
MOX–Report No. 21/2012

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

Guglielmi, A.; Ieva, F.; Paganoni, A.M.; Ruggeri, F.; Soriano, J.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

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

Alessandra Guglielmi[a], Francesca Ieva[a], Anna Maria Paganoni[a], Fabrizio Ruggeri[b] and Jacopo Soriano[c]

April 30, 2012

[a] Politecnico di Milano
alessandra.guglielmi@polimi.it, francesca.ieva@mail.polimi.it, anna.paganoni@polimi.it
[b] IMATI-CNR, Milano
fabrizio@mi.imati.cnr.it
[c] Duke University, NC
jacopo.soriano@duke.edu

**Keywords**: Bayesian clustering; Bayesian nonparametrics; random-effects models; Unbalanced binary outcomes; Random partitions.

**AMS Classification**: 62F15, 62P10, 62J12.

**Abstract**

In this work, Bayesian semiparametric logit models are fitted to grouped data related to in-hospital survival outcome of patients hospitalised with 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), in order to provide a cluster analysis of the hospitals. The clustering structure is highlighted through the optimal random partition that minimises the posterior expected value of a suitable loss function. Two are the main goals of the work: to provide model-based clustering and ranking of the providers according to the similarity of their effect on patients' outcome, and to make reliable predictions on the survival outcome at patient's level, even when the survival rate itself is strongly unbalanced. The study is within a project, named Strategic Program of Regione Lombardia, and is aimed at supporting decisions in healthcare policies.

# 1 Introduction

Bayesian nonparametrics provides extremely flexible models for fitting a variety of datasets. One of its most popular use is in modelling distributions for random effects in hierarchical models for grouped data, as in the seminal paper [24]. With such grouped data, the aim is usually to find clusters among groups which are able to capture the latent structure in the data assigned to each group. In this context, a natural way to achieve model-based clustering via Bayesian nonparametrics 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 parameters values coincide. Dirichlet Processes (DPs), introduced by [15], are the most popular discrete random probability measures, 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 [13] and [20] among others). Models incorporating DP priors play an important role in Bayesian applied statistics, spanning a wide range of applications, i.e., density estimation, nonparametric regression, survival analysis, as recalled in [33].

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 [3], [31] and [32]. Applications or extensions of such priors include covariate-DDPs resembling traditional ANOVA models [12], DDP with an additional probability model for group classification for longitudinal data [13], and probit stick-breaking random probability measures [38]. 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 hospitalised because affected by STEMI) are grouped by hospital of admission. In particular, in such hierarchical framework we adopt nonparametric DDP priors for modelling random effect superimposed on the grouping factor, in order to provide a proper methodological approach to the problem of assessing hospitals performances and to profile hospitals according to their effects on patient's outcome. This topic is crucial within the context of healthcare planning, and proper methods for addressing such a problem are extremely of interest for people in charge of healthcare government (see [2] and [41] for details on recent discussions and developments). Since the outcome of interest (in-hospital survival, i.e., if a patient is discharged alive from hospital) is strongly unbalanced within the context of the disease we focus on, any model will perform poorly in predicting it. Therefore we propose a new method for classifying patients according to the whole predictive distributions of their outcome, based on the posterior predictive

credibility intervals.

We adopt a Bayesian semiparametric approach since it has a twofold advantage. First of all, as mentioned before, Bayesian semiparametric models allow for a great flexibility in modelling data, avoiding critical dependence on parametric assumptions. Moreover, they robustify parametric models and define model diagnostics and sensitivity analysis for parametric models by embedding them in a larger encompassing semiparametric model (see [33]). Secondly, Bayesian nonparametric priors selecting discrete probability measures yield a "natural" clustering of the groups (hospitals in our application), according to the grouping provided by the random-effects parameters sampled from the discrete random probability measure. In this way, the nonparametric prior component leads to a random partition of the group indices set; cluster estimates will be based on the posterior distribution of the random partition itself. A common way for estimating the unobserved true random partition is to observe the Maximum A Posteriori (MAP), i.e., the mode of the posterior distribution. However, since the number of partitions is large even for moderate sizes of the indices set, 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 are discussed in [29] and [35].

One of the main focuses of the paper is to exploit model-based clustering of groups provided by the semiparametric Bayesian models considered in Section 2. We will pursue this issue providing a Bayesian estimate, as proposed in [29], looking for a posteriori clustering structure, optimal with respect to a specified loss function. Specifically we focus on a loss function based on pairwise coincidences, that is, whether pairs of items are clustered in the same group or not, as in [5] and [6]. In this case, [29] shows that the problem of estimating the optimal random partition can be formulated as a binary integer programming problem. On the other hand, 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, based on the posterior credibility intervals of patients' survival probability, instead of point estimates, discussing how the classification obtained in such way depends on the choice of a reference threshold, according to what was suggested in [9]. A discussion on performances of threshold criteria for binary classification based on pointwise outcome estimates is presented in [16].

Finally, as we mentioned before, we apply these methods to a dataset arising from a clinical registry (the STEMI Archive, see [11] and [25]) on patients affected by STEMI and admitted to any hospital of Regione Lombardia, a northern Italian region whose capital is Milano. Specifically, the binary outcome of interest is measured at patients' level, and patients are grouped according to the hospital of admission. Then, there is a hierarchical structure in the dataset: providers (i.e., hospitals) at a higher level and the patients at a lower one. Bayesian generalised linear mixed models provide a natural framework for such

data. It is known from the literature (see [8], among others) that STEMI is characterized by a strongly unbalanced share of success in terms of in-hospital survival; in our dataset, in fact, 97% of patients are discharged alive from the hospital. It is also known (see [7] and [26] for instance) that, for such disease, reducing treatment times and optimizing pre- and intra-hospital patterns of care strongly improve patients' prognosis. Among the number of variables available in the STEMI Archive at patient's level, in the models considered in Section 2 we introduced the following covariates: age, total ischemic time (Symptom Onset to Balloon time, denoted by $OB$), presence of Chronic Kidney Disease ($CKD$ - equal to 1 if the patient had loss in renal function, 0 otherwise) and Killip class (an ordinal variable indicating the severity of infarction, from 1 - lowest severity to 4 - highest severity). Moreover, since hospital-dependent covariates are also present in the registry, we included in the models hospital exposure, i.e., the number of patients treated with primary angioplasty per year, and a binary variable ($Milano$) indicating if the hospital is in ($Milano = 1$) or outside ($Milano = 0$) the city of Milano. In particular, we are interested in profiling healthcare providers, investigating whether any clustering of the hospitals where patients are admitted has a meaning. Since clustering is obtained through estimates of the posterior distribution of the random partition of the hospital index set, we will be able to assess the effect of groups of healthcare providers with "similar" behaviour on patients' outcome, as well as to evaluate the quality of their performances in treating STEMI patients, adjusting for case-mix and all other known sources of variability that induce overdispersion in the outcomes distribution.

We will 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 we consider, the random-effect parameters are given a nonparametric prior, similarly to [24], while lower level covariates are treated parametrically. Specifically, the random-effect parameters are assumed as a sample from a Dirichlet Process (DP), in order to exploit the discreteness of its trajectories to carry out a cluster analysis. In our case, since a random effect is superimposed on the grouping factor represented by the hospital of admission of patients, we model the dependence across random distributions through the hospitals' covariates, so that priors can be interpreted as DDP densities. The two Bayesian models differ for the choice of covariates included in the likelihood and for the nonparametric components of the random-effect parameters (see Section 2).

