A semiparametric Bayesian joint model for multiple mixed-type outcomes: an application to acute myocardial infarction

Alessandra Guglielmi[1] · **Francesca Ieva**[2] · **Anna Maria Paganoni**[3] · **Fernando A. Quintana**[4]

Abstract We propose a Bayesian semiparametric regression model to represent mixed-type multiple outcomes concerning patients affected by Acute Myocardial Infarction. Our approach is motivated by data coming from the ST-Elevation Myocar-dial Infarction (STEMI) Archive, a multi-center observational prospective clinical study planned as part of the Strategic Program of Lombardy, Italy. We specifically consider a joint model for a variable measuring treatment time and in-hospital and 60-day survival indicators. One of our main motivations is to understand how the various hospitals differ in terms of the variety of information collected as part of the study. To do so we postulate a semiparametric random effects model that incorpo-rates dependence on a location indicator that is used to explicitly differentiate among hospitals in or outside the city of Milano. The model is based on the two parameter Poisson-Dirichlet prior, also known as the Pitman-Yor process prior. We discuss the resulting posterior inference, including sensitivity analysis, and a comparison with the particular sub-model arising when a Dirichlet process prior is assumed.

Keywords Bayesian clustering · Bayesian nonparametrics · Two parameter Poisson-Dirichlet process prior · Random-effects models · Random partition models · Unbalanced binary outcomes

✉ Anna Maria Paganoni
anna.paganoni@polimi.it

Alessandra Guglielmi
alessandra.guglielmi@polimi.it

Francesca Ieva
francesca.ieva@unimi.it

Fernando A. Quintana
quintana@mat.uc.cl

[1] Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133 Milan, Italy

[2] ADAMSS Center and Department of Mathematics “F. Enriques”, Università degli Studi di Milano, via Saldini 50, 20133 Milan, Italy

[3] MOX-Department of Mathematics, Politecnico di Milano, via Bonardi 9, 20133 Milan, Italy

[4] Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile

1 Introduction

Studies with multiple outcomes that are used to properly characterize an effect of interest are becoming increasingly more common nowadays. In the particular case of clinical studies, multiple outcomes are often used to characterize the patient's status or the performances of health care service with respect to patients' management (see, for example, Normand 2008; Parekh et al. 2011; AHRQ 2015).

We are concerned with the analysis of data collected in a clinical registry named STEMI Archive (see Lombardia 2009; Ieva 2013), which is a result of a wider compre-hensive project, namely The Strategic Program "Exploitation, integration and study of current and future health databases in Lombardy for Acute Myocardial Infarction" (for additional information, visit http://ima.metid.polimi.it). This project is funded by the Italian Ministry of Health. Its main goal is to enhance the integration of different sources of health information (clinical registries and administrative databases) so as to automate and streamline clinicians' work flow, and that all the data collected can be generally used. We specifically consider outcomes of patients with ST segment elevation myocardial infarction (STEMI) diagnosis admitted to a hospital. STEMI is caused by an occlusion of a coronary artery which causes an ischemia that, if untreated, can damage heart cells and make them die (infarction). It is fundamental for the patient's recovery to do a reperfusion therapy (i.e. restoration of the blood flow to the ischemic tissue) as quickly as possible, since its benefits decrease highly non-linearly with treatment delay. All patients in the study were treated with Percuta-neous Transluminal Coronary Angioplasty (PTCA). Data were recorded in a registry collecting clinical outcomes, process and time indicators measuring the way the health care structures manage the patients, and personal information on patients with STEMI diagnosis admitted to hospitals of Lombardy. These data were combined with informa-tion coming from the standard administrative database, so as to obtain out of hospital mortality (i.e., mortality for any reason). It is important to point out that the STEMI Archive is not linked to the Emergency Room (ER) database, so we are discarding the deaths occurred in the ER. This fact could limit the validity of our findings to this particular subpopulation. However, there is a specific reason for selecting the ana-lyzed population: the STEMI Archive was primarily designed to assess the impact of hospital response times and organization on in-hospital survival after treatment. This implies that the cohort of interest is the one undergoing angioplasty, and then surviving once entering the ER. Data in the survey are grouped by hospital of admission. This automatically induces a policy issue about the effect that such grouping may have on patients' outcome, which is the main motivation for this work. In fact, the problem of profiling hospitals according to their effects on patients' outcomes is crucial within the context of healthcare planning. Proper methods for addressing such a problem are of great interest to healthcare policymakers.

We propose a Bayesian nonparametric hierarchical model that includes a cluster analysis, aimed at identifying profiles or hospital behaviors that may affect the outcome at patient level. In particular, we introduce a multivariate multiple regression model, where the response has three mixed-type components. The components are, respectively: (1) the door to balloon time (DB), i.e. the time between the admission to the hospital and the PTCA; (2) the in-hospital survival; and (3) the survival after 60 days from admission. The first response (continuous) is essential in quantifying the efficiency of health providers, since it plays a key role in the success of the therapy; the second is the basic treatment success indicator, while the third concerns a 60-days period, during which the treatment effectiveness, in terms of survival and quality of life, can be truly evaluated. Note that the last two responses are binary, so that, as a whole, the multivariate response is of mixed type. It is worth noting that the information on patients' survival after 60 days is obtained from the linkage between STEMI archive and a further administrative database concerning patient-specific vital statistics such as date of birth and death for general causes. The linkage between the different data sources was carried out by Lombardia Informatica S.p.A, the agency managing regional datawarehouses. We do not have direct access to the data sources so as to construct different outcomes of potential interest. Moreover we work with a singly imputed data set and we could not identify data preprocessing tools used by the Lombardia Informatica S.p.A. agency, in particular the technique used to impute the missing data. The modeling of multiple outcomes from data collected in STEMI Archive was previously discussed in Ieva et al. (2014), under a semiparametric fre-quentist bivariate probit model. Their aim was to analyze the relationship among in-hospital mortality and a treatment effectiveness outcome in the presence of con-founders, that is, variables that are associated with both covariates and response. This is a problem that poses serious limitations to covariate adjustment since the use of classical techniques may yield biased and inconsistent estimates. In this context, Ieva et al. (2014) proposed the use of a semiparametric recursive bivariate probit model, as an effective way to estimate the effect that a binary regressor has on a binary outcome in the presence of nonlinear confounder response relationships. In contrast, we focus on a joint model for the grouped outcomes. As discussed below, our aim is to find relevant groups of hospitals in terms of patient-specific characteristics, which may assist in further planning and policy making.

In recent years, there has been a considerable interest on developing models that overcome the challenges posed by the modeling of outcomes of mixed types. Sammel et al. (1997) discuss a model for mixed discrete and continuous outcomes where the multiple outcomes correlate through subject-specific latent variables. The observed outcomes are thus manifestations of unobserved latent variables. Conditionally on these, the outcome components are assumed independently distributed according to the exponential family, whose parameters are allowed to be a function of the latent variables as well as other component-specific covariates. Dunson and Herring (2005) proposed a Bayesian latent variable model for clustered mixed outcomes that allows nonlinear relationships between covariates and latent variables, and that uses multiple latent variables for different types of outcome as well as covariate-dependent modifications of these relationships. In contrast, a single linear combination of the covariates is used to predict multiple outcomes simultaneously in the Bayesian multivariate model by Weiss et al. (2011), where correlations among outcomes are modeled by latent variables. Bello et al. (2012) present a hierarchical Bayesian extension of bivariate generalized linear models whereby functions of the variance-covariance matrices are specified as different linear combinations of fixed and random effects.

A somewhat different approach for bivariate outcomes of mixed type arises by factorizing the joint distribution of outcomes and introducing latent variables to model the correlation among the multiple outcomes. The main idea of this method is to write the likelihood as the product of the marginal distribution of one outcome and the conditional distribution of the second given the previous one. In particular Cox and Wermuth (1992) discuss two factorization models for a continuous and a binary outcome as functions of covariates. In Catalano and Ryan (1992) and Fitzmaurice and Laird (1995) the factorization approach is extended to clustered data.

Our approach is based on factorizations. In particular, we factorize the patient-specific likelihood factor for the three responses as the product of (1) the marginal likelihood of the continuous response (DB time); (2) the distribution of the in-hospital survival given DB time; and of (3) the 60-days survival, given the previous two. All these conditional distributions lie within the class of univariate generalized linear mixed models, with random-effects given by hospital intercepts. Covariates corre-sponding to the other regression parameters include those related to hospital admission, patient's clinical status at hospital admission, and patient's general health status. A full description of available covariates is given in Sect. 2. Of course, other factor-izations could be adopted, but we find this one easy to define and explain. To deal with differences across hospitals, we adopt a nonparametric random effects approach, with a random distribution function that is allowed to vary with an indicator that explicitly differentiates among hospitals in or outside the city of Milano. We adopt an ANOVA-Dependent Pitman-Yor process prior for hospital effects, that is, a family of distributions of dependent random probability measures with (marginal) almost surely discrete trajectories that generalize the Dirichlet process (DP). Such priors induce a random partition of the hospital labels. As we discuss later, the Pitman-Yor process (PY) process includes two parameters that allow for increased flexibility in the prior clustering structure compared to the DP. This is particularly useful to achieve one of our main goals, that is, to estimate a latent clustering among hospitals from the dataset, identifying groups of care providers affecting outcomes at patient level in a similar way. In this context, a cluster analysis of the hospitals is straightforward, based on posterior estimates of the induced random partition parameter itself. Besides marginal posterior inference on all relevant parameters, we discuss predictive inference for new hospitals, and hospitals clustering. Moreover, some competitor models are considered and compared to our proposal through predictive goodness-of-fit tools.

