\left[\bigvee_{ij} \varepsilon_{ij}\right]}{\mathbb{E}\left[\bigvee_{ij} \mu_{ij}\right]} = 1 - \frac{\operatorname{var}(\varepsilon)}{\mathbb{E}\left[\bigvee_{ij} \mu_{ij}\right]}, \tag{11}$$

where '$\bigvee$' represents the sample variance operator

$$\bigvee_{ij} x_{ij} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2 \Big/ \left(\sum_{j=1}^{J} n_j - 1 \right).$$

The Bayesian R^2 provides an index of the explained variability at the latent variable level. It is close to 1 when μ_{ij}s approximate well the conditional mean of Y_{ij}s and close to 0 when the sample variance of the εs is approximately equal to the variance of the μ_{ij}s. Whereas the frequentist R^2 ranges from 0 to 1, the Bayesian R^2-index could also be negative.

2.3. Random partitions for model-based cluster analysis

As we mentioned in Section 1, one of the main aims of the work is to exploit the clustering that is induced by the random-effects prior to investigate the effects of groups of 'similar items' on the outcomes of interest. In particular, the idea is to carry out a model-based clustering, in which labels are exchangeable, and items are also exchangeable, possibly up to covariate effects.

In a Bayesian formulation of a clustering procedure, the partition ρ of item labels into subsets depends on the probability model for the data, and therefore cluster inference is obtained from the posterior distribution of the partition itself. Specifically, here, ρ is a random partition of hospitals labels $\{1, 2, \ldots, J\}$ that is induced by the sampling from a DP or from any random probability measure which is discrete with positive probability. For model B, where the non-parametric prior component is an analysis-of-variance DDP, clusters of random-effect parameters occur both within the two groups (hospitals in Milan and outside Milan) as well as across the geographical location.

Our aim is to compute a suitable posterior estimator $\hat{\rho}$, representing the best estimate of the 'true' clustering of the random effects. Clinically speaking, we would like to estimate a latent clustering among hospitals of our data set, identifying groups of providers affecting outcomes at patients' level in a similar way. This could be of great interest for decision makers, to point out outliers with respect to a reference standard of quality, as well as to rank groups of structures according to suitable criteria, after adjusting for all confounding factors, due to both patients' covariates and hospital features. Choosing a partition ρ can be considered as a model choice problem, and different approaches to tackle it are available, as proposed in Dahl (2006), Gordon (1999), Heard *et al.* (2006) and Ray and Mallick (2006). A loss function approach avoids some criticisms that are related to the spread of the posterior distribution of random partitions. As in Lau and Green (2007), we concentrate on loss functions that rely on *pairwise coincidences* (see Binder (1978, 1981)) and count how many times a wrong labelling happens, assigning a different weight to the two types of misclassification. Specifically, we choose the loss function which assigns a positive cost u any time that two random effects are incorrectly assigned to different clusters, and a positive cost w any time that two random effects are incorrectly clustered together. The total loss is then obtained by summing over all pairs. Denoting by c_i the true allocation variable, then

$$L(\rho, \hat{\rho}) = \sum_{(i,j)\in\mathcal{M}} \{u\,\mathbb{I}(c_i = c_j, \hat{c}_i \neq \hat{c}_j) + w\,\mathbb{I}(c_i \neq c_j, \hat{c}_i = \hat{c}_j)\},$$

where $\hat{\rho}$ is the estimate and ρ is the current value of the partition, and $\mathcal{M} = \{(i, j) : i < j; i, j \in \{1, \ldots, J\}\}$. The proposed estimate of the random partition in this case is the estimate minimizing the posterior expected loss

$$\mathbb{E}[L(\rho, \hat{\rho})|\mathbf{Y}] = \sum_{(i,j)\in\mathcal{M}} \{u\,\mathbb{I}(\hat{c}_i \neq \hat{c}_j)\,\mathbb{P}(c_i = c_j|\mathbf{Y}) + w\,\mathbb{I}(\hat{c}_i = \hat{c}_j)\,\mathbb{P}(c_i \neq c_j|\mathbf{Y})\},$$

where $\hat{c}_i$ is the estimated allocating variable for the ith unit. Lau and Green (2007) showed that this is equivalent to maximizing

$$l(\hat{\rho}, K) = \sum_{(i,j)\in\mathcal{M}} \mathbb{I}(\hat{c}_i = \hat{c}_j)(\gamma_{ij} - K),$$

where $\gamma_{ij} = \mathbb{P}(c_i = c_j|Y)$, and $K = w/(u+w) \in [0,1]$. As a function of K, $l(\hat{\rho}, K)$ characterizes the quality of each possible $\hat{\rho}$, and the whole family of such functions determines for which K, if any, each partition is optimal, as well as defining the optimal $\hat{\rho}$ for each K. As will be clear in Section 3, we shall observe how the clustering that is induced by the random partition changes as different values of K are considered. This will lead to a sort of 'implicit ranking' of the hospitals in our data set.

The maximization of $l(\hat{\rho}, K)$ can be carried out through binary integer programming techniques, as explained in Lau and Green (2007). Since the total number of hospitals is not large, the computational effort that is required for solving the optimization problem can be carried out by using the R package `lpSolve` (Berkelaar *et al.*, 2004).

2.4. Outcomes classification and prediction

The second major goal of the present work is to make predictions for outcomes of interest starting from the posterior predictive distributions of our models. It is well known that the rarest event is difficult to predict, irrespective of the model considered, when the data set contains binary variables that are characterized by unbalanced shares of success. We propose a method for addressing this issue, enhancing the strength of the Bayesian approach.

The usual predictive method for binary data is based on point estimates of the posterior predictive distribution, i.e., since p_{ij} is the probability of observing a successful outcome for item i in group j, the outcome Y_{ij} will be predicted as a success whenever $\mathbb{E}[p_{ij}|Y]$ is bigger than a given threshold. In the application, we consider the in-hospital survival probability p_{ij} of patient i admitted to hospital j, and we are interested in correctly classifying the patients belonging to the current data set as well as in making prediction on the status of a new patient. Since the survival outcome is strongly unbalanced in this case (97% of in-hospital survival is observed), the models will provide poor results in predicting deaths, if the usual criteria based on pointwise estimates are adopted.