The novelty of this work consists of exploiting a model-based clustering, provided by the optimal partition of the random effects estimated through a Bayesian semiparametric hierarchical model, for carrying out providers profiling in a real clinical problem, i.e., the hospitals' performances evaluation aimed at cardiovascular healthcare planning. In fact, the method we propose in this paper yelds 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 organised as follows. In Section 2 we present the models and the methodology developped for hospital clustering and patients classification. Goodness-of-fit indices for comparing the models are also considered, and details on random-effects clustering 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 and comments are given in Section 4. All the analyses have been carried out with `R` [36] and `JAGS` [34].

# 2 Bayesian semiparametric models for random effects clustering

In this section, we present the two models we will use to analyse the data in Section 3, discussing different alternatives in terms of likelihood and priors on random effects. In what follows, the model formulation is already intended for the application of interest, where the outcome is the in-hospital survival after a STEMI event, and patients are grouped by hospital of admission. We will also explain why hyperparameters tuning was made in order to match the marginal effect of random components. Moreover, details on some goodness-of-fit tools will be given. Then the mathematical framework of loss functions for the evaluation of the optimal partition is briefly described, in order 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 we are considering, the random-effect parameters $\mathbf{b}_j$s, corresponding to the $j$-th 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 [32]). 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 rather elaborated forms, but here we analyse only two such specifications, which retain interpretability of model parameters.

## 2.1 DDP 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 $p_{ij}$. To our aims, $p_{ij}$ represents the probability that

the patient $i$ treated in hospital $j$ is discharged alive after a STEMI event. The $p_{ij}$'s are modelled through a multivariate logistic regression with fixed effects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and a random-effect $\mathbf{b}$ superimposed on the covariates referred to the grouping factor, i.e.,

$$Y_{ij}|p_{ij} \overset{\text{ind}}{\sim} Be(p_{ij}) \tag{1}$$

$$\log(\frac{p_{ij}}{1-p_{ij}}) = \sum_{l=1}^{4} \alpha_l u_{ijl} + \sum_{k=1}^{5} \beta_k x_{ijk} + b_{0j} + b_{1j} z_j. \tag{2}$$

Within the context of the application motivating this study, $\boldsymbol{u}_{ij} = (u_{ij1}, \ldots, u_{ij4}) = (\text{Killip1}, \ldots, \text{Killip4})_{ij}$ is a vector of dummies, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ij5}) = (\text{age}, \text{logOB}, \text{CKD}, \text{exposure}, \text{Milano})_{ij}$ and $z_j$ is the exposure of the $j$-th hospital. All continuous covariates have been centred and standardised (so that their range is between $-1$ and 1) to get 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, too. In what follows, we will refer to such patient as a "standard reference", and will compare hospitals effects once adjustments for all fixed effects has been carried out in the "standard reference" setting. The prior distributions assumed for the parameters of the model are

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4) \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \sigma_\alpha^2 \mathbb{I}_4), \qquad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_5) \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \sigma_\beta^2 \mathbb{I}_5) \tag{3}$$

$$(b_{0j}, b_{1j})'|P \sim P \qquad j = 1, \ldots, J \tag{4}$$

$$P|a, P_0 \sim DP(a, P_0) \tag{5}$$

Independence among $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $P$ is assumed. By $P \sim DP(a, P_0)$ we mean that $P$, the (conditional) distribution of the bivariate random-effects parameter $\boldsymbol{b}_j$, has Dirichlet process prior with total mass parameter $a > 0$ and base probability measure parameter $P_0$; see [15] for details on definition and standard notation of DPs. The base probability measure on $\mathbb{R}^2$, $P_0$, for this model will be chosen as the product measure $N(0, \sigma_0^2) \times N(0, \sigma_1^2)$, being $\sigma_0$ and $\sigma_1$ 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 (2), the random-effect parameter of hospital $j$ appears linearly as $b_{0j} + b_{1j} z_j$. Moreover, each $\mathbf{b}_j$, given $P$, has distribution

$$P = \sum_{h=1}^{+\infty} w_h \delta_{\boldsymbol{\theta}_h} \tag{6}$$

where $\boldsymbol{\theta}_h$ are i.i.d. according to $P_0$ and $\{w_h\}$ are the weights in the stick-breaking representation (see [40]). It is straightforward to see that $b_{0j} + b_{1j} z_j$, given $\tilde{P}$ is

distributed as $\tilde{P}$, where

$$\tilde{P} = \sum_{h=1}^{+\infty} w_h \delta_{\tilde{\theta}_h(z_j)}. \tag{7}$$

Here $\tilde{\theta}_h$s are i.i.d. according to $\tilde{P}_0$ which is the distribution of $b_{0j} + b_{1j} z_j$ if $\mathbf{b}_j = (b_{0j}, b_{1j})$ is distributed according to $P_0$. Therefore, by (7), the random-effect contribution to the likelihood in (2) is distributed according to a DDP. This is a rather simple case of Dependent Dirichlet Process, called "single-p linear DDP" [32], 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 as follows:

$$Y_{ij}|p_{ij} \overset{\text{ind}}{\sim} Be(p_{ij}) \tag{8}$$

$$\log(\frac{p_{ij}}{1-p_{ij}}) = \sum_{l=1}^{4} \alpha_l u_{ijl} + \sum_{k=1}^{3} \beta_k x_{ijk} + b_{v_j j}. \tag{9}$$

being $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\mathbf{b}$ the parameters vectors corresponding to the fixed and random effects, respectively, as in the previous case. Referring to the motivating application, $\boldsymbol{u}_{ij}$ is the Killip dummy vector and $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, x_{ij3}) = (\text{age}, \log\text{OB}, \text{CKD})_{ij}$. Finally, $b_{v_j j}$ is the random intercept depending on values assumed by the location dummy *Milano* ($v_j = 0/1$). Notice 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 $j$-th hospital, which will be $b_{1j}$ if the $j$-th hospital is located in Milano, $b_{0j}$ otherwise. We assume the following priors for fixed and random effects respectively:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4) \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \sigma_\alpha^2 \mathbb{I}_4), \qquad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_3) \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \sigma_\beta^2 \mathbb{I}_3) \tag{10}$$

$$(b_{0j}, b_{1j})'|P \sim P \qquad j = 1, \ldots, J \tag{11}$$

$$P|a, P_0 \sim DP(a, P_0) \tag{12}$$

Independence among $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $P$, is assumed. For our scopes, we will 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,$\sigma_0$ and $\sigma_1$ will be assumed to be Uniformly distributed. Finally, Gaussian distribution will be considered for $\mu_1$ and truncated-exponential for $a$, respectively.

Observe that the number of random-effect parameters in (9) 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 $b_{1j}$; on the other hand, if $j$ is the index corresponding to a group with $v_j = 0$, the corresponding random-effects parameter is $b_{0j}$. This means that the marginal prior of the hospitals' effect is partially exchangeable instead of being exchangeable.