The rest of this paper is organized as follows. Section 2 gives a complete data description and states the main inference questions that drive the analysis. Section 3 describes the adopted model in detail, and posterior inference, implementation details and comparison among different models are discussed in Sect. 4. Final comments are given in Sect. 5.

2 Motivation and data description

We consider a dataset coming from the integration of a clinical registry named STEMI Archive (see Lombardia 2009; Ieva 2013), with data from the administrative health database. Our focus is on data from patients in any of the hospitals in Lombardy, and the analysis of their time to treatment, their in-hospital and 60-days survival outcomes. Our goals are (1) to understand the effect of other factors on the selected outcome variables; (2) to compare marginal posterior distributions of the different nonparametric components; (3) to compare hospital performances by means of a cluster analysis; and (4) to make predictions for new hospitals entering the study (e.g., hospitals outside the region, but in districts gravitating towards Lombardy).

A similar problem, with a related dataset, was already considered in Guglielmi et al.(2014). However, there are differences in the two statistical problems tackled, and in the two datasets analyzed. The specific focus in Guglielmi et al. (2014) was on building a model for predicting only the in-hospital survival after STEMI at patients' level, and to provide model-based clustering of the providers. These goals were achieved via a univariate regression model having patient's in-hospital survival as the response. On the other hand, it is known that one of the crucial factors influencing in-hospital survival for STEMI patients is treatment time (see, for example, De Luca et al. 2004; Antoniucci et al. 2002), which was considered as a fixed covariate in the latter paper. Here we also focus on the relationship among in-hospital survival and treatment time, but also aim at uncovering determinants (both logistic and environmental) jointly affecting times to treatment and the two survival outcomes. What strongly motivates the new dataset we analyze here is the statistical interest in survival beyond discharge time, which would be of great help to deepen our understanding of the disease progression and health recovery of STEMI patients.

The dataset at hand includes information about $n = 697$ patients treated with PTCA during the 6-month time period (January–June 2011) in $J = 33$ hospitals of Lombardy, 12 of these located in Milan. The number of patients per hospital ranges from a minimum of 5 to a maximum of 60, with mean 21. The available information about each patient are then the hospital of admission, the mode of admission (a binary variable indicating whether the patient was delivered by rescue units of 118, which is the Italian toll-free emergency number), demographic features such as age and gender, the severity of infarction, risk factors (such as diabetes, smoking and high cholesterol), times to treatment or intervention, and process indicators within the pre- and in-hospital phase. We resume all the information content of the dataset through the following list:

- **DB** (Y_1): the time between the admission to the hospital (*Door*) and primary angioplasty (*Balloon*);
- **ALIVEIN** (Y_2): the in-hospital survival;
- **ALIVE60** (Y_3): the survival after 60 days from admission.

These three variables represent the outcome. Observe that the dataset is strongly unbalanced: 96.84 % of patients are alive after the discharge and 98.37 % of these are alive after 60 days. The sample mean and standard deviation of *DB* in the log-scale are 4.452 and 0.551.

The available covariates are listed here:

- **ACCESS**: 0 if the patient came to hospital by any rescue unit, 1 otherwise (by own means). The sample mean is 0.597;
- **ECG**: time of the first electrocardiogram (minutes). The sample mean is 9.671 (std. dev. 18.296);
- **WE**: 1 if the admission was on holiday, weekend or between 6pm–8am, 0 otherwise. The sample mean is 0.469;
- **AGE**: age of the patient (years). The sample mean is 64.651 (std. dev. 13.122);
- **MALE**: gender of the patient; 1 when male, 0 female. The sample mean is 0.776;
- **RISK**: 1 if patient had at least four among the following risk factors: diabetes, smoking, hypertension, cholesterol, vasculopathy, 0 otherwise. The sample mean is 0.006;
- **KILLIP**: 1 if the infarction was severe (Killip class 3 or 4), 0 otherwise (Killip class 1 or 2). The sample mean is 0.060;
- **EF**: ejection fraction at admission to hospital, i.e. the volumetric fraction of blood pumped out of the ventricle with each heart beat (%). The sample mean is 47.858 (std. dev. 9.663);
- **COMP**: 1 if there were complications after the primary angioplasty, 0 otherwise. The sample mean is 0.386;
- **CKD**: 1 if the patient had chronic kidney disease, 0 otherwise. The sample mean is 0.080;
- **preMI**: 1 if there is a history of previous infarction, 0 otherwise. The sample mean is 0.113;
- **STres**: 1 if the treatment was not effective, 0 otherwise; this covariate is quantified by physicians as equal to 0 if there was a reduction of at least 70 % in the ST-elevation within one hour after the angioplasty. The sample mean is 0.198;
- **HOSPITAL**: hospital of admission of the patient;
- **MILAN**: 1 if the hospital is located in Milano, 0 otherwise; the sample mean is 0.445.

We note that treatment times (DB and ECG) are computed with respect to time of admission at the hospital. From the analysis viewpoint, hospital is the natural grouping factor here, since patients are delivered to hospitals by 118 rescue units. We have considered in the analysis only hospitals provided with intensive unit care and coronary unit. Covariate MILAN is all the information we have on the hospitals at this point. However, note that Guglielmi et al. (2014) did include hospital exposure, i.e. the number of patients who were treated with primary angioplasty per year, as a hospital covariate. This explanatory variable was not included in the original clinical database, so it was retrieved from a different administrative database. We remark that according to Section 3.1 in Guglielmi et al. (2014), this exposure variable has no effect on the in-hospital survival, which was the response for that case. Of course, since we have

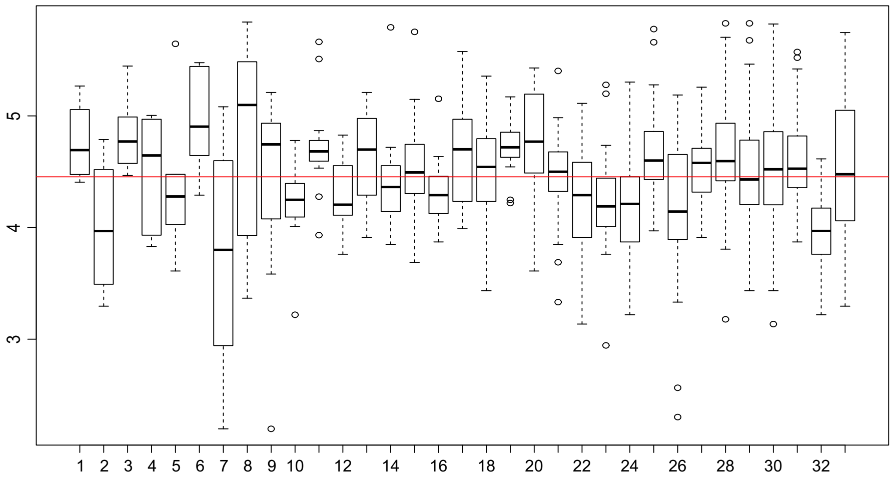

Fig. 1 *Boxplots*of the DB time (in the log scale) stratified by hospitals. The *red line* is the global median of all DB times (color figure online)

no extra knowledge on hospitals, we should assume an exchangeable prior for their effects, a common procedure that generally reflects lack of additional information.

It is important to stress here that in previous analyses on similar datasets, such as those reported in Ieva and Paganoni (2010), Grieco et al. (2012) and Guglielmi et al. (2014), the OD (Onset-to-Door time, i.e. the time from onset to admission to hospital) was considered in the dataset as a part of OB (Onset to Balloon time, that is the sum of the OD plus the DB, the Door to Balloon time, i.e. the total ischemic time from onset to angioplasty). Unfortunately, this information revealed to be very poor and misleading. In fact, OD times are usually subjectively recorded, since the patients tend to declare onset times based on their perception of the onset, i.e. strongly biased by the fact they are scared by the events, or not sufficiently prompt to recognize them. Moreover, as a confirmation, we checked that, at least for patients with no missing OD time, including this covariate in the regression did not improve the fit.

In Fig. 1 we report boxplots of the DB time (in logarithmic scale) stratified by hospitals. The large variability and overdispersion due to the grouped nature of the data suggests that it is reasonable to assume a random effect on the grouping factor.

Figure 2 shows the difference between in-hospital and 60-days survival rates per hospital. Observe that for most cases, the in-hospital and 60-days survivals are very similar. The two hospitals where this difference is the largest are 8 and 22.

3 A multi-response Bayesian semiparametric model with Pitman-Yor process prior

To achieve the goals described in Sect. 2, we propose a trivariate regression model of mixed types, according to the three outcome variables described before. We use a Bayesian semiparametric approach with a discrete random probability measure prior.