Many solutions to this problem have been proposed, since the classification is typically very sensitive to the choice of the threshold (see, for example, Freeman and Moisen (2008) for a review and comparison of such most popular criteria in the frequentist literature). In our opinion, classification rules based on pointwise estimates are not completely satisfactory. First, they are not robust with respect to the choice of the thresholds. Moreover, since a Bayesian approach is adopted for modelling data and Bayesian inference provides the whole posterior predictive distribution of outcomes, we would like to exploit the richer information that it provides. The posterior predictive distribution for a new patient i in hospital j can be easily simulated through a Markov chain Monte Carlo algorithm via the compositional parameter method, first generating a draw from the posterior distribution of the parameters characterizing the model, and then generating from the conditional distribution of Y_{ij}^{new} given the parameters and the corresponding covariates. We propose a new method for outcome predictions at a lower unit level. It is based on an interval estimate of the posterior success rate and can be considered as a generalization of the 'standard' estimate, based on pointwise estimates to be compared with given thresholds. We classify a patient as alive if the, say 90%, credibility interval (CI) of his or her survival rate is entirely above a given threshold (say 0.5), or as dead if the CI is entirely below the threshold; we do not decide on the status of the patient if the CI contains the threshold. In such cases we say that the patient belongs to the uncertainty class (UC). Of course, the higher the credibility level is, the larger is the number of patients belonging to the UC.

3. Data analysis

In this section we present the analysis of data arising from the motivating problem, according to the two models and techniques that were presented in the previous section. As we said before, the data that we consider come from a clinical registry, named the STEMI archive, gathering patients admitted with STEMI diagnosis in any hospital of Regione Lombardia district. A complete description of the registry is provided in Ieva (2012, 2013), where data are presented together with the clinical setting that motivated their collection. As mentioned in Section 1, information about both patients and hospitals is available. Among the most important patient information that is provided by the clinical registry there are mode of admission (a patient reaches the hospital on his or her own or is delivered by three different types of rescue units of 118, which is the national toll-free number for emergencies), demographic features (age and sex), clinical appearance (Killip), risk factors (diabetes, smoking, chronic kidney disease, ...), times to treatment and times to intervention, as well as all the process indicators concerned with pre- and in-hospital phase, and clinical outcomes. Some of these covariates have already been described. In this application we focus on in-hospital survival of patients whose data are contained in the STEMI archive. However, information about the hospital of admission—considered as the grouping factor—is also present (in particular, a dummy variable indicating whether the hospital is in or outside Milan and the hospital exposure).

The variability of the distribution of patients' outcomes is high between structures. The data set contains $n = 697$ patients, who were admitted in $J = 29$ hospitals of Regione Lombardia. Initial selection of patient covariates was done on a similar data set in Ieva and Paganoni (2011) by using clinical know-how and stepwise selection procedures, based on the Akaike information criterion index AIC, confirmed later by a Bayesian variable selection method, using Gibbs variable selection (as reported in Guglielmi *et al.* (2012)). As we said in Section 1, the most significant factors which explain survival probabilities are age, Killip, CKD and total ischaemic time on the log-scale from symptom onset to the primary angioplasty (balloon), i.e. $log(OB)$. Providers' covariates Milano and exposure are also included. In fact, we are interested in evaluating whether differences between the hospitals may be assessed and, in this case, whether such differences lead to a clustering of providers.

As far as posterior inference from the models that have been introduced so far is concerned, first we provide posterior estimates of the parameters for each model, focusing in particular on posterior interval estimates and cluster estimates of the hospital random effects; then we evaluate models' goodness of fit and classify patients according to the predictive rule that was proposed in Section 2. All estimates have been carried out by a Gibbs sampler algorithm, translated into a JAGS code which calculates the full conditional distributions automatically. For completeness some details on the full conditionals are shown in Appendix A. In the two models we implemented the truncated DP approximation that was suggested by Ishwaran and Zarepour (2000) to obtain a trajectory from P; we truncated (and normalized) the sums in expressions (4), (5) and (9) at $H = 30$. The code is available from the authors on request. We ran the two models for 200000 iterations, discarding the first 100000, and using a thinning of 20 to reduce auto-correlations, so that the final sample size was 5000. Trace plots, auto-correlations and Geweke diagnostics indicate that the Gibbs sampler algorithms could have converged.

A robustness analysis showed that inferences are quite sensitive to the choice of the fixed effects' hyperparameters and the variance of the non-parametric components σ_0^2 and σ_1^2. Concerning the former, we fixed them 'informatively' as the means of the posterior distributions obtained by fitting a parametric model with the same covariates and Gaussian-distributed errors on data arising from a previous data collection of the same registry (see Guglielmi *et al.*

(2012)). This enabled us to set informative values for the means of fixed effects α and β, as well as for their variances σ_α^2 and σ_β^2. In contrast, concerning the random-effect variance components we tested two classes of prior: the conjugate inverse gamma distribution on the variances and the uniform distribution on the standard deviations. We refer to Gelman (2006) for a discussion on priors of the variance components in hierarchical models. The estimates of the random effects are particularly sensitive to the choice of the inverse gamma hyperparameters, whereas they are more robust by using uniform priors chosen according to prior information derived from Guglielmi *et al.* (2012). The total mass parameter was assumed bounded away from zero owing to numerical instability of the posterior simulation algorithms, as implemented in JAGS. Finally, we tested an exchangeable prior for the Killip vector $(\alpha_1, \ldots, \alpha_4)$, instead of assuming them to be IID. The estimation is robust to these choices, but the mixing is better under the independence assumption.

Fig. 1 shows the survival posterior predictive distributions for a patient who was discharged alive (Fig. 1(a)) and for a patient who died (Fig. 1(b)) for model A (full curve) and model B (broken curve). Note that the two posterior predictive distributions in each panel do not differ too much, but they do differ from the corresponding prior predictive distributions (which are not displayed here, to make the graphs clearer). Concerning the patient who was discharged alive (Fig. 1(a)), he is a male, aged 66 years, with a less severe infarction (Killip class equal to 1), no chronic kidney disease (CKD $= 0$) and an acceptable total ischaemic time (OB $= 120$ min), according to guidelines indicating the limit of 120 min. In contrast, the patient who died (Fig. 1(a)) was a male, aged 59 years, with a severe infarction (Killip class equal to 4), no chronic kidney disease (CKD $= 0$) and a total ischaemic time (OB $= 72$ min) that is much lower than that indicated by guidelines. Both patients had been admitted to hospitals in Milan, although not the same.