In this case the nonparametric prior component assumed for the random-effects parameters can be interpreted as an ANOVA-DDP prior with one factor

and two levels (see [12]), 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

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

where, for $v$ equal to 0 or 1,

$$P_v = \sum_{h=1}^{+\infty} w_h \delta_{\theta_{vh}}, \quad (\theta_{0h}, \theta_{1h})' \overset{\text{iid}}{\sim} P_{00} \times P_{01} \tag{13}$$

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

Observe that the main difference between the priors of the two models described in the previouos section stands for the *Milano* covariate effect, which is included directly in the locations of the stick-breaking representation in (13) in the second model.

In what follows, we will refer to the model defined by equations (1)-(5) as "Model A", and to the model defined by equations (8)-(12) as "Model B".

### 2.2 Models comparison

Since the dataset we deal with in the motivating application is complex and rich in covariates, there is a number of Bayesian models that could be fitted to the data. In particular, the covariates dependency could be included in the DDP in many different ways. We focused on likelihoods containing the most significant covariates pointed out in previous works (see [22] and [26]) by some variable selection methods, and tried different way of combining hospitals covariates within the nonparametric priors. Some covariates (both at patients 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 (2), and assumed a univariate DP prior for the random intercept. The inference we obtained from the two models was similar, but we preferred to consider the likelihood as in (2), since it allowed us to draw conclusions on the relationship between goodness of performances and hospital exposure, as reported in Section 3. On the other hand, as a second alternative, we fitted a model with a DDP prior for the vector of the random-effects parameter $\boldsymbol{b}_{v_j j} = (b_{0j}, b_{1j})$ representing the effect of the intercept and the exposure for each hospital. The posterior inference we obtained was very similar to that given by Model B, reported with details in Section 3.

In order to compare the two different models 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 $\boldsymbol{\alpha}$, of the two models for an hospital located in or outside Milano.

Table 1: Random intercept parameters in Model A (first row) and Model B (second row).

| | Hospital location | |
|---|---|---|
| | in Milano | outside Milano |
| Model A | $\beta_5 + b_{0j}$ | $b_{0j}$ |
| Model B | $b_{1j}$ | $b_{0j}$ |

As we mentioned before, since we deal with standardised covariates, the random intercepts reported in Table 1 represent the in-hospital survival probability on the logit scale for a "standard reference" patient (without the Killip effect). As we will see in Section 3, we fixed hyperparameters so that the prior marginal distributions of random intercepts of hospitals located in Milano are equal, as well as that of random intercepts of hospitals located outside Milano. Anyway, even if denoted with the same symbols, the intercepts have a different interpretation, according to the different likelihoods 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 an hospital $h$ outside Milano and an hospital $l$ in Milano,

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

whereas for the Model B

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

To evaluate model goodness-of-fit, we compute an index introduced in [18], where the authors propose 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 explained by the linear model. Here we apply it to the first level of the logistic regression, that can be rewritten in terms of latent variables formulation (see [1]) as follows:

$$Y_{ij} = \begin{cases} 1 & \text{if } Z_{ij} \geq 0 \\ 0 & \text{if } Z_{ij} < 0 \end{cases},$$

and

$$Z_{ij} = \mu_{ij} + \epsilon_{ij}. \tag{14}$$

Here $\mu_{ij}$'s are the linear predictors, as in the right handsides of (2) or (9) and $\epsilon_{ij}$'s are i.i.d. standard logistic random variables, i.e., random variables with density function $f_\epsilon(t) = \mathrm{e}^{-t}(1+\mathrm{e}^{-t})^{-2}$, mean equal to zero and variance equal to $\pi^2/3$. We assume that, conditioning on the latent variables $Z_{ij}$'s, the $Y_{ij}$'s are independent.

Starting from the latent variable representation of the model provided in (14), a Bayesian generalization of the $R^2$ index for linear models can be defined as

$$\begin{aligned} R^2 &= 1 - \frac{\mathbb{E}[\bigvee_{ij} \epsilon_{ij}]}{\mathbb{E}[\bigvee_{ij} \mu_{ij}]} \\ &= 1 - \frac{Var[\epsilon]}{\mathbb{E}[\bigvee_{ij} \mu_{ij}]}, \end{aligned} \tag{15}$$

where $\bigvee$ represents the sample variance operator

$$\bigvee_{ij} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2}{\sum_{j=1}^{J} n_j - 1}. \tag{16}$$

Bayesian $R^2$ provides an index of the explained variability at the latent variable level. It is close to 1 when $\mu_{ij}$'s approximate well the conditional mean of $Y_{ij}$'s and close to zero when the sample variance of the $\epsilon$'s is approximately equal to the variance of the $\mu_{ij}$'s. While frequentist $R^2$ ranges from zero to one, 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 aim of the work is to exploit the clustering induced by the random-effect prior in order 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 covariates effects.

In a Bayesian formulation of a clustering procedure, the partition of the 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. As already recalled, the DP prior selects discrete distributions almost surely. Since there is a positive probability of coincident values, i.i.d. sampling from $P$ induces a random partition on the positive integers, and consequently a posterior distribution for the random partition itself.

More specifically, let $X_1, \ldots, X_J$ be a sample from a Dirichlet Process $P$ on $\mathbb{R}^s$, for some positive integer $s$, i.e., $X_1, \ldots, X_J | P \overset{i.i.d.}{\sim} P$. Since $P$ is almost

surely discrete, two sampled random variables $X_i$ and $X_j$ are equal with positive probability. We say that $X_i$ and $X_j$ share the same cluster if and only if $X_i = X_j$. In this case, the set of integers $\{1, 2, \ldots, J\}$ is partitioned into a finite number of sets $\{A_1, \ldots, A_{k(J)}\}$, where $k(J)$ is the number of different determinations among $(X_1, \ldots, X_J)$ and each $A_j$ contains the labels of the random variables $\{X_1, \ldots, X_J\}$ which coincide but are different from the others. Note that since $(X_1, \ldots, X_J)$ is a random vector, the partition $\{A_1, \ldots, A_{k(J)}\}$ of $\{1, 2, \ldots, J\}$ is random as well. Specifically, $\rho = \{A_1, \ldots, A_{k(J)}\}$ is a random partition induced by the sampling from a DP or from any random probability measure which is discrete with positive probability. For both models, clusters are defined by ties in the variables $\{(b_{0j}, b_{1j}), j = 1, \ldots, J\}$ in (4) and (11), respectively. Since both prior components can be interpreted as DDP, the posterior distributions under the two models will be similar. In [24], the full conditionals (and therefore the posterior distributions) of a Generalised Linear Mixed Effects (GLME) model with a DP prior for the random-effect parameters are computed. They consist of a mixture of a discrete distribution and a continuous one. The discrete component is the empirical distribution of the other $b_j$'s and the continuous one is the baseline distribution. The weights in such mixture depend on the conditional distribution of the data, given the parameters, and therefore they will be different under the two models. In particular, note that for Model B, where the nonparametric prior component is an ANOVA - DDP, clusters of random-effect parameters occur both within the two groups (hospitals in Mialno and outside Milano) as well as across the geographical location.

Concerning the application of interest, since the discrete nature of the prior of the random-effects parameters yields a random partition on the set of hospital labels, any inference on hospital clustering, aimed at quantifying the effect of each group of hospitals on outcomes at patient's level, must be based on the posterior of the random partition $\rho$ itself. Our aim is then to compute a suitable estimate $\hat{\rho}$ of this posterior distribution, representing the best estimate of the "true" clustering of the random-effects estimates. Clinically speaking, we would like to estimate a latent clustering among hospitals of our dataset, identifying groups of providers affecting outcomes at patients' level in a similar way. This could be of great interest for decision makers, in order 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, both due to patients' covariates and hospital features.