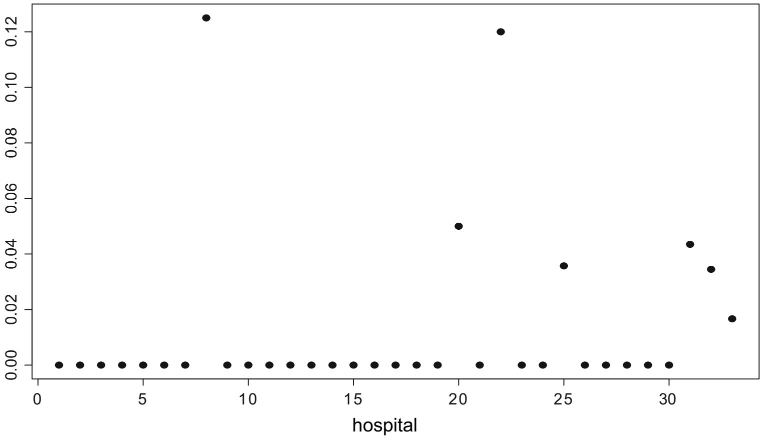

Fig. 2 Difference between sample survival proportions (at discharge and after 60 days) per hospital

This choice is also due to the flexibility they provide in modeling data, as well as the implied robustness against incorrect model specifications. See Müller and Quintana (2004) and Müller and Mitra (2013) for a thorough discussion on Bayesian Nonparametrics. Moreover, the discrete random measure model that we adopt as a prior for the hospital random-effects allows us to infer on a partition of hospital labels. The prior we set here is the Pitman-Yor process (Pitman and Yor 1997), which includes the (regular) DP chosen in Guglielmi et al. (2014) as a special case. Model details will be given below and in the next section.

In particular, we consider a generalized linear model for the response of patient i treated in hospital j, $\mathbf{Y}_{ji} := (Y_{ji1}, Y_{ji2}, Y_{ji3}) = (\log(\mathrm{DB}\ _{ji}\ ), \mathrm{ALIVEIN}\ _{ji}\ , \mathrm{ALIVE60}_{ji}\ )$, with $i = 1, \ldots, n\ _j\ , j = 1, \ldots,\ J$. Since patients are admitted to hospitals, and one of the aims is to compare the hospitals themselves, it is natural and straightforward to consider generalized linear models with random intercepts to account for hospital variability. As usual, we assume that observations, given parame-ters and covariates, are independent. To facilitate model specification, we consider a conditional specification of the joint sampling model, conditional on parameters and covariates, as

$$\mathcal{L}(Y_{ji1}|\text{par, cov}) \times \mathcal{L}(Y_{ji2}|Y_{ji1}, \text{par, cov}) \times \mathcal{L}(Y_{ji3}|Y_{ji2}, Y_{ji1}, \text{par, cov}). \quad (1)$$

Our assumptions in this case are the following: $\mathcal{L}(Y_{ji1}|\text{par, cov})$ is a Gaussian linear regression, $\mathcal{L}(Y_{ji2}|Y_{ji1}, \text{par, cov})$ and $\mathcal{L}(Y_{ji3}|Y_{ji2}, Y_{ji1}, \text{par, cov})$ are logistic regression models.

Before detailing the covariates at the three levels, we point out that we have carried out an extensive exploratory analysis which gave rise to a careful choice of covariates. First we consulted with our experts (cardiologists and health managers from different hospitals in Lombardy) to gain a better understanding of the covariates to include in a parametric regression model where $\mathbf{Y}_{ji}$ is the response, as in (1). We conducted a parametric variable selection procedure to determine which covariates to include in our model. The binary indicator MILAN was included among the covariates to be potentially selected, unlike the hospital label. We adopted two priors for selecting the variables, i.e. the Normal Mixture of Inverse Gamma (NMIG) distributions and the SSVS spike-and-slab prior in Rockova et al. (2012); see notation and details of the priors there. In particular, under both priors, we fixed the variance of the spike and of the slab components equal to 0.001 and 10, respectively; here we report the selection considering all the covariates selected by the four models with highest posterior probabilities under the NMIG prior. The other prior gave consistent results.

Let $\boldsymbol{x}_{ji\ell}$ denote the covariate vector for the sampling model at level $\ell = 1, 2, 3$ for patient i in hospital j. Our variable selection analysis determined the following subsets of covariates per level:

$$\begin{aligned}
\boldsymbol{x}_{ji1} &:= (\mathrm{ACCESS}_{ji}, \mathrm{ECG}_{ji}, \mathrm{WE}_{ji}, \mathrm{CKD}_{ji})\\
\boldsymbol{x}_{ji2} &:= (\mathrm{EF}_{ji}, \mathrm{COMP}_{ji}, Y_{ji1}, \mathrm{KILLIP}_{ji})\\
\boldsymbol{x}_{ji3} &:= (\mathrm{EF}_{ji}, \mathrm{MALE}_{ji}, \mathrm{STres}_{ji}, \mathrm{KILLIP}_{ji}).
\end{aligned}$$

Note that with a slight abuse of notation, Y_{ji1} was also added to the set of covariates $\boldsymbol{x}_{ji2}$.

It is interesting to note that at the first level, all the covariates ($\boldsymbol{x}_{ji1}$) related to logistic and organizational issues are retained: the way a patient is delivered to the emergency room, the time at which the first ECG is received, and the arrival time (recall this is coded as on/off hours), which are clearly related to the efficiency and promptness of the treatment received. Finally, the presence/absence of chronic kidney disease (CKD) is also meaningful, because this condition may influence the time to intervention. Indeed, it is likely that more complex procedures will be required before undergoing surgery, since the radiocontrast agent may harm kidneys. On the other hand, in-hospital survival in STEMI patients undergoing angioplasty in this dataset will be modeled as depending on the initial patient's heart condition (EF), treatment time DB, severity of infarction (as indicated by KILLIP) and the presence of complications after PTCA (COMP). Finally, our model for mid-term survival (here ALIVE60) includes the initial heart condition (EF), the severity of infarction (KILLIP), whether the treatment had been effective (STres) and the patient's gender (MALE). Note that the patient's age was never selected, even under different variances of the prior, or under the SSVS spike-and-slab prior. However, gender is usually highly correlated to age in STEMI patients (Trappolini et al. 2001; Vakili et al. 2001).

Observe that Y_1, i.e. the treatment time that mainly depends on the organizational issue of hospitals, is selected as significant in predicting the in-hospital survival. As we will later see from the posterior summaries, an elevated DB time decreases in-hospital survival. However the DB is not affecting the mid-term survival. This finding may seem counterintuitive and even disappointing, but it is not really surprising in the real life context. Despite the fact that DB is determinant on the effectiveness of in-hospital practices, there are many other factors that may affect the patient's quality of life and survival once discharged. Among these factors, we mention, for example, compliance to the prescribed therapy, comorbidities affecting the patient, her/his reaction to the disease, and the assistance received at home. All these factors are determinant and not always measurable, so it is perfectly coherent that the hospital impact, here measured by the efficiency in delivering patients to the coronary unit, decreases as the time from discharge increases. Finally, it is worth mentioning that the selection included the binary indicator MILAN for explaining the first and the third responses, but as a modeling choice we decided to include this information in the non-parametric component; see details below.

We introduce the model now. Recall that $i = 1, \dots, n_j$ indexes patients treated in hospital j, for $j = 1, \dots, J$. We assume that the conditional distributions in (1) are:

$$Y_{ji1} | \mu_{ji}, \sigma_j \sim \mathcal{N}(\mu_{ji}, \sigma_j^2), \quad \mu_{ji} = \boldsymbol{\beta}_1{}^T \boldsymbol{x}_{ji1} + b^1_{\phi_j j} \tag{2}$$

$$Y_{ji2} | p_{ji}, Y_{ji1} \sim Be(p_{ji}), \quad \text{logit}(p_{ji}) = \boldsymbol{\beta}_2{}^T \boldsymbol{x}_{ji2} + b^2_{\phi_j j} \tag{3}$$

$$Y_{ji3} | q_{ji}, Y_{ji1}, Y_{ji2} \sim \begin{cases} Be(q_{ji}) & \text{if } Y_{ji2} = 1 \\ \delta_0 & \text{if } Y_{ji2} = 0 \end{cases}, \quad \text{logit}(q_{ji}) = \boldsymbol{\beta}_3{}^T \boldsymbol{x}_{ji3} + b^3_{\phi_j j}. \tag{4}$$

Here, as usual, δ_0 denotes the degenerate distribution at 0.

Observe that in (2)–(4), parameters $b^1_{\phi_j j}, b^2_{\phi_j j}, b^3_{\phi_j j}, \sigma_j$ refer to hospital-specific random effects; the former three are random intercepts, while the latter is the standard deviation of the first response. Notation ϕ_j is a dummy variable indicating if the hospital is in Milano, or outside the city. In fact, the management of emergencies is pretty different inside or outside the city, due to the different concentration of providers on the territory and to the related accessibility. These are expected to affect times to intervention and consequently, patients' outcomes. On the other hand, Milano can be considered as a hub that is more attractive to patients, which may explain the wider spectrum of recorded cases in the city. Therefore, it seems reasonable to establish an explicit difference in the random effects, according to whether the hospital is in (b^1_1) or outside of (b^1_0) Milano. In fact, a feature of our model is that we allow the entire shape of the random effects distributions to change according to this geographical characteristic. This is exactly the reason why we consider a dependent nonparametric prior specification; see details below.