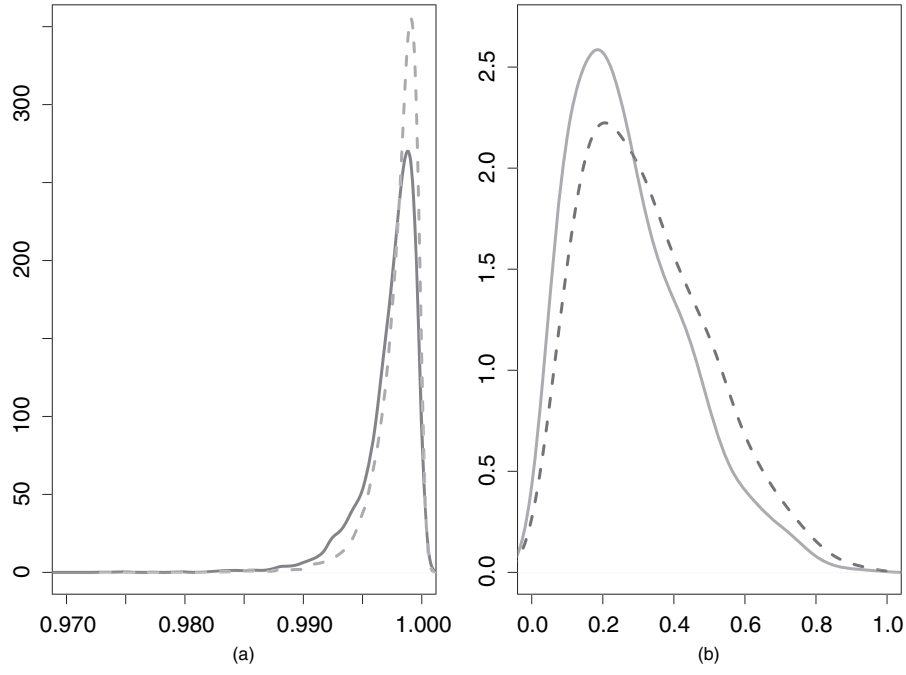

Fig. 1. Posterior predictive distributions for model A (———) and model B (- - - - - -) of the survival probability for two patients: (a) one who was discharged alive and (b) one who died

3.1. Fixed and random-effects estimates

In Table 2 we provide posterior 95% CIs of the fixed effects under models A and B. Hyperparameters in equation (2) were set informatively as we mentioned before, specifically $\mu_\alpha = (4.2, 4.2, 4.2, 4.2)'$, $\mu_\beta = (-1.7, -0.45, -1.7, 0.07, -0.45)'$, $\sigma_\alpha^2 = 4$ and $\sigma_\beta^2 = 4$. For model B, the same values are adopted, selecting only the fixed effects of interest for expression (7).

Note that the estimates are similar. In particular, the Killip index seems a good stratification parameter for both models, since the posteriors of the Killip 1 parameter concentrate on 'high' values (i.e. it leads to high survival probability), those of Killip 2 and 3 concentrate on 'average' values, and those of Killip 4 concentrate on 'small' values. As we might expect, as age, log(OB) and CKD increase, the survival probabilities decrease. Finally, the binary covariate Milano has a negative effect in model A, whereas the exposure is not significant. This was the reason why we decided not to include the exposure in model B, but we used the Milano-covariate to enrich the hospital random-intercepts prior distribution. Results about exposure and location influence have been investigated in detail by decision makers and physicians. Exposure not significant means that there is no evidence from data for concluding that hospitals that treat more patients are necessarily the best in terms of performance, contrary to what people who are in charge of healthcare government believed. However, as can be appreciated from Fig. 2, it seems that being treated in Milan results in a worse outcome, which is unexpected. We asked epidemiologists whether their data would confirm this finding, and they verified that, according to the evidence of our results, the epidemiology seems to be different between Milan and its neighbourhoods, expecially for people over 80 years old.

As we discussed in the previous section, we tuned the hyperparameters of the priors of the two models to match them in terms of marginal random-intercepts priors (see Table 1). In particular, the matching in Section 2.2 is achieved by fixing (informatively) both marginal distributions of the random intercepts in Milan as

$$\int \mathcal{N}(-0.45, 4+\sigma^2)\, \mathbb{I}_{[0,5]}(\sigma)\, \mathrm{d}\sigma = \int \mathcal{N}(\mu_1, \sigma^2)\, \mathbb{I}_{[0,5]}(\sigma)\, \pi(\mu_1)\, \mathrm{d}\sigma\, \mathrm{d}\mu_1,$$

with $\pi(\mu_1)$ being the normal distribution $\mathcal{N}(-0.45, 4)$, and outside Milan as

$$\int \mathcal{N}(0, \sigma^2)\, \mathbb{I}_{[0,5]}(\sigma)\, \mathrm{d}\sigma.$$

Table 2. Posterior 95% CIs of the fixed effects

Parameter	*Results for model A*			*Results for model B*		
	2.5%	*Median*	*97.5%*	*2.5%*	*Median*	*97.5%*
Killip1	4.81	6.59	8.49	4.17	6.04	8.10
Killip2	2.79	4.69	6.70	2.45	4.39	6.58
Killip3	2.10	4.22	6.42	1.61	3.70	6.07
Killip4	−0.24	1.57	3.43	−1.12	0.81	2.94
age	−3.41	−1.88	−0.50	−3.38	−1.77	−0.35
log(OB)	−3.33	−1.82	−0.22	−3.46	−1.91	−0.17
CKD	−3.00	−1.71	−0.41	−3.41	−2.09	−0.79
exposure	−2.34	0.19	2.79			
Milano	−3.68	−2.00	−0.26			

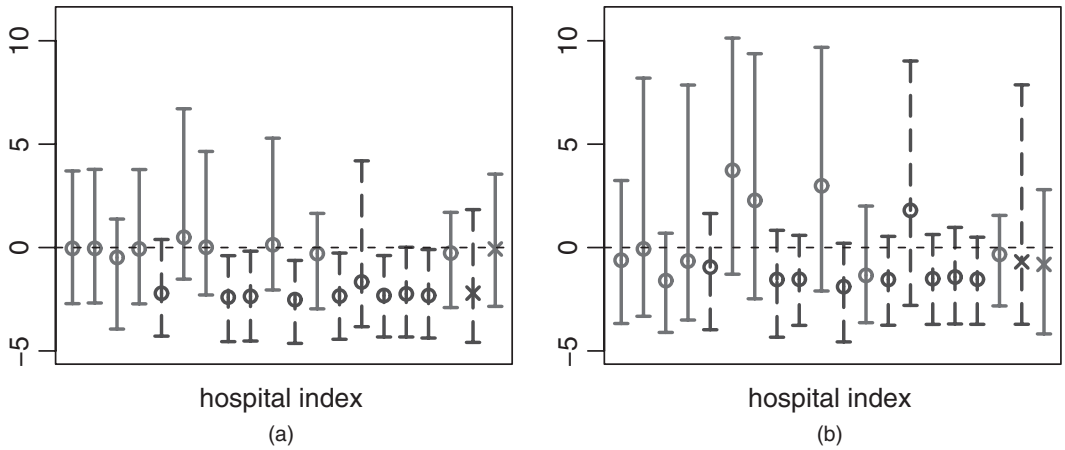

Fig. 2. Posterior 95% CIs of the random intercepts for hospitals with at least 10 patients for (a) model A and (b) model B, highlighting the Milano-effect (the estimates are in increasing order of number of patients per hospital and the last two intervals represent new random intercepts for a hospital in and outside Milan): ⊺, hospitals in Milan; ⊺, hospitals outside Milan

In Fig. 2 we provide posterior 95% CIs of the hospital random intercepts with at least 10 patients, highlighting the Milano-effect, for the two models.