Choosing a partition $\rho$ can be considered as a model choice problem, and different approaches to tackle it are available (see [10], [19], [23] and [37]). The most naive solution would be to choose the maximum a posteriori (MAP) partition, but it may not be a good choice if the posterior distribution of the random partition is very spread out, as it is usually the case. A loss function approach avoids some criticisms related to sparsity of random partitions, which are common also to the selection methods based on marginal likelihood and Bayes factor.

As in [29], we concentrate on loss functions that rely on *pairwise coincidences* (see [5]), penalising pairs of items that are assigned to different clusters when they should be in the same one, and vice-versa. Specifically, we choose the loss function which assigns a positive cost $u$ any time two random effects are incorrectly assigned to different clusters, and a positive cost $w$ any time two random effects are incorrectly clustered together. The function counts how many times a wrong labelling happens, assigning a different weight to the two types of misclassification. The total loss is then obtained by summing over all pairs. Denoting by $c_i$ the true allocation variable, which is $c_i = j$ if and only if $i \in A_j$, we define

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

where $\hat{\rho}$ is the estimate and $\rho$ 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 one minimising the posterior expected loss

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

where $\hat{c}_i$ is the estimated allocating variable for $i$-th unit. If we define $\gamma_{ij} = \mathbb{P}[c_i = c_j|\mathbf{Y}]$, the previous formula can be written as

$$\mathbb{E}[L(\rho, \hat{\rho})] = u \sum_{(i,j)\in\mathcal{M}} \gamma_{ij} - (u + w) \sum_{(i,j)\in\mathcal{M}} \mathbb{I}[\hat{c}_i = \hat{c}_j](\gamma_{ij} - K)$$

being $K = w/(u + w) \in [0, 1]$. Minimising the posterior expected loss is equivalent to maximising

$$l(\hat{\rho}, K) = \sum_{(i,j)\in\mathcal{M}} \mathbb{I}[\hat{c}_i = \hat{c}_j](\gamma_{ij} - K) \tag{17}$$

over all possible choices of $\hat{\rho}$ (see [29]). The right-hand side of (17), as a function of $K$, characterises the quality of each possible $\hat{\rho}$, and the whole family of such functions determines in particular for which $K$, if any, each partition is optimal, as well as defining the optimal $\hat{\rho}$ for each $K$. The approach proposed in [29], therefore, is to consider all values of $K$ simultaneously. As it will be clear in Section 3, we will observe how the clustering induced by the random partition changes as long as different values of $K$ are considered. This will lead to a sort of "implicit ranking" of the hospitals in our dataset, in the sense clarified in Section 3.

The maximisation of (17) can be carried out through binary integer programming techniques, as explained in [29]. Since the total number of hospitals is not large, the computational effort required for solving the optimisation problem can be carried out using the R package *lpSolve* [4].

## 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 hard to predict, aside from the model considered, when the dataset contains binary variables characterised 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., being $p_{ij}$ the probability to observe a successful outcome for the item $i$ in the group $j$, the outcome $Y_{ij}$ will be predicted as a success whenever $\mathbb{E}(p_{ij}|\boldsymbol{Y})$ is bigger than a given threshold. In the application setting we are interested, we consider the in-hospital survival probability $p_{ij}$ of patient $i$ admitted in hospital $j$, and we are interested in correctly classifying the patients belonging to the current dataset 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.

Quite a large number of solutions to this problem have been proposed, since the classification is typically very sensitive to the choice of the threshold (see for example [16] for a review and comparison of such most popular criteria in the frequentist literature). Anyway, in our opinion, classification rules based on pointwise estimates are not completely satisfactory. First of all, they are not robust with respect to the choice of the thresholds. Moreover, since 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 it provides. The posterior predictive distribution for a new patient $i$ in hospital $j$ can be easily simulated through MCMC 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 interval estimate of posterior success rate and it can be considered as a generalisation of the "standard" one, based on pointwise estimates and thresholds. Concerning the application of interest, we classify a patient as alive if the Credibility Interval (CI) of his/her survival rate is entirely over a given threshold, or as dead if the CI is entirely below the threshold; rather, we do not classify it if the threshold lies within the CI. Of course, the higher the credibility level is, the larger is the number of patients belonging to this latter Uncertainty Class (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 presented in the previous section. As we said before, the data we consider come from a clinical registry, named STEMI Archive, gathering patients admitted with STEMI diagnosis in any hospital of Regione Lombardia district. A complete description of the registry is provided in [25] and [27], where data are presented together with the clinical setting that motivated their collection. As mentioned in Section 1, information about both patients and hospitals are available. Among the most important patients' information provided by the clinical registry there are mode of admission (a patient reaches the hospital on his/her own or delivered by three different types of rescue units of 118, the national free-toll number for emergencies), demographic features (age, sex), clinical appearance (Killip), risk factors (diabetes, smoke, Chronic Kidney Disease (CKD), ...), 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 Section 1. In this application we focus on in-hospital survival of patients whose data are contained in the STEMI Archive. On the other hand, information about the hospital of admission - considered as the grouping factor - are also present (in particular, a dummy variable indicating if the hospital is in or outside Milano and the hospital exposure).

The variability of the distribution of patients' outcome is high between structures. The dataset contains $n = 697$ patients, admitted in $J = 29$ hospitals of Regione Lombardia. A first patient covariates' selection was done in [26] according to clinical know-how and stepwise selection procedures, based on the AIC index, confirmed later on by a Bayesian variable selection method, using Gibbs variable selection (as reported in [22]). As we said in Section 1, the most significant factors which explain in-hospital survival probabilities are age, Killip, CKD and total ischaemic time in log-scale from symptom Onset to the primary angioplasty (Balloon), i.e., *logOB*. Providers' covariates *Milano* and exposure are also included. In fact, we are interested in evaluating if differences among the hospitals may be assessed and, in this case, if such differences lead to a clustering of providers.

As far as posterior inference from the models 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 proposed in Section 2. All estimates have been carried out by a Gibbs sampler algorithm, translated into a JAGS code. In the two models we implemented the truncated DP approximation suggested by [28] to obtain a trajectory from $P$; we truncated (and normalised) the sums in (6), (7) and (13) at $H = 30$. We ran the two models for $200,000$ iterations, discarting the

first $100,000$, and using a thinning of 20 to reduce autocorrelations, so that the final sample size was $5,000$. Traceplots, autocorrelations 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 nonparametric components $\sigma_0^2$ and $\sigma_1^2$. Concerning the former, we fixed them "informatively" as the means of the posterior distributions obtained fitting a parametric model with the same covariates and Gaussian-distributed errors on data arising from a previous data collection of the same registry. This enabled us to set informative values for the means of fixed effects $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, as well as for their variances $\sigma_\alpha^2$ and $\sigma_\beta^2$. On the other hand, concerning the random-effect variance components we tested two classes of priors: the conjugate inverse-gamma distribution on the variances and the uniform distribution on the standard deviations. The estimates of the random-effects are particularly sensitive to the choice of the inverse-gamma hyperparameters, while they are more robust using the uniform prior. We refer to [17] for a discussion on priors of the variance components in hierarchical models. Finally, the lower bound of the support of the prior distribution for the total mass parameter was set equal to 1 to avoid computational problems. This choice does not affect too much the total number of clusters a priori. Finally, we tested an exchangeable prior for the Killip vector $(\alpha_1, \ldots, \alpha_4)$, instead of assuming them i.i.d.. The estimation is robust to these choices, but the mixing is better under the independence assumption.