Consequently, our inference will mainly focus on parameter

$$\boldsymbol{\theta} = \Big(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, (b^1_{0j}, b^1_{1j}, b^2_{0j}, b^2_{1j}, b^3_{0j}, b^3_{1j}, \sigma_j, j = 1, \dots, J)\Big),$$

where J is the number of hospitals in the dataset. We assume a priori independence of all components of $\boldsymbol{\theta}$ and:

$$\boldsymbol{\beta}_1 \sim \mathcal{N}_4(\mathbf{0}, 100\mathbb{I}_4), \quad \boldsymbol{\beta}_2 \sim \mathcal{N}_4(\mathbf{0}, 100\mathbb{I}_4), \quad \boldsymbol{\beta}_3 \sim \mathcal{N}_4(\mathbf{0}, 100\mathbb{I}_4), \tag{5}$$

and for $j = 1 \dots, J$,

$$\sigma_j \overset{\text{iid}}{\sim} U(0, 10), \tag{6}$$

$$(b^1_{0j}, b^1_{1j}, b^2_{0j}, b^2_{1j}, b^3_{0j}, b^3_{1j}) | P \overset{\text{iid}}{\sim} P, \quad P \sim PY(a, b, P_0). \tag{7}$$

By $P \sim PY\ (a,\ b,\ P_0)$ we mean that P is a draw from the Pitman-Yor process (Pitman and Yor 1997), sometimes known as the two-parameter Poisson-Dirichlet process, with parameters $0 \leq a < 1$ and $b > -a$, and where P_0 is a probability measure on $\mathbb{R}^6$. When $a = 0$, the DP case is recovered. Note that the nonparametric specification (7) together with the sampling model (2)–(4) results in a generalization of the ANOVA-DDP prior in De Iorio et al. (2004).

For ease of computation it is useful to introduce the stick-breaking representation for P (Pitman 1995):

$$P = \sum_{i=1}^{\infty} V_i \delta_{\boldsymbol{\tau}_i}, \quad \text{where } \{V_i\} \perp \{\boldsymbol{\tau}_i\}, \tag{8}$$

the $\boldsymbol{\tau}_i'$s are iid according to P_0 and $\{V_i\}$ are stick-breaking weights, i.e.

$$V_1 = Z_1, \quad V_j = Z_j \prod_{i=1}^{j-1} (1 - Z_i) \ \ j \geq 2, \quad Z_i \overset{ind}{\sim} \text{Beta}(1 - a, b + ia), \quad i = 1, 2, \ldots . \tag{9}$$

It is well-known that a random sample from a distribution F that is assigned a discrete random prior probability measure such as (7) induces a random partition ρ of corresponding labels. In this case, the partition is induced on the hospitals labels $\{1,\ 2, \ldots,\ J\}$ by (7). Thus, hospitals would be grouped by identifying those with iden-tical random intercept, according to (2)–(4). In this way, we will be allowed to carry out model-based clustering by computing a summary estimate of the posterior distri-bution of the random partition ρ. The induced partition structure is more general than that coming from the particular DP case.

Next, we assume for the locations $\boldsymbol{\tau}_i \in \mathbb{R}^3$ the parametrization usually adopted for the ANOVA-DDP prior, i.e.

$$\boldsymbol{\tau}_i = \boldsymbol{\tau}_{0i} + \boldsymbol{\tau}_{1i} \eta_i,$$

where η_i is 1 if the patient i was admitted to an hospital in Milano, and 0 otherwise. On the whole, the location parameters are identified by $(\tau_{01i}, \tau_{11i}, \tau_{02i}, \tau_{12i}, \tau_{03i}, \tau_{13i})$, and we assume they are iid from the base probability measure P_0 on $\mathbb{R}^6$ given by the product measure of six independent Gaussian distributions with random means and variances:

$$\begin{gathered} P_0 = \mathcal{N}(m_1, \lambda_1^2) \times \mathcal{N}(m_2, \lambda_2^2) \times \mathcal{N}(m_3, \lambda_3^2) \times \mathcal{N}(m_4, \lambda_4^2) \times \mathcal{N}(m_5, \lambda_5^2) \times \mathcal{N}(m_6, \lambda_6^2) \\ (m_1, \ldots, m_6, \lambda_1, \ldots, \lambda_6) \sim \pi_m \times \pi_\lambda. \end{gathered} \tag{10}$$

The prior clustering is controlled by hyperparameters a and b in (7). For fixed a, the number of clusters is stochastically increasing with b. This can be seen as a "rich gets richer" property of the PY that is also shared by the DP. In fact, when considering a sample of size n from a PY process with k ($\leq n$) distinct values in the sample, the probability that the next observation coincides with the j-th sampled distinct value is equal to $(n_j - a)/(b + n)$, where n_j is the number of observations equal to the j-th sampled distinct value. Therefore, it is clear that the increase in number of clusters is attenuated by a, which can be thought of as a discount parameter. See, for instance, Müller and Mitra (2013). Therefore, the prior distribution of K_J, the number of clusters among hospitals, depends on those parameters; later on in the paper, we determine (a, b) by matching the first two moments of K_J with prior information. The model specification is completed by assuming

$$m_i \overset{\text{ind}}{\sim} \mathcal{N}(m_{i0}, \sigma_{i0}^2), \quad \lambda_i \overset{\text{iid}}{\sim} \mathcal{U}(0, \lambda_0) \quad i = 1, \ldots, 6, \tag{11}$$

but other weakly informative choices could be considered as well.

Before discussing specific results, we point out that we tried extensive posterior simulation experiments under several types of priors for the PY parameters a and b. These experiments found a number of posterior simulation problems, such as poor mixing of certain parameters, but also, none of the models we tried produced a better fitting to the data than what we will describe in Sect. 4.

4 Posterior inference

4.1 Model details

It is quite common in practice to consider a truncated version of the stick-breaking representation (8)–(9) so as to work with a finite mixture model (see Ishwaran and Zarepour 2002). This is achieved by considering a number of components, say H, and setting $Z_H = 1$ in (9). Posterior inference can then be implemented through a standard Gibbs sampler algorithm, which we coded in JAGS (Plummer 2003) with the aid of R (R Development Core Team 2012). In what follows, we always use the first 50,000 iterations as burn-in, and saving every 75-th iteration after that, to complete a Monte Carlo posterior sample of size 5000. Standard convergence diagnostics criteria such as those available in the R package CODA (Plummer et al. 2006) were applied to all parameters, indicating that convergence had been achieved.

To fit the model, we selected hyperparameter values that reflect lack of prior information, in other words, a vague yet proper prior distribution. Specifically, we chose

$$m_{i0} = 0, \quad \sigma_{i0}^2 = 25, \quad i = 1, \ldots, 6, \quad \lambda_0 = 5.$$

The prior for the Pitman-Yor process was specified as follows. We fixed $a = 0.3$ and $b = 0.5$ so that the prior mean and variance of K_J are equal to 5.285 and 6.113, respectively.

4.2 Posterior summaries

Table 1 reports the 95 % posterior credible intervals, as well as the posterior marginal probability on the negative reals, for the fixed-effects parameters at all the three levels. From the reported inference, it is clear that patients who were not delivered by the

Table 1 Posterior 95 % credibility intervals for the fixed-effects parameters; p^- is the posterior probability that the parameter is negative

Level 1				Level 2				Level 3			
Parameter	2.5 %	97.5 %	p^-	Parameter	2.5 %	97.5 %	p^-	Parameter	2.5 %	97.5 %	p^-
ACCESS	−0.015	0.133	0.059	EF	0.692	1.732	0	EF	0.708	1.578	0
ECG	0.106	0.161	0	COMP	−5.543	−1.229	1	MALE	−1.120	0.851	0.604
WE	0.012	0.141	0.013	Y_1	−1.288	0.588	0.764	STres	−1.520	0.284	0.912
CKD	0.032	0.318	0.007	KILLIP	−2.382	0.151	0.963	KILLIP	−2.491	−0.301	0.994