The plots of the hospitals' slope (exposure) for both models show no appreciable variability, and for this reason we do not include them here. Note that under model A (Fig. 2(a)) all the hospitals outside Milan have a higher median than Milan hospitals and the intervals are shorter. Model B, in contrast, gives higher variability within each subpopulation. This variability is reasonable because of the greater flexibility of the prior of the second model.

3.2. Hospital clustering

As mentioned in Section 2.3, the non-parametric prior component induces a random partition of the hospital labels. Therefore we analyse the posterior of the process P to investigate hospital clustering. In Grieco *et al.* (2012) we pointed out that few groups could be detected among hospitals. The same conclusion holds under a parametric Bayesian mixed effects model (see Guglielmi *et al.* (2012) for details). We tuned hyperparameters of the prior for the total mass a in our models according to this prior information, i.e. $a \sim$ trunc-exp(1) on the interval $[1, \infty)$ (equivalently, $a = 1 + X$, where $X \sim \exp(1)$), which *a priori* leads to $\mathbb{E}[a] = 2$. The *a priori* number of groups in this case is 5.8. The mass parameter a is *a posteriori* concentrated around small values under both models: mean 1.61 (standard deviation 0.62) in model A and 1.65 (standard deviation 0.65) in model B. We observe a slight reduction in the expected number of groups, going from a prior mean of 5.8 to a posterior mean of 4.24 in model A and 4.58 in model B. Furthermore, we ran the algorithm, varying the prior specification for a (degenerate on different values or fixing different hyperparameters for its prior) but still yielding a small prior number of clusters. The posterior estimates of the number of clusters are robust and consistent across the two models (Table 3).

As far as the fixed effects estimates are concerned, we obtained very robust estimates as well, with absolute differences, with respect to those in Table 2, of the order of 0.1. The estimates like those reported in Tables 4–8 later are robust as well.

Even if Bayesian semiparametric models allow a model-based clustering without making

Table 3. Prior and posterior expected number of clusters under models A and B

Mass parameter	*E(K)*	*E(K|data)*	
		Model A	*Model B*
$a=1$	3.96	3.32	3.74
$a=2$	6.00	4.80	5.03
$a=5$	10.03	8.00	7.99
$1+\text{gamma}(0.5,2)$	4.50	3.67	4.04
$1+\text{gamma}(1,2)$	4.99	3.94	4.28
$1+\text{gamma}(1,1)$	5.80	4.24	4.58

any extra assumption, the results that are provided in this sense by such models may not be straightforward to interpret. The precise estimation of the true number of clusters is, in general, a very difficult task. As explained in Section 2, the estimated grouping is the optimal partition defined by the maximization of $l(\hat{\rho}, K)$. Two hospitals belong to the same cluster j if their labels are in the same subset of the hospital indices' partition. In model A, this is equivalent to saying that two hospitals belong to the same cluster if the observed effects are equal. In model B two different observed effects can share the same cluster because they could be generated from the two different subpopulations (i.e. in and outside Milan). Since for any choice of u and w the optimal partition can be determined, we consider different values for the pair (u, w), enabling K to range from the maximum value allowing all hospitals to be clustered together and the minimum value allowing all hospitals to be singletons. Note that low values of K penalize separation of items more than aggregation, whereas high values of K do the opposite.

Fig. 3 shows how the clustering that is induced by the optimal partition evolves as K increases, for model A (Fig. 3(a)) and model B (Fig. 3(b)). Hospitals on the abscissa are sorted to allow a more effective visualization. On the vertical axis we retain only K-values corresponding to relevant changes in hospital grouping.

As can be seen from Fig. 3, model B starts to distinguish groups for lower values of K and it reaches the setting where all items are singletons for higher values of K. In fact, when model B is fitted to data, for $K=0.21$ the best partition minimizing the expected loss is that where all the hospitals are clustered together. As K increases, some hospitals progressively exit the cluster and disaggregate, until all hospitals are singletons, which occurs for $K=0.81$. Analogous considerations hold for model A, with a smaller range from $K=0.39$ to $K=0.66$. Note that disaggregation provided by the model B case is more gradual than that provided by model A. Observing how the partition evolves as long as K increases, we obtain a sort of implicit ranking of the providers. In general, starting from low values of K (hospitals clustered together) up to the high values (hospitals all singletons), the two models give similar results: in fact, hospitals 6 and 11, and, then, 7 and 10 are in both cases among the first that are distinguished from others. In particular, in model B they are also aggregated in a different cluster for almost all K. Moreover, during the progressive splitting of the initial group, we observe similar groups appearing and disappearing in partitions that are generated by both models. Finally, hospitals 9, 15, 21, 23 and 29 are the last to become singletons, and they are grouped together in both models.

Tables 4 and 5 show 95% CIs of the posterior distribution of the random effects for model A and model B respectively. It can be observed that estimates concerning hospitals 11, 7, 6