Figure 1 shows the survival posterior predictive distributions for a patient who was discharged alive (left panel) and who died (right panel), respectively for Model A (solid line) and Model B (dashed line).
Note that the two posterior predictive distributions in both panels do not differ too much, but they do differ from the corresponding prior predictive distributions (not displayed here in order to make the graphs clearer). Concerning the patient who was discharged alive (left panel of Figure 1), he is a man, aged 66, 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 minutes. On the other hand, the dead patient (right panel of Figure 1) was a man, aged 59, 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 the one indicted by guidelines. Both patients have been admitted to hospitals located in Milan, although not the same.

### 3.1 Fixed and random-effects estimates

In Table 2 we provide posterior 95% credibility intervals (CIs) of the fixed effects under the Model A and Model B. Hyperparameters in (3) were set informatively as we mentioned before, and, as a consequence, $\boldsymbol{\mu}_\alpha = (4.2, 4.2, 4.2, 4.2)'$, $\boldsymbol{\mu}_\beta =$

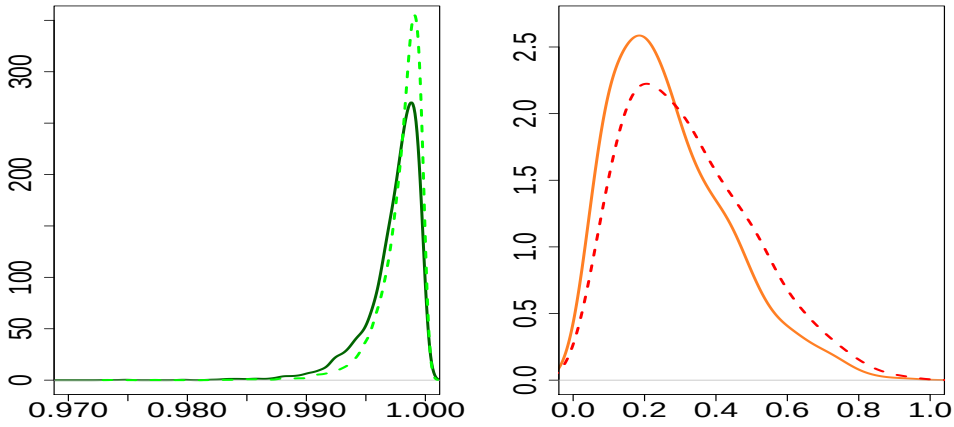



Figure 1: Posterior predictive distribution, respectively for Model A (solid line) and Model B (dashed line), of survival probability for two patients: one discharged alive (left panel) and one who died (right panel).

$(-1.7, -0.45, -1.7, 0.07, -0.45)'$, $\sigma^2_\alpha = 4$ and $\sigma^2_\beta = 4$. For Model B, the same values are adopted, selecting only the fixed effects of interest for (10).

Table 2: Posterior 95% CIs of the fixed effects.

| | Model A | | | Model B | | |
|---|---|---|---|---|---|---|
| Parameter | 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 | | | |

Notice that the estimates are similar. In particular, the Killip 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, while those of Killip 4 concentrate on "small" values. As we could expect, as long as age, logOB and CKD increase, the survival probabilities decrease. Finally, the binary covariate *Milano* has a negative effect in Model A, while the exposure is not significant. For this reason we decided to omit the exposure from Model B, but we used

*Milano* covariate to enrich the hospital random intercepts prior distribution. Results about exposure and location influence have been deeply investigated by decision makers and physicians. The exposure being not significant means that there is no evidence from data for concluding that hospitals that treat more patients are necessarily the best ones in terms of performances, countrary to what people in charge of healthcare government sustained. On the other hand, it seems that being treated in Milano results in a worse outcome, that is pretty unexpected. We asked epidemiologists if their data would confirm this finding, and they verified that, according to the evidence of our results, the epidemiology seems to be different between Milano and neighbourhoods, expecially for elder people over 80s.

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

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

being $\pi(\mu_1)$ the Normal distribution $\mathcal{N}(-045, 4)$, and outside Milano

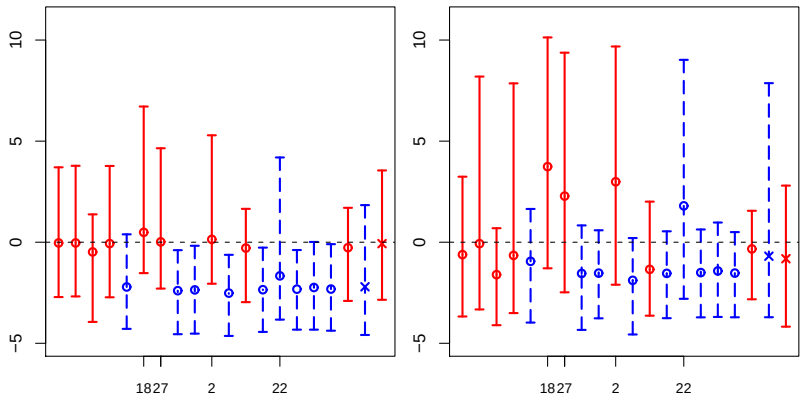
$$\int \mathcal{N}(0, \sigma^2)\mathbb{I}_{[0,5]}(\sigma)d\sigma,$$

respectively.

In Figure 2 we provide posterior 95% CIs of the hospital random intercepts with at least ten patients, highlighting the *Milano* effect, for the two models.
The plots of 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 (left panel) all the hospitals outside *Milano* have a higher median than *Milano* ones, and intervals are shorter. Model B , on the other hand, gives higher variability within each subpopulations. This variability is reasonably due to the greater flexibility of the prior of the second model.

### 3.2 Hospital Clustering

As mentioned in Section 2.3, the nonparametric prior component induces a random partition of the hospitals labels. Therefore we analyse the posterior of the process $P$ to obtain an insight on the clustering among the hospitals. In [21], investigating the clustering structure of the random-effect estimates arising from different frequentist techniques implemented on a similar database, we pointed out that few groups could be detected among hospitals. The same conclusion holds under a parametric Bayesian mixed effects model (see [22] for details). We tuned hyperparameters of the prior for the total mass $a$ in our models according




(a) Model A (b) Model B

Figure 2: Posterior 95% CIs of the random intercepts for hospital with at least ten patients, highlighting *Milano* effect. The hospitals located in Milano are depicted in blue dashed lines, those outside Milano in red solid lines. The estimates are in increasing order of number of patients per hospital. The last two intervals represent new random intercepts for a hospital in and outside Milano, respectively.