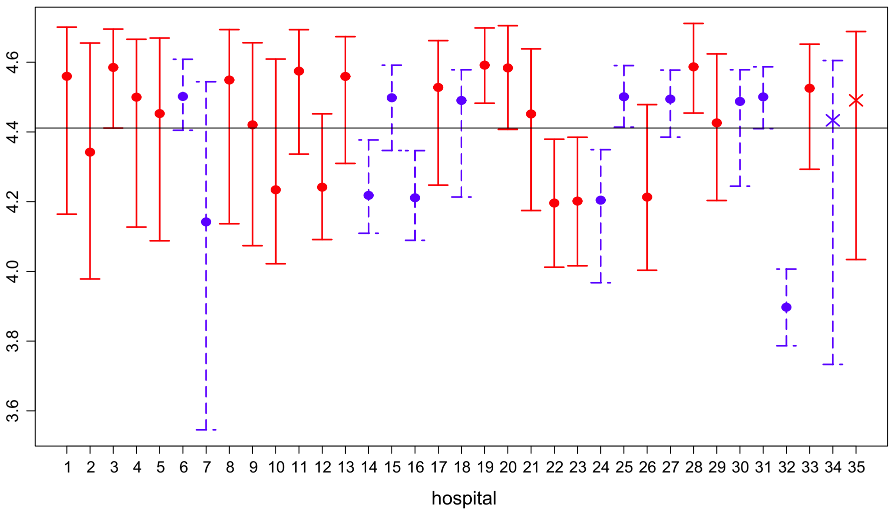

Fig. 3 Posterior CI of b_j^1, the random effect parameters at the first likelihood level. Hospitals located in Milano are depicted in *dashed* (*blue*) *lines*, those outside Milano in *solid* (*red*) *lines*, and *bullets* represent the corresponding posterior medians. The last two intervals show predictions (medians are marked by *crosses*) for random intercepts corresponding to two hypothetical new hospitals, located in (*red*) and outside (*blue*) Milano. For reference, the horizontal line represents the mean of all displayed means (color figure online)

118 service and/or arrived at weekends or nights are penalized in terms of DB time. Furthermore, as expected, an increase of ECG time yields an increase of DB, testifying the importance of executing promptly ECG to the patients when infarction diagnoses are suspected. The presence of CKD is also significant: as we said before, this makes sense since complications may arise when treating a patient whose kidneys do not work properly.

For the in-hospital survival probability (level 2), patients with a more severe infarction (KILLIP equal to one) are penalized. The presence of complications after primary angioplasty is a negative prognostic factor too. In addition, an elevated DB time (Y_1) decreases significantly the survival probability. On the other hand, the ejection fraction at admission (EF) has a positive effect on in-hospital survival.

Similarly, EF and KILLIP have positive and negative effect, respectively, on mid-term survival, while the non efficacy of the PTCA, quantified by the STres, plays a negative role as expected. Even if it is clear that gender (i.e. the male indicator) has a negative effect, this is rather moderate.

In Figs. 3 and 4 we provide posterior 95 % CIs of the hospital random intercepts; in all these figures, as before, the hospitals are ordered from left to right by increasing number of patients available in the sample. The last two intervals in each panel represent predictions for random intercepts corresponding to two hypothetical new hospitals, located in (continuous red) and outside (dashed blue) Milano. It is clear that there is a hospital effect in the first DB times (see the variability of the estimates in Fig. 3). In particular, hospitals located in Milano show a lower variability than those located outside. On the other hand, there is much more homogeneity in the random

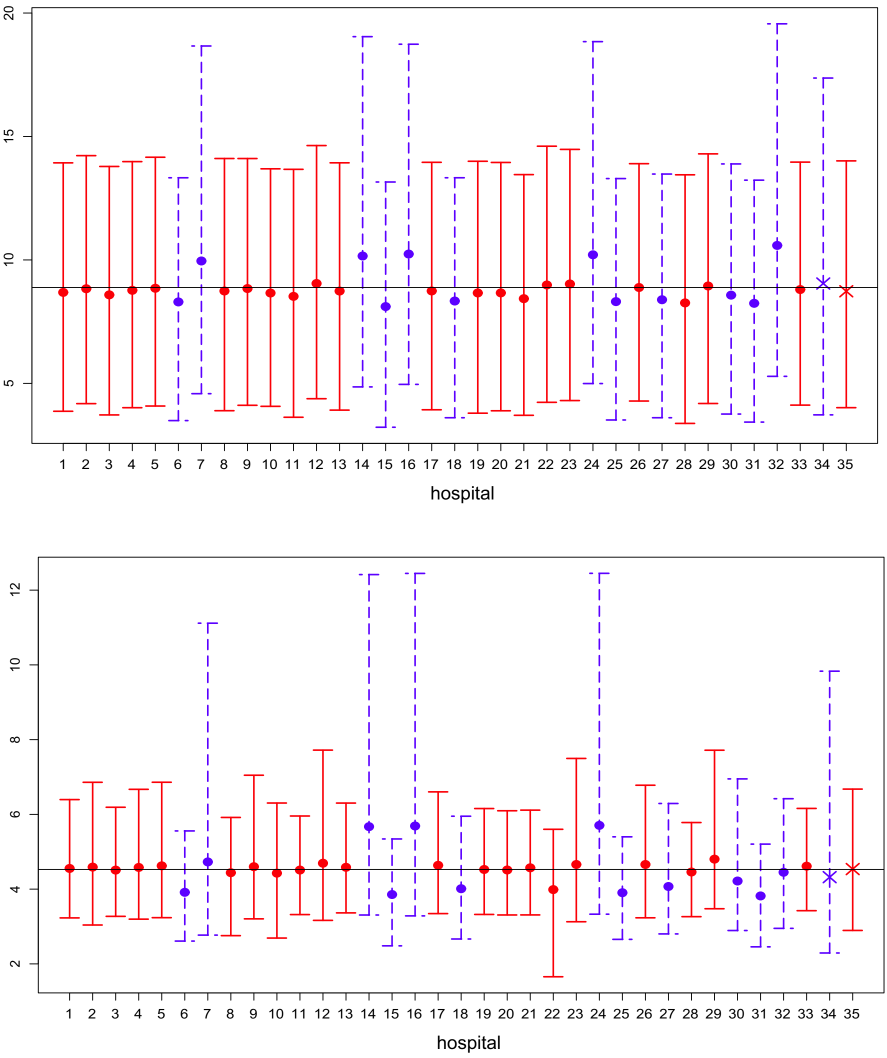

Fig. 4 Posterior 95 % CIs of the random intercepts b^2 (*top*) and b^3 (*bottom*): hospitals located in Milano are depicted in *dashed* (*blue*) *lines*, those outside Milano in *solid* (*red*) *lines*, and *bullets* are the posterior medians. The last two intervals represent new random intercepts for a hospital in and outside Milano, with crosses representing posterior medians (color figure online)

intercepts at the second and third level. This behavior can be explained because all the coronary units treat patients according to general standards, which yields rather uniform hospital performances in terms of in-hospital survival.

We have also computed posterior predictive estimates of the different nonparametric component of the mixing measure. Figure 5 displays posterior estimates of the components of $P = \sum_{i=1}^{H} V_i \delta_{\boldsymbol{\tau}_i}$. In particular, the first row shows the first (left) and second (right) predicted components (level one in the likelihood, i.e. the posterior of

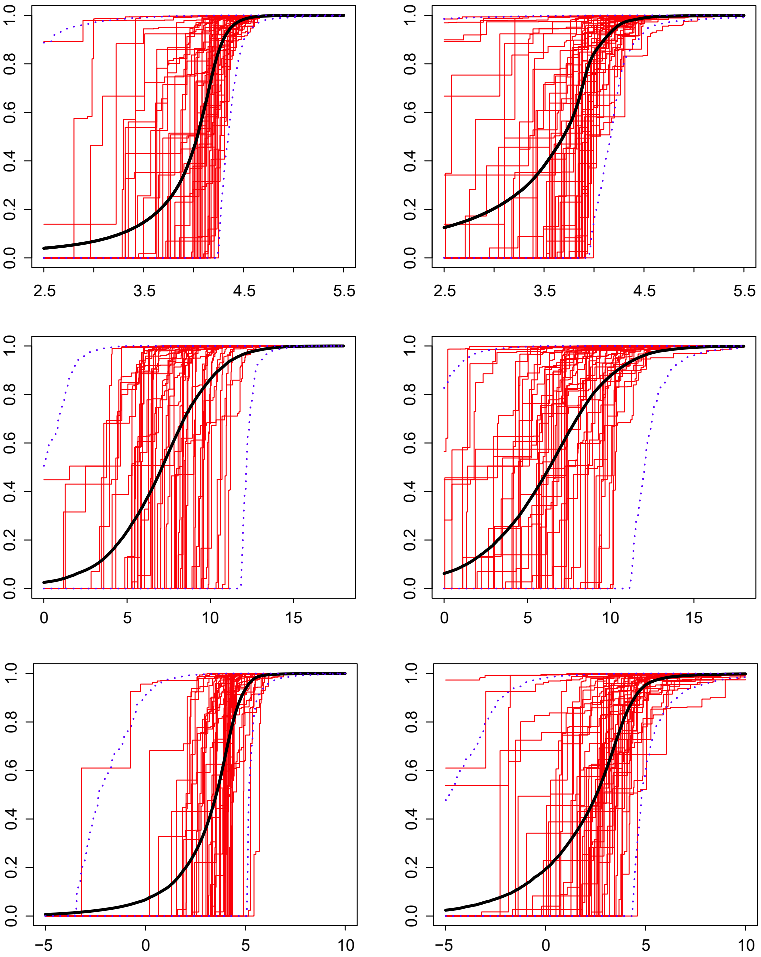

Fig. 5 Posterior predictive 95 % CIs of the components of P. See the text for an explanation. *Dashed* (*blue*) *lines* correspond to quantiles, the *solid central* is the mean (the Bayesian estimates). The last 50 iterations are superimposed (in *red*). The left (*right*) column displays trajectories of the three random components related to hospitals outside (*in*) Milano (color figure online)

$\sum_1^H V_i \delta_{\tau_{01i}}$ and of $\sum_1^H V_i \delta_{\tau_{01i}+\tau_{11i}}$), and similarly for the third and fourth (level two in the likelihood) and fifth and sixth (third level in the likelihood) components. Dashed (blue) lines correspond to 0.025 and 0.925 quantiles, while the solid central lines represent the respective means (i.e. the Bayesian estimates). The last 50 iterations are superimposed (in red). The picture shows a difference in the variability of the sampled trajectories, at least at the second and third level.