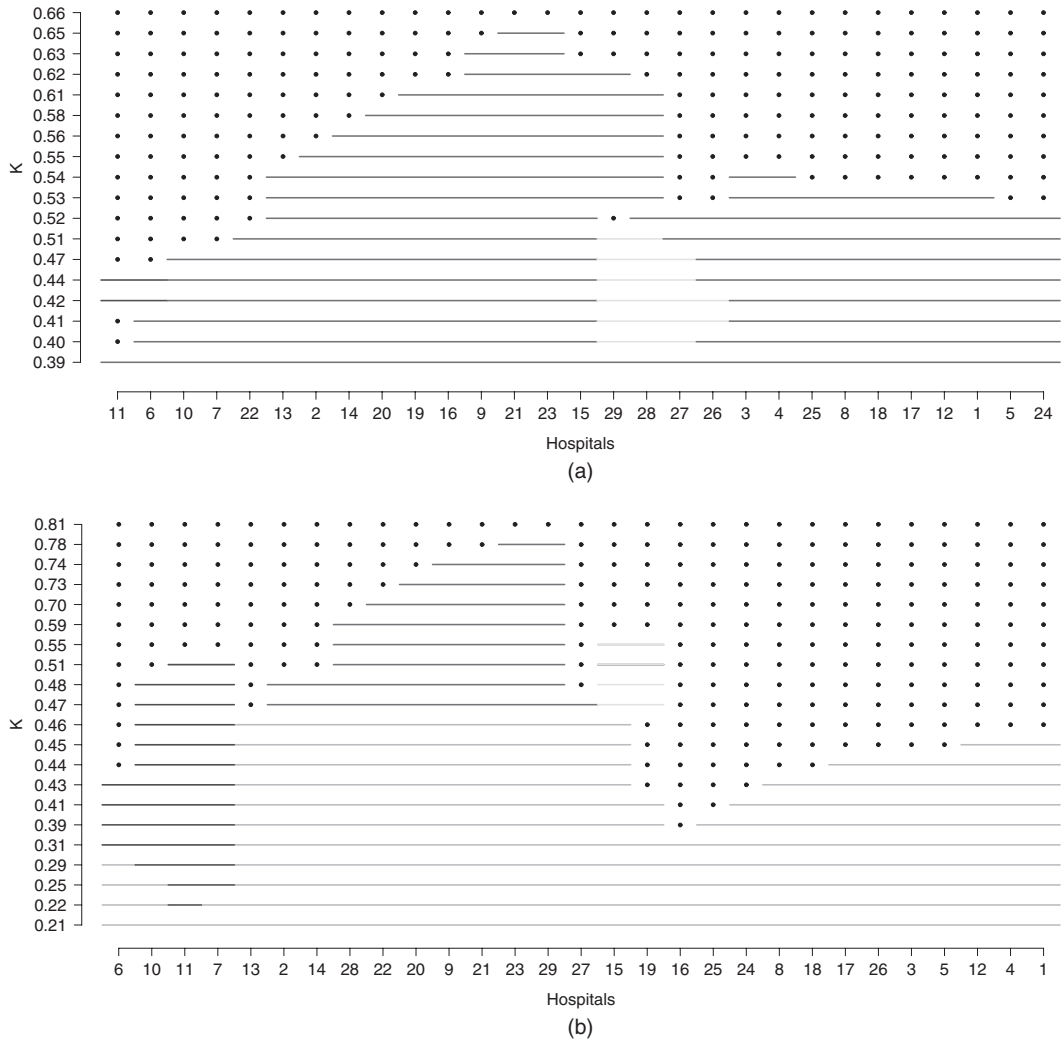

Fig. 3. Optimal partition of hospitals as K increases for (a) model A and (b) model B

and 10, which were highlighted as similar and detected early as different from all the others by both models, are concentrated on higher values than the others. Moreover estimates concerning hospitals 9, 15, 21, 23 and 29, which were grouped together by both models for almost all K-values, are concentrated on lower values than the others. In conclusion, the first items that are discarded by the initial group are those with the most favourable contribution to the patient's survival and the last are those with the less favourable contribution to the patient's survival; for this reason we could say that the 'evolving partition' is yielding a ranking between hospitals.

Thus, as long as values of K are far from 0.5 (i.e. (u, w) far from (1,1)), partitions tend to point out outliers with respect to a 'reference' group, in the sense of Shotwell and Slate (2011). The discriminating power is determined by K, which is problem driven. We conclude that model B is better at distinguishing different cases; this is probably due to its greater flexibility.

Table 4. Posterior 95% CIs of the random effects of model A

Hospital	*2.5%*	*Median*	*97.5%*	*Hospital*	*2.5%*	*Median*	*97.5%*
11	−1.53	0.49	6.71	29	−2.86	−0.19	1.90
6	−2.18	0.31	5.84	28	−2.96	−0.29	1.65
10	−2.29	0.02	4.65	27	−2.88	−0.06	3.62
7	−2.05	0.14	5.29	26	−2.79	−0.05	3.88
22	−3.94	−0.48	1.38	3	−2.68	−0.04	3.70
13	−2.85	−0.06	3.52	4	−2.81	−0.06	3.31
2	−2.68	−0.04	3.62	25	−2.68	−0.04	3.78
14	−2.79	−0.06	2.41	8	−2.70	−0.04	3.60
20	−2.86	−0.09	2.11	18	−2.82	−0.06	3.52
19	−3.03	−0.38	1.52	17	−2.71	−0.04	3.70
16	−2.90	−0.27	1.70	12	−2.72	−0.07	3.77
9	−2.89	−0.20	1.88	1	−2.79	−0.04	3.41
21	−2.90	−0.21	1.95	5	−2.88	−0.07	3.61
23	−2.86	−0.17	1.96	24	−2.70	−0.03	3.89
15	−2.96	−0.26	1.75				

Table 5. Posterior 95% CIs of the random effects of model B

Hospital	*2.5%*	*Median*	*97.5%*	*Hospital*	*2.5%*	*Median*	*97.5%*
6	−2.80	1.80	9.02	15	−4.34	−1.54	0.83
10	−2.48	2.28	9.37	19	−4.56	−1.89	0.21
11	−1.29	3.74	10.13	16	−2.82	−0.33	1.55
7	−2.10	2.99	9.68	25	−3.33	−0.07	8.19
13	−4.18	−0.83	2.99	24	−3.41	−0.47	8.23
2	−3.98	−0.73	3.12	8	−3.57	−0.64	7.89
14	−3.97	−0.95	1.65	18	−3.48	−0.60	7.86
28	−3.63	−1.34	2.01	17	−3.67	−0.61	3.24
22	−4.11	−1.60	0.69	26	−3.52	−0.66	7.81
20	−3.70	−1.42	0.97	3	−3.86	−0.68	3.26
9	−3.76	−1.54	0.54	5	−3.60	−0.71	7.79
21	−3.76	−1.53	0.59	12	−3.51	−0.65	7.86
23	−3.71	−1.53	0.50	4	−3.71	−0.70	7.78
29	−3.72	−1.50	0.63	1	−4.09	−0.74	3.27
27	−3.56	−0.72	8.01				

3.3. Model fit and patients' classification

In this section we estimate the variability that is explained by our models by using the Bayesian R^2 that is defined in equation (11) and evaluate their performance by predicting the in-hospital survival probability for each patient. In particular, we compare two different predictive methods: the usual method based on point estimates summarizing the posterior predictive distributions, and the new method that we proposed, based on interval estimates.

The Bayesian R^2 of the two models is provided in Table 6. Observe that model B seems to fit the data better (higher value of the Bayesian R^2), as we expected according to its greater flexibility.