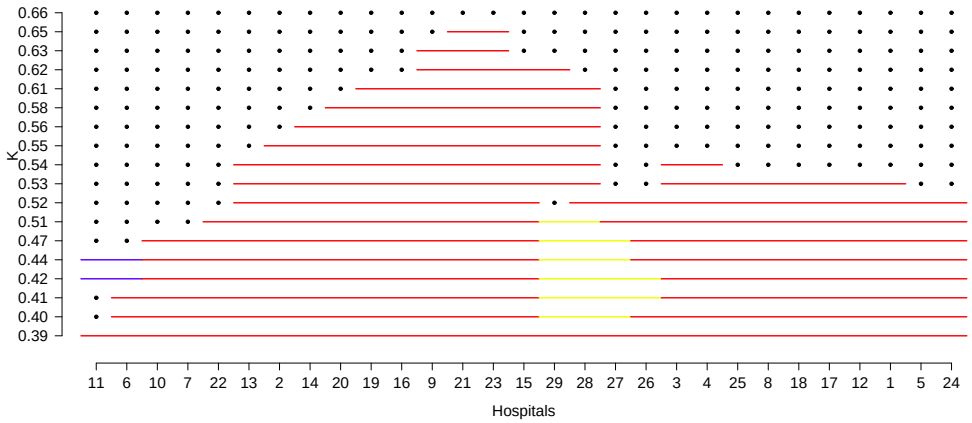
to this prior information, i.e., $a \sim Exp - trunc(1)$ on the interval $[1, \infty)$, which a priori leads to $\mathbb{E}[a] = 2$. The a priori number of groups in this case is 5.8. The mass parameter $a$ is a posteriori concentrated around small values under both models: mean 1.61 (SD 0.62) in Model A and 1.65 (SD 0.65) in Model B. We observe a slight reduction of the expected number of groups, going from a prior mean of 5.8 to a posterior one of 4.24 in Model A and 4.58 in Model B.
Finally, we run the algorithm fixing the mass parameter $a$ equal to one (doing so, the expected number of cluster is 4) and we obtained similar posterior estimates; hence we can conclude that the inference is quite robust to the prior specification of the mass parameter $a$.

Even if Bayesian semiparametric models allow a model-based clustering without making any extra assumption, the results provided in this sense by such models may not be straightforward to be interpreted. The precise estimation of the true number of clusters is, in general, a very hard task. As explained in Section 2, the estimated grouping is the optimal partition defined by the maximisation problem in (17). Two hospitals belong to the same cluster $j$ if their labels are in the same set $A_j$. In Model A, this is equivalent to say 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 since we have two sub-populations. Since for any choice of $u$ and $w$ the optimal partition can be determined, we consider different values for the couple $(u, w)$, 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. Notice that low values of $K$ penalise

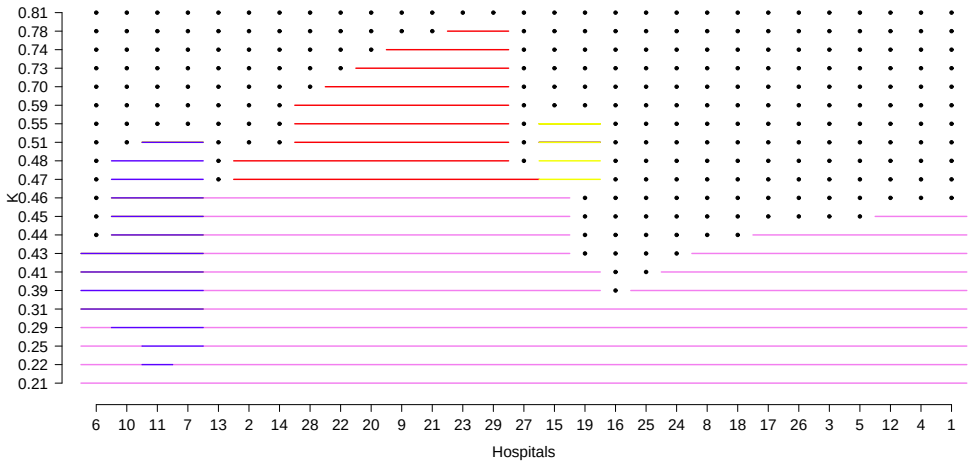
separation of items more than aggregation, whereas high values of $K$ do the opposite.

Figure 3 shows how the clustering induced by the optimal partitions evolves as long as $K$ grows up, for Model A (upper panel) and Model B (lower panel) respectively. Hospitals on the abscissa are sorted so that a more effective visualisation is allowed. On the other hand, on the vertical axis we retain only $K$ values corresponding to relevant changes in hospitals grouping.
As it can be seen from the picture, 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, in the case where Model B is fitted to data, for $K = 0.21$ the best partition minimising the expected loss is the one where all the hospitals are clustered together. As long as $K$ increases, some hospitals progressively exit the cluster and disaggregate, up to the case where all hospitals are singletons, that comes out for $K = 0.81$. Analogous considerations hold for Model A, with a smaller range from $K = 0.39$ to $K = 0.66$. Notice that disaggregation provided by Model B is more gradual than the one provided by Model A. Observing how the partition evolves as long as $K$ increases, we obtain a sort of implicit ranking of the providers (see also [39]). In general, starting from low values of $K$ (hospitals clustered together) up to the high values (hospitals all singletons), the two models point out similar results: in fact, hospitals 6 and 11, and, then, 7 and 10 are in both cases among the first items 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 generated by both models. Finally, hospitals 9, 15, 21, 23 and 29 are the last ones becoming singletons, and are grouped together in both models.

Tables 3 and 4 show the 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 and 10, highlighted as similar and early detected 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, grouped together by both models for almost all $K$s, are concentrated on smaller 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 ones are those with the less favourable contribution to the patient's survival, for this reason we may say that the "evolving partition" is pointing out a ranking among hospitals.
According to the previous comments, we may say that, as long as values of $K$ are far from 0.5 (i.e., couples $(u, w)$ far from $(1, 1)$), partitions tend to point out outliers with respect to a "reference" group, in the sense of [39]. The discriminating power is determined by $K$, which is problem-driven. Summing up, we conclude that Model B is better in distinguishing different cases; this is probably




(a) Model A




(b) Model B

Figure 3: Hospitals' optimal partition as long as $K$ increases, for Model A (upper panel) and Model B (lower panel) respectively.

due to the higher flexibility it allows for.

### 3.3 Model fit and patients classification

In this section we estimate the variability explained by our models using the Bayesian $R^2$ defined in (15) and evaluate their performance by predicting in-hospital survival probability for each given patient. In particular, we compare

Table 3: 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 4: 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 | | | | |

two different predictive methods: the usual one based on point estimates summarising the posterior predictive distributions, and the new one we proposed, based on interval estimates.

In Table 5 we provide the Bayesian $R^2$ of the two models. Observe that Model B seems to better fit the data, as we expected according to the greater flexibility it allows for. As we said before, the Bayesian $R^2$ provides an index of the explained variability at the latent variable level; however we would evaluate also the predictive performance at the outcome level.

Table 5: Bayesian $R^2$ defined in (15) for the two models.

| | Model A | Model B |
|---|---|---|
| Bayesian $R^2$ | 0.35 | 0.57 |

In our application, since the share of outcome success in the dataset 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 rarest 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 aimed at improving the capability of the model in predicting deaths is based on adopting a threshold equal to the empirical rate of success, as suggested in [9]. Table 6 displays the results of the patient classification under Model A (left) and Model B (right), using a threshold equal to the sample survival rate ($\bar{p} = 0.97$). The posterior predicted rates of survival and death respectively are more balanced than using a threshold of 0.5. On the other hand, we obtain a worse overall misclassification rate (around 10% for all models). This is because the overall misclassification rate is less dependent on the unbalance of shares, as explained in [9].

Table 6: Predictive tables of survival outcome when the classification rule is based on the comparison between survival posterior means and $\bar{p} = 0.97$.