The results in Fig. 5 are in agreement with the information conveyed by the marginal posterior distributions of m_2, m_4 and m_6 (not reported here), which represent the average difference in the random effect parameters between hospitals in or out of Milano. In fact, the marginal posterior distribution of m_2 is concentrated around 0 (posterior mean and variance are −0.152 and 0.153, respectively), denoting that on average there is no Milano effect on the log *DB* response. The marginal posterior distributions of m_4 and m_6 are much more spread out.

4.3 Posterior inference on clustering

As pointed out earlier, the discrete trajectories of the nonparametric prior assumption imply a clustering of the hospitals. We found that the posterior mean and variance of K_J, the number of groups among hospitals, are 5.602 and 3.795, respectively, with a posterior mode of 4 (but 5 has a posterior probability very close to that of 4). Figure 6 displays the whole estimated posterior distribution.

The Bayesian cluster estimate was here computed as the random partition of the hospital labels $\{1, 2, \ldots, 33\}$ that minimizes the posterior expectation of Binder's loss function, as proposed in Lau and Green (2007); this function assigns cost w when two elements are wrongly clustered together and cost u when two elements are erroneously assigned to different clusters. For equal misclassification costs w and u, we obtained 9 clusters in total, but only 4 with sizes larger that 1. Table 2 reports the four non-singleton groups in the cluster estimate. We underline that this estimate agrees with the least squares estimate of Dahl (2006).
We have computed sample means of responses and covariates per hospital clusters in Table 2, averaging over all patients of all hospitals in each group. Table 3 reports those values. Cluster A_{PY} (the most populated) could be characterized as grouping patients

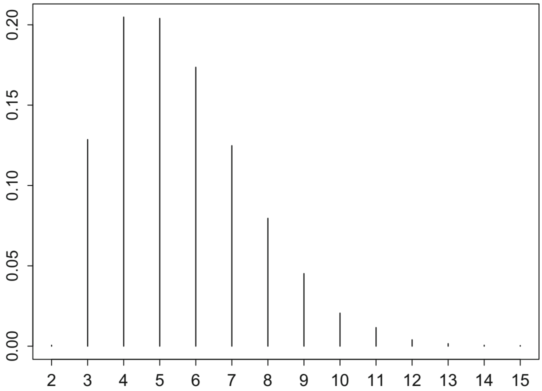

Fig. 6 Posterior distribution of the number of groups among the hospitals

Table 2 Hospital clusters with sizes larger than 1 from the proposed model under the Pitman-Yor process prior when $a = 0.3$ and $b = 0.5$

Cluster A$_{PY}$	{1, 3, 4, 6, 8, 11, 13, 15, 17, 18, 19, 20, 25, 27, 28, 30, 31, 33}
Cluster B$_{PY}$	{12, 16, 24, 29}
Cluster C$_{PY}$	{9, 22, 26}
Cluster D$_{PY}$	{5, 7, 10}

Table 3 Responses and covariate summaries by clusters in Table 2, for $a = 0.3$, $b = 0.5$

Groups	A_{PY}	B_{PY}	C_{PY}	D_{PY}
No. hospitals	18	4	3	3
No. patients	398	106	72	29
Y_1 (DB)	115.455	87.953	83.514	80.966
Y_2 (ALIVEIN)	0.955	1.000	0.986	0.966
Y_3 (ALIVE60)	0.940	1.000	0.944	0.966
MILANO	0.475	0.387	0.000	0.276
ACCESS	0.555	0.642	0.597	0.414
ECG	10.487	10.708	6.764	8.448
WE	0.440	0.481	0.417	0.448
CKD	0.090	0.028	0.111	0.103
EF	47.621	49.519	49.208	47.310
COMP	0.465	0.236	0.333	0.586
KILLIP	0.083	0.000	0.056	0.000
AGE	64.500	65.104	64.931	68.276
MALE	0.802	0.708	0.819	0.724
STres	0.176	0.226	0.250	0.069

with the highest DB times and the lowest survival rates. This suggests that the hospitals in this group may have procedures that could be improved to achieve a better performance. Note also that the KILLIP rate in cluster A_{PY} is the highest among the groups, and higher than the sample grand mean (0.06). In contrast, clusters B_{PY} and D_{PY} group hospitals with patients having less severe infarction. In addition, clusters B_{PY} and D_{PY} differ in the associated complications exhibited by patients and in the way patients accessed the hospitals. Finally, cluster C_{PY} contains only hospitals outside Milano and the corresponding patients have on average the lowest ECG values in the sample.

4.4 Predictive goodness-of-fit

We consider now predictive checks for the proposed model. We first computed the log pseudo-marginal likelihood (LPML) statistic (Geisser and Eddy 1979) for this model; see also Gelfand and Dey (1994). This corresponds to the product, of the conditional predictive density of the responses, expressed in log-scale, i.e.

$$LPML = \sum_{i=1}^{n} \log(CPO_i),$$

where CPO_i, the conditional predictive ordinate for the ith patient, represents the conditional density (evaluated at $\boldsymbol{y}_{ji}$), of $\boldsymbol{Y}_{ji}$, given all the other observations. We also computed the mean squared error MSE of the prediction errors, i.e. the mean of SE_i over hospitals, given by

Table 4 Predictive goodness-of-fit measures when the prior of the random effects is the Pitman-Yor process with parameters $a = 0.3$, $b = 0.5$, or $a = 0$, $b = 0.5$ (i.e. the Dirichlet process with parameter $b = 1.53$), or parametric

Random effects prior	PY $(a = 0.3, b = 0.5)$	PY $(a = 0, b = 1.53)$	Parametric
LPML	−594.40	−593.037	−596.053
MSE	0.282	0.282	0.280
$WAIC_1$	−581.847	−581.082	−582.096
$WAIC_2$	−589.613	−588.974	−591.105

$$SE_i = (Y_{ji1} - \hat{\mu}_{ji})^2 + (Y_{ji2} - \hat{p}_{ji})^2 + (Y_{ji3} - \hat{q}_{ji})^2,$$

where the hat denotes the posterior expectation of the corresponding parameters. Fol-lowing Gelman et al. (2014), we also considered the Watanabe-Akaike information criterion (WAIC), computed as the log pointwise predictive density, incorporating bias corrections. Specifically, we computed

$$WAIC_1 = lppd - p_{WAIC_1} \quad \text{and} \quad WAIC_2 = lppd - p_{WAIC_2},$$

where $lppd$ is the log pointwise predictive density, i.e. the product (in the log scale) of the conditional densities (evaluated at $\boldsymbol{y}_{ji}$), of $\boldsymbol{Y}_{ji}$, given all the data, and then adding the two alternative corrections p_{WAIC_1} and p_{WAIC_2} for effective number of parameters to adjust for overfitting. The bias correction p_{WAIC_1} is similar to the bias correction in the definition of the DIC, while p_{WAIC_2} is the sum of the posterior variances of the conditional density of the data. For further details, see Section 3 of Gelman et al. (2014). The computed predictive goodness-of-fit measures are in Table 4.

4.5 Comparison with competitor models

When introducing the proposed model (2)–(7), we aimed at justifying all the choices we made. However it is natural to wonder whether simpler models could give similar inference. While we are pretty satisfied about the conditional distribution of data, given parameters (see (2)–(4)), we acknowledge that other simpler priors could be considered here. We examine some alternatives next.

First, let us consider the same prior as before, but now setting $a = 0$, which reduces nonparametric prior to a Dirichlet process (DP). In this case we fixed $b = 1.53$ to match the prior mean under the Pitman-Yor process prior component. In particular we now have $\mathbb{E}(K_j) = 5.302$ and $\text{Var}(K_J) = 3.130$. The estimated fixed effects under the DP prior are very similar to those in Table 1 (data not shown).

We have also computed the posterior predictive estimates of the different nonpara-metric components in the mixing measure. Comparing Figs. 5 and 7 we find that the means are almost identical, but the quantile curves in the DP case are a bit more separated from the mean, suggesting a slightly increased posterior uncertainty in the corresponding posterior distributions.