In our application, since the share of outcome success in the data set is particularly unbalanced, if we consider the standard threshold equal to 0.5, we would obtain a very low overall misclassification rate (around 2% for all models), but a bad result in the prediction of the rare

Table 6. Bayesian R^2 defined in expression (11) for model A and model B

Model	R^2
A	0.35
B	0.57

Table 7. Predictive tables of survival outcome when the classification rule is based on the comparison between survival posterior means and $\bar{p} = 0.97$

$\hat{Y}$	$Y=1$	$Y=0$
(a) Model A		
1	599	3
0	75	20
(b) Model B		
1	596	3
0	78	20

Table 8. Predictive tables of survival outcome when the classification rule is based on survival posterior 90% CIs and threshold equal to 0.5

	$Y=1$	$Y=0$
(a) Model A		
$\hat{Y}=1$	661	8
$\hat{Y}=0$	0	3
UC	13	12
(b) Model B		
$\hat{Y}=1$	661	8
$\hat{Y}=0$	0	2
UC	13	13

outcome (death). In this case, more than 50% of deaths were misclassified. For this reason, it is important to keep the death misclassification rate as low as possible. A first attempt at improving the ability of the model in predicting deaths is based on adopting a threshold equal to the empirical rate of success, as suggested in Cramer (1999). Table 7 displays the results of the patient classification under model A (part (a)) and model B (part(b)), using a threshold equal to the sample survival rate ($\bar{p} = 0.97$). The posterior predicted rates of survival and death are more balanced than using a threshold of 0.5. However, we obtain a worse overall misclassification rate

(around 10% for all models). This is because the overall misclassification rate is less dependent on the imbalance, as explained in Cramer (1999).

As far as our new classification rule is concerned, in Table 8 we report 90% posterior predictive CIs and assume equal misclassification costs, i.e. the threshold is set equal to 0.5. With our data set, only around 4% of the patients belong to the UC and the total misclassification rate, based only on classified patients, is less than 3% for both models. Considering the number of patients in the UC as an index of the predictive performance of the model, the two models provide similar results. Of course there is a trade off between the length of the UC and the misclassification rate, whose setting is problem specific.

Furthermore, the number of patients who were classified in the UC depends on the lengths of the CIs of the posterior predictive distributions, which in turn are sensitive to the prior variances of the fixed effects. Therefore we suggest fixing prior components for the fixed effects informatively, using previous data and/or expert opinions. Moreover, the Bayesian R^2 can be also used '*a priori*' to verify whether the prior specification reflects what we expect in terms of explained variability.

In Fig. 4 we provide the 90% posterior predictive CIs for all patients under model B (the corresponding plot of model A is quite similar and we do not report it here). Note that most of the interval lengths of the patients who survived are quite small, whereas there is more uncertainty on the negative outcomes, as expected, since the unsuccessful outcome is rare. As an example, in Fig. 5 we focus on a smaller set of patients (those 29 who were treated in hospital 19, under model B). Note that predictive distributions with very large and very low mean have small width, whereas those with mean around 0.5 have wider interval estimates. There are five unclassified patients and only one is misclassified.

Finally, even if we fit a different model including the exposure random effect through a DDP (as mentioned in Section 2.2), the posterior inference does not differ from those reported here. Nevertheless, a comparison of the exposure parameter CIs shows that including the exposure non-parametrically through a DDP leads to more variability between hospitals than we observed by fitting model A.

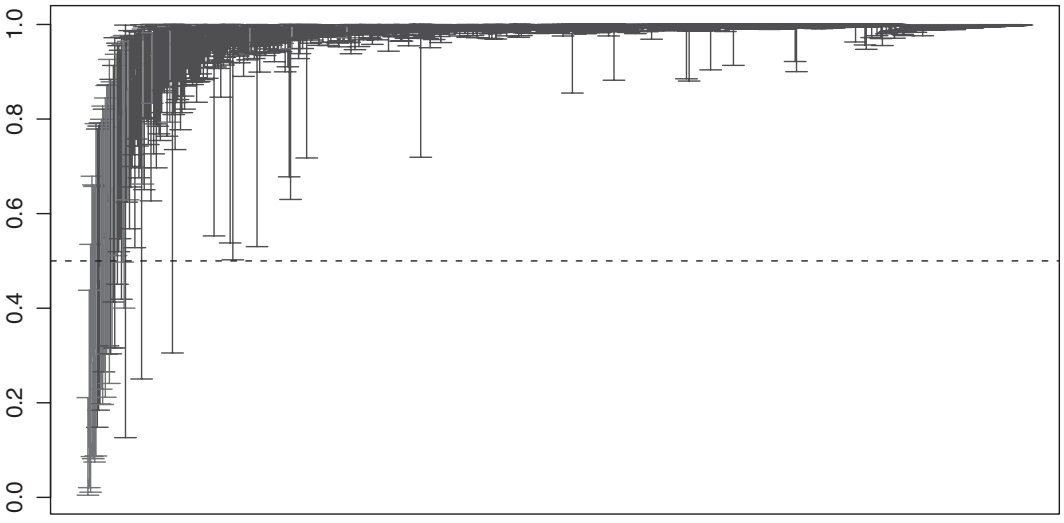

Fig. 4. 90% posterior predictive CIs of all the patients (ordered by increasing median) under model A: I, positive outcomes; I, negative outcomes

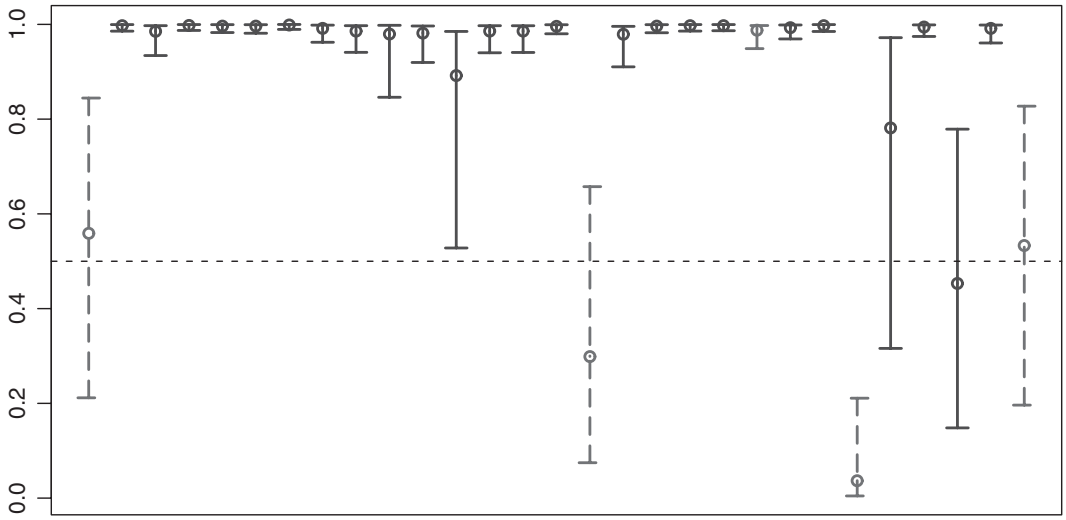

Fig. 5. 90% posterior predictive CIs of all the patients from one of the hospitals belonging to the STEMI archive, obtained by fitting model B to the data (there are five unclassified patients and only one was misclassified): I, alive patients; I, dead patients

4. Conclusions

In this work, two different Bayesian semiparametric logit models are fitted to grouped data related to the in-hospital survival outcome of patients hospitalized with STEMI diagnosis. DDP priors are considered for modelling the random-effect distribution of the grouping factor (the hospital of admission), with the aim of studying their clustering through the optimal partition minimizing a posterior pairwise coincidence loss function.