(a) Model A.

| | $Y = 1$ | $Y = 0$ |
|---|---|---|
| $\hat{Y} = 1$ | 599 | 3 |
| $\hat{Y} = 0$ | 75 | 20 |

(b) Model B.

| | $Y = 1$ | $Y = 0$ |
|---|---|---|
| $\hat{Y} = 1$ | 596 | 3 |
| $\hat{Y} = 0$ | 78 | 20 |

Since the overall misclassification rate represents a goodness of fit index, as mentioned in Section 2, we developed a rule in order to improve performances of our models in predicting the unsuccessful outcome, enhancing the information provided by the Bayesian approach.

In Table 7 we report 90% posterior predictive CIs and assume equal misclassification costs, i.e., the threshold is set equal to 0.5. With our dataset, only around 4% of the patients belong to the Uncertainty Class (UC) and the total misclassification rate, based only on classified patients, is less than 3% for both models. Considering the number of patients in 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.

Table 7: Predictive tables of survival outcome when the classification rule is based on survival posterior 90% CIs and threshold equal to 0.5.

(a) Model A.

| | $Y=1$ | $Y=0$ |
|---|---|---|
| $\hat{Y}=1$ | 661 | 8 |
| $\hat{Y}=0$ | 0 | 3 |
| UC | 13 | 12 |

(b) Model B.

| | $Y=1$ | $Y=0$ |
|---|---|---|
| $\hat{Y}=1$ | 661 | 8 |
| $\hat{Y}=0$ | 0 | 2 |
| UC | 13 | 13 |

Of course, the number of patients 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 to fix the prior components for the fixed effects informatively, i.e., using previous data and/or expert opinions.
In Figure 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 did not report it here).

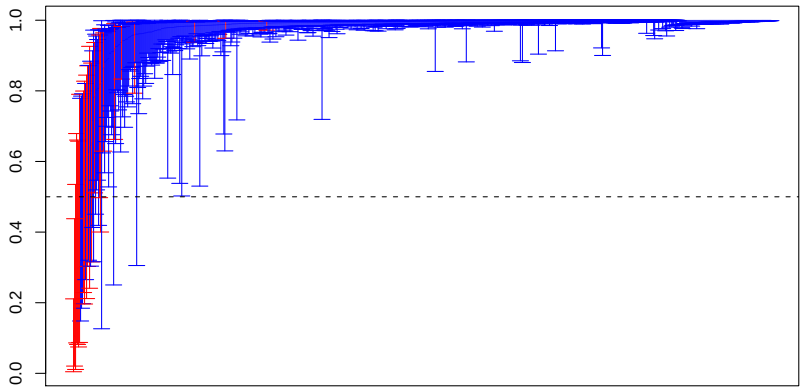



Figure 4: 90% posterior predictive CIs of all the patients (ordered by increasing median) under Model B. The positive outcomes are in blue and the negative ones in red.

Notice that most of the interval lengths of the survived patients are quite small, while there is more uncertainty on the negative outcomes, as expected since the unsuccessful outcome is rare.

As an example, in Figure 5 we focus on a smaller set of patients (those 29 treated in hospital 19, under Model B). Notice that predictive distributions with very large and very low mean have small width, while those with mean around

0.5 have wider interval estimates. There are five unclassified patients and only one is misclassified.

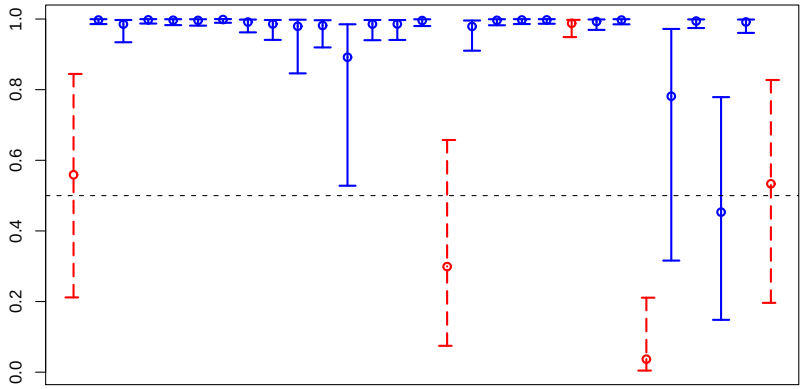



Figure 5: 90% posterior predictive CIs of all the patients from one of the hospital belonging to STEMI Archive, obtained fitting Model B to data. The CIs corresponding to alive patients are in blue solid line, while those corresponding to dead patients are in red dasehd. There are five unclassified patients and only one was misclassified.

Finally, we would like to mention that, 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. Neverthless, a comparison of the exposure parameter CI's shows that including the exposure nonparametrically through a DDP leads to more variability among hospitals than we observed fitting Model A.

# 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 hospitalised with STEMI diagnosis. Dependent Dirichlet Process 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. The study is within a project, named Strategic Program of Regione Lombardia and aimed at supporting decisions in healthcare policies.

We fitted two models to the data, matching the marginal distributions of corresponding random effects, and we compared them in terms of the Bayesian $R^2$ index proposed in [18]. Then we studied the evolution of the estimated partition as long as the proportion $K$ of incorrect clustering cost increases. A sort of "implicit ranking" among 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.

In general, random partitions may be considered a powerful tool to investigate the latent grouping structure among random effects in grouped data, without making any further assumption. Finally, we pointed out a classification rule for patients' survival (a strongly unbalanced outcome in our application) based on the posterior credibility intervals instead of pointwise estimates. This rule introduces the Uncertainty Class, which collects patients whose credibility intervals includes the reference threshold adopted for classification. This classification rule proved to be less sensitive to the choice of the threshold with respect to classification criteria based on pointwise estimates.

Further developments of this work will be focused on taking advantage of physicians' expertise in priors elicitation. Moreover, it would be of interest to develop a dynamic update of DDP priors, generalizing frameworks such those proposed in preliminar works like [14] and [30]. Finally, methods aimed at monitoring the evolution of the clusters over time, trying to identify the causes of the changes, are definitively of interest for a proper monitoring of hospital performances, since only a structured and systematic monitoring of the care-delivery process may lead to an improved healthcare process.

We think that the methods adopted in this paper properly and effectively tackle the problem of supporting decision makers in assessing hospitals performances, enhancing interactions among physicians and statisticians.

**Acknowledgments**

This work is within the Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction. The authors wish to thank Regione Lombardia - Healthcare division for having funded and sustained the project, Lombardia Informatica S.p.A. for having provided data and all the physicians who collaborated to STEMI Archive planning and data collection.

# References

[1] Albert, J.H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 669-679.

[2] Ash, A.S., Fienberg, S.E., Louis, T.A., Normand, S.T., Stukel, T.A., Utts, J. (2012). Statistical Issues in Assessing Hospital Performance, Commissioned by the Committee of Presidents of Statistical Societies. Original report submitted to CMS on November 28, 2011. Revised on January 27, 2012.

[3] Barrientos, A.F., Jara, A., Quintana, F.A. (2012). On the support of MacEachern's Dependent Dirichlet Process and extensions. *Bayesian Analysis*, *Forthcoming*.

[4] Berkelaar, M., Eikland, K., Notebaert, P. (2004). Open source (mixed-integer) linear programming system, version 5.1.0.0. `http://lpsolve.sourceforge.net/`

[5] Binder, D.A. (1978), Bayesian Cluster Analysis, *Biometrika*, **65**, 1, 31–38.