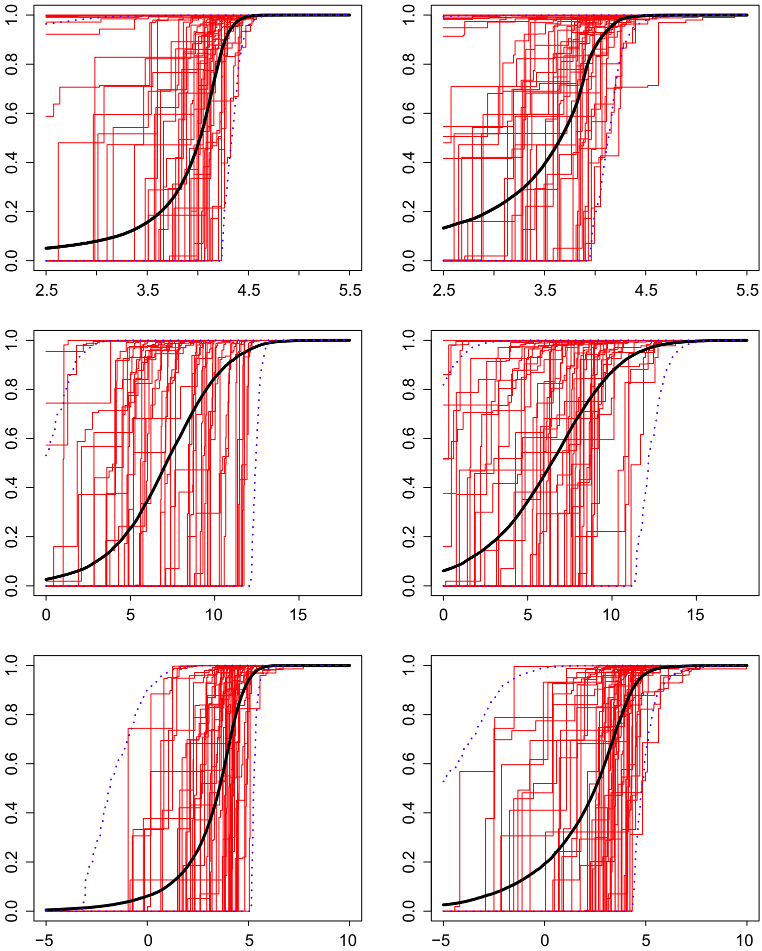

Fig. 7 Posterior predictive 95 % CIs of the components of P when $a = 0$, $b = 1.53$. *Dashed* (*blue*) *lines* correspond to quantiles, the *solid central* is the mean (the Bayesian estimates). The last 50 iterations are superimposed (in *red*). The left (*right*) column displays trajectories of the three random components related to hospitals outside (*inside*) Milano (color figure online)

The cluster estimate we obtained in this case contains 8 groups, but only 5 have sizes larger that 1. Table 5 describes the largest groups in the cluster estimate. Again, this results agrees with what is obtained when using the least squares method of Dahl (2006). The partitions are now a bit different, with the big cluster containing less points than the PY case.

Table 6 reports sample means of responses and covariates per hospital clusters, averaging over the entire set of patients in all hospitals forming each group in Table 5. The largest cluster here is similar to that in Table 3, but with two less members. However the other groups cannot be as clearly interpreted in terms of responses as in the case of the proposed PY process model. For instance, the mean values of DB are ordered in Table 3, unlike in Table 6. Under the previous case, the hospitals with

Table 5 Hospital clusters with size larger than 1 from the proposed model under the Pitman-Yor process prior when $a = 0$ and $b = 1.53$; this corresponds to assuming a Dirichlet process prior with parameter $b = 1.53$

Cluster A_{DP}	$\{1, 3, 6, 8, 11, 13, 15, 18, 19, 20, 25, 27, 28, 30, 31, 33\}$
Cluster B_{DP}	$\{4, 7, 17, 21, 29\}$
Cluster C_{DP}	$\{9, 14, 16, 24\}$
Cluster D_{DP}	$\{10, 23, 26\}$
Cluster E_{DP}	$\{5, 22\}$

Table 6 Responses and covariates summaries by clusters in Table 5, when $a = 0$, $b = 1.53$; this corresponds to a Dirichlet process prior for the random effect parameters

Groups	A_{DP}	B_{DP}	C_{DP}	D_{DP}	E_{DP}
No. hospitals	16	5	4	3	2
No. patients	375	99	73	72	37
DB	115.648	101.374	86.521	74.625	85.351
ALIVEIN	0.952	0.980	1.000	0.958	1.000
ALIVE60	0.936	0.980	1.000	0.958	0.919
MILANO	0.504	0.081	0.753	0.000	0.000
ACCESS	0.544	0.697	0.534	0.583	0.595
ECG	9.899	13.818	12.356	5.556	5.108
WE	0.445	0.434	0.466	0.514	0.486
CKD	0.093	0.091	0.027	0.056	0.189
EF	47.211	49.869	51.068	47.181	48.946
COMP	0.469	0.313	0.329	0.417	0.108
KILLIP	0.083	0.061	0.014	0.056	0.000
AGE	64.485	66.293	64.178	63.181	66.541
MALE	0.795	0.687	0.753	0.819	0.865
STres	0.176	0.232	0.288	0.208	0.270

highest averages of both in-hospital and 60-days survival were in the second largest cluster. In contrast, here the best hospitals in terms of in-hospital survival are split in the third and fifth largest clusters, while those with highest average 60-days survival are in the largest third and second groups.

We remark here that, according to Table 4, the DP prior is slightly superior. The differences in values between the several criteria reported there are minimal though, and we still prefer the Pitman-Yor model because of a clearer interpretation of the selected partition, as discussed earlier.

We also considered a non-dependent version of the proposed Pitman-Yor model, which eliminates the in/out of Milano indicator in the random effects. Specifically, we change $b^{\ell}_{\phi_j j}$ to b^{ℓ}_{j} for $\ell = 1, 2, 3$ and $j = 1, \ldots, J$ in (2)–(4). Doing so we obtained predictive check values comparable to those already reported in Table 4, but miss the subtle yet relevant differences in the predictive curves (left versus right panels) in Figs. 5 or 7.

Another natural comparison involves a model with a parametric prior for random effects. Under this alternative model, the likelihood is the same as in (2)–(4), but now the random effects are assumed to be i.i.d. draws from the baseline distribution (10). Fitting this model produced fixed effects estimates quite similar to those in Table 1. Regarding random effects, the CIs for the b_1 parameters are similar to those in Fig. 3, but differences arise for the b_2 and b_3 terms (data not shown). The predictive check measures for this model are also presented in Table 4. Note that, except for the mean squared error (MSE), the non-parametric alternatives produce better fits to the data. In this case, we lose the multimodality of the marginal prior of the random effect parameters, and this has an impact on the inference.

As a final comparison, we considered also finite mixture models, for which the stick-breaking representation (8)–(9) is replaced by a simple vector of weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_H)$ and a Dirichlet prior on $\boldsymbol{\pi}$ with joint prior density $p(\boldsymbol{\pi}) \propto 1$ on the H-dimensional simplex. The finite mixture weights $\boldsymbol{\pi}$ are no longer stochastically ordered as the stick-breaking weights. An immediate concern of this approach is how to choose H. One option is to put a prior on H and consider reversible jumps technology. We considered the pragmatic approach consisting in fitting the resulting model for $H = 1, \ldots, 10$, computing predictive checks for each case, and choosing the best value. Doing so with the same measures presented in Table 4, we found that $H = 9$ gave the best results, with $MSE = 0.281$ and $LPML = -593.19$, $WAC_1 = -580.42$ and $WAC_2 = -588.76$. These numbers are quite similar to those in Table 4 so that we do not find big differences in these models. A further comparison involved a cross-validation study. At this stage, the following procedure was repeated 100 times: we randomly divided the data into two parts, training (of size 356) and testing (of size 341), with the restriction that data from all hospitals was in each subset. Specifically, we split the data from each hospital into two equal parts using one part for the training subset, and the other for testing (if the total number of data points was odd, we rounded the training part up). Next, we fitted three models to the training subset: the finite mixture model with $H = 9$ components, and our proposed PY model with each of $(a, b) = (0, 1.53)$ and $(a, b) = (0.3, 0.5)$; and finally, we predicted all the responses in the testing subset. At every repetition, we computed the MSE for both, training and testing. The average of these MSEs are given in Table 7. As we can see the numbers are again quite similar, with finite mixtures giving a slightly better fit and a slightly worse prediction than the proposed PY. In summary, our model and finite mixtures perform very similar for these data, but our proposal automatically takes care of the number of components to be considered, saving the additional computational complexity and cost of implementing reversible jumps. See further comments in next Section.

Table 7 Cross-validation study results

MSE	Finite mixture $H = 9$	PY (0,1.53)	PY (0.3,0.5)
Training	0.27132	0.27481	0.27598
Testing	0.34533	0.34530	0.34449

We display the average mean squared errors (MSE) for training and testing subsets over 100 randomly split subsets, and for each of finite mixtures and proposed model with the indicated parameters

5 Conclusions

We have presented a framework for semiparametric Bayesian modeling of mixed-type multiple outcomes for Acute Myocardial Infarction patients admitted to hospitals in Lombardy; we considered patients with STEMI diagnosis and treated with PTCA. Specifically, we have proposed a Bayesian nonparametric hierarchical model for clus-ter analysis, aimed at identifying hospital behavior that may affect the outcome at patient level. We have considered a conditional specification of the joint model for three responses: the door to balloon time (DB), the in-hospital survival and the sur-vival after 60 days from admission. The information on survival is available as binary outcomes. A different study based on survival times or in general time-to-event data might be of interest, and could possibly lead to different results. Nevertheless the analysis of the binary outcome of in-hospital and mid-term survival was one of the main study targets, as requested by clinicians. In fact, performance assessment in clin-ical practice is usually based on binary survival data, which is probably the reason why we have been asked to focus on these outcomes. Each conditional specification is a generalized linear model with random intercepts to account for hospital variability. We postulated a nonparametric prior for the random effects that incorporates dependence on a location indicator, which is used to explicitly differentiate among hospitals in or outside the city of Milano. The random effects are a sample from the Pitman-Yor process, more flexible than, yet encompassing the Dirichlet process prior. We have provided Bayesian estimates of the random effect parameters, predictive inference for the nonparametric components of the prior, and cluster estimates for the grouping of the hospitals as well.