We fitted two models to the data, matching the marginal distributions of corresponding random effects, and compared them in terms of the Bayesian R^2-index that was proposed in Gelman and Pardoe (2006). Then we studied the evolution of the estimated partition as the proportion K of incorrect clustering cost increases. A sort of implicit ranking between hospitals or groups of hospitals can be sustained, since low values of K identify better performing hospitals in terms of influence on patient's survival, whereas high values of K retain worse performing hospitals. Random partitions provide a powerful tool to investigate latent grouping structure among random effects in grouped data, without making any further assumption. Finally, we developed a classification rule for patients' survival (a strongly unbalanced outcome in our application) based on the posterior CIs instead of pointwise estimates. This rule introduces the UC, which collects patients whose CIs include the reference threshold that is adopted for classification. However, it proved to be less sensitive to the choice of the threshold than classification criteria based on pointwise estimates.

Further developments of this work will be focused on taking advantage of physicians' expertise in prior elicitation. Moreover, it would be of interest to develop a dynamic update of DDP priors, generalizing frameworks such those proposed by Dunson *et al.* (2012) and Lin *et al.* (2010). Finally, methods aimed at monitoring the evolution of the clusters over time, trying to identify the causes of the changes, are of great interest for a proper monitoring of hospital performance, since only a structured and systematic monitoring of the care delivery process may lead to an improved healthcare process. We think that the methods that were adopted in this

paper properly and effectively tackle the problem of supporting decision makers in assessing hospitals' performance, enhancing interactions between physicians and statisticians.

Acknowledgements

This work is within the strategic programme 'Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction'. The authors thank Regione Lombardia—Healthcare Division for having funded and sustained the project, Lombardia Informatica SpA for having provided the data and all the physicians who contributed to the STEMI archive planning and data collection.

Appendix A

Section 2.1 shows that both models can be interpreted as generalized linear mixed models where the random-effect distribution is a DP (depending on covariates). In particular, in both models, the conditional distribution of the data is the natural exponential family

$$p(y_{ij}|\theta_{ij}) = \exp[y_{ij}\theta_{ij} - \log\{1+\exp(\theta_{ij})\}], \qquad \theta_{ij} = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_j. \tag{12}$$

In this appendix, to simplify the notation, β represents the vector of all fixed effects, whereas $\mathbf{b}_j$ is the random-effect parameters for the jth hospital as in the rest of the paper.

In particular, when describing model A through expression (12), β has $p=9$ components, and $\mathbf{b}_j$ is scalar and represents $b_{0j} + b_{1j}z_j$ in model (1), for fixed z_j. The value $\tilde{\mathbf{z}}_{ij}$ in expression (12) is 1, whereas $\tilde{\mathbf{x}}_{ij} = (\mathbf{u}_{ij}, \mathbf{x}_{ij})$, where $\mathbf{u}_{ij}$ and $\mathbf{x}_{ij}$ are as in Section 2.1. The prior for β is $\mathcal{N}(\mu_0, \Sigma_0)$, where $\mu_0 = (\mu_\alpha, \mu_\beta)$ and Σ_0 is a diagonal matrix with the first four values equal to σ_α^2 and the last five equal to σ_β^2 (see expression (2)). Finally, the prior for $\mathbf{b}_1, \ldots, \mathbf{b}_J$ is

$$\mathbf{b}_j \overset{\mathrm{IID}}{\sim} \mathrm{DP}(a, P_0), \qquad P_0 = \mathcal{N}(0, \sigma_0^2 + \sigma_1^2 z_j^2).$$

Therefore, the corresponding full conditionals in the Gibbs sampler are those described in Kleinman and Ibrahim (1998), section 4. In particular, in this setting we have

$$p(\beta|\mathbf{b}_1, \ldots, \mathbf{b}_J, \mathbf{y}) \propto \exp\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_j) - \log\{1+\exp(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_j)\} - \frac{1}{2}(\beta-\mu_0)^{\mathrm{T}}\Sigma_0^{-1}(\beta-\mu_0)\right]$$

$$\begin{aligned} p(\mathbf{b}_j|\beta, b_{-j}, \mathbf{y}) \propto & \sum_{k=1, k\neq j}^{J} \exp\left[\sum_{i=1}^{n_j} y_{ij}(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_k) - \log\{1+\exp(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_k)\}\right]\delta_{\mathbf{b}_k} \\ & + a\left(\int \exp\left[\sum_{i=1}^{n_j} y_{ij}(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_j) - \log\{1+\exp(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_j)\}\right] f_0(\mathbf{b}_j)\,\mathrm{d}\mathbf{b}_j\right) \\ & \times f_0(\mathbf{b}_j)\prod_{i=1}^{n_j}\exp[y_{ij}(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_j) - \log\{1+\exp(\tilde{\mathbf{x}}_{ij}^{\mathrm{T}}\beta + \tilde{\mathbf{z}}_{ij}^{\mathrm{T}}\mathbf{b}_j)\}], \end{aligned}$$

where f_0 is the density of P_0.

As far as model B's description is concerned, β has $p=7$ components, and the vector $\tilde{\mathbf{x}}_{ij}$ can be easily recovered from expression (6). However, here $\mathbf{b}_j = (b_{0j}, b_{1j})$ and $\tilde{\mathbf{z}}_{ij} = (0, 1)$ if hospital j is in Milan, and is (1,0) otherwise. From an analytical point of view, the prior for $(\beta, \mathbf{b}_1, \ldots, \mathbf{b}_J)$ remains unchanged, since β is still Gaussian distributed (it is straightforward to derive the mean and covariance matrix), and here P_0 is $\mathcal{N}(0, \sigma_0^2) \times \mathcal{N}(\mu_1, \sigma_1^2)$. Hence, the full conditionals have the same analytic expressions displayed above. The inferences were computed by using JAGS; the code for both models is available from the authors on request.