[6] Binder, D.A. (1981), Approximations to Bayesian Clustering Rule, *Biometrika*, **68**, 1, 275–285.

[7] 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 England Journal of Medicine*, **355**, 2308-2320.

[8] 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. *Journal of the American Medical Association*. **273**, 2941-2947.

[9] Cramer, J.S. (1999). Predictive performance of the binary logit model in unbalanced samples. *The Statistician*, **48**, 85-94.

[10] Dahl, D.B. (2006), Model-Based Clustering for Expression Data via Dirichlet Process Mixture Model. In *Bayesian Inference for Gene Expression and Proteomics*, chap. 10, eds. Kim-Anh Do, Peter Muller and Marina Vannucci, Cambridge University Press.

[11] Decreto No 10446, 15/10/2009, Direzione Generale Sanità - Regione Lombardia (2009), Determinazioni in merito alla Rete per il trattamento dei pazienti con Infarto Miocardico con tratto ST elevato (STEMI).

[12] De Iorio, M., Müller, P., Rosner, G.L. and MacEachern, S.N. (2004). An ANOVA Model for Dependent Random Measures. *Journal of the American Statistical Association*, **99**, 465.

[13] De la Cruz-Mesia, R., Quintana, F. A., Muller, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *Journal of the Royal Statistical Society - Series C*, **56**, 119–137.

[14] Dundson, D.B., Ren, L., Carin, L. (2012). The Dynamic Hierarchical Dirichlet Process. Under review by the *International Conference on Machine Learning (ICML)*.

[15] Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 4, 615–629.

[16] Freeman, E.A., Moisen, G.G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, **217**, 1-2, 48-58.

[17] Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian analysis*, **1**, 515–533.

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

[19] Gordon, A.D. (1999). Classification, Chapman & Hill, New York (2nd ed.)

[20] Green, P.J., Richardson, S. (2001). Modelling Heterogeneity With and Without the Dirichlet Process. *Scandinavian Journal of Statistics*, **28**, 355–375.

[21] Grieco, N., Ieva, F., Paganoni, A.M. (2011). Performance assessment using mixed effects models: a case study on coronary patient care. *IMA Journal of Management Mathematics*, **23**, 2, 117–131.

[22] Guglielmi, A., Ieva, F., Paganoni, A.M., Ruggeri, F. (2012), A Bayesian random effects model for survival probabilities after Acute Myocardial Infarction, *Chilean Journal of Statistics*, **3**, 1, 1–15.

[23] Heard, N.A., Holmes, C.C., 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. *Journal of the American Statistical Association*, **101**, 18–29.

[24] Kleinman, K.P., Ibrahim, J.G. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, **17**, 2579–2596.

[25] Ieva, F. (2012), Designing and mining a multicenter observational clinical registry concerning patients with Acute Coronary Syndromes. *Submitted*.

[26] Ieva, F., Paganoni, A.M. (2011). Process Indicators for Assessing Quality of Hospital Care: a case study on STEMI patients, *JP Journal of Biostatistics*, **6**, 1, 53-75.

[27] Ieva, F. (2012). Statistical methods for classification in cardiovascular healthcare. PhD Thesis

[28] Ishwaran, H. and Zarepour, M.(2000). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, **30**, 269–283.

[29] Lau, J.W., Green, P.J. (2007). Bayesian Model-Based Clustering Procedures. *Journal of Computational and Graphical Statistics*, **16**, 3, 526–558.

[30] Lin, D., Grimson, E., Fisher, J. (2010). Construction of Dependent Dirichlet Processes based on Poisson Processes. *Forthcoming*.

[31] MacEachern, S.N. (1999). Dependent Nonparametric Processes. *ASA proceedings* of the Section on the Bayesian Statistical Science, Alexandria VA. American Statistical Association.

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

[33] Müller, P., Quintana, F.A. (2004). Nonparametric Bayesian Data Analysis. *Statistical Science*, **19**, 1, 95–110.

[34] Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 20–22.

[35] Quintana, F.A., Iglesias, P.L. (2003). Bayesian Clustering and Product Partition Models. *Journal of the Royal Statistical Society - Series B*, **65**, 2, 557–574.

[36] R Development Core Team (2009), R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [Online] http://www.Rproject.org

[37] Ray, S., Mallik, B. (2006). Functional Clustering by Bayesian Wavelet Methods. *Journal of the Royal Statistical Society B*, **68**, 557–574.

[38] Rodriguez, A., Dunson, D. B.(2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, **6**, 1, 145–178.

[39] Shotwell, M., Slate, E.H. (2011). Bayesian Outlier Detection with Dirichlet Process Mixtures. *Bayesian Analysis*, **6**, 2, 1-22.

[40] Sethuraman, J. (1994), A Constructive Definition of Dirichlet Priors, *Statistica Sinica*, **4**, 639-650.

[41] Spiegelhalter, D., Sherlaw-Johnson, C., Bardsley, M., Blunt, I., Wood, C., Grigg, O. (2012). Statistical methods for healthcare regulation: rating, screening and surveillance. *Journal of the Royal Statistical Society A*, **175**, 1, 1-47.

# MOX Technical Reports, last issues

**Dipartimento di Matematica "F. Brioschi",**
**Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)**

**21/2012** Guglielmi, A.; Ieva, F.; Paganoni, A.M.; Ruggeri, F.; Soriano, J.
*Semiparametric Bayesian models for clustering and classification in presence of unbalanced in-hospital survival*

**20/2012** Ieva, F.; Paganoni, A. M.; Zanini, P.
*Statistical models for detecting Atrial Fibrillation events*

**19/2012** Faggiano, E.; Antiga, L.; Puppini, G.; Quarteroni, A.; Luciani G.B.; Vergara, C.
*Helical Flows and Asymmetry of Blood Jet in Dilated Ascending Aorta with Normally Functioning Bicuspid Valve*

**18/2012** Formaggia, L.; Vergara, C.
*Prescription of general defective boundary conditions in fluid-dynamics*

**17/2012** Manzoni, A.; Quarteroni, A.; Gianluigi Rozza, G.
*Computational reduction for parametrized PDEs: strategies and applications*

**16/2012** Cutri', E.; Zunino, P.; Morlacchi, S.; Chiastra, C.; Migliavacca, F.
*Drug delivery patterns for different stenting techniques in coronary bifurcations: a comparative computational study*

**15/2012** Mengaldo, G.; Tricerri, P.; Crosetto,P.; Deparis, S.; Nobile, F.; Formaggia, L.
*A comparative study of different nonlinear hyperelastic isotropic arterial wall models in patient-specific vascular flow simulations in the aortic arch*

**14/2012** Fumagalli, A.;Scotti, A.
*An unfitted method for two-phase flow in fractured porous media.*

**13/2012** Formaggia, L.; Guadagnini, A.; Imperiali, I.; Lever, V.; Porta, G.; Riva, M.; Scotti, A.; Tamellini, L.
*Global Sensitivity Analysis through Polynomial Chaos Expansion of a basin-scale geochemical compaction model*

**12/2012** Guglielmi, A.; Ieva, F.; Paganoni, A.M.; Ruggeri, F.
*Hospital clustering in the treatment of acute myocardial infarction patients via a Bayesian semiparametric approach*