In our data analysis we have considered a number of competitor models, either purely parametric, or with a nonparametric component. Though all the models provide similar results in terms of the fixed effects estimates, the most flexible model (where the random effects are modeled from the Pitman-Yor process) seems the one able to better explain the underlying clustering structure. JAGS code to fit the proposed model is available from authors upon request.

A referee asked why the Bayesian nonparametric (BNP) approach is needed in this case, suggesting instead maximum likelihood techniques in finite mixtures. One fundamental reason to prefer the BNP approach is the wide support it provides for the unknown dependent random effects distribution. See Barrientos et al. (2012). Of course, alternative approaches are plausible, but we stick to the Bayesian context to model the prior information we have (e.g. exchangeability of the hospitals), and to get richer inference (e.g. whole distribution of "parameters", instead of estimates). Flexibility in models, beyond simple standard parametric distributions, can be added in a number of various principled ways. Nonparametric mixture models (or their truncated versions, as in our model) is one of these ways. Finite mixture models, such as mixtures of experts (e.g., see McLachlan and Peel 2000) is yet another one. There is already experience in comparing such approaches for some specific models, and we have presented some in the previous Section. See further discussion about such comparisons in, for instance, Richardson and Green (1997) and in Müller et al.(2011). Part of the flexibility of the nonparametric approach lies in the fact the number of imputed mixture terms is random. For mixture models a similar setting requires reversible jump methodology, which may not be easy to implement, even when fixing a maximal number of components. The computational cost of the resulting model is comparable to that of the nonparametric one, the most consuming part being the updating of configurations. See, for instance, the recent paper Malsiner-Walli et al. (2016), where sparse finite mixture models have been proposed as an alternative to infinite mixtures in the context of (Bayesian) model-based clustering. It is apparent that the computational effort there is as heavy as in BNP infinite mixture models. We argue that the BNP approach is more natural, since the number of components in the mixture has a prior which is spread out on the number of "items" we have considered.

Finally, we value a comment by one of the referees on more general random parameter priors that could have been used here. Specifically the comment suggested to allow the effect of covariates vary with clusters of hospitals. This is certainly a very sensible option. One possibility here is to adopt the PPMx prior of Müller et al. (2011), where clusters are encouraged to group individuals that are homogeneous with regard to covariate values. In other words, individuals are more likely to co-cluster if their covariate values are closer. Adopting this method and comparing to the current analysis is part of future work.

Acknowledgements This work is part of the strategic programme "Exploitation, integration and study of current and future health databases in Lombardy 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.The first author wishes to thank people at Departamento de Estadística at PUC, Chile, for their kind hospitality. The fourth author was partially funded by Grant FONDECYT 1141057.

References

AHRQ (2015) Agency for healthcare research and quality. http://www.ahrq.gov/professionals/prevention-chronic-care/decision/mcc

Antoniucci D, Valenti A, Migliorini A et al (2002) Relation of time to treatment and mortality in patients with acute myocardial infarction undergoing primary angioplasty. Am J Cardiol 89:1248–1252

Barrientos AF, Jara A, Quintana FA (2012) On the support of MacEachern's dependent Dirichlet processes and extensions. Bayesian Anal 7:277–309

Bello NM, Steibel JP, Tempelman RJ (2012) Hierarchical Bayesian modeling of heterogeneous cluster- and subject-level associations between continuous and binary outcomes in dairy production. Biom J 54:230–248

Catalano P, Ryan L (1992) Bivariate latent variable models for clustered discrete and continuous outcomes. J Am Stat Assoc 87:651–658

Cox D, Wermuth N (1992) Response models for binary and quantitative variables. Biometrika 79:441–461

Dahl DB (2006) Model-based clustering for expression data via a Dirichlet process mixture model. In: Do KH, Müller P, Vannucci M (eds) Bayesian inference for gene expression and proteomics. Cambridge University Press, Cambridge, pp 201–218

De Iorio M, Müller P, Rosner G, MacEachern S (2004) An anova model for dependent random measures. J Am Stat Assoc 99:205–215

De Luca G, Suryapranata H, Ottervanger JP, Antman EM (2004) Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: every minute of delay counts. Circulation 109:1223–1225

Dunson DB, Herring AH (2005) Bayesian latent variable models for mixed discrete outcomes. Biostatistics 6:11–25

Fitzmaurice G, Laird N (1995) Regression models for a bivariate discrete and continuous outcome with clustering. J Am Stat Assoc 90:845–852

Geisser S, Eddy WF (1979) A predictive approach to model selection. J Am Stat Assoc 74:153–160

Gelfand AE, Dey DK (1994) Bayesian model choice: asymptotics and exact calculations. J R Stat Soc B 56:501–514

Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. Stat Comput 24:997–1016

Grieco N, Ieva F, Paganoni A (2012) Performance assessment using mixed effects models: a case study on coronary patient care. IMA J Manage Math 23(2):117–131

Guglielmi A, Ieva F, Paganoni A, Ruggeri F, Soriano J (2014) Semiparametric Bayesian modeling for the classification of patients with high observed survival probabilities. J R Stat Soc C 63:25–46

Ieva F (2013) Designing and mining a multicenter observational clinical registry concerning patients with acute coronary syndromes. In: Grieco N, Marzegalli M, Paganoni AM (eds) New diagnostic, therapeutic and organizational strategies for patients with Acute Coronary Syndromes, Springer, pp 47–60

Ieva F, Paganoni A (2010) Multilevel models for clinical registers concerning stemi patients in a complex urban reality: a statistical analysis of momi2 survey. Commun Appl Ind Math 1(1):128–147

Ieva F, Marra G, Paganoni AM, Radice R (2014) A semiparametric bivariate probit model for joint modeling of outcomes in stemi patients. Comput Math Methods Med. doi:10.1155/2014/240435 (**in press**)

Ishwaran H, Zarepour M (2002) Exact and approximate sum representations for the Dirichlet process. Can J Stat 30:269–283

Lau JW, Green PJ (2007) Bayesian model based clustering procedures. J Comput Graph Stat 16:526–558

Lombardia (2009) Determinazioni in merito alla rete per il trattamento dei pazienti con infarto miocardico con tratto ST elevato (STEMI)

Malsiner-Walli G, Frühwirth-Schnatter S, Grüun B (2016) Model-based clustering based on sparse finite Gaussian mixtures. Stat Comput 26:303324

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

Müller P, Mitra R (2013) Bayesian nonparametric inference—why and how. Bayesian Anal 8:269–302

Müller P, Quintana FA (2004) Nonparametric Bayesian data analysis. Stat Sci 19:95–110

Müller P, Quintana F, Rosner GL (2011) A product partition model with regression on covariates. J Comput Graph Stat 20:260–278

Normand SLT (2008) Some old and some new statistical tools for outcomes research. Circulation 118:872–884

Parekh A, Goodman R, Gordon C, Koh H (2011) Managing multiple chronic conditions: a strategic framework for improving health outcomes and quality of life. Public Health Reports 126:460–471

Pitman J (1995) Exchangeable and partially exchangeable random partitions. Probab Theory Relat Fields 102:145–158

Pitman J, Yor M (1997) The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann Probab 25:855–900

Plummer M (2003) Jags: a program for analysis of Bayesian graphical models using Gibbs sampling

Plummer M, Best N, Cowles K, Vines K (2006) Coda: convergence diagnosis and output analysis for MCMC. R News 6:7–11

R Development Core Team (2012) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/, ISBN 3-900051-07-0 Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components. J R Stat Soc B 59:731–792

Rockova V, Lesaffre E, Luime J, Löwenberg B (2012) Hierarchical Bayesian formulations for selecting variables in regression models. Stat Med 31:1221–1237

Sammel M, Ryan LM, Legler JM (1997) Latent variable models for mixed discrete and continuous outcomes. J R Stat Soc Ser B (Methodol) 59:667–678

Trappolini M, Chillotti F, Rinaldi R, Trappolini F, Coclite D, Napoletano A, Matteoli S (2001) Sex differences in incidence of mortality after acute myocardial infarction. Ital Heart J Suppl 3:759–766

Vakili B, Kaplan R, Brown D (2001) Sex-based differences in early mortality of patients undergoing primary angioplasty for first acute myocardial infarction. Circulation 104:3034–3038

Weiss R, Jia J, Suchard MA (2011) A Bayesian model for the common effects of multiple predictors on mixed outcomes. Interface Focus 1:886–894