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Ash, A. S., Fienberg, S. E., Louis, T. A., Normand, S. T., Stukel, T. A. and Utts, J. (2012) Statistical issues in assessing hospital performance. *Report*. Committee of Presidents of Statistical Societies, Yale University, New Haven.

Barrientos, A. F., Jara, A. and Quintana, F. A. (2012) On the support of MacEachern's Dependent Dirichlet Process and extensions. *Baysn Anal.*, **7**, 277–310.

Berkelaar, M., Eikland, K. and Notebaert, P. (2004) Open source (mixed-integer) linear programming system, version 5.1.0.0. Eindhoven University of Technology, Eindhoven. (Available from `http://lpsolve.sourceforge.net/`.)

Binder, D. A. (1978) Bayesian cluster analysis. *Biometrika*, **65**, 31–38.

Binder, D. A. (1981) Approximations to Bayesian clustering rule. *Biometrika*, **68**, 275–285.

Bradley, E. H., Herrin, J., Wang, Y., Barton, B. A., Webster, T. R., Mattera, J. A., Roumanis, S. A., Curtis, J. P., Nallamothu, B. K., Magid, D. J., McNamara, R. L., Parkosewich, J., Loeb, J. M. and Krumholz, H. M. (2006) Strategies for reducing the door-to-balloon time in acute myocardial infarction. *New Engl. J. Med.*, **355**, 2308–2320.

Cannon, C. P., Gibson, C. M., Lambrew, C. T., Shoultz, D. A., Levy, D., French, W. J., Gore, J. M., Weaver, W. D., Rogers, W. J. and Tiefenbrunn, A. J. (2000) Relationship of symptom-onset-to-balloon time and door-to-balloon time with mortality in patients undergoing angioplasty for acute myocardial infarction. *J. Am. Med. Ass.*, **273**, 2941–2947.

Cramer, J. S. (1999) Predictive performance of the binary logit model in unbalanced samples. *Statistician*, **48**, 85–94.

Dahl, D. B. (2006) Model-based clustering for expression data via Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* (eds K.-A. Do, P. Müller and M. Vannucci), ch. 10. Cambridge: Cambridge University Press.

De la Cruz-Mesía, R., Quintana, F. A. and Müller, P. (2007) Semiparametric Bayesian classification with longitudinal markers. *Appl. Statist.*, **56**, 119–137.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An ANOVA model for dependent random measures. *J. Am. Statist. Ass.*, **99**, 205–215.

Direzione Generale Sanità—Regione Lombardia (2009) Determinazioni in merito alla Rete per il trattamento dei pazienti con Infarto Miocardico con tratto ST elevato (STEMI). *Decreto 10446*. Direzione Generale Sanità—Regione Lombardia, Milan.

Dunson, D. B., Ren, L. and Carin, L. (2012) The dynamic hierarchical Dirichlet process. In *Proc. 25th Int. Conf. Machine Learning*, pp. 824–831.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.

Freeman, E. A. and Moisen, G. G. (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecol. Modllng*, **217**, 48–58.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Baysn Anal.*, **1**, 515–533.

Gelman, A. and Pardoe, I. (2006) Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, **48**, 241–251.

Gordon, A. D. (1999) *Classification*, 2nd edn. New York: Chapman and Hall.

Green, P. J. and Richardson, S. (2001) Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Statist.*, **28**, 355–375.

Grieco, N., Ieva, F. and Paganoni, A. M. (2012) Performance assessment using mixed effects models: a case study on coronary patient care. *IMA J. Mangmnt Math.*, **23**, 117–131.

Guglielmi, A., Ieva, F., Paganoni, A. M. and Ruggeri, F. (2012) A Bayesian random effects model for survival probabilities after Acute Myocardial Infarction. *Chil. J. Statist.*, **3**, 1–15.

Heard, N. A., Holmes, C. C. and Stephens, D. A. (2006) A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. *J. Am. Statist. Ass.*, **101**, 18–29.

Ieva, F. (2012) Statistical methods for classification in cardiovascular healthcare. *PhD Thesis*. Politecnico di Milano, Milano. (Available from `http://hdl.handle.net/10589/56803`.)

Ieva, F. (2013) Designing and mining a multicenter observational clinical registry concerning patients with Acute Coronary Syndromes. In *Identification and Development of New Diagnostic, Therapeutic and Organizational Strategies for Patients with Acute Coronary Syndromes* (eds N. Grieco, A. M. Paganoni and M. Marzegalli). New York: Springer.

Ieva, F. and Paganoni, A. M. (2011) Process indicators for assessing quality of hospital care: a case study on STEMI patients. *JP J. Biostatist.*, **6**, 53–75.

Ishwaran, H. and Zarepour, M. (2000) Exact and approximate sum representations for the Dirichlet process. *Can. J. Statist.*, **30**, 269–283.

Kleinman, K. P. and Ibrahim, J. G. (1998) A semi-parametric Bayesian approach to generalized linear mixed models. *Statist. Med.*, **17**, 2579–2596.

Lau, J. W. and Green, P. J. (2007) Bayesian model-based clustering procedures. *J. Computnl Graph. Statist.*, **16**, 526–558.

Lin, D., Grimson, E. and Fisher, J. (2010) Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Proceeding Systems 23* (eds J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zerel and A. Culotta).

MacEachern, S. N. (1999) Dependent nonparametric processes. *Proc. Baysn Statist. Sci. Sect. Am. Statist. Ass*.

MacEachern, S. N. (2000) Dependent Dirichlet Processes. *Technical Report*. Department of Statistics, Ohio State University, Cleveland.

Müller, P. and Quintana, F. A. (2004) Nonparametric Bayesian data analysis. *Statist. Sci.*, **19**, 95–110.

Plummer, M. (2003) JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In *Proc. 3rd Int. Wrkshp Distributed Statistical Computing* (eds K. Hornik, F. Leisch and A. Zeiles), pp. 20–22. Vienna.

Quintana, F. A. and Iglesias, P. L. (2003) Bayesian clustering and product partition models. *J. R. Statist. Soc.* B, **65**, 557–574.

Ray, S. and Mallick, B. (2006) Functional clustering by Bayesian wavelet methods. *J. R. Statist. Soc.* B, **68**, 305–332.

R Development Core Team (2009) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rodriguez, A. and Dunson, D. B. (2011) Nonparametric Bayesian models through probit stick-breaking processes. *Baysn Anal.*, **6**, 145–178.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, **4**, 639–650.

Shotwell, M. and Slate, E. H. (2011) Bayesian outlier detection with Dirichlet process mixtures. *Baysn Anal.*, **6**, 1–22.

Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C. and Grigg, O. (2012) Statistical methods for healthcare regulation: rating, screening and surveillance (with discussion). *J. R. Statist. Soc.* A, **175**, 1–47.